ANALYSIS OF PRESIDENTIAL ELECTRONIC RECORDS: FINAL REPORT

William E. Underwood

September 1999

Computer Science and Information Technology Division
Georgia Tech Research Institute
Atlanta, Georgia

The National Archives and Records Administration (NARA) sponsored the Presidential Electronic Records Project under a contract to the Army Research Laboratory (ARL) and ARL through a subcontract to the Georgia Tech Research Institute. The findings in this paper should not be construed as an official NARA or Department of Army position unless so indicated by other authorized documentation.

EXECUTIVE SUMMARY

Transfers of historically valuable electronic records to the National Archives and Records Administration (NARA) have increased in recent years. In the near future, NARA anticipates receiving an ever-increasing number of both Federal and presidential electronic records. Many of these records have been and will be created on personal computers (PCs) outside of any records management regime. NARA needs to gain archival control of such records. This project was established with the objective of identifying and evaluating automated methods for establishing archival control over such records.

Two important collections of files created on PCs that have been received by NARA are those created in the Executive Office of the President during the Bush administration and those from PCs used by Office of Independent Counsel di Genova. Both collections were transferred to NARA on the original hard drives. These collections were used as test cases in this project. The content of the Bush hard drives and the di Genova hard drives were analyzed to identify issues that need to be addressed to gain intellectual and physical control of the user-created files.

The procedures of the Bush Presidential Library were analyzed to understand the processes and information currently used to gain archival control of paper textual records. An object-oriented model is described that characterizes the functions and types of information that an archival system should provide to support an archivist in gaining archival control of user-created PC files. Information technologies needed to implement these requirements include information filtering, object-oriented software development, natural language processing, knowledge-based systems, and document retrieval. These technologies are described and alternative technologies evaluated as to their utility in supporting archival processing requirements.

The primary results of this project are:

- The development of a prototype document file filter to provide archivists support in distinguishing user created files from operating system and office application software files which are nonrecords; and experimental confirmation that, with some refinements, the document file filter can support this task.
- Evaluation of viewer technology for displaying legacy document files created on personal computers.
- Text interpretation technology can be used to extract, generalize, categorize and index information in PC files, and thus to support archivists in gaining intellectual control of these files.
- Finite-state methods can be used to determine document types such as memoranda, letters, agenda, and schedules.
- A model is provided for the types of information objects and physical (digital) objects that must be created and updated in an archival system for digital records.
- Some software requirements for archival processing of electronic records are derived.

- Concept-based document retrieval with ranking of results by relevance to query is the preferred information technology to meet the need to determine which unprocessed electronic records are relevant to Freedom of Information Act (FOIA) requests.
- Knowledge-based system and text interpretation technologies can be combined to check whether text in an electronic document might be restricted under provisions of the Presidential Records Act or exempt from release under the FOIA. The feasibility of this approach is demonstrated with a prototype PRA checker.

Object-oriented software development offers the National Archives the greatest opportunity for cost-effective system implementation and maintenance due to its support of software reuse. It is suggested that as part of the process of accessioning and processing the PC records from the Bush hard drives, software that employs these information technologies should be acquired, and new software developed

- To support accessioning of digital record series and collections;
- To preserve records by identifying nonrecords;
- To convert records in proprietary file formats to current or standard formats;
- To arrange PC record series and files within a digital archives;
- To check digital documents for FOIA/PRA restrictions and exemptions;
- To support redaction and withdrawal of documents restricted or exempted from disclosure;
- To support archival description and creation of finding aids; and
- To support search of unprocessed and open archival documents in response to FOIA requests.

The result of such an effort would be an archival toolkit that could be applied to future accessions of Presidential and Federal Records. Such a toolkit would support gaining archival control of electronic records, support increased productivity, and reduce the time from accession to opening of Presidential and Federal records for Citizen access.

TABLE OF CONTENTS

Table of Figures

1. INTRODUCTION

1.1 Background

Transfers of historically valuable electronic records to the National Archives and Records Administration (NARA) have increased in recent years. In the near future, NARA anticipates receiving an ever-increasing number of both Federal and presidential electronic records. In the period from 1989-1998, the National Archives accessioned approximately 80,000 electronic files, primarily data files. In calendar year 2000 it is estimated that the Archives will accession 1.25 million electronic files. In 2001, NARA expects to receive 5-50 million electronic files.[1] Many of these records have been and will be created on personal computers (PCs) outside of any records management regime. NARA needs to gain archival control of such records. This project was established with the objective of identifying and evaluating automated methods for establishing archival control over such records.

