12-1996

Digital Rosetta Stone: A Conceptual Model for Maintaining Long-Term Access to Digital Documents

Steven B. Robertson

DIGITAL ROSETTA STONE: A CONCEPTUAL MODEL FOR MAINTAINING LONG-TERM ACCESS TO DIGITAL DOCUMENTS

THESIS

Steven B. Robertson, Captain, USAF

AFIT/GIR/LAR/96D-8

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 Department of Defense or the U. S. Government.

AFIT/GIR/LAR/96D-8

DIGITAL ROSETTA STONE: A CONCEPTUAL MODEL FOR MAINTAINING LONG-TERM ACCESS TO DIGITAL DOCUMENTS

THESIS

Presented to the Faculty of the Graduate School of
Logistics and Acquisition Management of the
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

Steven B. Robertson, B.S.
Captain, USAF

December 1996

Acknowledgments

I would like to take this opportunity to acknowledge the people who have made this research effort possible. I am deeply indebted to my thesis advisor, Dr. Alan Heminger. His insight and guidance during the research for this project were invaluable. He provided motivation and direction while giving me the freedom to explore, learn, and develop in my own way. I would also like to thank my reader Lt Col James Wedertz. In addition to providing moral support and a genuine interest in my work, he contributed valuable information that guided my research efforts.

Additionally, I would like to thank the entire Air Force Institute of Technology library staff. During the research phase of this project, their assistance in obtaining research materials was invaluable.

Finally, I would like to thank the most important person, Marcia A. Robertson, my wife. Without her understanding and patience this effort would not have been possible. During the past 18 months she has provided me with the support, motivation, and encouragement I needed to complete this thesis and the AFIT program.

Steven B. Robertson

Table of Contents

Page

List of Figures

List of Tables

Abstract

Due to the rapid evolution of technology, future digital systems may not be able to read and/or interpret the digital recordings made by older systems, even if those recordings are still in good condition. This thesis addresses the problem of maintaining long-term access to digital documents and provides a methodology for overcoming access difficulties due to technological obsolescence. A review was conducted to determine the long-term access methods that have already been suggested by other researchers. These previously suggested methods are then combined with other ideas that were encountered and conceived while performing research for this project. The combination of these methods and ideas led to the creation of a model, the Digital Rosetta Stone, that provides a methodology for maintaining long-term access to digital documents. The hypothesis for the model is that knowledge preserved about different storage devices and file formats can be used to recover data from obsolete media and to reconstruct the digital documents. The Digital Rosetta Stone model describes three processes that are necessary for maintaining long-term access to digital documents in their native formats--knowledge preservation, data recovery, and document reconstruction. Finally, recommendations are made for the evaluation and implementation of the Digital Rosetta Stone.

DIGITAL ROSETTA STONE: A CONCEPTUAL MODEL FOR MAINTAINING LONG-TERM ACCESS TO DIGITAL DOCUMENTS

I. Introduction

Background

Any large organization has the need to retain and, on occassion, refer to various documents. Until recently, documents were generally paper based and stored as paper copies or on microfilm. However, modern data storage methods have evolved to include digital storage of documents.

Due to the rapid evolution of technology, future digital systems may not be able to read and/or interpret the digital recordings made by older systems, even if those recordings are still in good condition (OASD, 1995). This has become a major concern within the United States Air Force (USAF) and continues to grow in importance as the use of digital documents on Air Force information systems increases.

Digital documents that are official records must be categorized and managed in accordance with approved records schedules (OASD, 1995). This means that digital documents that are official Government records must be retained and accessible throughout their life cycle in accordance with the same laws and standards that govern paper records. Additionally, the law dictates that an official Government record must be classified into one of 26 retention periods set forth by the Archivist of the United States (USAF, 1995). These retention periods range from 30 days to Permanent storage and include time periods of 30 years, 50 years, and 75 years.

Digital documents that require long retention periods face accessibility problems due to the technology obsolescence of hardware and software. The National Research Council (1986) best describes this problem in the following statement:

> The fact that most electronic hardware is expected to function for no more than 10 to 20 years raises very serious problems for long-term (more than 20 years) archival preservation. Even if the operating systems and documentation problems somehow are dealt with, what is the archivist to do when the machine manufacturer declares the hardware obsolete or simply goes out of business? Will there be an IBM or Sony in the year 2200? If they still exist, will they maintain a 1980-1990 vintage machine? Moreover, it must be realized that no archival organization can hope realistically to maintain such hardware itself. Integrated circuits, thin film heads, and laser diodes cannot be repaired today, nor can they be readily fabricated, except in multimillion-dollar factories.

The Task Force on Archiving of Digital Information (1996) made a similar observation in their report *Preserving Digital Information* in which they stated:

> Reading and understanding information in digital form requires equipment and software, which is changing constantly and may not be available within a decade of its introduction. . . We cannot save the machines if there are no spare parts available, and we cannot save the software if no one is left who knows how to use it.

As information technology, including both systems software and hardware, continues to evolve, the ability to access documents in their "native formats" becomes a constant challenge. In order to maintain access to these documents over the period of their life cycle, it is necessary to overcome technology obsolescence. Furthermore, the ability to access documents in their native formats can prevent the need to reformat millions of documents in long-term storage on government information systems. The reformatting of these documents from their native formats to evolving formats can be resource intensive in areas such as manpower, money, and time. Additionally, the DoD has stated that, "When records are converted to an electronic (digital) medium, that medium becomes an electronic records management system. That records system must be capable of printing, presenting, and storing records in their native (original) format" (OASD, 1995). Therefore, the capability to access digital documents in their native

formats is not only a beneficial option but also necessary as the USAF upgrades its information systems.

A Call For Action

Unlike document preservation in the past when documents could be stored on microfilm, modern digital documents have evolved to a point where they cannot all be stored in such a manner. Today's digital documents, in addition to text and graphics, can contain items such as audio and video clips. These types of data cannot be printed in a way as to convey the same meaning as their audio and video representations. For example, what if a video and audio clip of Martin Luther King, Jr.'s famous "I Have a Dream" speech is embedded in a digital document on civil rights. While preserving the digital document for our posterity, instead of including the video and audio clip it is replaced with a photograph and a text version of King's speech. Will the photograph and words convey the same meaning to future generations as that of seeing King's charismatic nature and hearing his dynamic delivery of the speech? Probably not. In many cases it takes more than words and a still picture to convey the significance of something. Therefore, to preserve the context of such documents, we must create and implement a method to preserve them as they were originally presented.

Additionally, as digital technology continues to evolve at a rapid pace, superseded technologies are quickly discarded and new technologies are embraced in the hopes of gaining improved efficiency, effectiveness, or a competitive advantage. Because of the haste to discard superseded digital technologies "action is urgently needed to ensure that documents, software products and other digital information objects that document the early digital age from 1945 to 1990 are preserved before they slip irrevocably away" (TFADI, 1996).

Investigative Purpose

The purpose of this thesis is to undertake a systematic, scientific study of the issues concerning long-term digital document access and the possible alternatives available to the USAF as its information systems are upgraded. This study will concentrate on maintaining access to digital documents in their native formats without converting them to emerging digital format standards. This study is largely exploratory and prescriptive in nature because of the relative newness of this subject area and the lack of previous studies. The use of secondary data analysis techniques will be used to develop a model that can be used to recover and reproduce digital documents from their native file formats.

Scope of Research

The scope of this research will be based on a review of secondary data sources concerning the access of digital documents in their native formats as systems are upgraded. The sources will include the review of reports, academic and professional journals, government documents, case studies, electronic and digital sources, and the advice and opinions of digital preservation experts. This procedure has been chosen because: (1) this area of research is so new that it is necessary to explore the issues in order to develop hypotheses and questions for future research; and (2) it is "inefficient to discover anew through primary data collection or original research what has already been done (Cooper and Emory, 1995)."

Assumptions

Several assumptions have been made during the course of this research. The first assumption is that today's digital storage media may not be accessible in the future--maybe even in the near future (OASD, 1995; Dollar, 1992). Digital media are vulnerable to the problems of hardware obsolescence, and software and documentation loss that can render data unreadable even if the bit-streams remain preserved on the primary medium

(National Research Council, 1986). For example, current technology has overridden the need for most organizations to maintain punched card readers and the future likely holds the same fate for current CD-ROM, magneto-optical disks, and other storage technologies.

The second assumption is that the resource and financial burdens of converting an entire archival collection every 10 to 20 years is "likely to be out of the question except for relatively small collections that have great historical importance, sustain heavy use, or require rapid access (National Research Council, 1986)." Therefore, those digital documents requiring immediate continued access will be transferred to advanced storage media as information systems are upgraded (National Research Council, 1995; Rothenberg, 1995; Curle, 1993; Mohlhenrich, 1993; Michelson and Rothenberg, 1992; Willis, 1992). However, at some point in the future, space savings alone may result in the periodic transference of digital documents onto new media types (Lynn, 1994). Additionally, all digital data that requires long-term storage will be transferred to a new media before the superseded, aging media deteriorates to a point that its stored information is lost. An example that illustrates the necessity for this type of migration, is the large volume of weather and climate satellite information from the 1960s and 1970s that the United States has lost because the digital tapes where the data were stored have become unreadable (National Research Council, 1995).

The third and final assumption foresees that future information technology may not be compatible with current technology (OASD, 1995). For example, currently a personal computer (PC) that implements Pentium technology has the capability to execute software written for older 8086 and 8088 PCs. This compatibility is built into Pentium systems because there is a huge installed base in this technology that the public is not willing to part with until a reasonable return on investment has been recognized. However, at some point in the future, in order to take advantage of advanced

technologies, manufacturers will sacrifice some level(s) of compatibility and organizations will be forced to upgrade their information technology.

Significance of Research

A major potential benefit of this research is the cost savings to the USAF by eliminating the need to convert millions of digital documents on Air Force information systems to evolving digital file formats. This will result in continuous resource savings since standards change constantly and the costs to convert files to current formats may be incurred as often as every 10-20 years. Conversely, it will prevent the loss of many potentially valuable digital documents that become inaccessible because of advancing technology and/or the deterioration of the original media access equipment.

Another benefit of this research is the ability to continuously access a digital document once it has been created and stored. This access will be provided without the need to modify the digital document's original structure by a system administrator or user.

Furthermore, this research explores a method of accessing digital documents which will prevent the Air Force from being locked into a proprietary digital format. This is a requirement set forth in the *Automated Document Conversion Master Plan* (OASD, 1995).

Preview

The next chapter analyzes the existing literature exploring the issues surrounding long-term digital document access. Additionally, it will examine the different techniques that maybe used to access outdated digital document formats. In Chapter III, the research methodology will be discussed. Chapter IV will examine and analyze the data obtained. Finally, Chapter V will provide conclusions and additional recommendations for research.

II. Literature Review

Rapid Pace of Technological Advances

As information technology (IT) continues to evolve at a rapid pace, it becomes more important than ever that organizations devise methods that allow necessary long-term access to digital documents. An unfortunate consequence of the rapid evolution of digital technology is the short time span in which technology passes from state-of-the-art into obsolescence. For example, digital media may have the ability to store information for 100 or more years, but the technology to access it may become obsolete long before the media deteriorates (OASD 1995; Dollar, 1992; Willis, 1992).

Information Technology Research

Another distressing consequence of the fast pace of technological advancement is the lack of studies which can be performed and published on an existing technology before a newer and more advanced technology becomes available. Information technology is changing "so fast that by the time we suffer journal lead times and publish our descriptive science, new methodologies, technologies, and approaches are being practiced" by information management professionals (Cule and Grover, 1994).

Additionally, there are no pre-set guidelines to follow when evaluating new technology and systems based on that technology (Peterson, 1991). Unfortunately, "the technological catalyst of our field is moving too fast for us to hold it long enough in our grasp to study its impacts. And so we capture transient concepts, fads and describe organizational case experiences" (Cule and Grover, 1994).

A Growing Problem

Along with many other organizations, the Department of Defense (DoD) has recognized the benefits offered by digital documents and has established a policy leading

to a digital document environment. By using automated procedures to store and retrieve digital documents, the DoD anticipates that:

> documents in existing repositories could enhance the repositories' value beyond their original intended purposes by making those documents available throughout the DoD, thereby fostering increased access to important digital information. (OASD, 1995)

For the DoD and others, these digital documents can be used to move information, quickly and efficiently, to locations where value-added processes can be performed as necessary (Beatty, 1995).

However, as information systems are upgraded the ability to view digital documents in superseded formats becomes a problem due to the technological obsolescence of the hardware and software systems needed to access them. This is because most digital documents contain information that is only meaningful to the software and hardware systems that were used to create, edit, and access them (Rothenberg, 1995). These hardware and/or software dependent digital documents are the result of the information industry's failure to standardize the structure of digital documents. While digital document formats may become standardized in the future, a large volume of digital documents exist today in a variety of formats and organizations are continually creating more of these non-standard documents (Task Force on Archiving of Digital Information, 1996). Therefore, because of these non-standard digital document formats, organizations that archive digital documents must develop a method that will allow them to maintain continual access to digital documents in their native formats.

Information Technology Investment

Why maintain access to digital documents in their native formats? Because organizations do not need and can ill afford the additional burden of having to migrate digital documents to emerging formats every five to ten years in order to maintain access to their information. This type of migration expends an organization's financial,

physical, and human resources needlessly. Additionally, as organizational working budgets continue to shrink, organizations cannot afford the high costs of migrating/converting large volumes of superseded files to advanced information systems. Therefore, it is important that an effective method of accessing digital documents in their native formats is developed using minimal investment.

Strategies for Preserving Digital Documents

With technology advancing at such a rapid pace, the need to develop a method to maintain long-term access to digital documents increases with each new technological generation. There have been several strategies identified for maintaining access to digital documents. However, because of the relative newness of this subject area there is a lack of specific information on how to implement these strategies. For this reason, no formal methodology exists for maintaining long-term access to digital documents.

Charles Dollar (1992) identified two strategies in which technological obsolescence could be mitigated. The first strategy suggested that customers should demand that vendors provide cost-effective migration paths to advancing hardware and software systems. Many vendors already provide this capability in that an advanced version of their hardware or software system will provide the ability to migrate a customer's operations from the superseded system to the advanced system so the customer can continue operate with minimum interruption. However, this type of conversion is generally limited to the previous generation of the hardware or software system and therefore, it is a one-time fix that must be repeated with each successive system upgrade.

Additionally, the translation of a digital document into successive short-term standards over its life cycle may result in the loss of the document's original content (Rothenberg, 1995). Rothenberg's example of a digital document containing a treasure map depicted by a visual pattern of words and line spaces demonstrates the importance of

a document's content. Without the original document and the original software to accurately interpret the document, then the format and content of the document may be compromised and the original meaning lost. Furthermore, "old documents cannot always be translated into unprecedented forms in meaningful ways, and translating a current file back into a previous form is frequently impossible" (Rothenberg, 1995).

Another disadvantage of the migration strategy is that the resources consumed during the conversion of thousands, and perhaps millions, of software-dependent digital documents can be astronomical. An organization may not have the personnel nor the time necessary to migrate all documents to a new system. Additionally, this type of migration may tie a customer to a proprietary system.

