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

Alan R. Heminger, Ph.D.
Air Force Institute of Technology
Error! Bookmark not defined.

Steven B. Robertson, Captain
United States Air Force
steve.robertson@worldnet.att.net

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 paper addresses the problem of maintaining long-term access to digital documents and provides a methodology for overcoming access difficulties due to technological obsolescence. The result of this effort led to the creation of a model, which we call the Digital Rosetta Stone, that provides a methodology for maintaining long-term access to digital documents. The underlying principle of 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.

1. Background

Any large organization has the need to retain and, on occasion, refer to various stored documents. Until recently, documents were generally paper or microfilm based. However, modern data storage methods have evolved to include digital storage as well. Due to the rapid evolution of storage technologies, 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 [24].

Within many organizations today, digital documents that are official records must be categorized and managed in accordance with approved records schedules. This means that these records must be retained and accessible throughout their life cycle in accordance with the same laws and standards that govern paper records. In the case of government documents for instance, 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. 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. As time goes on, we can expect the problems to become worse, as more and more documents are stored in digital format. The National Research Council [20] 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.

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. It is crucial that, in the haste to adopt newer and generally better technologies, we don't lose the ability to access historical digital documents.

The purpose of this paper is to respond to this call for action by developing a model with which we can assure ongoing access to the ever-increasing repository of digitally stored information. We will concentrate on maintaining access to digital documents in their native formats without converting them to emerging digital format standards. 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 [20]."

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 and explain a model that can be used to recover and reproduce digital documents from their native file formats.

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. 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.

2. Current Strategies for Preserving Digital Documents

A number of strategies for preserving digital documents have been discussed in the literature. Charles Dollar [9] suggested that customers should demand that vendors provide cost-effective migration paths to advancing hardware and software systems. Many vendors do provide a limited form of 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. However, this type of conversion is generally limited to the previous generation of the hardware or software system and therefore, it is a short-term 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. 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.

Dollar and Rothenberg also suggested promoting 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" [9]. While these open system standards would make digital documents accessible through any software system that conform to the standards, 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.

Rothenberg [27] suggested a means of maintaining long-term access to the information contained within digital documents by 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 might be achievable in principle, 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 [20]. These problems make it unrealistic to expect that any organization could effectively and efficiently maintain multiple, aging information technologies in order to maintain access to superseded digital documents.

To overcome the problems associated with maintaining aging hardware, Rothenberg [27] 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. 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.

3. Which Strategy to Use?

None of the strategies discussed above is entirely satisfactory by itself. Therefore, as information systems and their operating environments continue to evolve it may be necessary to use some combination of one or more 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 [26].

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 [20]. To recapture the information in the myriad digital documents that will be an increasingly large proportion of our information storage may require something comparable to the Rosetta Stone that opened up the writings of ancient Egypt to scholars of today.

4. The Rosetta Stone

At some point during the fourth century, all knowledge of ancient 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 to Egypt in 1799, Napoleon's army discovered an artifact which has become known as the Rosetta Stone. This stone contained the inscription of a decree issued in 196 BC by Ptolemy V Epiphanes. The decree was repeated three times in two languages, Greek and Egyptian, with the Egyptian version appearing twice, 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.

5. A Digital Rosetta Stone (DRS)

We draw on the strategies discussed above and add others to create a model for maintaining long-term access to digital documents. We call this model the Digital Rosetta Stone (DRS) because it offers a way for those in the future to be able to gain access to the information stored in the digital documents that we have stored, and will continue to store, in increasing numbers. The DRS will contain multiple levels of knowledge about specifications and processes by which information is stored on various types of storage media. It will also contain archives of knowledge about how to meaningfully interpret that information so that the original meaning can be recovered.

Rothenberg [27] 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" [5]. 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.

6. 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.

6.1 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 . 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 [14]. 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 medium is not available or no longer exists. 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 requires only that the techniques in which the bit patterns are stored and accessed on the media needs to be preserved.

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 of 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.

6.2 Data recovery

Data recovery is the process of extracting digital data from an obsolete medium and migrating it to a medium that is accessible to current information systems. 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.

6.3 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 [15]. 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.

7. Knowledge Preservation

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. 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).

7.1 Media 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.

When trying to recover data, recovery personnel must know where to look in order to find it. Media storage geometry defines where on a medium 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 a spiraling track with adjacent tracks 1.6 micrometers apart for a track density of 16,000 tracks per inch [23]. 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 medium.

After the medium's storage geometry has been identified, data recovery personnel must know the method used to store the data. The data storage method refers to how data are physically recorded on a medium 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 platter. 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 [29]. 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 [29]. 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. 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

7.2 File format 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.

The second step is to identify and catalog the objects that are supported by these applications. 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.

