Presidential Electronic Records PilOt System (PERPOS): Phase I Report

William E. Underwood
Mark R. Kindl
Matthew G. Underwood
Sandra L. Laib

August 2001

Computer Science and Information Technology Division
Information Technology and Telecommunications Laboratory
Georgia Tech Research Institute
Georgia Institute of Technology

The sponsors of this research are the Army Research Laboratory (ARL) and the National Archives and Records Administration (NARA). The findings in this paper should not be construed as an official ARL or NARA position unless so indicated by other authorized documentation.

Executive Summary

The primary objective of the Presidential Electronic Records Pilot System (PERPOS) project is to apply information technology to support the decisions that archivists must make when processing digital records, particularly those created on personal computers.

A Filing System Filter (FSF) was prototyped that separates the files of a DOS or Windows filing system into user-created files and system and software application program files. Experimental results show the filter to be effective and risk-free—there is no risk that a user-created file will be lost. This tool substantially reduces the workload in cases that user-created files must be separated from other files in a filing system.

Digital record series from the Bush hard drives are packaged in a JAR files. Attributes associated with the files by the record creator, such as author, subject and archival date, and attributes associated with the file by archivists, such as file format, document type and the date a file was closed, opened or redacted, are also stored in the JAR. This is a highly efficient method of storage, retrieval and transmission.

Presidential Libraries must be able to ensure the authenticity of digital records preserved in their archives. It was demonstrated how the message digests and digital signature services of JAR files supported verification of the integrity and authenticity of digital records stored in a JAR.

An Archival Review Tool (ART) was designed and prototyped that enables an archivist to separate files that can be opened to public access from those that must be withdrawn or redacted due to FOIA or PRA exemptions. Information extraction technology combined with models of archivists knowledge of FOIA and PRA review are being investigated as means of facilitating the review process.

A software tool was designed and prototyped to support archival arrangement and description. The tool allows an archivist to arrange records series within the organizational structure of the records creator. It also allows the archivist to describe the scope and content of the record series and its filing arrangement. Information extraction and machine learning technologies are being applied to support archivists in describing digital record series, folders and items.

Experiments are being conducted to evaluate document retrieval technologies that could support identification of files in large unprocessed collections that might be relevant to FOIA requests. The objective of these experiments is to determine the relative effectiveness and scalability of document retrieval technologies.

Keywords: digital government, digital archives, information filtering, information extraction, text summarization, record integrity, record authenticity

Introduction

Background

The National Archives and Records Administration (NARA) is responsible for preserving the records of the Executive, Congressional, and Judicial branches of the Federal Government. It ensures continuing access to essential evidence that documents the rights of American citizens, the actions of Federal officials, and the national experience.

NARA's holdings of Federal records, which are greater than 20 million cubic feet, include textual materials, reels of motion picture film, maps, charts, and architectural drawings, sound and video recordings, aerial photographs and still pictures. They also include over 100,000 electronic files (500 gigabytes). NARA also administers the Nixon Presidential Materials Staff and 10 Presidential libraries, which preserve the papers and other historical materials of all past Presidents since Herbert Hoover.

NARA faces the problem that many digital records created today may not be readable in the future due to the obsolescence of the computer and software systems on which they depend. NARA also faces the prospect of being overwhelmed by the sheer quantity of email, word-processing documents, and other electronic records in multiplying formats being generated or received by the Government.

To gauge the magnitude of the electronic records challenge facing NARA, consider the Clinton administration's email. According to the Office of the Inspector General of NARA [1], about 40 million email messages were to be transferred to NARA at the end of the Clinton administration. This would be NARA's largest acquisition of electronic records. According to that report, five archivists will be initially assigned to process these email messages. Archival processing consists of arrangement preservation, review and description activities. Presidential records must be reviewed page-by-page for restrictions on release to the Public under the Freedom of Information Act and the Presidential Records Act. Let us assume that, on average, an email message and possible attachments is one page in length, and that it takes an archivist one minute to review one page. Five archivists working 2000 hours per year, or 600,000 minutes per year can review 600,000 pages per year. It will take 66.67 years for five archivists to review the Clinton email. Archivists need tools to support their processing decisions, to increase processing performance, and decrease human workload.

The growth of web access and digital government, and the availability of electronic access under the Freedom of Information Act, as amended by the Electronic Freedom of Information Act further increase demands for online records and services. NARA must preserve electronic records in a way the makes them available in systems through which users can locate needed records and retrieve and read them.

NARA must also be able to ensure the authenticity of electronic records. It must assure that provisions are made for controlling the transmission, storage and maintenance of

electronic records to guard against tampering and that preservation actions do not compromise the integrity of records in the digital archives.

NARA has partnered with other Federal agencies and academic research institutions to develop an Electronic Records Archive (ERA) that will enable NARA to preserve and make available millions of records born digital in the Federal Government. NARA expects to have a pilot version of the ERA in operation by 2004 or 2005 [2].

Project Objectives

Two important collections of digital records that have been acquired by NARA are the personal computer files created in the White House Offices during the Bush administration and those from PCs used by Office of Independent Counsel (OIC) di Genova. Archivists do not have a set of tools to support archival processing of digital records created on personal computers. The general objective of first year research was to prototype and pilot test tools for gaining intellectual and physical control of the Bush Presidential personal computer records. Specific research objectives include: (1) definition of the functional and information requirements for archival processing and storage of presidential electronic records, (2) identification of information technologies that can be used to meet these requirements, (3) application of these technologies in software prototypes and evaluation of their performance, and (4) demonstration of technologies for meeting non-functional requirements such as scalability, reliability, portability, and availability.

Archival Processing

Managing Presidential Library textual record collections involves the activities of accessioning, processing and accessing records. *Accessioning* is the procedure by which a Presidential Library takes physical and legal custody of records or papers and establishes initial intellectual control over the material. *Access* is the process of searching for archival documents, retrieving them from their storage location, checking them out to a researcher, and returning them to their stack location. It also includes the procedure for responding to FOIA requests. Systematic processing of textual records involves the activities of arranging, preserving, reviewing and describing records as shown in Fig. 1.