In Dollar's second strategy he promoted a "trend toward non-proprietary standardized open systems environments, which are designed to overcome compatibility between computer systems and applications and are reflected in international standards" (Dollar, 1992; Rothenberg, 1995). These open system standards would make digital documents accessible through any software system that conform to the standards. While maintaining documents in a standardized format does not tie a customer to a proprietary system, there is still the problem that even the open system standards will change as information systems technologies continue to advance. Thus, over time, as hardware and software systems continue to evolve, it will still be necessary to either migrate digital documents to an updated standardized format or to provide some other method to maintain continual access to these documents.

In addition to Dollar's strategies, Jeff Rothenberg (1995) described a strategy that archivists have identified as a method for maintaining long-term access to the information contained within digital documents. The method entails extending the life of the original computer hardware and software systems on which the digital documents were created. These life-cycle extensions involve the operation and maintenance of antiquated

hardware systems and the archiving of the software needed to access digital documents in their native formats.

While maintaining a depository of antiquated hardware is achievable, it is also plagued with problems. The main drawbacks being the cost of operating multiple information systems and the difficulty in acquiring antiquated hardware system components (Kendall and Kendall, 1995; Rothenberg, 1995; National Research Council, 1986). These problems make it unrealistic to expect that any organization could effectively and efficiently maintain multiple, aging information systems in order to maintain access to superseded digital documents. An example of this problem was identified by the NRC (1986):

> Large on-line digital data storage systems have existed within the government since the 1960s. . . These data storage systems have a maximum useful life of 10 to 20 years, after which the system is no longer maintainable because parts and service are difficult to obtain and the system figuratively crumbles to dust. The important point of this revelation is that the data may still exist on the media but the equipment to retrieve the data does not exist.

To overcome the problems associated maintaining aging hardware, Rothenberg (1995) has suggested the creation and use of system emulators that can imitate the behavior of antiquated hardware systems. This method would allow the operation of superseded software on advanced systems as a way to view digital documents in their native formats. However, in order to emulate an antiquated information system this method requires exhaustive specifications on the original system's hardware (Rothenberg, 1995). Therefore, this method may require extensive participation by hardware manufacturers. Many manufacturers may be reluctant to supply all of the specifications to software developers because some of the technology may still be in use in advanced systems they have developed. This method would also require the archiving of superseded software necessary to access digital documents in their native formats.

Additionally, this method would require an organization to maintain an extensive training program in order to educate their personnel on the many different software programs utilized.

The use of algorithms to emulate software and hardware was also a recommendation of the Task Force on Archiving of Digital Information (TFADI) to the Commission on Preservation and Access (CPA) and the Research Libraries Group (RLG). In the TFADI's report (1996), *Preserving Digital Information*, the task force stated that there was a need to:

> Foster practical experiments or demonstration projects in the archival application of technologies and services, such as hardware and software emulation algorithms, transaction systems for property rights and authentication mechanisms, which promise to facilitate the preservation of the cultural record in digital form.

Another strategy that can be used to maintain long-term access to digital documents is to interpret the documents by using a description of how the original software interpreted its data files (Michelson and Rothenberg, 1992). This method would allow access to digital documents without the need to maintain depositories of superseded software and hardware systems. However, it is also difficult for computer scientists to describe the ways in which complex software performs different functions (Rothenberg, 1995, Michelson and Rothenberg, 1992). Additionally, as with emulation, this method requires exhaustive specifications on the original software systems used to create and interpret digital documents and so, this method requires extensive participation by software manufacturers. Unfortunately, because of incomplete documentation and other factors, a complete description of how a software-dependent digital document is interpreted may only be contained within the proprietary software that it is dependent upon (Michelson and Rothenberg, 1992).

Which Strategy to Use?

Several strategies for migrating and maintaining digital documents have been identified. However, no single strategy can be applied to all digital document formats and none of the identified strategies is entirely satisfactory (TFADI, 1996). Therefore, as information systems and their operating environments continue to evolve it may be necessary to use several of these strategies in order to maintain access to digital documents in superseded formats. The strategies chosen will need to evolve from organizational requirements and conform to the limits of its financial, physical, and human resources (Peterson, 1991).

Unfortunately, the National Archives and Records Administration (NARA) does not have the capability nor the expertise to handle the many forms of digital data that are being collected by governmental agencies (NRC, 1995; NRC, 1995a; National Academy of Public Administration, 1989). However, it has been stated that the NARA must take the lead in developing a long-term strategic plan for maintaining access to digital documents (National Academy of Public Administration, 1989).

Because a long-term strategic plan may call for a conglomerate of the methods mentioned here, it is conceivable that no existing organization can afford the financial, physical, and human resources necessary to carry out such a tremendous task. Therefore, it may be necessary to establish organizations or processing centers that specialize in maintaining long-term access to digital documents (TFADI, 1996; NRC, 1995; NRC, 1995a).

Summary

Maintaining long-term access to digital documents is a relatively new field of study. To maintain continued access to digital documents several methods have been identified in existing literature:

1. Vendors providing cost-effective migration paths to advanced hardware and software systems,
2. The development of non-proprietary open system digital document standards,
3. Maintaining depositories of antiquated hardware and software systems,
4. Creating and using system emulators that imitate the behavior of antiquated hardware and software systems, and
5. Interpreting documents by using a description of how the original software interprets its data files

Each of these methods are achievable but because this field is so new, very little scientific and academic study has actually been completed within any one of these areas. Because of this there is a lack of specific information on how to implement these strategies.

III. Methodology

As discussed in the literature review, the ability to maintain long-term access to digital documents has been identified as a serious problem facing organizations that archive digital materials. Furthermore, many pieces of the puzzle have been suggested to deal with this problem but no overall plan has yet been developed. Additionally, there is little research that has been accomplished and published concerning the complete nature of this subject. Because of the lack of current research, a recent report commissioned by the Commission on Preservation and Access and the Research Libraries Group stated the need to "[f]oster practical experiments or demonstration projects in the archival application of technologies and services, . . . which promise to facilitate the preservation of the cultural record in digital form" (TFADI, 1996).

In response to this recommendation, a conceptual model will be designed for retaining long-term access to digital documents in their native formats. The model being developed will synthesize techniques suggested by other researchers along with other principles and ideas that were encountered or developed during the course of this research project. These techniques, principles, and ideas will be integrated through a series of processes and, upon completion, will provide a general model for maintaining long-term access to digital documents. This model can be implemented as a demonstration project for organizations which archive digital documents.

IV. Model Development

The Rosetta Stone

At some point during the fourth century, all knowledge of Egyptian scripts was lost leaving no method available to decipher the language of hieroglyphics which had been richly preserved on ancient Egyptian monuments, stone tablets, and sheets of papyrus. Fortunately, while on an expedition in Egypt in 1799 CE, Napoleon's army discovered an artifact which has become known as the Rosetta Stone. The Rosetta Stone contained the inscription of a decree issued in 196 BC by Ptolemy V Epiphanes. The decree was repeated three times in two languages, Egyptian and Greek, with the Egyptian version appearing twice. The Egyptian version was written once in hieroglyphics and once in demotic, a cursive form of the hieroglyphic script. Fortunately, there is an abundance of information on ancient Greek dialects and therefore, the stone's Greek version of the decree contained the key to decipher the meaning of the ancient Egyptian texts. Today, because of the Rosetta Stone, we can interpret many ancient texts and inscriptions of Egyptian hieroglyphic and demotic scripts found on sheets of papyrus and monuments throughout Egypt.

The meanings of many ancient scripts have been lost over the centuries. Fortunately, archeological discoveries, such as the Rosetta Stone, are providing modern civilizations with the necessary means to translate ancient manuscripts. This is because ancient cultures preserved their histories the only way they knew how--using written scripts.

Digital documents do not have the same visual properties associated with them as ancient clay tablets, papyrus, and other forms of documentation. In an article by Jeff Rothenberg (1995) he showed a picture of the Rosetta Stone and stated "Besides being legible after 22 centuries, the Rosetta Stone owes its preservation to the visual impact of

its content--an attribute absent in digital media." In other words, ancient text maintain their original form simply because they have been physically preserved in a visual form. Unfortunately, digital documents do not share this characteristic and cannot be physically read by humans without the necessary digital equipment to interpret them. It is simply impossible to read the contents of a harddrive, CD-ROM, or any other type of digital media without the proper equipment to access the media, interpret its bit streams, and display its contents.

Digital History and Digital Knowledge

A major obstacle to preserving our digital history is not just a shortage of money, but also a shortage of knowledge (Conway, 1996; Darling 1981). Many of the issues for preserving our written history have been dealt with in the past simply because humans have been writing things for thousands of years. However, digital documents and digital preservation are relatively new phenomenons in the course of human history. Digital technology as we know it has existed since the 1940s and only now is the need to preserve our digital history becoming a critical factor because of its impending loss. The foundation to maintaining our digital history is the preservation of digital knowledge. This relationship is depicted in Figure 1.

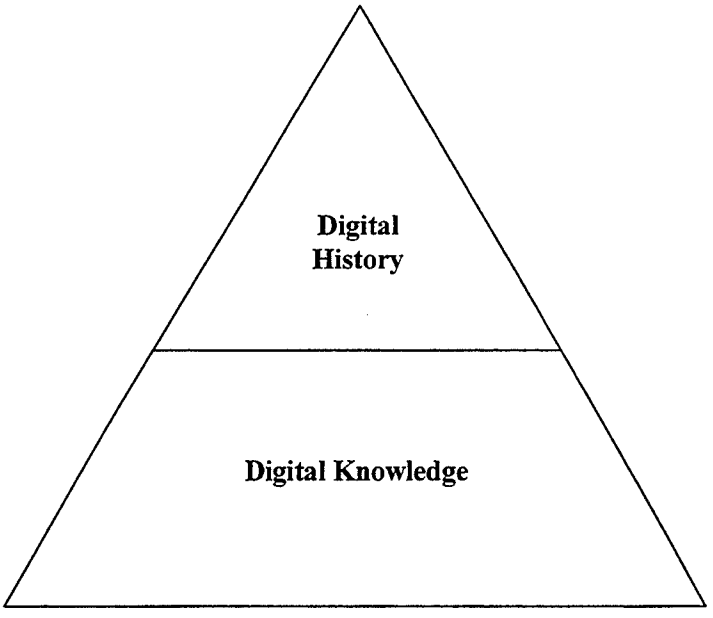


Figure 1. Digital Knowledge is the Foundation for Digital History

A Digital Rosetta Stone

To prevent the loss of our digital history, I propose that digital knowledge be preserved in a manner that I call the Digital Rosetta Stone (DRS). The data so preserved would be a collection of the knowledge and processes necessary to recover and reconstruct digital documents maintained in their original file formats. The data would be used to create or emulate the hardware and software necessary to recover data from obsolete storage media and reconstruct the digital documents.

Rothenberg (1995) stated that if the behavior of an information system could be sufficiently described, then future generations could re-create that behavior and reproduce digital documents without the need for the original systems. However, he also said that currently, information science cannot sufficiently describe this type of behavior in a way that will allow this strategy to succeed. One way to describe and preserve the behavior of information systems for our posterity is to create a DRS that can be used to reconstruct digital documents.

The processes and metadata maintained by the DRS will catalogue the many different aspects of digital technologies. After all, "in the digital world, preservation must be concerned with entire technology systems, not one or another component, such as a film or a storage disk" (Conway, 1996). In digital equipment each component is dependent upon other components of the digital systems in order to perform a specific task. A simplified example of this interdependence can be demonstrated by the process of viewing a file created by a word processor. The file must be interpreted by the application program which is dependent upon the operating system which is further dependent upon the system's hardware. Each layer of digital technology involved in this process contributes some form of information necessary to view the digital document.

DRS Components

Unfortunately, creating a DRS is not as simple as the creation of the original Rosetta Stone that held the key to Egyptian hieroglyphics. Instead, 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. The knowledge preservation process supports the data recovery and document reconstruction processes. These processes are depicted in Figure 2.

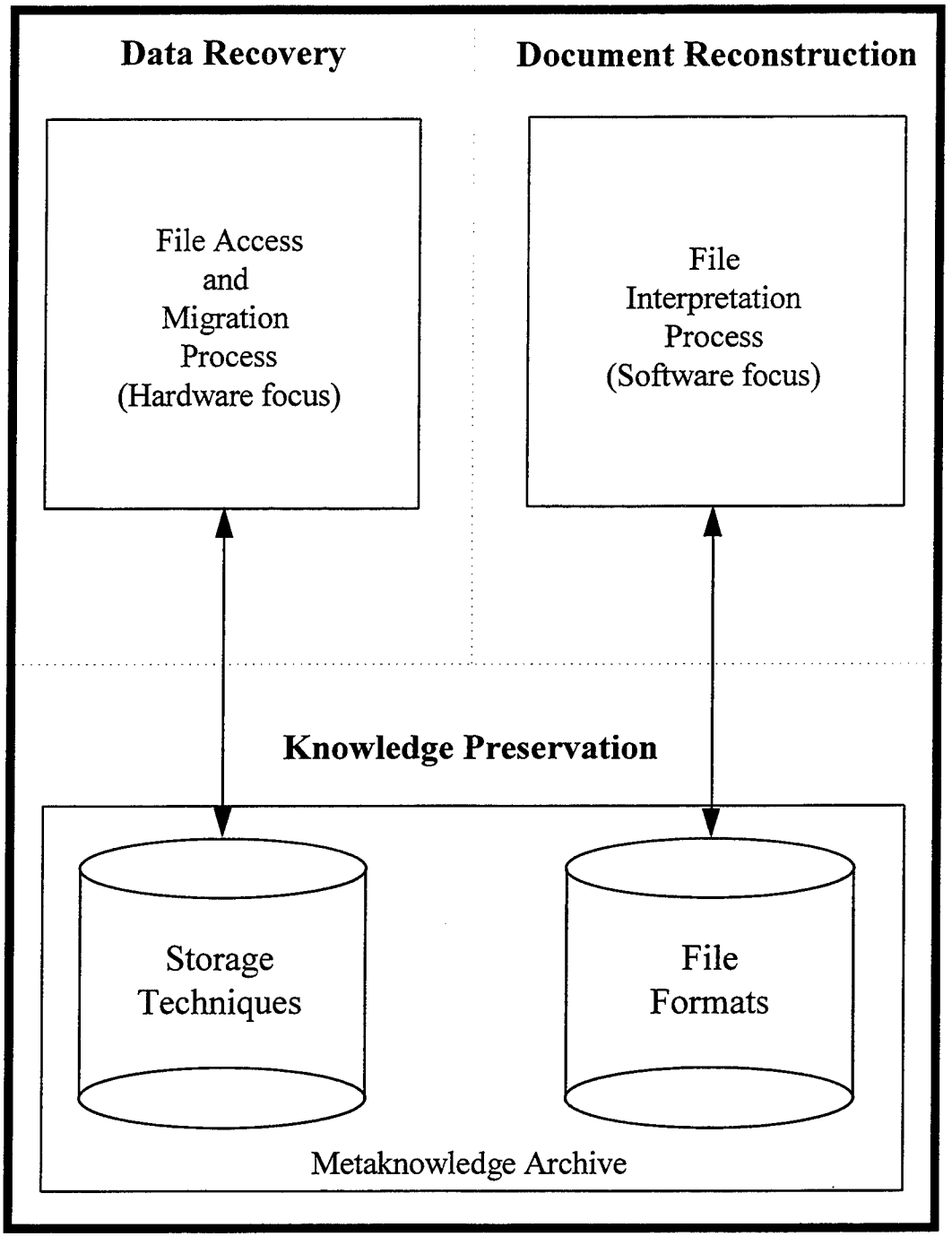


Figure 2. Digital Rosetta Stone Processes

Knowledge Preservation. Knowledge preservation is the process of gathering and preserving the vast amounts of knowledge needed to recover digital data from a superseded media and to reconstruct digital documents from their original formats. In a

DRS, the preservation of knowledge of media storage techniques and file formats will be maintained in a metaknowledge archive. Metaknowledge is the knowledge or awareness of facts, heuristics, and rules, and the context in which they are used and manipulated (Mockler and Dologite, 1992). The creation of standardized data dictionaries will be the tools used to store the metaknowledge necessary to aid document recovery personnel. The data dictionaries will contain the names and descriptions of the data items and processes necessary to recover a digital document (Martin, 1989). The metaknowledge archive (MKA) is the foundation upon which the DRS is dependent and it must extensively preserve the knowledge in two key areas--media storage techniques and file formats.

