The Presidential Electronic Records PilOt System: Results of Laboratory Experiments and Use by Archivists[1]

William E. Underwood
Georgia Tech Research Institute
Atlanta, Georgia

TR ITTL/CSITD 03-01

November 2003

[1] The National Archives and Records Administration (NARA) and the Army Research Laboratory (ARL) sponsored this research under ARL Contract No. DAKF11-97-D-001, Task Order 64. The findings in this paper should not be construed as an official NARA or ARL position unless so indicated by other authorized documentation.

Introduction

The original objective of the PERPOS research project was to conduct research that would aid archivists in gaining intellectual and physical control of the user-created files contained on the Bush PC hard drives. The Bush hard drives are the file systems of more than 500 hard drives removed from PC's in the White House offices at the end of the Presidential Administration of George H. W. Bush.

The approach to this problem began with a study to understand the archival processing activities of the Bush Presidential Library and Museum that were applied to paper records. The next step was to design and develop software prototypes that aided or supported archivists in performing these activities on the PC file systems from the Bush hard drives. The prototypes were developed via use case analysis and object-oriented design. The prototypes were used in laboratory experiments to process file systems from the Bush PC hard drives. Results of these experiments were used to refine the prototypes for further experiments.

The initial prototypes supported separate activities, for example, a File System Filtering Tool, an Archival Description Tool, and an Archival Review Tool. To support pilot use of the tools by archivists to actually process the file systems, a common user interface was designed, support for other activities provided (accession, arrangement, preservation) and the existing tools integrated into a prototype system called the Archival Processing Tool (APT).

Pilot use of the APT (version 2.04) by two archivists began at Archives II in the fall of 2002 and continued through the winter. Version 2.05 of the APT has been installed at the Bush Presidential Library and Museum in College Station, Texas and evaluation of its operations will begin during fall-winter 2003.

This report summarizes new functional requirements for archival processing of file systems of PC records that have been identified using the experimental prototypes and technologies that can be used to satisfy these requirements. It also describes advanced technologies that can be used to satisfy pre-existing requirements for processing electronic records.

New Requirement and Technologies

This section states what are believed to be new requirements for processing records created on legacy personal computer platforms. The requirements are identified based on the results of laboratory experiments and use of the APT prototype by archivists. With each requirement is described the experimental or practical situation in which the requirement was identified, risks that will be encountered if the requirement is not satisfied, and the technological alternatives that have been, or are being, explored to meet the requirement and mitigate the risks.

Requirement 1: *An effective procedure is needed for separating user-created records from operating system and software applications in a file system.*

This is the original problem faced in processing the file systems from the Bush hard drives. The file systems transferred include operating system and software applications as well as the user-created files. Archivists who began to process these records did not have the knowledge, experience, or tools to reliably separate records from non-records.

An effective procedure for filtering operating system and software application files out of file systems that also include user-created files that might be records has been developed. It is based on hash codes of the operating system and software applications files. The procedure compares the hash code of a file in the file system against the hash codes of operating system and software application files in a reference set. The hash code algorithm used is the Secure Hash Algorithm-1 (SHA-1) [6]

The reference set can be built interactivity by a computer technician who has knowledge of operating and software application files. An even better alternative is to use the Reference Data Set (RDS) being created by the National Software Reference Library (NSRL) at NIST. Version 2.1 of the RDS includes the SHA-1 hash codes of files from 2400 products.

There is a remote risk of hash code collisions with the resultant risk that a file will have been concluded to be a non-record, when it is actually a record. However, according to the NSRL "It is *computationally infeasible* to find two different files less than 2^{64} bits in size producing the same SHA-1. 2^{64} bits is one-million terabytes."

Requirement 2: *When personal computer records assumed to be unclassified are being accessioned into a computer system, there should be an automatic check that the records do not actually include security classified files, and this check must occur before the records are stored in the archival storage of the computer system.*

An archivist at Archives II used the Archival Processing Tool to view PC files from the Bush hard drives. During the process, two electronic documents were found with security markings. The two file systems that contained the documents with security markings were stored on a hard disk of an uncertified computer system and thus resulted in a security infraction. The storage of these files on the hard disk of an uncertified computer system compromised work that had already been accomplished and stored on the hard disk of the same computer system [4].