Arrangement is the proper ordering of materials within a collection and the placement of materials in archival storage areas. *Preservation* consists of refoldering, reboxing, annotation, replacing fasteners, and preservation photocopying. *Review* is the process of identifying and segregating materials that are to be temporarily closed to researchers. *Description* is a procedure for providing intellectual control over the Library's holdings and for facilitating reference services.

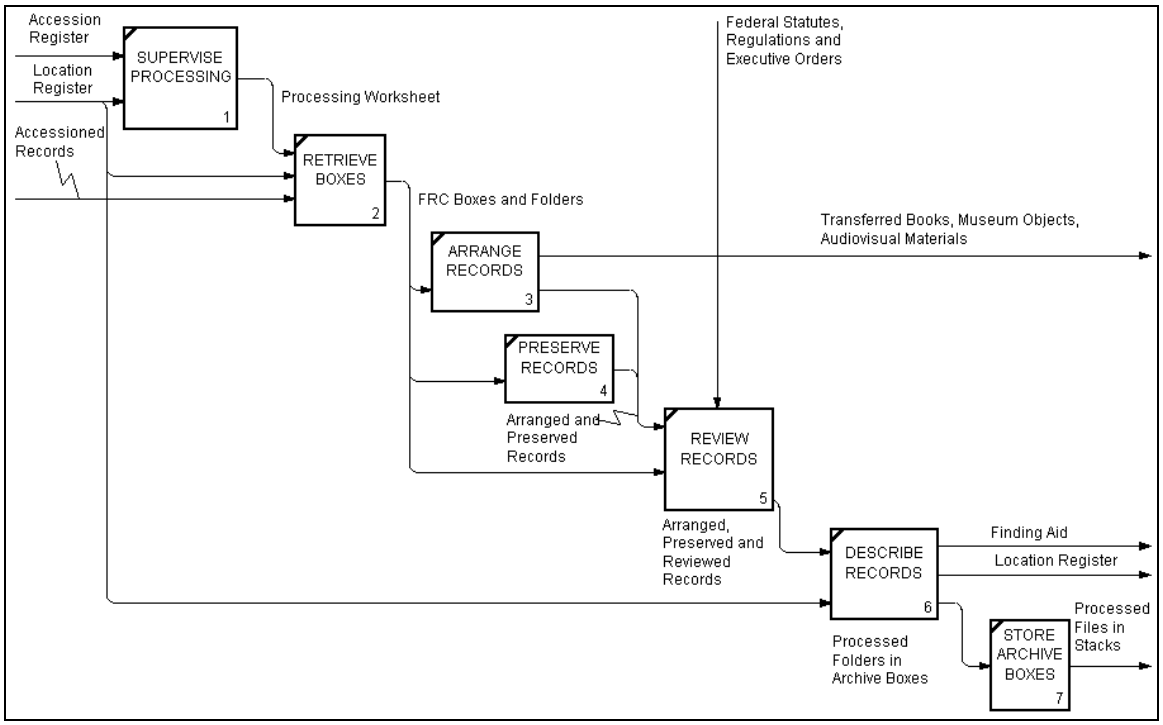

Figure 1. The Activities involved in Systematic Processing of Textual Records.

To achieve the project objectives, current archival and processes were analyzed. Opportunities were identified for reengineering archival processes to support archival decision-making and to improve workflow. Software tools were designed using use case analysis and software prototypes were developed. The performance of the prototypes is experimentally evaluated. The tools will be deployed, used and evaluated in a pilot system for processing of Presidential electronic records.

Research Progress and Results

File System Filtering to Support Preservation Decisions

The user-created PC files from the Bush hard drives must be preserved. An archivist must decide which files are system or software application files used to create records, and which are user-created files that need to be preserved. For an archivist who is not well acquainted with the DOS and Windows operating system files and the legacy computer applications of the period 1988-1992, this would be a formidable and time-consuming task.

A filing system filter (FSF) was prototyped that separates the files of a DOS or Windows filing system into user-created files and system and software application program files. Figs. 2 and 3 shows the user interface of the filing system filter.