The knowledge of media storage techniques is a collection of the way data are defined and stored on specific media. While it is expected that some data will be migrated to new storage devices for archival purposes, it is likely that some data will not be migrated. Therefore, it is necessary to maintain a record of the methods in which bit patterns are used to represent data on storage devices. The knowledge of the location and meaning of these bit patterns will be necessary to recover data if equipment to access a storage media is not available or no longer exist. This is not to say that all specifications for storage devices must be accurately preserved so engineers can manufacture them in the future. Instead, it means that only the techniques in which the bit patterns are stored and accessed on the media needs to be preserved. After all, the purpose is to read the data and migrate it to a new storage device. Not read it, manipulate it, and write it back to the media. Storage techniques knowledge will be stored in the metaknowledge archive. This will be further discussed in the knowledge preservation section.

Just as the knowledge of the techniques used to store data on a digital media must be preserved, so must the information on file formats be collected on data files created using different software applications. The knowledge of file formats is a collection of the

techniques used by specific software applications to define formatting operations within digital documents. Software applications that create digital documents use data located in specific positions and predefined character sequences to define the digital document's appearance. Interpretation software is necessary to view a digital document whether it is simply stored in an ASCII text format or in a complex database format. Software products, commercial-off-the-shelf (COTS) and Non-COTS, store digital data using a variety of techniques. Therefore, every data file is dependent upon some form of software to properly interpret and display the data file's contents. Character sequences embedded within a digital document inform the interpretation software how the document's data is to be interpreted. For example, in order to bold a section of text using the Hypertext Markup Language (HTML), all characters following the character sequence "<B>" are bolded until the character sequence "</B>" is encountered. Any software capable of interpreting an HTML document must recognize these character sequences and all other format character sequences that are characteristic in HTML documents. Likewise, any software capable of interpreting a digital document must recognize the formatting character sequences unique to the application that was used to create that digital document. File formats knowledge will also be stored in a metaknowledge archive. This will be further discussed in the knowledge preservation section.

Data Recovery. Data recovery is the process of extracting digital data from an obsolete media and migrating it to a media that is accessible to current information systems. As stated earlier, it is hoped that data on superseded media will be migrated to advancing media technologies for archival purposes so that digital document bit streams are always readily accessible. However, over time it is likely that some data will not be migrated to advanced storage media. Therefore, it is necessary to maintain a process by which data can be recovered from obsolete media. The recovery will, of course, depend on the cost effectiveness of recovering the data. That is, if the need for the knowledge in

the digital document(s) is greater than the cost of recovery, then the cost of the recovery method(s) may be justified.

Document Reconstruction. Document reconstruction is the process of interpreting digital documents from their original data files by using file format information gathered during the knowledge preservation process. Interpreting digital documents by describing how the original software interpreted the documents is a strategy that was suggested by Michelson and Rothenberg (1992). The file format information describes the formatting information used by specific software applications. In other words it is a template that can be used to describe the way data is formatted and displayed by word processing, graphics, and other applications that create digital documents. This does not mean that the algorithms used to produce the documents are preserved so programmers can replicate them in the future. Instead, it means that the bit or character sequences and other formatting information are preserved as a template for document interpreters to use to reconstruct and view documents in their original forms. When the reconstruction process is complete the document should appear in its original form. As in the data recovery process, the methods used during document reconstruction are dependent upon the cost effectiveness of reconstructing the document.

Knowledge Preservation

The Metaknowledge Archive. The metaknowledge archive is the foundation upon which the DRS is built. It contains templates which can be used to extract and display data in the form prescribed by the information systems used to create digital documents. To insure the success of the DRS the metaknowledge archive must develop a standardized format to preserve media storage techniques so engineers can extract data from the many different types of media. Likewise, it must also develop a standardized format to preserve digital document formatting information for the different types of digital documents that may need to be recovered.

As long as there is a template that can be used to interpret a document, then a document can be displayed in its original form. No matter which software or hardware system is used to create a digital document the end product is the same--a document that contains text, graphics, and various other objects that communicates information to humans. After the creation of a digital document, its interpretation is dependent upon the hardware and software systems that were used to create it. However, most modern computer systems have the ability to process and display the multitude of objects that appear in digital documents. Therefore, on any given hardware system routines can be designed to interpret and present the contents of digital documents that were created on another system (even if the systems themselves are incompatible).

Media Storage Techniques Metadata. Media metadata is probably the easiest type of data to gather for the DRS. This is because the standards for most storage media are rigidly defined before a media is brought to market. For example, ISO9660 is the standard that specifies how data are stored on a CD-ROM. This standard defines the volume structures, file structures, and all other attributes associated with a CD-ROM. This type of data must be gathered for each type of media to be included in the metaknowledge archive.

The first step in gathering media storage techniques metadata is to identify all of the storage media which may need to be accessed in the future. Once a media has been identified it is necessary to collect and maintain information on its data storage geometry, storage method, encoding scheme, and file allocation method.

When trying to recover data the first thing that data recovery personnel must know is where to look in order to find it. Media storage geometry defines where on a media data are stored. In order for data recovery personnel to find the data they must know the geometric shape of the data's path and the locations of those paths. For example, on a CD-ROM data are stored on spiraling tracks that are spaced at 1.6

micrometers apart for a track density of 16,000 tracks per inch (Norton and others, 1995). Furthermore, the tracks are divided into sectors containing 2048 bytes of data and each sector has an address that is used during the file allocation. This type of geometric storage information must be collected for each type of media.

After the media's storage geometry has been identified, data recovery personnel must know the method used to store the data. The data storage method is how data are physically recorded on a media and this information must be known so a device can be engineered to read the digital patterns. In the past data have been stored on media using a variety of methods. Early storage media stored data as a series of holes punched into lengths of paper tape or punched cards. Hard and floppy disks store data as a series of magnetic patterns stored on a layer of magnetic particles. More recent optical technologies, such as the CD-ROM, store data as a series of lands and pits (0.12 micrometers deep and 0.6 micrometers in diameter) burned into a plastic media. There are many other storage methods that have been used, that are in use, and that will be used in the future. Knowing these storage methods tells data recovery personnel what to look for to identify the digital data stored on the media.

After data recovery personnel have identified where the data are stored and the data storage method they must determine how the data are encoded. Encoding techniques define how the data's bit patterns are stored on the media. The encoding information will be used to decode the data and restore the data bit stream to its original form. Encoding schemes may be fairly simple with one setting identifying a 0 bit and another setting defining a 1 bit. Or encoding schemes may implement coding algorithms to encrypt and compress recurring bit patterns. Two popular encoding schemes used today are multiple frequency modulation (MFM) and run length limited (RLL). Multiple frequency modulation is a method of encoding analog signals into magnetic pulses or bits. Run

length limited is another method of encoding data into magnetic pulses but its encoding scheme allows 50 percent more data to be stored on a disk than MFM.

During the next step it is necessary to determine the file allocation method used on a media. File allocation is how storage space is assigned to files so that storage space is effectively utilized and files can be accessed (Silberschatz and Peterson, 1988). Once data recovery personnel can locate, read, and decode the information on a media , they must know the file allocation method in order to properly reassemble the files. Descriptions on items such as volume and file structures are identified in media standards, such as ISO9660 for the CD-ROM. The operating system also controls a media's file allocation method and therefore, it is necessary to access operating system specifications to gather data on file allocation methods. There are several file allocation methods in use and each operating system and media combination uses a specific allocation method. Examples of some popular allocation schemes are the contiguous, linked, and indexed allocation methods. The contiguous allocation method requires each file to occupy a set of contiguous addresses on a disk. With linked allocation each file is a linked list of sectors and the sectors may be scattered anywhere on the disk, and with the indexed allocation method each file has its own index block which is an array of disk block addresses (Silberschatz and Peterson, 1988). The allocation method may also provide other valuable information such as distinguishing between the locations of data bytes and error detection/correction bytes.

Collecting and maintaining metadata on these four entities, data storage geometry, storage methods, encoding schemes, and file allocation methods will provide the keys to recover data once an access system is no longer available to access that media type. This metadata is used to support the data recovery process represented in Figure X. As hardware and software systems become obsolete this metadata is used to develop hardware and software systems to recover data and migrate it to currently accessible

storage media. It is necessary to state again that the intent of this information is to aid in the extraction of the data stored on the media; not to reproduce the original equipment used to read and write the media. There is no need to engineer a device to write data to the media because there is no desire to change the data. The need is only to read data from an obsolete storage media and migrate it to a currently accessible storage medium.

File Formats Metadata. The first step in gathering file format information is to identify all of the applications used to create the digital documents which may need to be reconstructed in the future. This includes both commercial-off-the-shelf (COTS) and non-COTS applications. Gathering and cataloging metadata to reconstruct digital documents created with COTS and non-COTS applications is going to be a time intensive and difficult task. However, it is necessary because many organizations use these applications to create and store digital documents. It is important to note that many systems implemented within the USAF and other DoD agencies have been developed because of special needs. These systems produce a vast number of documents that contain valuable information. Archives may be required to maintain these documents because of the value of their contents. Therefore, archived documents created by non-COTS applications require the cataloguing of their metadata so that they can be properly interpreted in the future.

The second step is to identify and catalog the objects that are supported by these applications. An *object* is a structure that encapsulates attributes and methods that operate on those attributes. An *object class* is a logical grouping of objects that have the same or similar attributes and behavior (McFadden and Hoffer, 1994). An object in a digital document can be text, graphics, audio, video, and any number of other structures that have been included by the document's creator. It is necessary for an interpreter to have the ability to identify the objects embedded in a digital document before the

interpretation process begins. If an object is not properly identified then the document is uninterpretable. The process of identifying data objects is demonstrated in Figure 3.

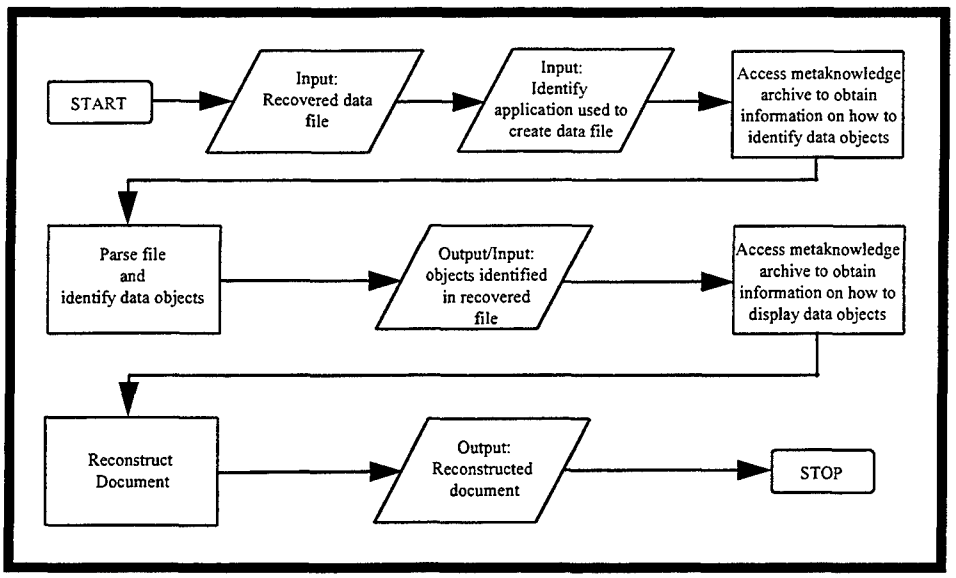


Figure 3. Object Identification Process

Once the objects are identified, interpretation routines are created to present these objects in their original form on the current information system. Since objects are utilized over and over again by different applications, it is only necessary to create a routine to interpret and display that object once. A routine can be used to display an object regardless of the application used to create the digital document. For instance, most digital documents support the use of text objects. Since text is used in multiple applications, it is only necessary to create a routine to handle a text object once. That routine can then be used to interpret and display text on the current system regardless of the software and hardware systems that were used to create the original document.

The final step is to identify and catalog the formatting structures implemented within each application. These formatting structures describe how objects are identified, formatted, and arranged within a digital document. Additionally, this information describes how to determine such things as page size, margins, line spacing, tabs, fonts,

footnotes, and a multitude of other page layout information. This information must be maintained in a standardized form so that an interpreter can easily access it and switch between digital documents that were created by different applications. Fortunately, modern database technology provides the ability to create lookup tables that can be used to map formatting techniques across different applications.

The formatting process may be made more difficult because there is no standardized way in which applications store formatting information. Applications disperse formatting information (1) throughout the document, (2) in designated locations within the document, or (3) in combinations of 1 and 2. Additionally, some applications store document files in an ASCII format while others opt for a binary format. Defining a standardized method to describe these currently non-standardized procedures is one of the goals of the DRS metaknowledge archive.

Data Recovery

Once it is no longer economically feasible to maintain antiquated hardware systems, it is necessary to implement an alternate method in order to maintain the ability to recover data from superseded media. If data are stored on an obsolete media that is not accessible by current systems then the data must be migrated to a currently accessible media before document reconstruction can begin. That is, the data must be recovered.

Data recovery involves the retrieval of the storage techniques information gathered during the knowledge preservation process and using it to recover data from an obsolete media. This information is used to modify or construct the equipment needed to migrate digital data from an obsolete media to a currently accessible media.