This use of the pilot system identified the risk that accessioned record series that are assumed to be sensitive but unclassified may actually contain classified materials. While the source of the problem was in the commingling of sensitive records with classified records on a PC file system in the White House, there must be a check before transfer to the archives and at the time of accession into the archives to ensure that the assumption is correct.

In response to this new requirement, a technique has been formulated that may satisfy it. The technique will be implemented and evaluated in a future version of the experimental system.

Requirement 3: *When personal computer and email records are being accessioned, the computer system should be configured to automatically check that the files do not contain computer viruses, and this check must occur before the records are stored in the archival storage of the computer system.*

During preparation for pilot use of the system by archivists, over five hundred file systems from the Bush hard drives were copied to a hard drive of the APT system. During this process a previously undetected computer virus was detected and isolated. The media from which the file systems were being copied had been previously processed to produce a preservation copy. Those procedures did not include configuration of the copy procedure to check for viruses, so the computer virus had not been detected.

The risk of infection of archival holdings by computer viruses was identified. This risk implies the risk of potential damage to other record series in archival storage and reintroduction of a computer virus to the wild by future distribution of electronic records to researchers. Accessioned PC record series from Federal or Presidential offices must always be checked for computer viruses.

Requirement 4: *PC files that are transferred, accessioned, and processed should be stored in archive (package, container) files.*

In the process of copying the PC file systems to the hard drive of the APT system, in loading such file systems for filtering, preservation, review, arrangement and description, and in storing these files back to archival storage, it was found that there was substantial delay in transferring the files between media.

The risk is that archival processing throughput will be reduced by slow input/output operations on small files. This delay can be substantially reduced by storing many files as one, in other words, in an archive file, e.g., a Zip or TAR file. Compression of the archive file also substantially reduces this delay and decreases the amount of archival storage required.

In the APT, all accessioned file systems are stored in a TAR file. Results of archival processing steps are stored back to a TAR file [4].

Requirement 5: *It must possible to distinguish self-extracting archive (SFX) files, which are executables from other executables, and to extract files from SFX files without executing the SFX file.*

During laboratory experiments, archive files (e.g., ARC and Zip archives) were found in some of the file systems. Such files might contain electronic records of the Bush

Administration. Self-extracting archives include a copy of the program for extracting the files prefixed to the archive itself. These are executable files that contain the archive as data.

There is a risk that a self-extracting archive file commingled with operating system and software application executable files will not be recognized as a file that might contain user-created files. There is a further risk that a self-extracting archive transferred to the National Archives a decade or more after its creation may not execute on existing technology. Finally, there is a risk that a self-extracting archive might contain in the executable portion malevolent code such as a computer virus or Trojan that might infect the archival system or otherwise damage the system.

To address this issue the capability has been developed to automatically identify a dozen of the 75 or so known self-extracting archive formats from their internal format, rather than from their file extension (.exe). Viewers and archive extractors are provided for the self-extracting archive file types, that can display the filenames and other attributes of the files contained in the archive, and then extract the files from the self-extracting archive without actually executing it. This capability can be extended defining the criteria necessary to identify other self-extracting file types, and by adding programs for viewing and extracting files from those self-extracting file types [3,4].

Requirement 6: *A repository of specifications for legacy file formats is needed to ensure that legacy file formats can be converted to current file formats that are viewable.*

During laboratory experiments, electronic records were encountered that are represented in obsolete file formats. Display of these files is currently dependent on the existence of DOS or Windows programs that must be executed from the DOS command line. Future generations of Microsoft operating system are expected to abandon DOS support. Then it may be necessary to write programs for current computer platforms to display the electronic records.

There is a risk that electronic records represented in legacy file formats may not be viewable (renderable) on future computer platforms. There is a further risk that the specifications of the file formats in which records are represented will not be available, substantially increasing the risk of extraordinary costs to empirically determine these formats for conversion to current file formats for which there are viewers.