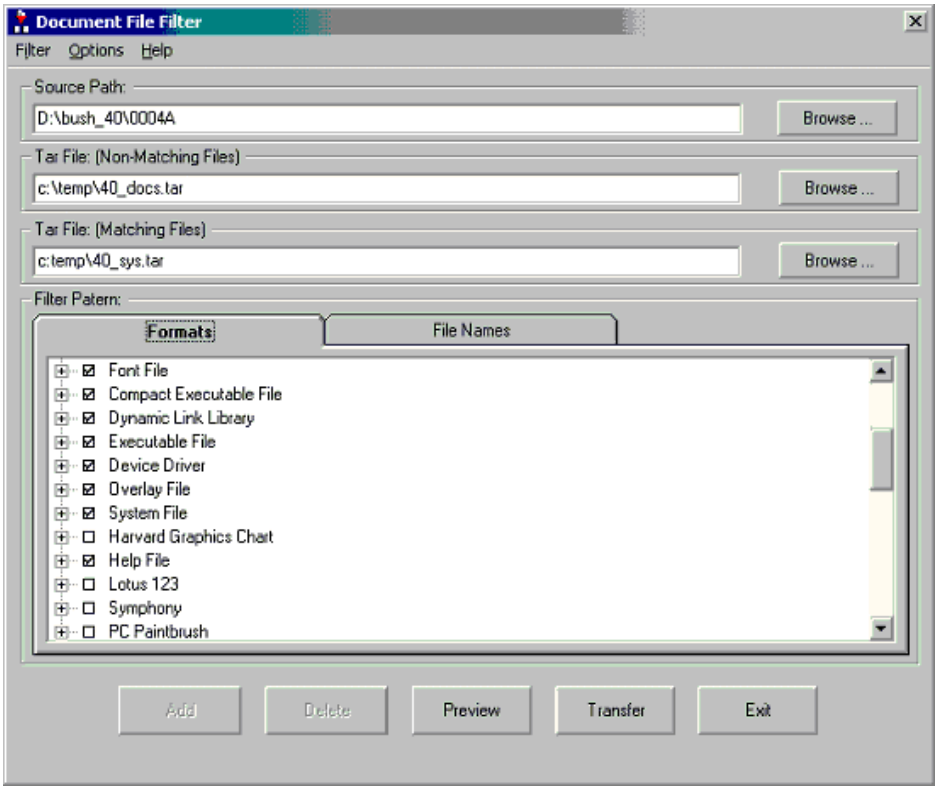

Figure 2. File Formats for the Filter Pattern

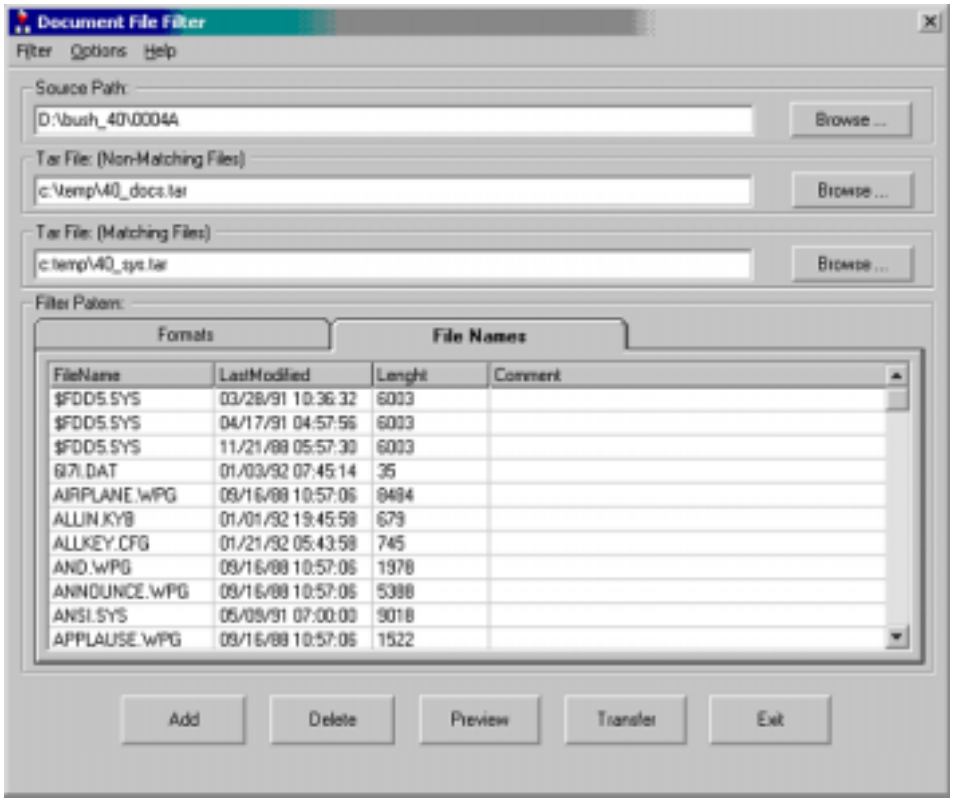

Figure 3. File Attributes of Filter Pattern.

Fig. 2 shows the possible file formats that can be used to define the filing system filter. Fig. 3 shows attributes of files that are in the filing system filter. A *filter* is a pattern that is matched against input data. Only data that matches the pattern is passed through the filter. A *filing system filter* sorts the files in a filing system into two sets, one defined by file format or file attributes an/or bitstring, the other set, being the complement of the first, is said to be *stopped*. The filing system directory structure (paths to the files) is maintained for each of the sets.

The filter checks file formats, and in the cases that a file type cannot be determined from its format, checks the file name, date of creation and length. An archivist can view the results of filtering as shown in Fig. 4.

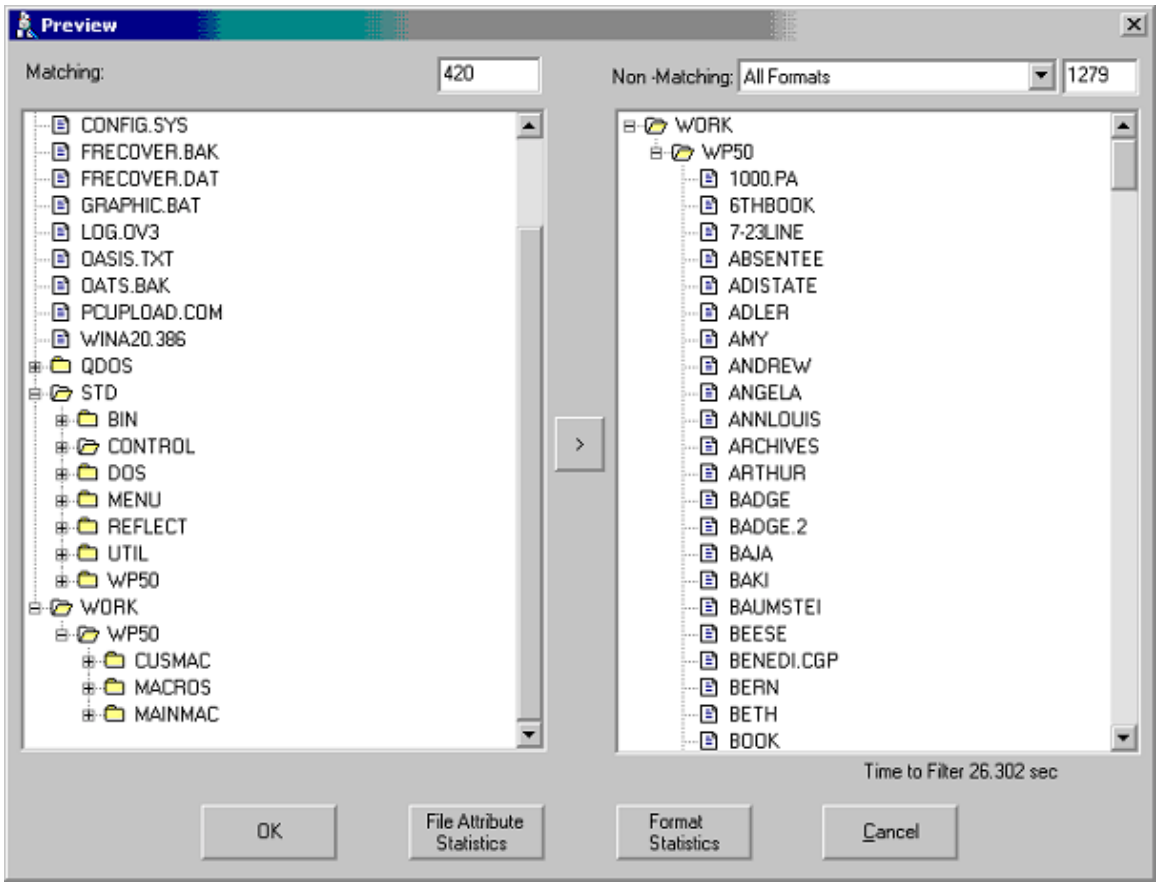

Figure 4. Passed and Stopped Files.