Two important collections of files created on PCs that have been received by NARA are those created in the Executive Office of the President during the Bush administration and from PCs used by Office of Independent Counsel (OIC) di Genova. Both collections were transferred to NARA on the original hard drives. These collections were used as test cases in this project. The content of the Bush hard drives and the di Genova hard drives were analyzed to identify issues that need to be addressed to gain intellectual and physical control of these document files. Information technologies to support gaining and maintaining control of these electronic records were identified and demonstrated.

1.1 Purpose

One purpose of this paper is to report the results of an analysis of the contents of the Bush hard drives and the OIC di Genova hard drives to identify issues that need to be addressed to gain archival control of user created files on these disks. A second purpose is to specify the functions and information structures that an archivist requires to support processing of digital record series. Another purpose is to identify and evaluate information technologies that support these archival processing functions.

1.2 Scope

The first step was to analyze the contents of the Bush hard drives and the di Genova hard drives. As a part of this process a software tool was developed to filter user created files from the system software and office application software files. Some of the collections from the hard drives were experimentally accessioned. The analysis, tools and experiments are described in the next section of this report.

The second step was to analyze the archival processing activities of the Bush Presidential Library and to identify software and information requirements for supporting the processing of PC records. This analysis is described in the third section.

The third step was to identify and evaluate software tools and information technologies that support the processing of these files. Text interpretation based on natural language processing technology can support the accession and description activities. Knowledge-based text interpretation can support PRA and FOIA review. Text-based conceptual search is a document retrieval technology that can support response to FOIA requests. The results of this task are described in the third section of this report.

2. Analysis of the Bush and Di Genova Hard Drives

During the Bush Administration, Independent Counsel di Genova began an investigation of the possible illegal use of personal information on William Clinton obtained from official State Department passport files. On January 15, 1993, a grand jury subpoena was served on the White House for paper documents, material maintained on the hard drives of personal computers used by various White House personnel, and computer back up tapes of electronic mail. On January 19, 1993, agents of the Federal Bureau of Investigation (FBI), acting on this subpoena, removed and inventoried 516 hard drives from computers in various White House and Executive Office of the President (EOP) components including the Office of the Vice President and the Office of Policy and Development. On January 29, the White House Office of Administration and the Office of Management and Budget transferred additional hard drives to the FBI.[2]

The FBI inventoried and boxed the "Bush hard drives" and stored them in a FBI facility in Alexandria during the OIC investigation. The di Genova staff identified about 200 of the drives that might contain records relevant to their investigation. The FBI copied files that were user created from the identified hard drives. The FBI used a commercial text retrieval product to index the files. All resulting files and associated indexes were loaded onto three personal computers and transferred to the di Genova staff. . The hard drives from these three computers are referred to hereafter as the "di Genova drives."

On February 25, 1995, Judge Richey ordered in *American Historical Association v. Trudy Peterson and George Bush* that NARA "shall retain custody of, control, preserve, and provide access to the backup tapes and hard drives identified in the Memorandum of Agreement, including hard drives deposited with the National Archives and Records Administration at the conclusion of the Independent Counsel's investigation, in accordance with the Presidential Records Act and the Federal Records Act."

After the conclusion of the OIC investigation, the FBI transferred custody of the hard drives to NARA. NARA contracted with Raytheon E-Systems to migrate the information on the hard drives to SyQuest drives, removable hard disks. The contractor's report

indicates that the complete contents of each source disk were migrated to a directory on a SyQuest drive that had the same label as the hard disk in the FBI Inventory. It also indicates that information from 25 of the hard disks could not be recovered without opening the head disk assembly, a very expensive process. [3]

Files of paper records that are created by staff members and offices are periodically transferred to the White House Office of Records Management (WHORM). Electronic copies of documents on the personal computers of staff members and offices were not transferred to the WHORM. Record copies of correspondence or other documents were printed and sent to the central files or preserved in an office file drawer. The WHORM regularly transfers paper files to NARA at the end of a presidential administration.