An example, of this usage can be depicted by data stored on punched cards. Punched card pass through a punched card reader at the rate of approximately 1,000 cards per minute. As the cards pass between a light source and a row of solar cells the location of the holes are detected and the pattern is transformed into electric signals which are sent

to the computer and translated into machine instructions (Downing and Covington, 1986). Because of advances in storage technologies, punched cards are seldom used as a storage media because they are slow, bulky, and cumbersome compared to modern storage media. Additionally, few organizations maintain punched card readers because it is an inferior and superseded technology. So, if a stack of punched cards were to be found and there were no punched card readers available to read the data, how could the data be read? First, the punched card storage techniques information that was gathered during the knowledge preservation process is retrieved. Once the information is analyzed and engineers understand the way information is stored on a punched card, they may find that it is a simple task to reprogram a modern scanning device, such as those used in supermarkets or on assembly lines, to read the patterns of holes on a punched card. Therefore, a device can be modified to read, translate, and migrate the data on punched cards to a modern storage device without the need for an original punched card reader. There is no need to engineer a device to write punch cards because there is no desire to change the data. The need is only to read the data and migrate it to a currently accessible storage media.

While this is a relatively simple example of how the storage techniques information can be used, it demonstrates how easily yesterday's digital technologies can be more easily reproduced using today's digital technologies. Likewise, this same method could be used to manufacture readers for paper tapes, CD-ROMs, and other storage devices. If someone finds a CD-ROM disk in the year 2222, perhaps he or she will be able to take it to a DRS processing center to recover the data. Instead of building a CD-ROM drive, the processing center may simply use a high-tech scanner to scan the disk and identify the patterns of lands and pits burned into the disk's surface. Using the data gathered about CD-ROM storage techniques (ISO9660 standards) during the knowledge preservation process, an information system analyzes the location and patterns

of lands and pits, identifies the file allocation system, processes the data, and then writes the files to a twenty-third century storage device. See Figure 4 for a visual representation of this process. Once the files are recovered, they are ready for document reconstruction.

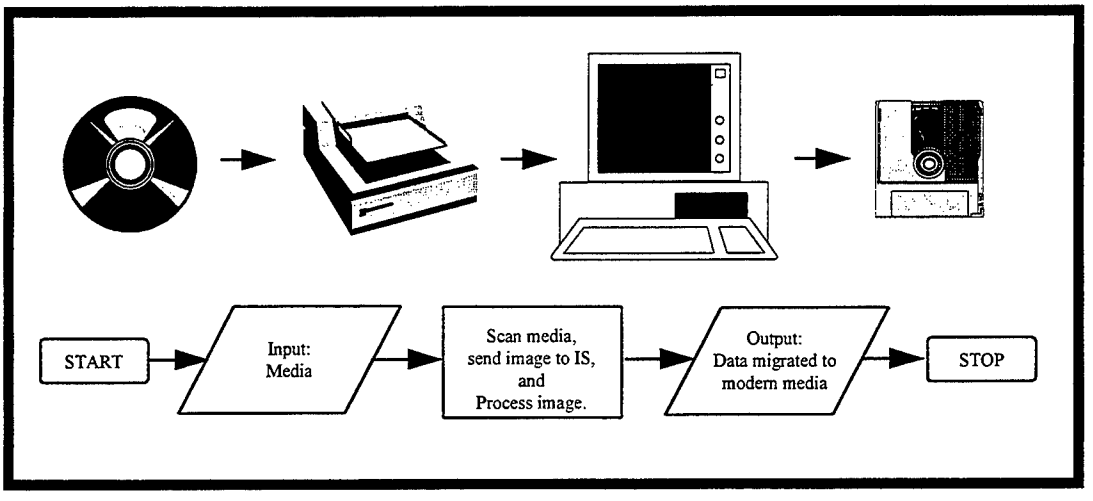


Figure 4. Data Recovery Process

Document Reconstruction

If digital documents are stored in superseded formats then they must go through an interpretation process in order to restore them to their original forms. That is, the documents must be reconstructed. Reconstruction is accomplished by document interpreters. Document interpreters are either (1) trained technicians or (2) software applications that use file formatting information to reconstruct digital documents.

The DRS relies upon the file format descriptions gathered during the knowledge preservation stage to describe how the original software interpreted files. These file format descriptions identify the information, such as character sequences (and their locations if they are position sensitive), that identify data objects and specify formatting operations within a digital document.

Table 1 contains examples of the character sequences used by three different applications to perform **bolding** operations on text.

Table 1. Example of Character Sequences for Bolding Text

Software Application	**Begin Bold** (in Hexadecimal)	**End Bold** (in Hexadecimal)
Wordstar® 6.0	02	02
Ami Pro® 3.1	3C 2B 21 3E	3C 2D 21 3E
HyperText Markup Language	3C 42 3E	3C 2F 42 3E

Therefore, when an interpreter is reconstructing an Ami Pro® 3.1 document, the character sequence (hexadecimal values) "3C 2B 21 3E" specifies to the interpreter that all characters following this sequence need to be bolded. Likewise, the character sequence (hexadecimal values) "3C 2D 21 3E" signals the interpreter to stop the bolding process.

This is a simplified view of how file format information can be used, but it demonstrates the types of information that need to be collected and stored to aid document interpreters in the reconstruction of all types of digital documents. In addition to identifying text-based objects and operations, character sequences are used to identify other objects imbedded within digital documents.

The Digital Rosetta Stone Model

The DRS model can be represented in three stages. The first stage of the model represents the knowledge preservation process. This is the foundation upon which the DRS is dependent. During this process the data needed to support the data recovery and document reconstruction processes is gathered and stored in the metaknowledge archive.

The second stage of the model is the data recovery process. The data recovery processes uses the knowledge of storage techniques to extract a digital document's bit stream from an obsolete storage device and then migrates the bit stream to a currently accessible storage device.

Once a digital document's bit stream has been recovered the bit stream is advanced to third stage. The third stage of the model is the file reconstruction process. The document reconstruction process uses the knowledge of file formats to interpret the bit stream and display the document in its original form. Upon completion of the reconstruction process, the final product is a reconstructed digital document that appears in its original form. The complete DRS model is depicted in Figure 5.

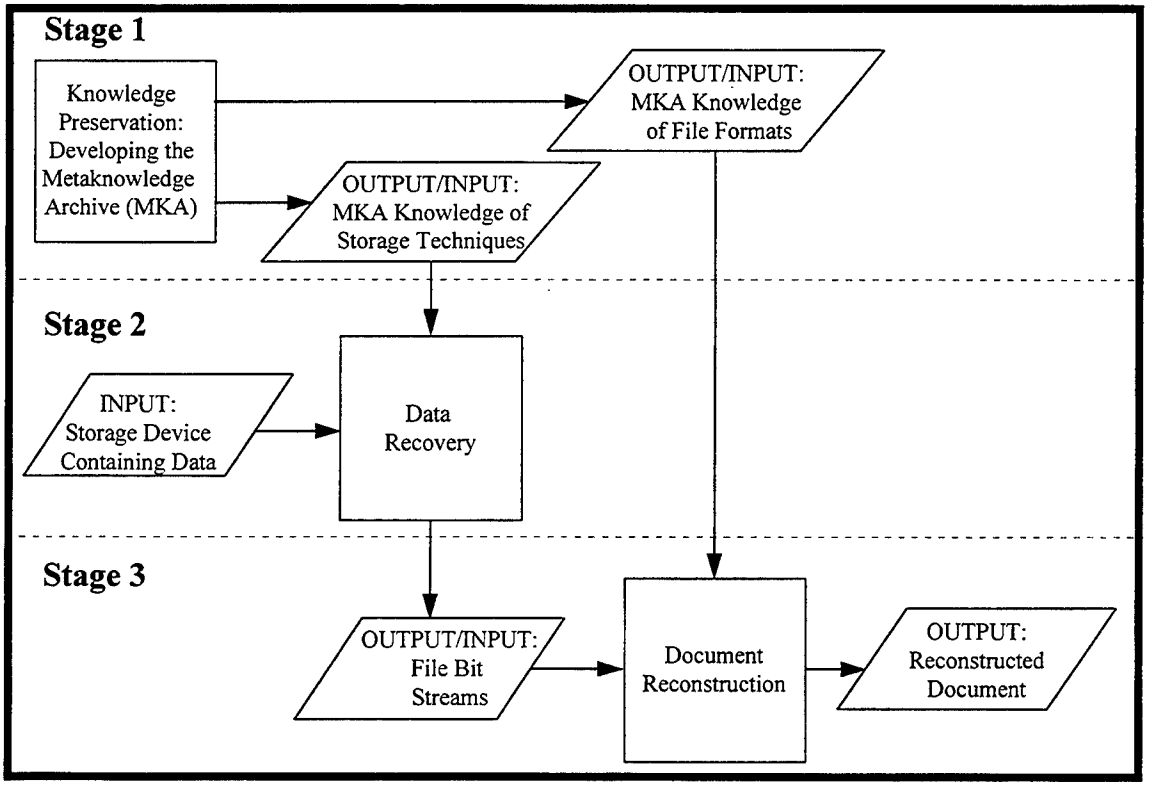


Figure 5. Digital Rosetta Stone Model

Metadata Mapping Example. Because of the complexity of many objects, such as graphics, video, and audio, the examples used here will deal with text objects for the sake of simplicity. However, the theory behind the methods used to interpret textual objects also applies to the more complex objects--only on a grander scale.

A common object supported by most applications is text. Text objects are generally represented in a binary character set (BCS). A byte is a group of consecutive

bits that is treated as a single unit or character that has a constant meaning for a specific group of people. In a BCS each character is represented by a binary code. For example, in 8-bit American Standard Code for Information Interchange (ASCII) the character "A" is represented by the binary bit pattern of 01000001.

The bit patterns of a BCS represent alphabetic characters, numeric characters, special characters, and control characters. Alphabetic characters are the letters 'A' to 'Z' and may be represented in both an upper case (A..Z) and a lower case (a..z). Numeric characters are the digits 0 to 9. Special characters are other printable characters such as punctuation marks, foreign language characters, and mathematics and graphics symbols. Control characters are codes that represent non-printable characters that specify actions such as a line return or a backspace. These control characters are also used to perform specialized operations during word-processing, telecommunications, and other information system operations. For example, during word-processing, control characters may be used to specify document formatting operations, such as centering, bolding, and italicizing. Likewise, telecommunications programs use control characters to perform synchronizing, transmitting, and acknowledgment functions.

Tracking information codes may seem like a trivial matter. However, as information systems have evolved so have binary character sets. Since the beginning of the modern computer age there have been several different BCSs. The American Standard Code for Information Interchange has undergone several evolutionary changes to reach the 8-bit form it has today. In addition to the ASCII BCS, IBM's extended binary coded decimal information code (EBCDIC) is also in use today. A BCS known as UNICODE is currently under development in 16-, 31-, and 32-bit versions. UNICODE will have the capability to store all of the characters from all known languages since the dawn of recorded history. If a version of UNICODE is standardized and adopted

throughout the information industry then even the 8-bit ASCII character set will be superseded.

Fortunately, binary character sets are defined and maintained by standards organizations such as the American National Standards Institute (ANSI). This makes it easy to gather data on standardized BCSs. However, while character sets have been defined and standardized, many applications implement modified BCSs for various reasons. Therefore, it is necessary to track both standard and non-standard BCSs. For example, Wordstar® used 7-bit ASCII with 8-bit bytes. The eighth bit was used to store formatting information.

Modern database technology makes it simple to build a table that maps the values of a character in one BCS to the value of that same character in another BCS. The main attributes that need to be tracked are the BCS's byte length, bit code, and symbolic meaning. An example of a table that maps symbolic coding across different BCSs is found in Table 2.

Table 2. Example of a Binary Character Set Mapping Table

Symbol	7-bit ASCII	8-bit ASCII	8-bit EBCDIC
A	1000001	01000001	11000001
B	1000010	01000010	11000010
. . .	. . .	. . .	. . .
1	0110001	00110001	11110001
2	0110010	00110010	11110010
. . .	. . .	. . .	. . .
#	0100011	00100011	01111011
@	1000000	01000000	01111100
. . .	. . .	. . .	. . .

After a data file has been recovered, an interpreter can use a BCS mapping table to convert text-based characters from a recovered data file to characters that are

understood by the current information system. The BCS interpretation process is depicted in Figure 6.

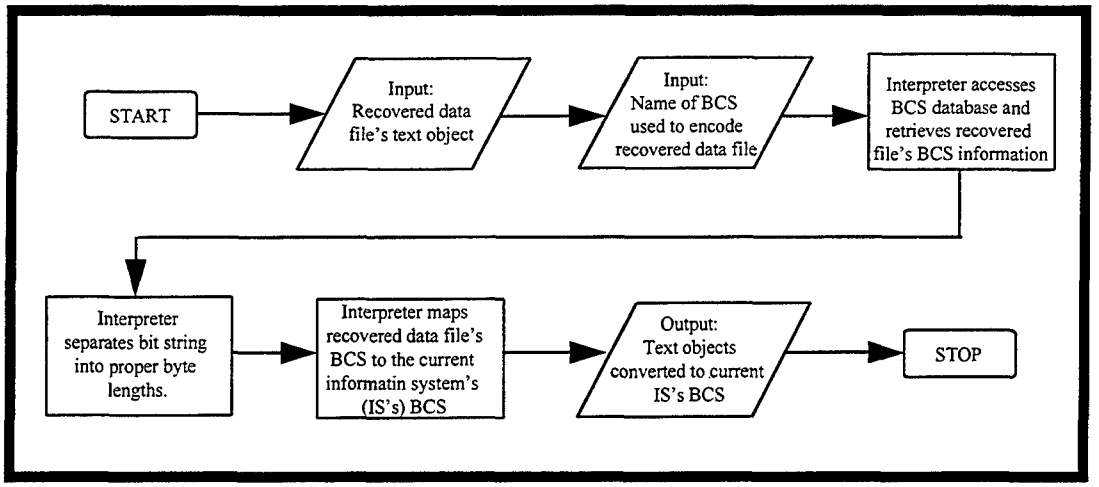


Figure 6. Binary Character Set Interpretation Process

Once text is identified an interpretation routine must be called to present it in its original form. However, before the text can be displayed or printed there are some other factors that must be addressed--text font and font size. In the metaknowledge archive other tables (similar to tables 1 and 2) representing how software-dependent digital documents indicate font and font size must exist in order to present the information in its original form. These tables are used to map a digital documents font and font size information to the drivers needed to properly display the text. These drivers are then used by the interpretation routine to present the text in its proper font and size.

The interpretation process of a text object is described in Figure 7.