The names of files (or directories on the path to the files) that passed the filter are shown in the left frame. They are system and software applications files whose file format was selected in defining the filter as well as files whose file format was not definable, but which could be defined via file attributes or bitstring. The name of files that were stopped by the filter are shown in the right frame. They are user-created files that need to be preserved.

An archivist can highlight any of the file names and display the file using a set of software viewers. If there is a file in that was passed in the right frame that is actually a system or software application, it's file name can be highlighted and the button (< >) between the frames selected to move the file to the set of passed files shown in the left frame. The file name's path will also be moved with the file. Similarly, files that were passed that should have been stopped can be moved from the left frame to the right.

This mechanism accounts for how the file filter itself is defined. When a file in the right frame is determined by viewing it to be a system or software application that should not be preserved, and the archivist moves it to the left frame, then its attributes are added to the filter so it will be passed in future filtering. Alternatively, if the file is determined by viewing it to be a system or software application, an archivist can return to the file format screen and select its file format to be a part of the filter, so it will be passed in the future.

The file format statistics button on the preview screen shown in Fig. 4 shows the kinds of file formats occurring the stopped files, e.g., WordPerfect 5.1, Lotus 1-2-3, and the number of occurrences of each file format.

The system and software application files and the user-created files can be saved to different locations. However, this will not occur until the archivist returns to the previous screen and selects transfer. Copies of all files are maintained. A copy of all unique system and software application files is created for review to ensure that no user-created files that should be preserved were incorrectly identified as system or software applications that do not need to be preserved.

The filter was run on an 800 Mhz Intel/Windows platform and applied to the contents of 93 of the 500 Bush hard drives. There were about 35 file formats in the filter, as well as attributes of about 3000 files that could not be defined by file format. The filter correctly identified 51,541 system and application files (1.126 GB) and 30,897 user-created files (279 MB). The time for reading and filtering the filing system of a drive ranged between 8.8 and 185.5 seconds and averaged 43.4 seconds. The filter processed approximately 21 files/second [3]. This tool substantially reduces the workload in cases that user-created files must be separated from other files in a filing system.

Digital files are preserved in their original file formats and a commercial-off-the-shelf (COTS) software program is used to view the files in their legacy (obsolete), proprietary file formats. As computer and software systems evolve, it will be necessary to migrate the software viewers to new systems, or explore other preservation strategies, e.g., conversion to standard formats or hardware emulation.

Preserving Authentic Records in Archival Information Packages

Within the decade, NARA's Electronic Records Archives will need to store a billion or more digital records. For efficiency in storage and transmission, it is necessary to store many digital files as a single file. Furthermore, it is desirable to store attributes associated

with the files by the record creator, such as author, subject and archival date, and attributes associated with the file by archivists, such as file format, document type and the date a file was closed, opened or redacted, in the same single file. It is not practical to store such information about every file in an on-line database.

The problem of deciding whether archived digital records are authentic is dependent on being able to verify the provenance (origin and custodial history) and integrity of the digital records since the time of their creation. It is complicated by the possible need to convert files from their original formats to standard formats with an assurance that the documentary form and content of the digital record is not compromised. Hence, record authenticity cannot be dependent on data integrity alone. Without a procedure for preserving authentic digital records that is provably correct, an archivist would have to decide on a case-by-case basis whether a record was authentic.

The Consultative Committee for Space Data Systems (CCSDS) has developed a Reference Model for an Open Archival Information System (OAIS) that is being considered as an ISO standard [4]. It includes a logical model for the information in an open archival information system. The logical model for supporting preservation of information is an information package. An *information package* is a container that contains two types of information objects, content information and preservation description information. There are three subtypes of information packages in OAIS: the submission information package (SIP) used to transport archival information from the producer to OAIS, the Archival Information Package (AIP) used to structure and store the OAIS holding, and the Dissemination Information Package (DIP) used to transport requested information from OAIS to the consumers.

An AIP is an information object that contains other information objects. It contains packaging information, a package description, content information, and preservation description information (PDI). The PDI contains reference provenance, context and fixity information.

The Java Archival (JAR) file format was investigated as a method for implementing an AIP.The JAVA archive (JAR) format is a platform-independent file format that aggregates many files into one. JAR was developed so that JAVA applets and their components could be bundled into a single file (package) and quickly downloaded to a browser in an http transaction. The JAVA application launcher can launch one of the files, e.g., a class with a method, in the package. A JAR provides the capability to verify the origin of the components in the JAR so that only programs authored by persons or organizations trusted by the user will be executed. JAR is an open industry standard [5].

The procedure for preserving a series of digital records in a JAR is shown in Fig. 5. The procedure shown in Fig. 6 is used to verify the authenticity of the files preserved in a JAR.

1. Create a JAR file that contains the path/filenames and files of a record series and a manifest file that contains the path/filename of each of the files.

2. Create a message digest for each file and in the manifest file associate it with the path/filename of the file.

3. In the manifest file, associate the name of the record creator and archival date of each file with its path/filename.

4. Create a message digest for the entire manifest file (the message digests of each of the files in the JAR and the metadata stored with the message digests) and store it in a signature file.

5. Sign the manifest file using an archival private key and the message digest for the manifest file. Insert the archival public key certificate file in the META-INF directory.

Figure 5. Procedure for Preserving an Authentic Record Series in a JAR.

1. The public key in the certificate in the Signature File is used to verify that the digital signature applied to the message digest of the manifest is that of an archival authority of the record creator's organization.

2. The message digest for the manifest file is re-computed and compared against the message digest in the signature file.