Digital copies of specifications for many of the file types discovered on the Bush hard drives have been collected. They are being stored in a reference repository on the PERPOS secure Web Portal [6].

Requirement 7: *A specification language for file formats is needed to support conversion from legacy to current or standard file formats.*

To support preservation of electronic records, legacy file format specifications need to be in a form that they can easily be used to support transformations to current or standard

formats. With such specifications there is the possibility of generic software for translating files from one format to another. Without such specifications, and with an increasing number of software applications and new file formats each year, there is a risk of increasing software cost for developing conversion software for new computer platforms [4, 6].

There is a need for a file format specification language. The Data Description Language-EAST is a possibility [7]. CNES is developing a tool called EXTRA that converts EAST descriptions to XML. BinX, Binary Exchange Format is also a possibility [8]. It attempts to describe the structure of binary files in XML. Experiments are needed to determine whether these or other file format specification languages are expressive enough to describe the structure and semantics of legacy PC file formats and standard file formats and to support the transformation of files in legacy formats to current or standard formats.

Requirement 8: *Password recovery tools are needed to recover passwords that were used to encrypt Bush PC files. Legacy software applications that were used to create and encrypt these files are needed to decrypt the files.*

During the use of the APT, an archivist encountered files that are password encrypted. Many PC software applications provide a capability to encrypt files. While sensitive, these files did not contain security-classified records. Presidential records are of long-term value and must be preserved by the National Archives and made available to the public qualified by FOIA or PRA access restrictions. But password encrypted files cannot be reviewed and made available unless the password is recovered and the encrypted file decrypted using the original software application.

A commercial-off-the-shelf password recovery tool kit (PRTK) from Access Data was interfaced to the APT to support this function. Copies were acquired of the legacy PC software that can decrypt the files given the password. These were linked to the PRTK to decrypt the encrypted files [4].

Requirement 9: *A system that supports archival processing must be able to repair a variety of corrupted files.*

During laboratory experiments, files were discovered that could not be viewed due to various forms of file corruption. There is a risk of loss of digital records due to media deterioration or digital transmission errors. Such corruption can occur at any time during the maintenance of records by their creator or during long-term preservation of records by a central archives.

There are media refreshment policies that reduce the likelihood of file corruption due to media deterioration. Cyclic redundancy codes (CRCs), hash codes, and error-correcting codes are among the technologies for detecting and correcting such errors.

Further investigation is needed to characterize the types of file corruption that can occur, and to determine those types that can be easily repaired, and those that are unrepairable.

The APT supports further experimentation by including an extensible collection of PC file repair tools [3].

Requirement 10: *A system that supports archival processing must support the rearrangement of records that are not in logical order and to record the fact of their rearrangement.*

During pilot use of the APT, it was discovered that records saved in DOS file systems were often not stored in a folder representing a business activity of their creator. For instance, they were often saved in the root directory or in the folder containing the software application used to create them. Furthermore, they are often not arranged by date or filename, but in DOS directory order. While an archivist works according to the Principle of Original Order—preserve records in the original order in which they were created—it may be desirable to arrange misplaced records in an appropriate folder or to reorder files by date or filename. If records are not in a logical order, it makes it difficult to understand the relationships of records in a folder or records series. If an archivist determines that rearrangement is needed, the APT supports this operation [3].

Requirement 11: *A system to support accession and description of record systems stored as PC file systems should identify the number of bytes (Kilobytes, Megabytes, Gigabytes) in a file system as well as the number of files.*

During pilot use of the APT, it was discovered that archivists must determine the approximate volume of records accessioned. Archivists measure the approximate volume of records in a series of paper records in linear feet. Linear feet of paper are directly translatable into number of pages (1 linear foot = 2000 pages). The volume of records is used as a measure of archival processing work to be done, work accomplished, or as a measure of the volume of records in a series or file unit that is available to a researcher.

Archival processing work in the Center for Electronic Records is measured in number of files. However, due to the large variation in number of bytes per file, bytes (Kb, Mb, Gb) may be a better measure. The APT determines the number of bytes (Kb, Mb, Gb) and the number of files per file system and on request displays this for the archivist [3].