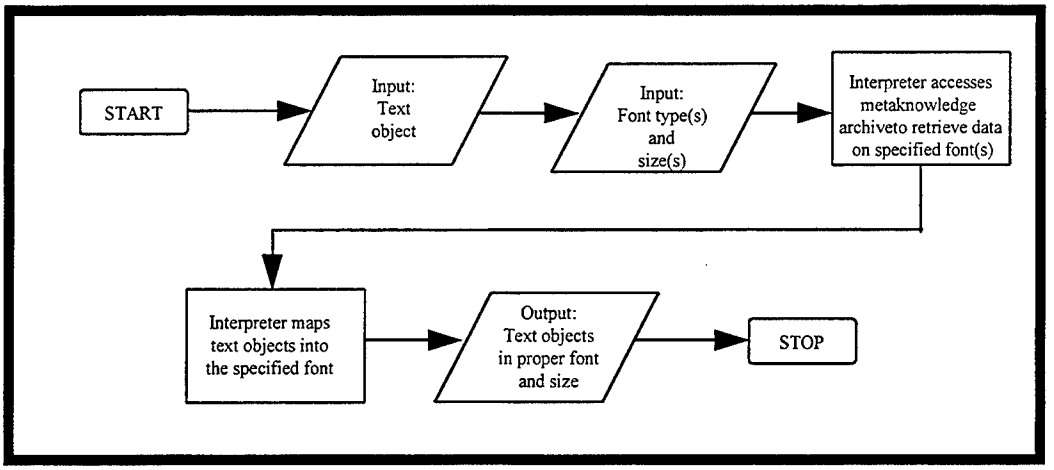


Figure 7. Text Interpretation Process

Object Identification Mapping. The same method used to map text objects can also be used to map other objects. Tables similar to those represented in Tables 1 and 2 can be used to map object identification codes across applications. Additionally, these tables can link these objects to the interpretation routines that present the information in the same form as the system used to create the original document. However, it is still necessary to develop a standardized coding scheme to be used in the mapping tables so document interpreters can reconstruct documents using this data.

DRS Demonstration

The theory behind the Digital Rosetta Stone (DRS) can be demonstrated using an 8-track punched paper tape (8-TPPT). The 8-TPPT technology was widely used during the 1960s and 1970s. This technology was developed before industry standards were the norm and therefore, this technology is primarily proprietary. Finding information on the 8-TPPT coding scheme was very difficult. While doing research for this thesis, I contacted the technical support and archive sections of International Business Machines (IBM) to get some information on 8-TPPT equipment. Unfortunately, I was told that IBM no longer supported this technology and does not maintain any information in its archives on it. However, some functional 8-TPPT readers still exist.

After being unable to locate a listing of the character coding scheme, several aging data processing books were consulted to find the information. While much of the coding scheme was obtained from these books, the set is far from complete. The books used to compile this information were written by Awad (1971), Nashelsky (1972), Langenbach (1968), and Williams (1965). All of the information concerning the 8-TPPT used in this example was compiled from these sources.

The 8-TPPT stores data sequentially along the length of the tape. Individual characters are stored vertically on the tape in eight channels. The eight channels represent seven data channels and one check (or parity) channel. From the least significant bit to the most significant bit these channels are identified as 1, 2, 4, 8, Check, "O", "X", and the End of Line (EL). An example of 8-TPPT can be seen in Figure 8. Notice that unlike today, the check bit is not the most significant bit, but instead is in the fifth bit position.

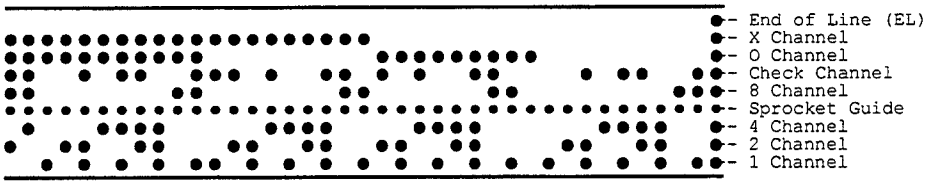


Figure 8. Example of 8-Track Punched Paper Tape

Data are stored in the eight channels as follows:

- A punch or combination of punches in channels 1, 2, 4, and 8 represent numeric characters
- A punch in the Check channel is only to be used as a parity check (odd parity is generally used)
- A punch in the "O" and "X" channels are used in combination with channels 1, 2, 4, and 8 to define alphabetic characters, symbols, and other functions such as shift to upper case, shift to lower case, or stop
- A punch in the EL channel represents the end of a line and performs the same function as the return key on a typewriter

The character set represented by vertical patterns of holes in the paper tape are depicted in Table 3. Notice that the patterns for upper cased and lower cased alphabetical characters are identical. This is because the equipment used to print documents stored on 8-TPPT operated in a fashion similar to typewriters. That is, shift keys were used to define the difference between upper and lower case characters. However, once a shift to upper case symbol was encountered, the type basket was shifted to the upper case position, and all of the characters that followed were typed in the upper case mode. Likewise, once the shift to lower case symbol was encountered, the type basket was shifted to the lower case position, and all of the characters that followed it were typed in the lower case mode. This ability to shift from upper case to lower case mode, and vice versa, provided the ability to use an identical bit pattern for two separate symbols.

Table 3. 8-TPPT (Partial) Character Set

1	2	4	S	8	C	0	X	EL	Bit Pattern	Decimal Value	Upper Case Symbol	Lower Case Symbol
								●	100 0 0000	64	End of Line	End of Line
●			●			●	●		011 0 0001	49	A	a
	●		●			●	●		011 0 0010	50	B	b
●	●		●		●	●	●		011 1 0011	51	C	c
		●	●			●	●		011 0 0100	52	D	d
●		●	●		●	●	●		011 1 0101	53	E	e
	●	●	●		●	●	●		011 1 0110	54	F	f
●	●	●	●			●	●		011 0 0111	55	G	g
			●	●		●	●		011 0 1000	56	H	h
●			●	●	●	●	●		011 1 1001	57	I	i
●			●		●		●		010 1 0001	33	J	j
	●		●		●		●		010 1 0010	34	K	k
●	●		●				●		010 0 0011	35	L	l
		●	●		●		●		010 1 0100	36	M	m
●		●	●				●		010 0 0101	37	N	n
	●	●	●				●		010 0 0110	38	O	o
●	●	●	●		●		●		010 1 0111	39	P	p
			●	●	●		●		010 1 1000	40	Q	q
●			●	●			●		010 0 1001	41	R	r
	●		●		●	●			001 1 0010	18	S	s
●	●		●			●			001 0 0011	19	T	t
		●	●		●	●			001 1 0100	20	U	u
●		●	●			●			001 0 0101	21	V	v
	●	●	●			●			001 0 0110	22	W	w
●	●	●	●		●	●			001 1 0111	23	X	x
			●	●	●	●			001 1 1000	24	Y	y
●			●	●		●			001 0 1001	25	Z	z
			●			●			001 0 0000	16	0	)
●			●						000 0 0001	1	1	!
	●		●						000 0 0010	2	2	@
●	●		●		●				000 1 0011	3	3	#
		●	●						000 0 0100	4	4	$
●		●	●		●				000 1 0101	5	5	=
	●	●	●		●				000 1 0110	6	6	¢
●	●	●	●						000 0 0111	7	7	?
			●	●					000 0 1000	8	8	*
●			●	●	●				000 1 1001	9	9	(
●			●		●	●			001 1 0001	17	/	:
			●		●	●	●		011 1 0000	48	&	;
●	●		●	●	●		●		010 1 1011	43	%	_ (underscore)
			●				●		010 0 0000	32	- (dash)	"
	●		●	●			●		010 0 1010	42		'
●	●		●	●	●	●			001 1 1011	27		,
●	●		●	●		●	●		011 0 1011	59		.
	●		●	●	●	●	●		011 1 1010	58	Shift Lower	Shift Lower
		●	●	●	●	●	●		011 1 1100	60	Shift Upper	Shift Upper
			●		●				000 1 0000	0	Space	Space
●	●		●	●					000 0 1011	11	Stop	Stop
	●	●	●	●	●	●			001 1 1110	30	Tab	Tab
●	●	●	●	●	●	●	●		011 1 1111	63	Backspace	Backspace

The information in Table 3 has been gathered, described, and preserved on 8-TPPT and it essentially describes the types of data that need to be maintained in the metaknowledge archive. This data can be used to recover and reconstruct a document that is stored on an 8-TPPT.

However, there is some additional data that is needed before this data can be used to reconstruct a file. For example, when paper tape was in popular use, the practice was to store a document on its own piece of tape. Therefore, each document was stored on a piece of tape according its size and that piece of tape could be 18 inches long, 25 feet long, or what ever length was required to store the document. Regardless of the length of the tape, each document was stored on a separate piece of tape and only entered into a machine when needed (Nashelsky, 1972).

Another important code is the STOP code (00001011). This code generally does not exist on information systems today. The STOP code halted the printing of a document so a typist could manually enter data. After the manual typing was completed, the typist would depress a "START READ" button on the equipment and the punched tape would continue to control the printing of a document (Langenbach, 1968).

To demonstrate how this data can be used, suppose a researcher needs some valuable information from a 1960s vintage research document that has only been preserved on a segment of 8-TPPT. Additionally, assume that no 8-TPPT reader is available to interpret the contents of the tape. The knowledge stored in the DRS on 8-TPPT can be used to recover and reconstruct the document. A short segment of the tape is depicted in Figure 9.

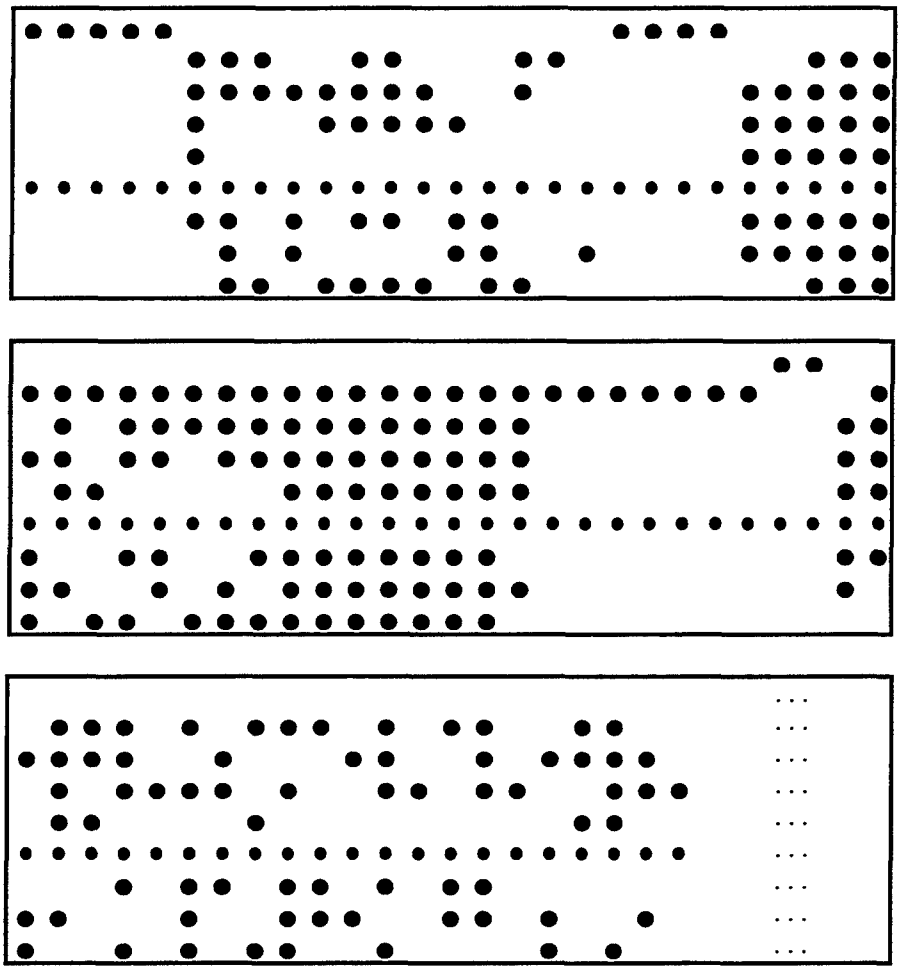


Figure 9. 8-TPPT Containing Digital Document

Before continuing, it is important to note that the document used in this demonstration was originally typed and not stored on an 8-TPPT. All data concerning this document has been retrofitted for the purpose of this demonstration. The document is from a thesis by Girard (1967). Additionally, due to a lack of documentation concerning 8-TPPT, some additional assumptions need to be made. These assumptions follow the basic keyboarding principles that were common during the time period when 8-TPPT was in common use. The assumptions are that (1) the first tab setting is for paragraph indentation (five characters), (2) the second tab setting is the center of the page, and (3) an elite type font was used (12 characters per inch).

Data Recovery Process. Upon examining the media storage techniques information on 8-TPPT in the DRS, engineers find that they can reprogram a modern day scanner to interpret the bit patterns represented by the series of holes and migrate the data to a modern storage device. As the 8-TPPT is scanned, it logically partitions the tapes horizontal tracks and vertical byte regions of the tape. This logical partitioning is depicted in Figure 10.

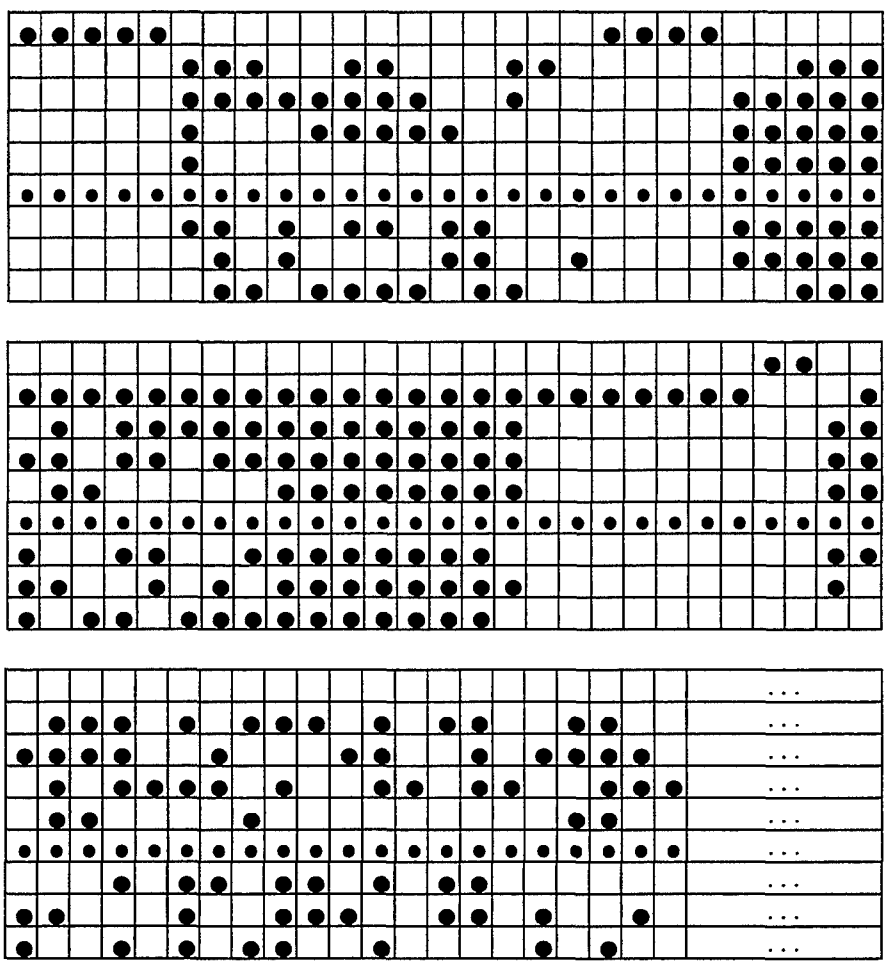


Figure 10. Logical Partitioning of 8-TPPT

An algorithm analyzes the data regions of the tape and converts the regions with no holes to a 0 and converts the regions containing holes to a 1. The analysis of the tape is depicted in Figure 11.