3. To ensure that the files in the JAR haven't changed since the JAR was signed, the message digest of each of the files in the JAR is re-computed and compared with the corresponding message digest in the manifest file.

Figure 6. Procedure for Verifying the Authenticity of a Record Series in a JAR.

The concepts of data, document, record and record series integrity and authenticity were formally defined. Formal methods were used to demonstrate how the integrity and authenticity of records stored in a JAR could be verified [6].

Information Extraction to Support Review

Prior to releasing records to the public, archivists at the Bush Presidential Library must review all records for FOIA exemptions and PRA restrictions. FOIA and PRA review is a bottleneck because the current procedures are manual and intellectually intensive. Many records in a Presidential Library remain unavailable, because there has not yet been sufficient time for page-by-page review to locate PRA restrictions or FOIA exemptions that may apply. A tool to help archivists find possible FOIA exemptions or PRA restrictions is needed. An Archival Review Tool (ART) was designed and prototyped that enables an archivist to separate files that can be opened to public access from those that must be withdrawn or redacted. Fig. 1 shows the user interface to ART [7].

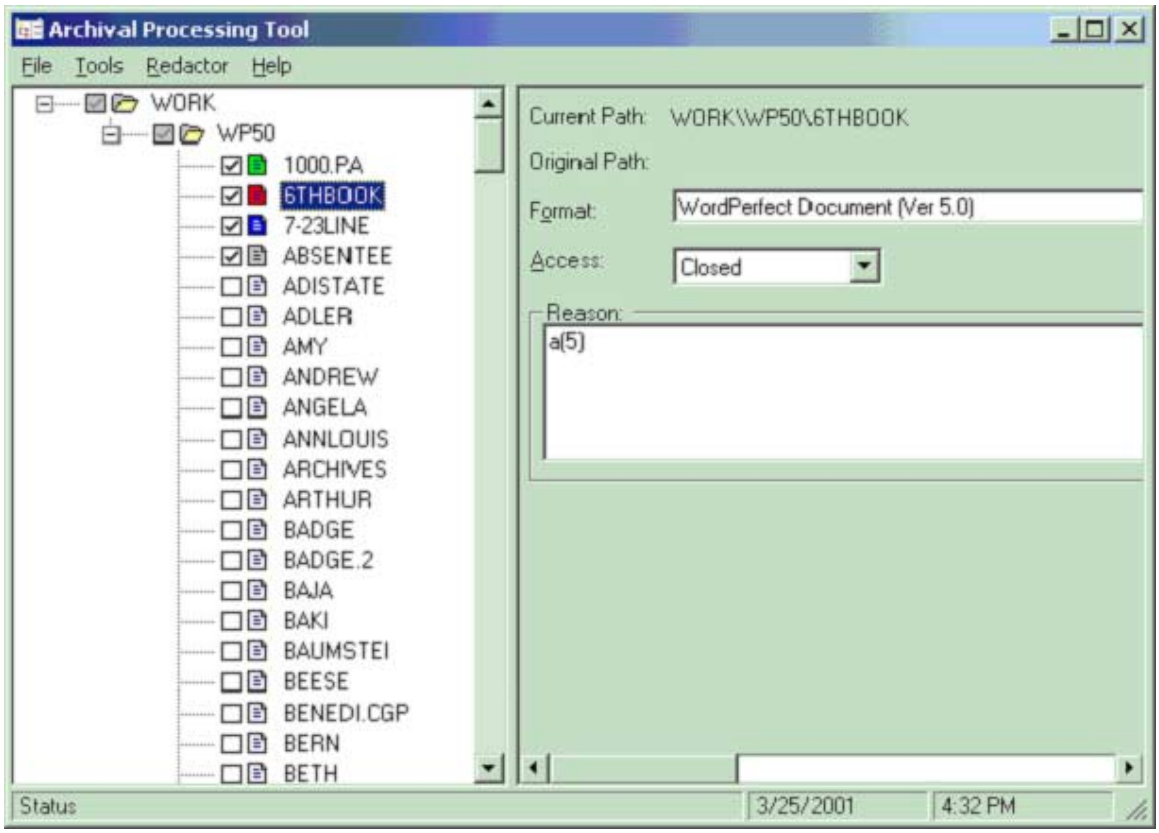

Figure 7. User Interface to the Archival Review Tool.

An archivist selects a collection (or series) of records to be reviewed. The collection may be in a directory or contained in a JAR file. The directory structure of the loaded collection is displayed in the left window of the user interface. An archivist can open or close directories and display files with a set of file viewers.

If the archivist determines that there are no FOIA exemptions and PRA, he can select *Open* from the Access field and the file will be marked as reviewed (with a check mark along side the name of the file) and marked as open (with a green document icon). If an archivist determines that FOIA exemptions or PRA restrictions on the contents of the document preclude release of the document, then he selects *Closed* from the pull down list of the Access fields. When the archivist indicates that the file is to be closed, the Reason Withdrawn window opens to show FOIA exemptions and PRA restrictions. The archivist selects the reasons for withdrawal (closing of the file). The file is marked as reviewed and marked as closed (with a red document icon).

For files withdrawn in their entirety, the following information will be saved: date of closure, the document type of the closed file (memo, letter, agenda), correspondent's titles, subject, restriction authority (reason withdrawn), and archivists initials.

An archivist may discover that while there are FOIA exemptions or PRA restrictions that apply to a displayed file, there are substantial portions of the text that could be released to

the public. An archivist can redact portions of the digital document, by selecting a redaction tool from the Redactor pull-down menu. For instance, if he pick an image redactor, a document image of the displayed file will be created and it will be displayed in the Imaging for Windows tool. Using this too, text can be redacted with a solid box. A rubber stamp can be used to indicate the reason for redaction. When the document is saved, it is saved as a tif file with an additional file extension of rdt. In the directory in which the original file is stored. One returns to the ART window. The original file is marked as closed (with a red icon). The redacted image is marked as redacted (with a blue icon). The reason for redaction will have been copied into the reasons for withdrawal.