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. 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.

8. Data Recovery

Once it is no longer economically feasible to maintain antiquated hardware systems, it is necessary to implement an alternate method to maintain the ability to recover data from superseded media. If data are stored on an obsolete medium 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 use of the storage techniques information gathered during the knowledge preservation process 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 medium to one that is currently accessible.

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 photo-electric 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 [10]. Because of advances in storage technologies, punched cards are seldom used as a storage medium today because they are slow, bulky, and cumbersome compared to modern storage media. Because of this, few organizations maintain punched card today. 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 medium.

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 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.

9. 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. **Example character sequences for bold**

Software Application	Begin Bold	End Bold
Wordstar®	02	02
Ami Pro®	3C 2B 21 3E	3C 2D 21 3E
HyperText Markup Language	3C 42 3E	EC 2F 42 3E

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

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.

10. The Digital Rosetta Stone Model

As described above, 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 1.

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 largely proprietary. Finding information on the 8-TPPT coding scheme was very difficult. While doing research for this paper, we contacted the technical support and archive sections of the IBM Corporation to get some information on 8-TPPT equipment. Unfortunately, we were 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.

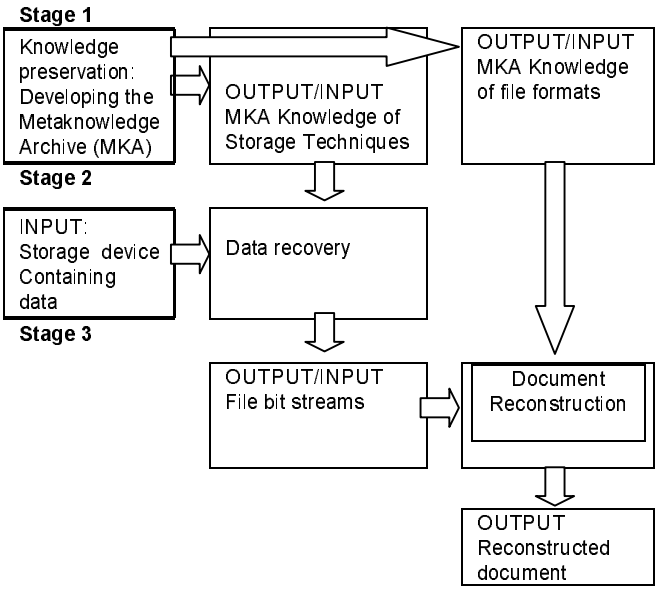

Figure 1. Digital Rosetta Stone Model

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 [2], Nashelsky [18], Langenbach [12], and Williams [32]. 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 2. Notice that unlike today, the check bit is not the most significant bit, but instead is in the fifth bit position.

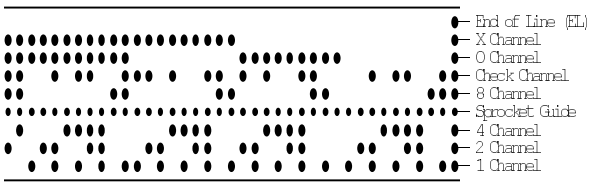

Figure 2. Example of 8-track paper tape

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 patterns for upper case and lower case 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. 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, until a symbol was encountered to shift back to lower case. 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. This along with other meta-information about 8-TPPT is necessary to recover information from 8 track paper tape. [18].

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. 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 bytes are then assembled into a bit stream, and migrated to a currently accessible storage medium. 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. 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.

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.

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.

After all of this knowledge is brought to bear, a document stored 40 years ago can be recovered and printed. In the future, we will have many more difficult tasks of digital document reconstruction. The DRS can be a significant agent in helping to ensure that we don't lose our ability to read our own history.

11. Conclusions

In this paper, we researched the problem of maintaining long-term access to digital documents. We reviewed the methods that have been suggested by others, and combined them with additional ideas to create a model we call the Digital Rosetta Stone. The Digital Rosetta Stone describes a method by which we will be able to maintain long-term access to our increasing repositories of digital documents.

The development of a DRS will be a time intensive and expensive task. Consider 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 technologies for capturing and storing digital documents is evolving faster than ever. 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 preservation of the knowledge and technology necessary to access these documents. The Digital Rosetta Stone presents a model for achieving that end.

References

[1] 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).

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

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

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

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

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

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

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

[9] 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.

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

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

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

[13] 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.

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

[15] 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).

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

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

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

[19] 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.

[20] 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.

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

[22] *Preservation of Historical Records*. National Academy Press, 1986.

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

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

[25] 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).

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

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

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

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

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

[31] 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).

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