1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0
0	0	0	0	0	1	1	1	0	0	1	1	0	0	0	1	1	0	0	0	0	0	0	0	1	1	1
0	0	0	0	0	1	1	1	1	1	1	1	1	0	0	1	0	0	0	0	0	0	1	1	1	1	1
0	0	0	0	0	1	0	0	0	1	1	1	1	1	0	0	0	0	0	0	0	0	1	1	1	1	1
0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1
0	0	0	0	0	1	1	0	1	0	1	1	0	1	1	0	0	0	0	0	0	0	1	1	1	1	1
0	0	0	0	0	0	1	0	1	0	0	0	0	1	1	0	0	1	0	0	0	0	1	1	1	1	1
0	0	0	0	0	0	1	1	0	1	1	1	1	0	1	1	0	0	0	0	0	0	0	1	1	1	1

0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	1
0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	1	1
1	1	0	1	1	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	1	1
0	1	1	0	0	0	0	0	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	1	1
1	0	0	1	1	0	0	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	1	1
1	1	0	0	1	0	1	0	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	1	0
1	0	1	1	0	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0

0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	...
0	1	1	1	0	1	0	1	1	1	0	1	0	1	1	0	0	1	1	0	0	0	...
1	1	1	1	0	0	1	0	0	0	1	1	0	0	1	0	1	1	1	1	0	0	...
0	1	0	1	1	1	1	0	1	0	0	1	1	0	1	1	0	0	1	1	1	0	...
0	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	...
0	0	0	1	0	1	1	0	1	1	0	1	0	1	1	0	0	0	0	0	0	0	...
1	1	0	0	0	1	0	0	1	1	1	0	0	1	1	0	1	0	0	1	0	0	...
1	0	0	1	0	1	0	1	1	0	0	1	0	0	0	0	1	0	1	0	0	0	...

Figure 11. 8-TPPT Analysis

The bytes are then assembled into a bit stream, as represented in Figure 12, and migrated to a currently accessible storage medium.

```
10000000100000001000000010000000100000000111110001100111011000010010011
00110001011101010111010100110001000101100000011101100001010000000000010
10000000100000001000000010000000001111100011111001111111011111101111111
01010111011110100100100101110101011101100110000101110011011101010111111
01111111011111110111111101111111011111110111111101111101001000000010000000
01000000010000000100000001000000010000000100000000011111001111100
00100011011110100110100001110101000100000101011100110100010010010101011
01000110001000100111010100010000010001100111011000010000001000110110100
01111001001100100001000 . . .
```

Figure 12. Digital Document's Bit Stream

Once the bit stream is transferred to an accessible medium, it can be interpreted using 8-TPPT file formatting data that has been preserved in the metaknowledge archive. The 8-TPPT binary character set and the character's binary codes and meanings are depicted in Table 3. This table is used to map the text in the 8-TPPT system's 7-bit binary character set to the current system's 8-bit binary character set. This allows the current system to interpret and display the digital documents text.

Document Reconstruction. Using information from the MKA, an interpretation algorithm reads the bit stream from the advanced media and breaks the bit stream into 8-bit bytes as depicted in Figure 13.

10000000	10000000	10000000	10000000	10000000	01111100
01100111	01100001	00100110	00110001	01110101	01110101
00110001	00010110	00000111	01100001	01000000	00000010
10000000	10000000	10000000	10000000	00111110	00111110
01111111	01111111	01111111	01010111	01111010	01001001
01110101	01110110	01100001	01110011	01110101	01111111
01111111	01111111	01111111	01111111	01111111	01111111
01111010	01000000	01000000	01000000	01000000	01000000
01000000	01000000	10000000	10000000	00111110	01111100
00100011	01111010	01101000	01110101	00010000	01010111
00110100	01001001	01010111	01000110	00100010	01110101
00010000	01000110	01110110	00010000	00100011	01101000
01111001	00110010	00010000	...	...	...

Figure 13. Document's Individual Bytes

The algorithm performs an error checking routine based on the fifth bit of the 8-bit byte to insure that the integrity of the data has not been compromised. Once error checking is complete, the 7-bit characters are mapped to the 8-bit character codes that can be displayed by the current system. When mapping the 8-TPPT 's 7-bit characters to the character codes used by the current system it is necessary to use a translation table which maintains two translation schemes--one for upper cased characters and one for lower cased characters. This is because the 8-TPPT character codes receive double use. That is, the same code used for the character "A" (0110001) was also used for the character "a" (0110001). The difference in character case was determined by the position of the type basket. Therefore, the algorithm translating the character set will have to track the position of the type basket and translate the characters appropriately. A translation table such as the one depicted in Table 4 can be used to translate characters.

Table 4. Translation Table

8-TPPT Bit Pattern	Upper Case Symbol	Upper Case Bit Pattern	Lower Case Symbol	Lower Case Bit Pattern
011 0001	A	0100 0001	a	0110 0001
011 0010	B	0100 0010	b	0110 0010
011 0011	C	0100 0011	c	0110 0011
. . .	. . .	. . .	. . .	. . .
001 0000	0	0011 0000	)	0010 1001
000 0001	1	0011 0001	!	0010 0001
000 0010	2	0011 0010	@	0100 0000
. . .	. . .		. . .	. . .
010 0000	- (dash)	0010 1101	"	1000 0100
000 0000	Space	0010 0000	Space	0010 0000
001 1110	Tab	0000 1001	Tab	0000 1001
. . .	. . .		. . .	

The translation of the 8-TPPT is depicted in Table 5 below.

Table 5. Character Set Mapping

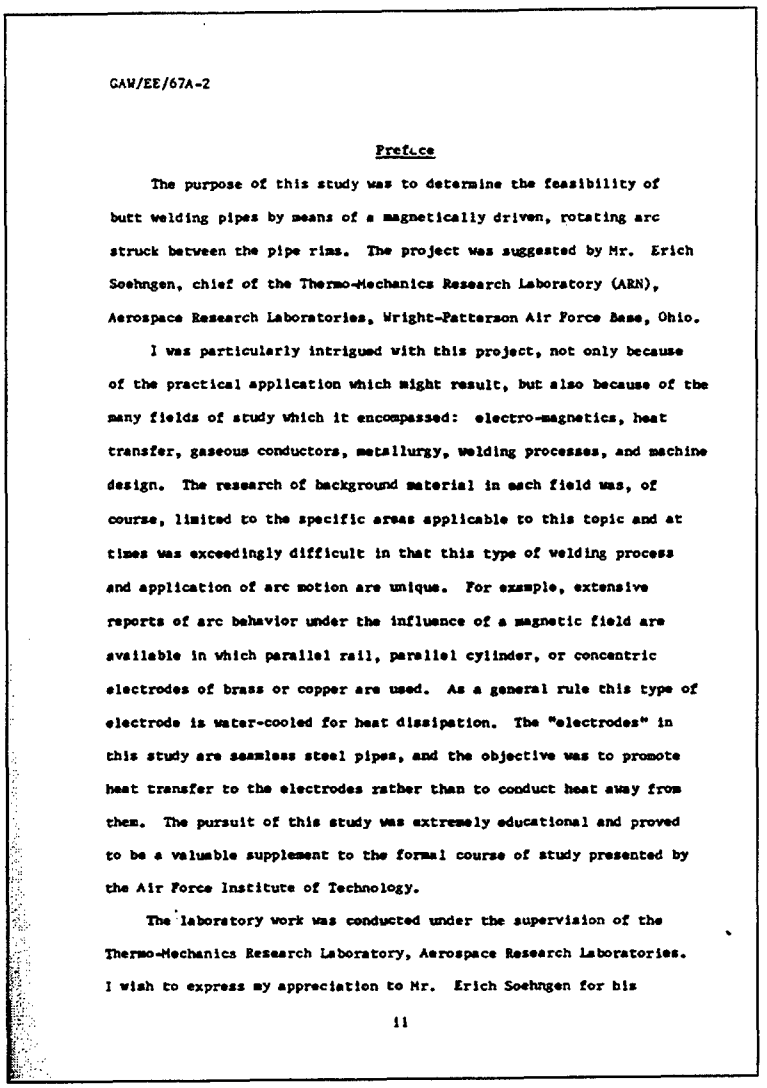
8-TPPT 7-BIT CODE	Current System's 8-BIT CODE	INTERPRETED CHARACTER
100 0000	0000 1101	Return
100 0000	0000 1101	Return
100 0000	0000 1101	Return
100 0000	0000 1101	Return
100 0000	0000 1101	Return
011 1100	?	Shift Upper (SU)
011 0111	0100 0111	G
011 0001	0100 0001	A
001 0110	0101 0111	W
001 0001	0010 1111	/
011 0101	0100 0101	E
011 0101	0100 0101	E
001 0001	0010 1111	/
000 0110	0011 0110	6
000 0111	0011 0111	7
011 0001	0100 0001	A
010 0000	0010 1101	- (dash)
000 0010	0011 0010	2
100 0000	0000 1101	Return
100 0000	0000 1101	Return
100 0000	0000 1101	Return
100 0000	0000 1101	Return
001 1110	0000 1001	Tab
001 1110	0000 1001	Tab
011 1111	0000 1000	Backspace (BS)
011 1111	0000 1000	BS
011 1111	0000 1000	BS
010 0111	0101 0000	P
011 1010	?	Shift Lower (SL)
010 1001	0111 0010	r
011 0101	0110 0101	e
011 0110	0110 0110	f
011 0001	0110 0001	a
011 0011	0110 0011	c
011 0101	0110 0101	e
011 1111	0000 1000	BS
011 1111	0000 1000	BS
011 1111	0000 1000	BS
011 1111	0000 1000	BS
011 1111	0000 1000	BS
011 1111	0000 1000	BS
011 1111	0000 1000	BS
010 0000	0101 1111	_ (underscore (US))
010 0000	0101 1111	_ (US)
010 0000	0101 1111	_ (US)
010 0000	0101 1111	_ (US)
010 0000	0101 1111	_ (US)
010 0000	0101 1111	_ (US)
100 0000	0000 1101	Return
100 0000	0000 1101	Return
001 1110	0000 1001	Tab
011 1100	?	SU
001 0011	0101 0100	T
011 1010	?	SL
011 1000	0110 1000	h
011 0101	0110 0101	e
000 0000	0010 0000	space
010 0111	0111 0000	p
001 0100	0111 0101	u
010 1001	0111 0010	r
010 0111	0111 0000	p
010 0110	0110 1111	o
001 0010	0111 0011	s
011 0101	0110 0101	e
000 0000	0010 0000	space
010 0110	0110 1111	o
011 0110	0110 0110	f
000 0000	0010 0000	space
001 0011	0111 0100	t
011 1000	0110 1000	h
011 1001	0110 1001	i
001 0010	0111 0011	s
000 0000	0010 0000	space
...	...	...

- 8-TPPT and current system's bit codes are shown without parity bits.
- ? signifies no direct translation but must use upper case or lower case column in translation table.

Once the character set has been translated the document can be printed. However, this is not as easy as it sounds. Many modern word processing operations, such as bolding, centering, and underlining are transparent to the document creator. However, the keyboarding techniques of the 1960s and 1970s were not as convenient. For example,:

- to bold text an individual had to type the text to be bolded, backspace to the beginning of that text, and then retype over the text.
- to center text an individual had to tab to the center of the page, backspace one-half of the total number of characters to be centered, and then type the text.
- to underline text an individual had to type the text to be underlined, backspace to the beginning of that text, and then use the underscore key to underline the text.

Therefore, to accurately reconstruct these documents, algorithms have to identify and translate these types of operations. Figure 14 shows how a page from the research document appears when the document reconstruction process is complete.



GAW/EE/67A-2

Preface

The purpose of this study was to determine the feasibility of butt welding pipes by means of a magnetically driven, rotating arc struck between the pipe rims. The project was suggested by Mr. Erich Soehngen, chief of the Thermo-Mechanics Research Laboratory (ARN), Aerospace Research Laboratories, Wright-Patterson Air Force Base, Ohio.

I was particularly intrigued with this project, not only because of the practical application which might result, but also because of the many fields of study which it encompassed: electro-magnetics, heat transfer, gaseous conductors, metallurgy, welding processes, and machine design. The research of background material in each field was, of course, limited to the specific areas applicable to this topic and at times was exceedingly difficult in that this type of welding process and application of arc motion are unique. For example, extensive reports of arc behavior under the influence of a magnetic field are available in which parallel rail, parallel cylinder, or concentric electrodes of brass or copper are used. As a general rule this type of electrode is water-cooled for heat dissipation. The "electrodes" in this study are seamless steel pipes, and the objective was to promote heat transfer to the electrodes rather than to conduct heat away from them. The pursuit of this study was extremely educational and proved to be a valuable supplement to the formal course of study presented by the Air Force Institute of Technology.

The laboratory work was conducted under the supervision of the Thermo-Mechanics Research Laboratory, Aerospace Research Laboratories. I wish to express my appreciation to Mr. Erich Soehngen for his

ii

Figure 14. Reconstructed Document

V. Conclusions and Recommendations

Chapter Overview

The final chapter of this thesis presents the conclusions and recommendations that evolved from this research effort. The conclusions section of this chapter discusses some important issues concerning the creation of a Digital Rosetta Stone, such as resources and practibility. The recommendations section suggest areas that require future study and provides a possible implementation scenario for the Digital Rosetta Stone model.

Conclusions

Resources. The development of a DRS will be a time intensive and expensive task. Look at the vast number of research projects, books, and museums that have been propagated in order to maintain access to our written history. The mechanics of the written language changes slowly over decades and centuries. However, "new devices, processes and software are replacing the products and methods used to record, store, and retrieve digital information on breathtaking cycles of 2- to 5-years" (TFADI, 1996). This rapid development calls for the preservation of the vast amounts of digital knowledge that has been and is being created. However, unlike written documents, the preservation of digital documents also requires the maintenance of hardware and software systems that can access these documents. This will be very costly in terms of money, manpower, and other resources.

Escalating Difficulty. As was seen in chapter 4, a document stored on primitive eight-track punched paper tape can be difficult to recover and reconstruct. As technology advances the problems seen with the paper tape become minuscule by comparison with the increasing complexity of the hardware and software systems used to create modern digital documents. Complex storage and format designs will make the recovery and reconstruction of documents even more difficult in the future.

Media Degradation and Cost. The concepts of the Digital Rosetta Stone appear practicable, but that does not mean that it is always possible or cost effective to recover and reconstruct every digital document. For example, the data stored on a media may degrade over time and result in the loss of the data that was stored on it. In this case the data is considered to be non-recoverable. Likewise, if there is no device available to read the media and it would be extremely expensive to build an access device then it may be worth the cost to recover the data that is stored on it. A representation of this basic form of the DRS recovery and reconstruction process is depicted in Figure 15.