If upon reviewing a file, an archivist determines that the file needs to be transferred to some other collection (for instance, a copyrighted software application might need to be transferred to the library, or a system or software application file not created by an individual or office, needed to preserved in another file), the archivist selects *Transfer* from the pull down list for the access field. The document icon will be colored to indicate that the file is to be transferred.

When the work is saved, it will be packaged in a JAR with metadata that was created during the session. This metadata will be reloaded in subsequent sessions. This copy is called the LICON or working copy. The original unreviewed copy is not replaced. It is the preservation copy.

The LICON copy is not available to the public because it contains closed files. For reference purposes, a copy of the collection is needed that corresponds to what for paper files is called the open files. It contains opened files and redacted files. An option on the file pull-down menu allows the archivist to create a reference copy.

The Archival Review Tool as described so far, supports manipulation of the digital files, but does not support archival decision-making. However, a decision support tool to support identification of possible FOIA exemptions and PRA restrictions is envisioned that identifies the kinds of information in text that archivist use to make review decisions.

This can be accomplished using information extraction technology. Information extraction (IE) is a procedure that selects, extracts and combines data from text in order to produce structured information. The Message Understanding Conferences (MUC) [8] have been the driving force for developing this technology. The MUC specifications for various IE tasks have become *de facto* standards in the IE research community. MUC divides IE into distinct tasks, namely, NE (Named Entity), TE (Template Element), TR (Template Relation), CO (Co-reference), and ST (Scenario Templates). The named entity task is to identify all named locations, named persons, and named organizations, time (dates and times) and numeric expressions (monetary amounts and percentages). Named entities can be marked up using XML. Performance in identifying named entities can be measured using the F-measure , which is a combination of precision and recall measures [9].

Fig. 8 shows an actual Presidential record in which a protype PRA checker has identified and highlighted passages of the document that could possibly be exempt under PRA restrictions a(2) appointments to Federal Office, and a(5) Confidential advice to the President.

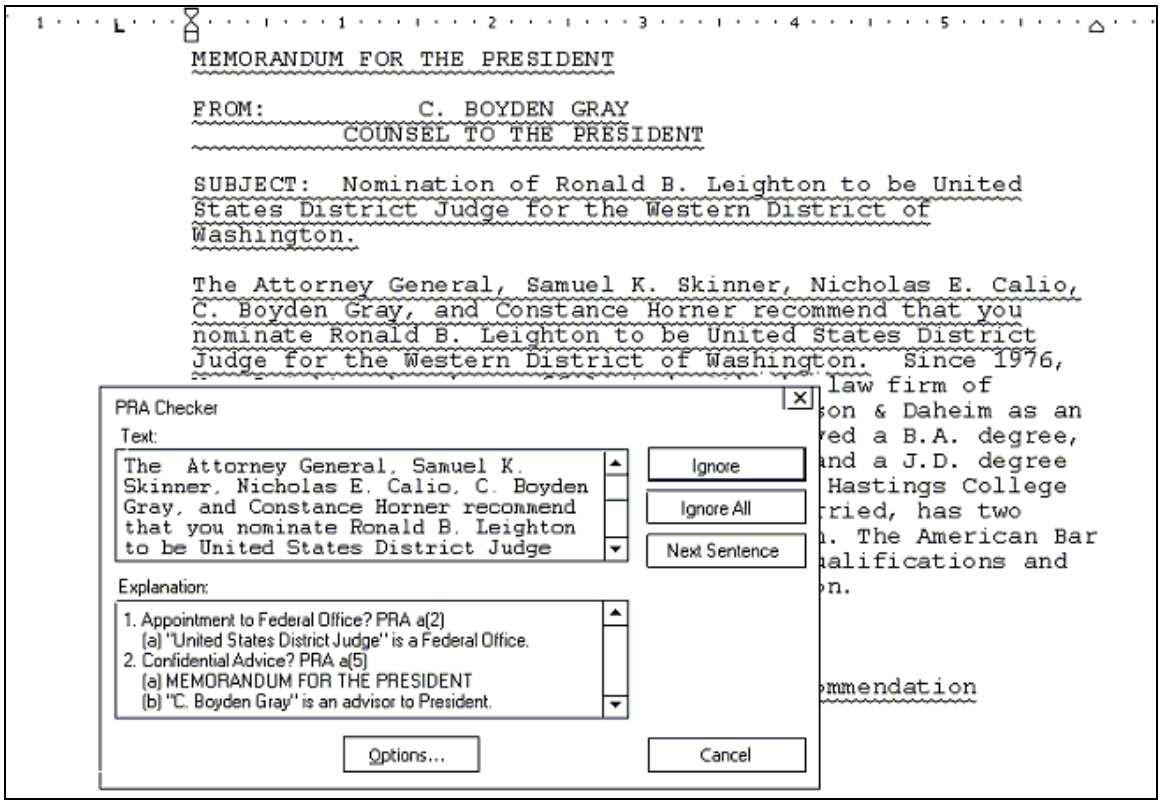

Figure 8. PRA Checker.

The statutory description of such restrictions is very general. The operational definition of the restrictions is based on case-by-case PRA and FOIA review experience. For instance, recommendations of someone for a Federal office may not be restricted, if not from a confidential advisor. Some advice is not confidential, e.g., "We think you should serve chicken." Such knowledge is being captured by recording reviewer's explanations of the their decisions. Rules are then formulated that refer to information extracted from the document and that reflect the case-based knowledge of its relevance to PRA restrictions and FOIA exemptions [10].

Information Extraction to Support Arrangement and Description

An archivist must ensure that the records in a record series are properly ordered and the series is described so that it can be located. A software tool was designed and prototyped to support archival arrangement and description. The tool allows an archivist to arrange records series within the organizational structure of the records creator. It also allows the archivist to describe the scope and content of the record series and its filing arrangement,

e.g., file folders in a hierarchical filing system. A tool for supporting archival arrangement and description is shown in Fig. 9 [11].