Requirement 12: *Tools are needed to inform archivists of files that contain information that is not part of the record and to support removal of this information.*

During laboratory experiments, it was discovered that some word processing and database files commonly contained information that was not part of the record. This is due in part to the storage of files in 512 byte blocks. Also, some database tables contained records that were marked for deletion, but had not been deleted. The result is that information that is not a part of the records is included in the records.

If an archivist is not notified of the existence of this information, there is a risk that the extra information will not be subject to review and so might include information that should be restricted. This can be addressed in the APT prototype by inclusion of software tools to remove this extra information from word processing files and to delete database records that have been marked for deletion.

Potential Process Improvements

This section describes archival processing requirements for PC records that have already been established. The risks encountered if these requirements are not satisfied are also discussed. Most importantly, technologies to satisfy these requirements that have been or are being developed during this research are discussed. Finally, additional knowledge that needs to be derived by archival use of the tools and additional laboratory experiments are outlined.

Requirement 13: *Archivists and the public must be ensured that an Archival System for electronic records maintains electronic record series authentic, that is, that the Archival System is trustworthy.*

This requirement is widely accepted. However, just what properties the system must have to be judged trustworthy, and what technologies are necessary to meet this requirement are not well understood.

A system that is used to process and store electronic records must be accessible only to those with the authority to process the records. It must be demonstrable that no records have been added, deleted, or modified in a file system by other than a person authorized to do so. The records must also be free from corruption (e.g., media deterioration, loss of bits during transmission) that might lead to the records not being viewable. Technology obsolescence may render the records unviewable unless they are converted to current formats. Such conversions must preserve the content and form of the original record.

The APT prototype includes a procedure for packaging PC file systems containing records as a JAVA Archive (JAR) file. This is a representation and procedure that uses message digests and Public/Private key technology to ensure the integrity of individual records, folders of records and file systems. A formal theory of record series authenticity has been formulated. With a theory of communication security can be used to prove that this procedures maintains the integrity and authenticity of the records [1,2].

Requirement 14: *An archival system must be able to convert files from obsolete file formats that cannot be viewed on current technologies to file formats that can.*

During laboratory experiments and use of the PERPOS tools by archivists, files were encountered that could not be viewed due to obsolescence of hardware, operating system, and or the software application used to create the files. There is a risk of loss of these files as records of the Bush administration unless viewers are constructed for them or they are converted to a current and/or standard format.

This is a universally recognized requirement for preservation of electronic records. The Bush PC records transferred to NARA were from computer platforms of just 10-15 years ago. There are already some digital objects created by White House staff members and offices that, short of emulation of the computer platform or writing of viewers for the legacy file format, cannot be viewed without conversion to some other format.

The APT supports conversion to current formats of those file formats that have been identified as not having viewers [3,4].

Requirement 15: *Metadata related to the activities of arrangement and preservation of record series must be preserved.*

This requirement ensures that there is a record of any modification to the order of the original file system and to the form and content of electronic records. Without this metadata there is a risk of losing the ability to conclude that an electronic record provided to a researcher is an authentic copy of the record.

The approach used in the APT differs from other approaches in that while some of the metadata is stored in a metadata catalog, most of the metadata is stored with the records in a manifest [3,4]. Thus, it is not necessary to store in the metadata catalog the properties of the electronic records and archival actions performed on the records.

Requirement 16: *Records closed due to PRA restrictions or FOIA exceptions and originals of redacted records must not be accessible to the public.*

In the Bush Presidential Library, opened and redacted paper records are stored in different locations than closed and originals of redacted records. So-called shadow folders relate the records in the different locations.

An alternative explored in the APT is to maintain opened, closed and redacted records in the same container, and create a reference copy for public use that contains only the open and redacted records [4].

Requirement 17: *An archival processing system must have the capability to execute files that present records, even when the computer platform on which the records were created and used is obsolete.*

During use of the APT by archivists, an executable file was discovered that is a record. It is a report from a government agency that was sent to White House offices, used and saved on PCs as a record. The contents of the report are data in a program and the executing program provides an interface to the reader so that they can navigate through the report and display its sections and subsections.