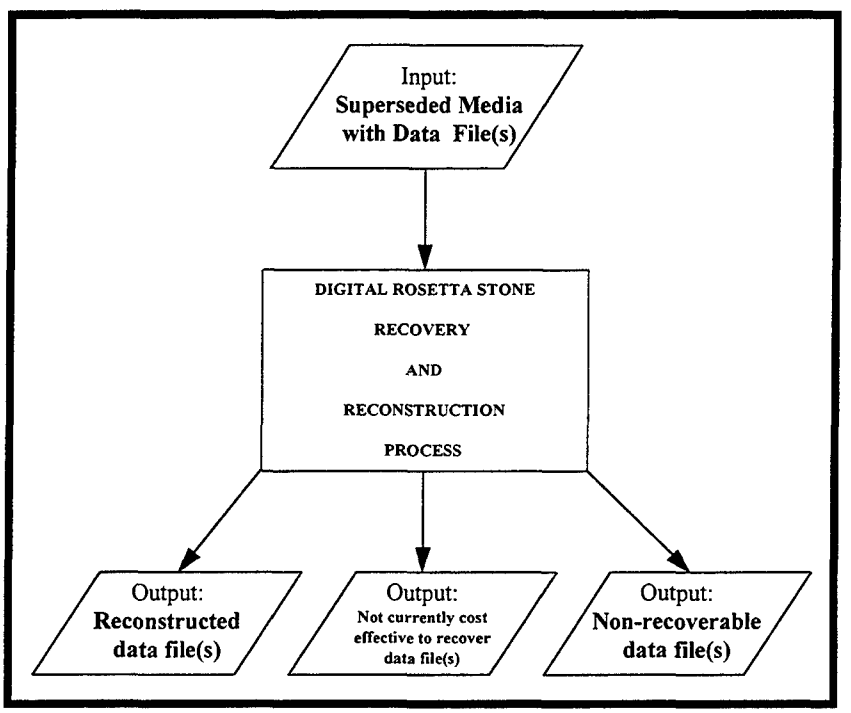


Figure 15. Digital Rosetta Stone Outputs

A detailed decision diagram describing the process represented in the DRS's recovery and reconstruction process block located in Figure 15 appears in Appendix A.

Recommendations

Suggested Implementation. To meet the high resource requirements of a rapidly

developing digital world, it is recommended that (1) a government organization be created to oversee the preservation of our digital history and (2) this organization be based on the concepts developed here under the name of Digital Rosetta Stone.

The creation of a separate governmental organization is recommended because the knowledge and capabilities of the DRS will be needed across most, if not all, government agencies, and the cost of maintaining such a capability rules out creating multiple versions of it. While it is not absolutely necessary for the organization to be under government control, a drawback of allowing this type of organization to be commercially operated is that commercial organizations will only pursue the data recovery and document reconstruction projects that are profitable to them. Therefore, it may be best for the organization to be under government control or at least partially funded by the government since the government is very likely to have special document recovery needs that require data to be recovered and reconstructed regardless of the costs.

Such an organization could be similar in concept to what has been presented here as the Digital Rosetta Stone Office (DRSO). In additional to maintaining a metaknowledge archive, the DRSO will be required to develop the hardware and software systems necessary to recover data and reconstruct documents as systems become economically infeasible to maintain. However, to offset the costs of maintaining a large, technically oriented staff, the government may only maintain the metaknowledge archive of such an operation. The government could then grant contracts for the development of the hardware and software necessary for data recovery and document reconstruction operations. Additionally, to offset the cost of maintaining a DRSO the government may offer recovery and reconstruction services to commercial organizations on a fee for service basis.

Because this would be an extensive operation, like the NARA, the DRSO may need to have offices across the country which service specific regions. Each regional

office could maintain the hardware and software to meet the most common recovery and reconstruction needs. However, specified sites will maintain the hardware and software to meet less demanding recovery and reconstruction needs. For example, some hardware and software systems may have received limited use and there may not be a large demand for services involving these systems. Therefore, it may not be economically feasible to maintain recovery and reconstruction capabilities at all DRSOs.

Limitations

A major limitation of this research is that it sets out a model that is based on what is currently known, but it does not test the model. The theory behind the model appears technologically feasible, but without a thorough evaluation its usefulness cannot be assured. Additionally, a test of the model may find that it is technologically and/or economically inferior to the other methods. Therefore, it is necessary that the model now be tested to determine its potential value. A call for an evaluation of the DRS model will be discussed further in the Recommendations section.

Also, this thesis does not address the need to physically monitor storage media for deterioration and obsolescence. An operation such as the Library of Congress' film preservation program, located at Wright-Patterson Air Force Base, may be helpful in preventing the need to recover data from obsolete storage media. The Library of Congress' program closely monitors aging movie films. If the films show signs of deterioration they are sent to a laboratory to be transferred to safety film. This process prevents the historical contents of aging films from becoming lost forever.

A similar program should be implemented in organizations that archive information on digital media. When a medium shows signs of deterioration it should be sent to a facility where its contents can be transferred to a medium that is accessible by modern information systems. Furthermore, such a facility should also track the types of media currently stored and the availability of equipment to access those media. As access

equipment for a specific medium approaches extinction, efforts should be made to transfer the contents of that medium to one that is accessible by modern information systems. This type of operation may reduce the need for large scale data recovery operations and therefore, warrants future study.

Recommendation for Further Research

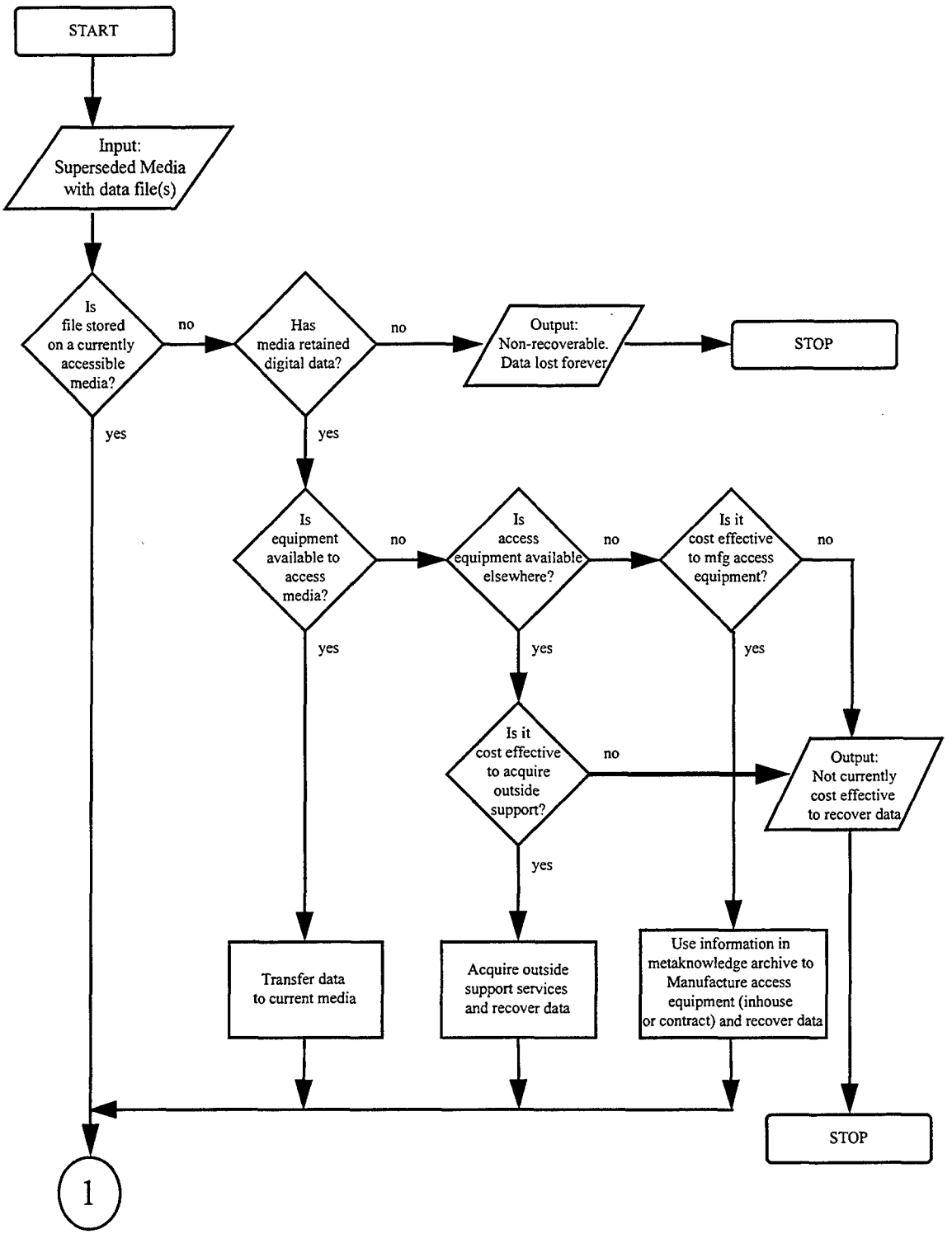
Model Presentation. The next step is to present this model to the archival community and other information management professionals to assess its validity. If the model is well accepted, then (1) this model may serve as a guide for the development of a working system, and (2) digital archeology may become an important component of the information management field. Personnel pursuing this field of study, digital archeologist, will have to determine how to develop and populate the metaknowledge archive so that the data can be easily accessible to personnel performing recovery and reconstruction operations. Additionally, hardware and software engineers with a knowledge of both antiquated and modern information systems will be needed for performing recovery and reconstruction operations.

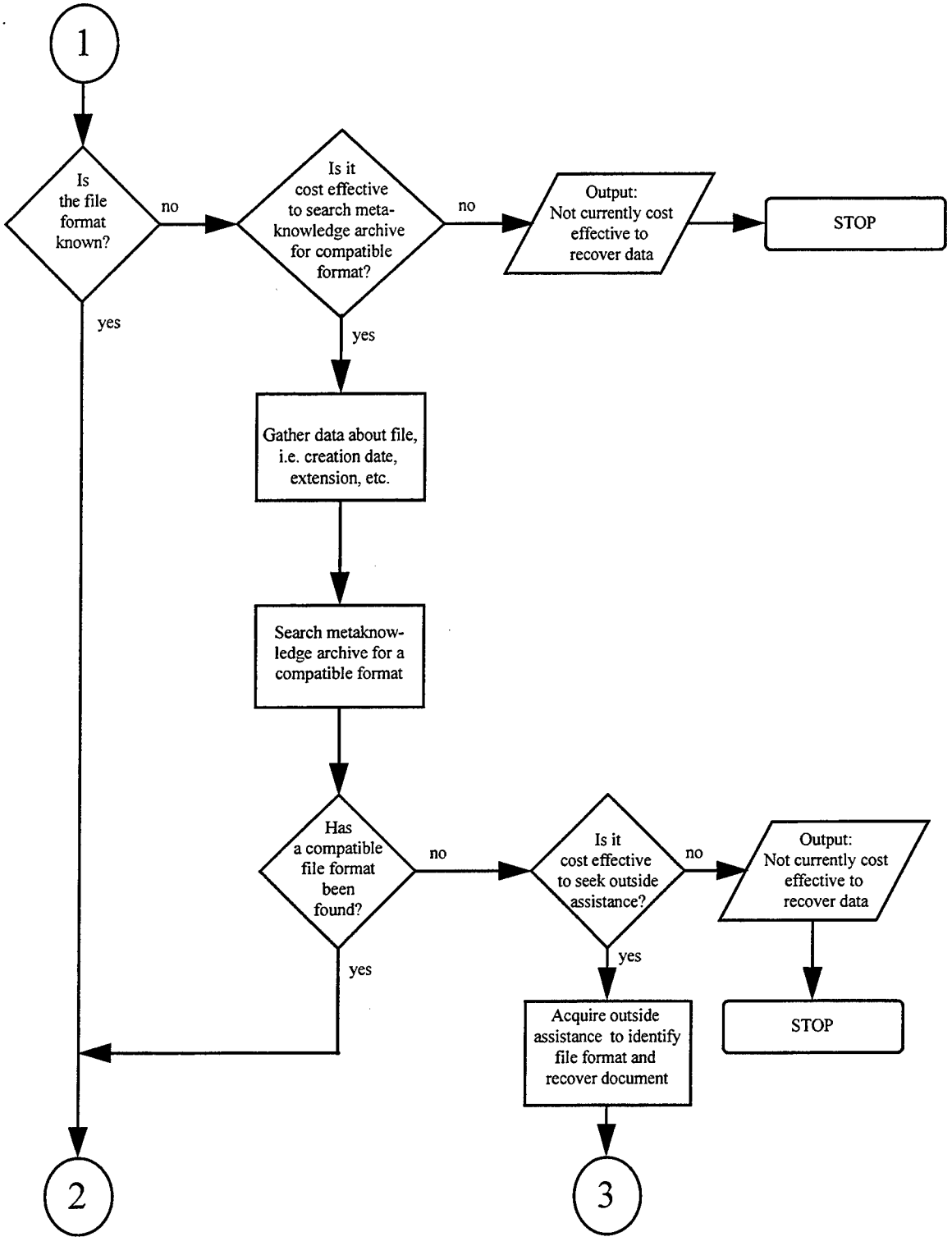
Summary

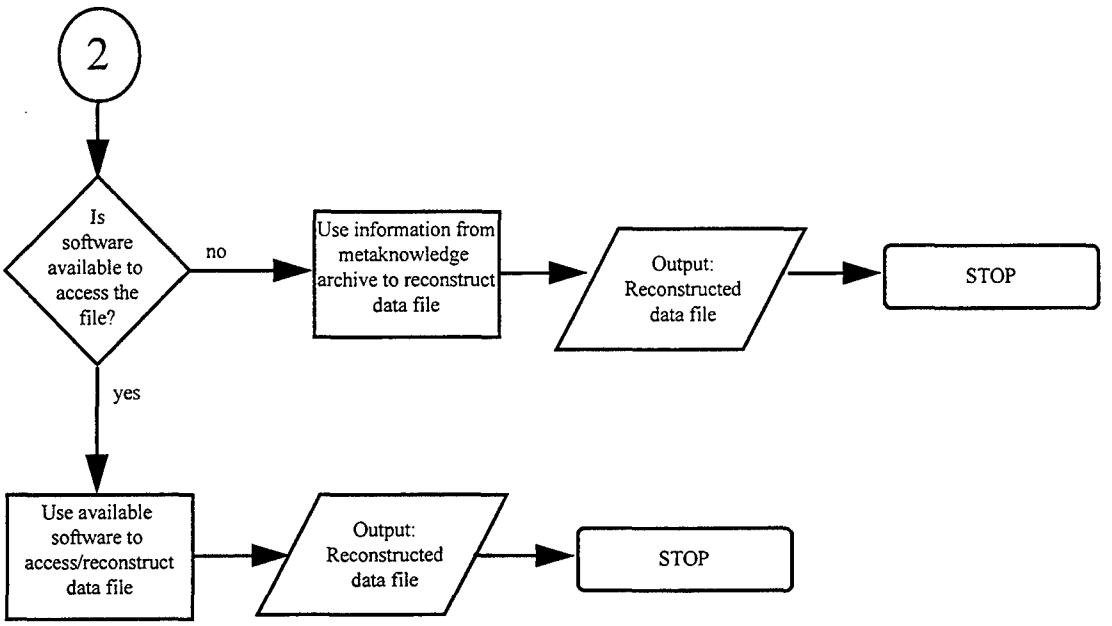
This thesis researched the problem of maintaining long-term access to digital documents. It reviewed the methods that have been suggested to maintain access to digital documents. These previously suggested methods were combined with other ideas that were encountered and conceived while performing research for this project. This led to the creation of a model, the Digital Rosetta Stone, that provides a method for maintaining long-term access to digital documents. Based on the model and other findings it was concluded that implementing the Digital Rosetta Stone may be an effective means for maintaining access to most digitally stored data as progress relegates increasing generations of storage technologies to obsolescence.