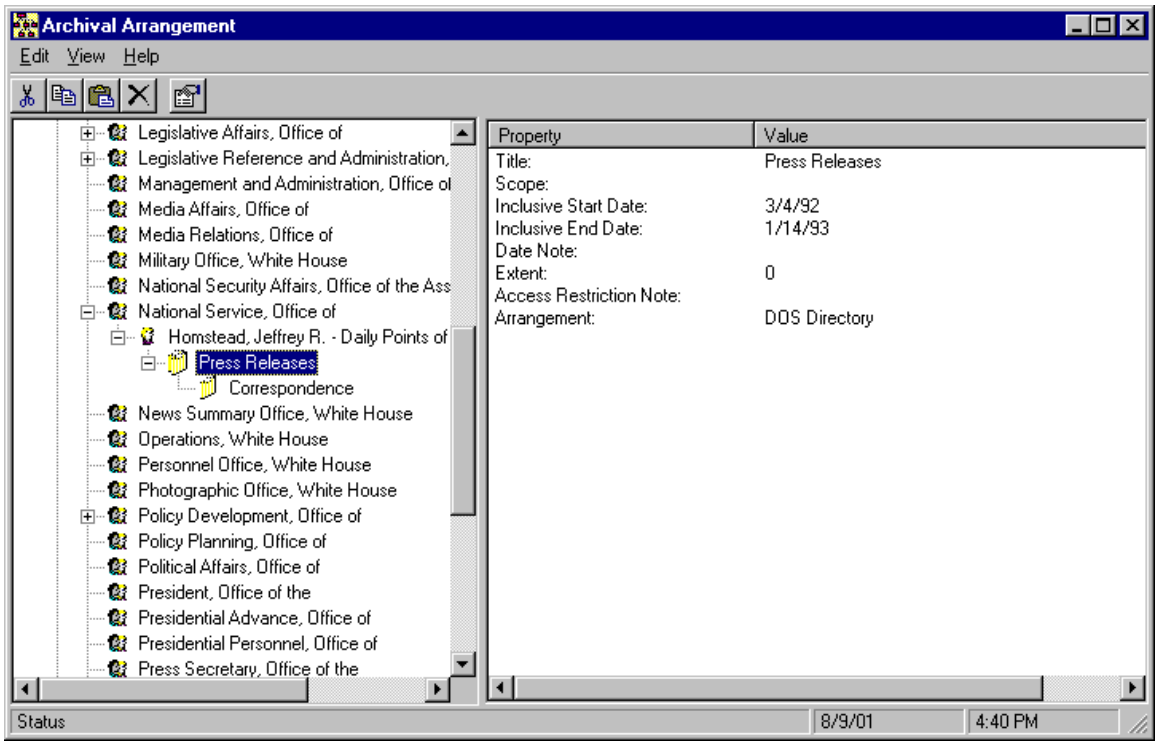

Figure 9. User Interface to Archival Arrangement and Description Tool.

By convention, the scope of a record series should provide information on the specific activity or activities generating the records and the period of time during which the records were generated. The description of the content of a record series includes information about the structure of the record series and documentary forms (e.g., memorandum, letter, transcript of press conference or interview with News Media, press release, newswire, President's remarks, name list, address list, calendar, agenda, schedule) of the records. An archivist must determine all of the documentary forms in the series.

A tool is being developed that can determine the documentary forms of a digital records, display these to the archivist and associate the name of the documentary form with the file in the manifest file of the JAR containing the records. It is based on method for identifying named persons, named organizations, dates, and locations described in the previous section and machine learning.

A sample of memoranda, letters, addresses, appointments, press conferences, etc. are manually marked up using XML tags to indicate the names of persons, organization names, dates, and locations. A program was developed that can take a sample corresponding to a particular document type and learn a regular expression that describes the structure of that document type.

Given a sample document files that have not been marked up, a program for named entity extraction automatically identify persons names, organization names, street addresses, dates, etc. and markups the documents. Another program then applies the regular expressions characterizing the document types to determine the most probable document type. These are then shown to the archivist, who can use the information in describing the content of a record series.

The next experiment will involve taking a sample of unmarked up documents of a particular documentary form, e.g., memoranda, and attempting to automatically mark up that sample using the information extraction method for named entities. This will reduce the effort required to learn new documentary forms.

Response to FOIA Requests for Unprocessed Digital Records

When a Presidential Library receives a request for unprocessed or closed records under provisions of the Freedom of Information Act, the Library must respond within the 20-day statutory time period as to the number of records that it believes are responsive to the request. To respond to FOIA requests for unprocessed paper records, the Bush Presidential Library created a database of file folder titles that could be searched for relevancy to a request. A similar approach could be used for PC files by creating a list of DOS directory names. However, the directory names are usually too cryptic (8 or less characters) or not indicative of the content of the directory, e.g. "Work". Automated finding aids are needed to support response to FOIA requests for records in unprocessed digital collections.

Experiments are being conducted to evaluate document retrieval technologies that could support identification of unprocessed files that might be relevant to FOIA requests. The objective of these experiments is to also to determine the relative effectiveness and scalability of document retrieval technologies from small to medium and large, heterogeneous collections such as the Clinton email files. This involves measures of time to index, size of index, response time to query, and average recall and precision.

In the first experiment, three document retrieval systems, representing three different information retrieval technologies were evaluated—WebGlimpse, Oracle Intermedia and Sun Laboratory's NOVA. WebGlimpse [12] was selected because it is the search engine currently used to search the online collection of Bush public papers. The search engine for WebGlimpse is Glimpse whose query language includes Boolean expressions, approximate matching, and regular expressions [13]. WebGlimpse does not provide relevance ranking.

The Oracle "word query" is a component of the interMedia Text search and retrieval tool within the Oracle8i database management system [14]. Using a standard inverted index to files stored or referenced within a database, a word query searches for the exact words and phrases requested. Boolean syntax, proximity constraints, value weighting, word stemming, wildcard matching, and an accumulation technique may be used to control the results of a query. Boolean syntax offers the user the greatest control by enabling the

presence of one or more words as a requirement, thus ruling out many of the indexed documents as non-pertinent. Proximity constraints require that certain words be NEAR to others or that they must all be WITHIN a sentence, paragraph, or definable section (e.g. <Title> <Date>). Value weighting allows the importance of a word or the importance of a part of the query to be modified, thereby percolating the more relevant documents to the top of a scored list. Accumulation allows the user to construct a list of words or phrases that are likely to occur within documents deemed as pertinent. InterMedia will include all documents that contain any of the elements of this list, and will place higher scores on documents that contain multiple elements of such.