It is more than a decade since the executable report was created. It was created to operate on a 386 computer in a DOS operating system—an obsolete computer platform. There is a risk that if the record is not converted to some other format, it will not be viewable in the future.

Fortunately, the executable report is still executable on the DOS operating system provided with Windows 2000. The APT provides tools for exploring preservation alternatives. It is planned to include an X86 emulator in the preservation toolset. This would allow the executable report to be run with the original DOD operating system. Also the capability can be provided to capture the screens of the executable report and save them as a multiple page TIFF file for which there are viewers.

Requirement 18: *A system to support archival processing of legacy PC records must be able to preserve and reconstruct records made up of multiple digital files.*

During laboratory experiments, database files were discovered that consist of multiple database tables, memo files, indexes and the program that relates the set of files. If these files are not all preserved, there is a risk that the electronic records cannot be completely reconstructed.

The Archival Processing Tool currently supports preservation of these files by identifying to the archivist the kinds of objects that make up a digital record, and providing the capability to put these components into a named folder. It also supports the display of the database program to determine what the relations are among the database tables and on which fields the tables are indexed. Since the Quick View Plus suite of viewers does not display the content of memo fields in a database tables, another viewer is provided that does. Experiments are needed to determine whether there is additional decision support that can be provided to improve this process [3].

Requirement 19: *An archival processing system needs the capability to support archival response to FOIA requests for unprocessed electronic records.*

Five years after the end of a Presidential Administration, a Presidential Library must respond to FOIA requests for unprocessed records. To accomplish this function for paper records, some sort of finding aid must be created for thousands of boxes of records. The Bush Presidential Library accomplished this by creating a table containing the folder titles for the Staff Member and Office Files indicating which box the folder was in. The White House Office of Records Management (WHORM) Subject Files List was also used as a finding aid in responding to FOIA requests.

With personal computer files, there is an opportunity to improve this process through use of text-based document retrieval technology. A laboratory experiment was conducted to determine which of Boolean search and retrieval, statistical retrieval with relevance ranking, and natural language-based retrieval had the best performance in identifying Presidential electronic documents relevant to actual FOIA requests. WebGlimpse was used as an example of Boolean search and retrieval technology, Oracle Text with word

query as an example of a statistical search & retrieval technology, and Precision Content Retrieval as an example of a natural language-based retrieval technology.

Oracle Text with word query had the best performance, followed by Precision Content Retrieval, followed by WebGlimpse. The document collection used in this experiment was the Bush Public Papers. It consists of approximately 5200 documents (files). To effectively evaluate text-based retrieval systems using ad hoc queries, one needs to use a much larger set of documents. For this, and other reasons, the results of the experiment are not conclusive. The experiments should be conducted again using a larger corpus such as the Bush administration's PC files systems, which are an order of magnitude larger than the Bush Public Papers, or the Bush administration's e-mail records, which are three orders of magnitude larger.

Requirement 20: *An archival processing system needs the capability to support archival review for FOIA and PRA access restrictions.*

Review of Presidential electronic records for FOIA and PRA access restrictions is an intellectually demanding task that requires page-by-page review of the records. Due to the increasing volume of electronic records from all branches of government, the requirement to review these records, and the cost of the limited human resources that can be applied to the review process, one should consider whether there are advanced technologies (e.g., information extraction, natural language processing, case-based reasoning and machine learning) that could increase throughput by supporting the review process.

In Phase II of the PERPOS project, tools are being developed to automatically analyze the content of text files, to extract information from the files and to mark up the extracted information in a copy of the text file. The markup includes identification of such information as persons' and organizations' names, locations, dates, addresses, telephone numbers and social security numbers. Events such as proposing legislation, enacting legislation, or signing legislation will also be automatically identified [9].

Members of the project staff are being trained in FOIA and PRA review. Those staff members will then analyze the kinds of knowledge and reasoning that are needed to review records for these types of restrictions. A prototype Review Assistant will be designed and developed that utilizes these kinds of knowledge and the information extraction technology to identify probable PRA and FOIA access restrictions. The Review Assistant will be incorporated into the Review function of the Archival Processing Tool. Experiments will be conducted to evaluate and refine the performance of the prototype Review Assistant.