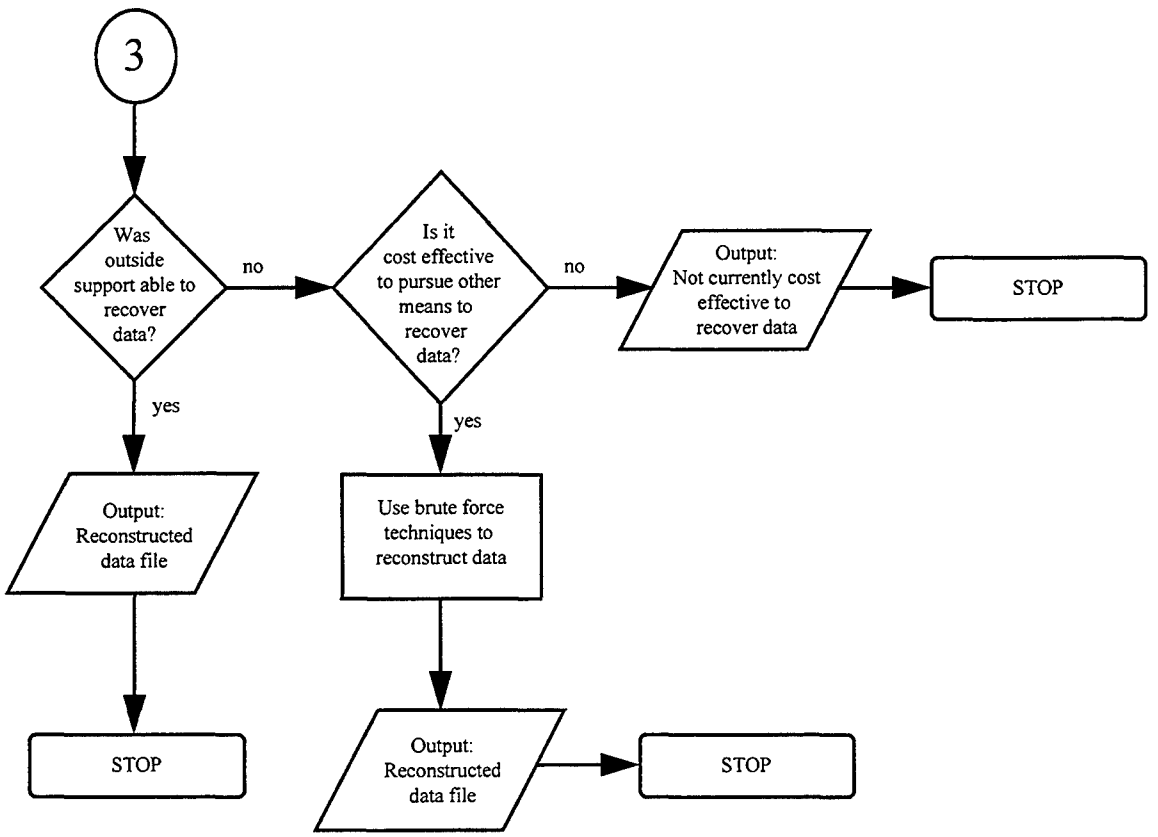
APPENDICES

Appendix A: Recovery and Reconstruction Decision Diagram




1
Is the file format known?
no
Is it cost effective to search meta-knowledge archive for compatible format?
no
Output: Not currently cost effective to recover data
STOP
yes
yes
Gather data about file, i.e. creation date, extension, etc.
Search metaknow-ledge archive for a compatible format
Has a compatible file format been found?
no
Is it cost effective to seek outside assistance?
no
Output: Not currently cost effective to recover data
yes
yes
Acquire outside assistance to identify file format and recover document
STOP
2
3


2
Is software available to access the file?
no
Use information from metaknowledge archive to reconstruct data file
Output: Reconstructed data file
STOP
yes
Use available software to access/reconstruct data file
Output: Reconstructed data file
STOP


3
Was outside support able to recover data?
no
Is it cost effective to pursue other means to recover data?
no
Output: Not currently cost effective to recover data
STOP
yes
yes
Output: Reconstructed data file
Use brute force techniques to reconstruct data
STOP
Output: Reconstructed data file
STOP

Appendix B: Acronyms

ANSI	American National Standards Institute
ASCII	American Standard Code for Information Interchange
BCS	Binary Character Set
COTS	Commercial-off-the-shelf
CPA	Commission on Preservation and Access
DoD	Department of Defense
DRSO	Digital Rosetta Stone Office
EBCDIC	Extended Binary Coded Decimal Information Code
DRS	Digital Rosetta Stone
HTML	Hypertext Markup Language
IT	Information Technology
MFM	Multiple Frequency Modulation
MKA	Metaknowledge Archive
NARA	National Archives and Records Administration
NRC	National Research Council
OASD	Office of the Assistant Secretary of Defense
PC	Personal Computer
RLG	Research Libraries Group
RLL	Run Length Limited
TFADI	Task Force on Archiving of Digital Information
USAF	United States Air Force

Bibliography

Adcock, Ken, Marilyn M. Helms, and Wen-Jang Kenny Jih. "Information Technology: Can It Provide a Sustainable Competitive Advantage?," *Information Strategy: The Executive's Journal, 9*: 10-15 (Spring 1993).

Awad, Elias M. *Business Data Processing, Third Edition.* Prentice-Hall, Inc., 1971

Beatty, Jeff. "State Office Streamlines Records," *Managing Office Technology, 40*: 58-61 (November 1995).

Boar, Bernard H. "Logic and Information Technology Strategy: Separating Good Sense from Nonsense," *Journal of Systems Management, 45*: 16-21 (May 1994).

Conway, Paul. *Preservation in the Digital World.* The Commission on Preservation and Access, March 1996.

Cooper, Donald R. and C. William Emory. *Business Research Methods, Fifth Edition.* Richard D. Irwin, Inc, 1995.

Cule, Paul E. and Varun Grover. "Into the Next Millennium: Some Thoughts on IS Practice and Research," *Data Base, 25*: 14-23 (May 1994).

Curle, Howard A., Jr. "Supporting Strategic Objectives: Building a Corporate Information Technology Structure," *Information Strategy: The Executives Journal, 10*: 5-12 (Fall 1993).

Darling, Pamela W. "Creativity vs Despair: The Challenge of Preservation Administration," *Library Trends, 30*: 179-188 (Fall 1981).

Dollar, Charles M. *Archival Theory and Information Technologies: The Impact of Information Technologies on Archival Principles and Methods.* Publications of the University of Macerata, 1992.

Downing, Douglas and Michael Covington. *Dictionary of Computer Terms.* Barron's, 1986.

Gehling, Robert G. and Michael L. Gibson. "Using Imaging to Reengineer Business," *Information Systems Management, 12*: 55-60 (Spring 1995).

Girard, Raymond F. *Automatic Pipe Welding Using a Magnetically-Driven Rotating Arch.* MS thesis, GAW/EE/67A-2. School of Engineering, Air Force Institute of Technology (AU), Wright-Patterson AFB OH, May 1967 (AD665128).

Kendall, Kenneth E. and Julie E. Kendall. *Systems Analysis and Design, Third Edition.* Prentice Hall, 1995.

Langenbach, Robert G. *Introduction to Automated Data Processing.* Prentice-Hall, Inc., 1968.

Lynn, M. Stuart. "Digital Imaging Technology for Preservation," *Proceedings from an RLG symposium held March 17 and 18, 1994 Cornell University, Ithaca NY.* 1-10. Research Libraries Group, 1994.

Martin, James. *Information Engineering Book I Introduction.* Prentice Hall, 1989.

McFadden, Fred R. and Jeffery A. Hoffer. *Modern Database Management, Fourth Edition.* The Benjamin/Cummings Publishing Company, Inc., 1994.

Michelson, Avra and Jeff Rothenberg. "Scholarly Communication and Information Technology: Exploring the Impact of Changes in the Research Process on Archives," *American Archivist, 55*: 236-315 (Spring 1992).

Mockler, Robert J. and D. G. Dologite. *Knowledge-Based Systems: An Introduction to Expert Systems.* Macmillian Publishing Company, 1992.

Mohlhenrich, Janice, editor. *Preservation of Electronic Formats & Electronic Formats for Preservation.* Highsmith Press, 1993.

Morell, Jonathan A. "The Organizational Consequences of Office Automation: Refining Measurement Techniques," *Data Base, 19*: 16-23 (Fall/Winter 1988).

Nashelsky, Louis. *Introduction to Digital Computer Technology, Second Edition.* John Wiley and Sons, 1972.

National Academy of Public Administration. *The Effects of Electronic Recordkeeping on the Historical Record of the U.S. Government. A Report for the National Archives and Records Administration.* January 1989.

National Research Council. *Study on the Long-term Retention of Selected Scientific and Technical Records of the Federal Government Working Papers.* National Academy Press, 1995.

-----. *Preserving Scientific Data on Our Physical Universe: A New Strategy for Archiving the Nation's Scientific and Technical Data.* National Academy Press, 1995a.

-----. *Preservation of Historical Records*. National Academy Press, 1986.

Norton, Peter, Lewis C. Eggebrecht, and Scott H. A. Clark. *Peter Norton's Inside the PC, Sixth Edition*. SAMS Publishing, 1995.

OASD (Office of the Assistant Secretary of Defense). *Automated Document Conversion Master Plan, Version 1*. April 1995.

Olson, Margrethe H. and Henry C. Lucas Jr. "The Impact of Office Automation on the Organization: Some Implications for Research and Practice," *Communications of the ACM, 25*: 838-847 (November 1982).

Peterson, Del. "Case Study: Improving Customer Service Through New Technology," *Journal of Information Systems Management, 8*: 28-35 (Spring 1991).

Schnitt, David L. "Reengineering the Organization Using Information Technology," *Journal of Systems Management, 44*: 14-20+ (January, 1993).

Settani, Joseph A. "Making the Jump from Paper to Image," *Managing Office Technology, 40*: 15-28 (April 1995).

Silberschatz, Abraham and James L. Peterson. *Operating System Concepts, Alternate Edition*. Addision-Wesley Publishing Company, 1988.

Smith, Milburn D. III. *Information and Records Management: A Decision-Maker's Guide To Systems Planning and Implementation*. Quorum Books, 1986.

United States Air Force (USAF). *Air Force Recordkeeping Requirements*. 17 March 1995.

van Nievelt, M.C. Augustus. "Managing With Information Technology--A Decade of Wasted Money?," *Information Strategy: The Executive's Journal, 9*: 5-17 (Summer 1993).

Williams, William F. *Principles of Automated Information Retrieval*. The Business Press, 1965.

Willis, Don. *A Hybrid Systems Approach to Preservation of Printed Materials*. The Commission on Preservation and Access, Washington DC, 1992.

Vita

Captain Steven B. Robertson [redacted] n. He graduated from Page County High School in 1979. He enlisted in the United States Air Force in May 1979 and entered active duty in October 1979. His first assignment was at Langley AFB as an Entomology Specialist.

In May 1985, he was reassigned to Myrtle Beach AFB where he served as the NCOIC, Pest Management. In April 1986, he married Marcia A. Robertson of Shenandoah, Virginia. He completed a Bachelor of Science degree in Computer Science at the University of South Carolina in June 1990. In August 1990, he was reassigned to Kunsan AB, Republic of Korea. During his tour in Korea he was selected to attend Officer Training School. He received his commission on September 25, 1991 upon graduation from Officer Training School.

His first assignment as an officer was at Grand Forks AFB as a Squadron Section Commander for the 321st Missile Security Squadron. In May 1995, he entered the Information Resource Management program in the School of Logistics and Acquisition Management, Air Force Institute of Technology.

[redacted]

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	December 1996	Master's Thesis

4. TITLE AND SUBTITLE
DIGITAL ROSETTA STONE: A CONCEPTUAL MODEL FOR MAINTAINING LONG-TERM ACCESS TO DIGITAL DOCUMENTS

5. FUNDING NUMBERS

6. AUTHOR(S)
Steven B. Robertson, Captain, USAF

7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(S)
Air Force Institute of Technology
2750 P Street
WPAFB OH 45433-7765

8. PERFORMING ORGANIZATION REPORT NUMBER
AFIT/GIR/LAR/96D-8

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Olthea Croom
SAF/AAIQ
1610 Air Force Pentagon
Washington DC 20330-1610

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for release; distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT (*Maximum 200 Words*)

Due to the rapid evolution of technology, future digital systems may not be able to read and/or interpret the digital recordings made by older systems, even if those recordings are still in good condition. This thesis addresses the problem of maintaining long-term access to digital documents and provides a methodology for overcoming access difficulties due to technological obsolescence. A review was conducted to determine the long-term access methods that have already been suggested by other researchers. These previously suggested methods are then combined with other ideas that were encountered and conceived while performing research for this project. The combination of these methods and ideas led to the creation of a model, the Digital Rosetta Stone, that provides a methodology for maintaining long-term access to digital documents. The hypothesis for the model is that knowledge preserved on different storage devices and file formats can be used to recover data from obsolete media and to reconstruct the digital documents. The Digital Rosetta Stone model describes three processes that are necessary for maintaining long-term access to digital documents in their native formats--knowledge preservation, data recovery, and document reconstruction. Finally, recommendations are made for the evaluation and implementation of the Digital Rosetta Stone.

14. SUBJECT TERMS	15. NUMBER OF PAGES
Digital Systems, Data Management, Information Retrieval, Document Access, Rosetta Stone	73
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

AFIT Control Number AFIT/GIR/LAR/96D-8

AFIT RESEARCH ASSESSMENT

The purpose of this questionnaire is to determine the potential for current and future applications of AFIT thesis research. **Please return completed questionnaire** to: AIR FORCE INSTITUTE OF TECHNOLOGY/LAC, 2950 P STREET, WRIGHT-PATTERSON AFB OH 45433-7765. Your response is **important.** Thank you.

1. Did this research contribute to a current research project? a. Yes b. No

2. Do you believe this research topic is significant enough that it would have been researched (or contracted) by your organization or another agency if AFIT had not researched it?
a. Yes b. No

3. **Please estimate** what this research would have cost in terms of manpower and dollars if it had been accomplished under contract or if it had been done in-house.

Man Years_____ $_____

4. Whether or not you were able to establish an equivalent value for this research (in Question 3), what is your estimate of its significance?

a. Highly Significant b. Significant c. Slightly Significant d. Of No Significance

5. Comments (Please feel free to use a separate sheet for more detailed answers and include it with this form):

_____ _____
Name and Grade Organization

_____ _____
Position or Title Address