Sun Laboratory's NOVA system determines the noun phrases and verb phrases in a document. It uses lexical subsumption rules, that is, rules that determine whether a phrase is more general that another phrase to create a conceptual index relating the noun phrases and verb phrases appearing in a collection of documents. It also parses the noun phrases and verb phrases occurring in a query and uses lexical subsumption to create a conceptual representation of a query. It uses a penalty scoring method to identify the passages most relevant to a query [15].

The collection used in this experiment was President Bush's Public Papers [16]. This small collection consists of about 5100 html documents (35 megabytes). Fifty-one statements of information need (topics) were constructed from actual FOIA requests processed by the Bush Presidential Library. One query for each topic was submitted to each of the three systems. The query constructed for each topic could only use the terms or phrases that occurred in the topic, the statement of information need. The results for each topic for all three systems were pooled. The documents relevant to a query were determined by reviewing the documents in the pool.

In this first experiment, the average recall and average precision measures for the three systems are unusually high as compared to TREC ad hoc document retrieval results. One reason for this is the relatively small collection of documents searched in this experiment. Another is that documents relevant to a query were determined from review of the results of just the three systems. There are documents that are relevant to the queries that are in the collection, that were not in the results set for any of the three systems [17].

Another experiment will be conducted in which a second query can be formulated for each topic based on information derived from passages or documents retrieved using the first query. This will in some cases increase the number of relevant documents in the pool for each topic. It also represents the strategy that an archivist will need to use to increase their confidence that have done their best to locate all relevant documents.

This experiment will be conducted again for the Bush PC files (estimated to be about 150,000 files and 1.5 gigabytes) when they have been filtered from the DOS filing systems. Additional document retrieval systems will be included in those experiments.

Summary of Progress and Results

A Filing System Filter (FSF) was prototyped that separates the files of a DOS or Windows filing system into user-created files and system and software application program files. Experimental results show the filter to be effective and risk-free—there is no risk that a user-created file will be lost. This tool substantially reduces the workload in cases that user-created files must be separated from other files in a filing system.

Digital record series from the Bush hard drives are packaged in a JAR files. Attributes associated with the files by the record creator, such as author, subject and archival date, and attributes associated with the file by archivists, such as file format, document type and the date a file was closed, opened or redacted, are also stored in the JAR. This is a highly efficient method of storage and transmission since attributes of files only need to be in online storage when they are being processed.

Presidential Libraries must be able to ensure the authenticity of digital records preserved in their archives. It was demonstrated how the integrity and authenticity of records stored in an archival information package (e.g., JAVA archival format) could be verified.

An Archival Review Tool (ART) was designed and prototyped that enables an archivist to separate files that can be opened to public access from those that must be withdrawn or redacted due to FOIA or PRA exemptions. Information extraction technology combined with models of archivists knowledge of FOIA and PRA review are being investigated as means of facilitating the review process.

A software tool was designed and prototyped to support archival arrangement and description. The tool allows an archivist to arrange records series within the organizational structure of the records creator. It also allows the archivist to describe the scope and content of the record series and its filing arrangement. Information extraction and machine learning technologies are being applied to support archivists in describing digital record series, folders and items.

Experiments are being conducted to evaluate document retrieval technologies that could support identification of files in large unprocessed collections that might be relevant to FOIA requests. The objective of these experiments is to also to determine the relative effectiveness and scalability of document retrieval technologies.

During the next phase of research, the PERPOS decision support tools will be evaluated in case studies at the Bush Presidential Library and NARA's Center for Electronic Records. During Phase I research, a number of additional opportunities were identified for supporting decisions during archival processing. These research issues and approaches to investigating them are described in another report [18].

References

1. NARA, Office of Inspector General, Semiannual Report to Congress April 1, 2000-September 30, 2000.
2. K. Thibodeau, Building the Archives of the Future: Advances in Preserving Electronic Records at the National Archives and Records Administration. *D-Lib Magazine*, Vol. 7 No. 2 (Feb. 2001)
3. W. E. Underwood and S. L. Laib, Archival File Utilities, PERPOS Technical Report 1, Georgia Tech Research Institute, Atlanta, GA (June 2001)
4. Consultative Committee for Space Data Systems (CCSDS). Reference Model for an Open Archival Information System (OAIS). (May 2000)
5. JavaTM 2 SDK, Standard Edition Documentation, Version 1.3.1 (2001)
6. W. E. Underwood, An Axiomatic Theory of Record Authenticity, PERPOS Technical Report 7, Georgia Tech Research Institute, Atlanta, Georgia (June 2001) (Submitted for publication)
7. W. E. Underwood and S. L. Laib, An Archival Review Tool, PERPOS Technical Report 2, Georgia Tech Research Institute (July 2001)
8. NIST, Message Understanding Conference Proceedings (MUC-7) (1998) Available online at www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_toc.html
9. W. E. Underwood, Information Extraction Technology for Archival Processing and Access. PERPOS Technical Report 6, Georgia Tech Research Institute (July 2001)
10. W. E. Underwood, FOIA and PRA Review, PERPOS Working Paper 5, Georgia Tech Research Institute (April 2001)
11. W. E. Underwood and S. L. Laib, An Archival Arrangement and Description Tool, PERPOS Technical Report 3, Georgia Tech Research Institute, Atlanta, GA (June 2001)
12. U. Manber, M. Smith and B. Gopal, WebGlimpse—Combining Browsing and Searching, Usenix Technical Conference (Jan 1997)
13. U. Manber and S. Wu, Glimpse: A Tool to Search Through Entire File Systems, Usenix Technical Conference (January 1994).
14. Oracle, Oracle8i interMedia Text, Reference, Release 2, (Dec. 1999)
15. W. Woods, NARA, Office of Inspector General, Semiannual Report to Congress April 1, 2000-September 30, 2000.
16. http://bushlibrary.tamu.edu/papers/
17. M. G. Underwood, Document Retrieval Technologies in Support of Response to FOIA Requests, PERPOS Technical Report 4, Georgia Tech Research Institute (July 2001)
18. W. E. Underwood, The Presidential Electronic Records PilOt System (PERPOS): Phase II Research Plan, Georgia Tech Research Institute (August 2001)