Requirement 21: *An archival processing system needs the capability to summarize record series and extend directory (folder) titles.*

The Archival Processing Tool enables an archivist to open, withdraw, or redact electronic records, to extend or create titles for directories, and to summarize the contents of record

series. However, until archivists find the time to process the files, the have a low degree of intellectual control over the records.

In Phase II of the PERPOS project, information extraction technology coupled with summarization technology is being investigated as a means to identify document types (e.g., memo, correspondence, press release, agenda, speech), extend 8-character directory names to more descriptive folder titles, and to create preliminary scope and content notes for record series [9]. This should provide archivists with improved intellectual control over archival holdings at an earlier date. When an archivist does process the record series, this same technology can support the archivist's description task, and thus potentially reduce archival workload and increase throughput.

Summary

When this project began the requirements for processing records created on Personal computers and saved in PC files systems was not well understood. Development of the PERPOS prototype(s) and laboratory experiments with them support the identification of new functional requirements for archival processing of PC file systems and the reduction of risks that would be encountered if these requirements were not satisfied. A number of advanced technologies, for example, the Secure Hash Algorithm 1, the NSRL Reference Data Sets, information extraction and case-based reasoning, are being applied to the satisfaction of some of these requirements.

The Archival Processing Tool prototype provides an experimental environment for further exploration of the research issues involved in processing PC records. Among the research issues under continued investigation is whether the techniques currently utilized are scalable to large volumes of PC records. It is anticipated that the use of the Archival Processing Tool by archivists at the Bush Presidential Library to process the contents of the Bush hard drives will result in a better understanding of the archival and technological issues involved in achieving archival control of records created on personal computers.

References

1. Underwood, W. E. (2002a) A step towards a logical theory of record integrity and authenticity. Georgia Tech Research Institute, Georgia Institute of Technology, Atlanta, March.

2. Underwood, W. E. (2002b) A formal method for analyzing the authenticity properties of procedures for preserving electronic records. *Proceedings of the 2002 International 'Conference on Digital Archive Technologies (ICDAT2002).* December 19-20, Academia Sinica, Taipei, Taiwan, pp. 53-64.

3. Underwood, W.E., Hayslett-Keck, M. and Laib, S. (2003) The Archival Processing Tool (APT): User's Guide Version 2.05. PERPOS Technical Report ITTL/CSITD 02-02, Georgia Tech Research Institute, Georgia Institute of Technology, Atlanta, August.

4. Underwood, W., Laib, S., Hayslett-Keck, M., Underwood, M. and Roberts, D. (2002) Study of the Use of the Presidential Electronic Records PilOt System (PERPOS): Final Report, PERPOS TR ITTL/CSITD 02-04, Georgia Tech Research Institute, Georgia Institute of Technology, December.

5. Underwood, W. E. and Underwood, M. G. (2002) Evaluation of Document Retrieval Technologies to Support Access to Presidential Electronic Records. Technical Report ITTL/CSITD 02-03, Georgia Tech Research Institute, Georgia Institute of Technology, Atlanta, December.

6. Underwood, W. E. and Laib S. L. (2003) Identifying File Types and Document Types and Filtering Records from Legacy PC File Systems. Working Paper, Georgia Tech Research Institute, Georgia Institute of Technology.

7. CCSDS (2002). The Data Description Language EAST Specification (CCSDS 644.0-b-2). Blue Book. Issue 2. November. Also ISO 15889:2000 Space data and information transfer systems – Data description language – EAST Specifications.

8. Baxter, R., Carroll, R, Eklund, D. J., Gibbins, B. Virdee, D., Wen, T. BinX—A tool for retrieving, searching, and transforming structured binary files.

9. Underwood, W. (2003) Presidential Electronic Records PilOt System—Phase II: Quarterly Report (July-Sept 2003), Georgia Tech Research Institute, Georgia Institute of Technology, October.