Because the PCs were used as word processors, the hard drives can be expected to include files that correspond to paper documents in the White House Office files. The hard drives may also contain files for which printed copies are not included in the WHORM files. In addition to word-processing files, there may be e-mail files on the hard drives for which there are paper copies that have been transferred to NARA.

2.1 EOP, FBI and NARA Inventories.

At the end of a presidential administration, the WHORM transfers the records of the administration to NARA. Traditionally, presidential records have been transferred principally in paper form. Hence, NARA does not have established procedures for processing or management of presidential records in electronic form. NARA regulations and guidelines regarding transfer of electronic records apply only to federal records. These directives require that only the records be transferred, not software. Federal records should be transferred in a format that does not depend on specific hardware or software. The records should be transferred on 3480 magnetic tape cartridges, nine-track tape or CD-ROMs, not fixed hard drives. Furthermore, the transfer of federal records to the National Archives is supposed to be accompanied by a records transmittal identifying the records, indicating who the creator of the records is and specifying any legal restrictions on access. However, in some cases, electronic records are transferred on other media, the Bush and di Genova hard drives, for example.

The inventory of hard drives removed from the White House Offices, the Office of the Vice President and the Office of Management and Budget[4] indicate that the labels affixed to the disks contained numbers between 0001 to 0806. However, the Office of Policy Development used labels numbered 1-75. Many of the labels in both of these sequences were not used. There were some hard drives that could not be read by either the FBI or Raytheon E-Systems. This was due to bad a boot sector, a common hard disk problem, or to undeterminable drive types.

2.2 Preservation: Eliminating Nonrecords

Routine archival processing of transfers of electronic records involves verifying that the records are those that should have been transferred, establishing intellectual and physical control over the records, and integrating them into existing holdings. In the case of the Bush and di Genova hard drives, preprocessing is necessary because the transfers did not consist of only collections of records, but of media that may contain records among other things. Preprocessing is required to segregate operating system or software application files, which do not need to be preserved as records, from the files that were created by EOP personnel and that may need to be preserved as records. In this paper, the files that were created by EOP personnel will be referred to as *document files*, to be distinguished from operating system or software application files which will be referred to as *system files* or *non-document files*.

Preprocessing separates the document files from the system files. The system files include operating system files, office software application files, such as word-processing or spreadsheet software, and related files, such as readme files, help files, tutorial files, and sample documents provided with office software applications. These files make up software that is not the property of the Federal Government or the EOP but is licensed for use on specific personal computers. It may be necessary to retain some software files in order to access document files that are in software dependent formats, but it is not necessary to preserve the operating system or office application software as records, because they are not records. On the basis of an experiment described later in this section, it is estimated with 95% confidence that there are on average 579 ± 109 system files per hard disk. To check each of these files manually to determine whether it is a document or system file would be time consuming and tedious. Furthermore, archivists would have to be trained to recognize the system and application files. An automated tool is needed to assist the archivist in this segregation.

Information filtering technology can be applied to satisfy this need. This technology is exemplified by e-mail software that allows the user to create rules to perform user-specified actions on e-mail messages based on keyword matches in the e-mail fields such as the sender field, the subject field and the text body. For instance, a rule may take the form of deleting all messages from a certain person or labeling messages with certain keywords as urgent.

One way of distinguishing DOS system files is by their filename extension.

Rule 1:

> If the filename of a file on the source device has an extension in a list of system or application program extensions,
> then copy the file's filename, extension and path to a directory of system files.

This rule would fail to recognize many sample files that were included with the operating system or software applications. A second list could be created to include the names of these sample files. Then the

Rule 2:

> If the filename and extension of a file on a source device is in a list of filenames and extensions of files known to be associated with a particular software application,
> then copy the file's filename, extension and path to a directory of system files,
> otherwise copy the filename, extension and path to a directory of document files.

A prototype tool for automating segregation of system files from document files was constructed. The tool included a procedure to identify all files on a drive and a filter to separate out document files. The procedure was written to traverse a directory structure and provide all files from a hard drive to the filter. The filtering segregates the system files and user-created document files as defined by rules 1 and 2.

The experimental filter first compares each file to a list of file extensions of operating system and application software, the list of *Excluded Extensions*. Any file that has one of the excluded file extensions is eliminated from further processing. A file that does not have an excluded extension is compared with the entries in a second list that includes file names with extensions. These *Excluded File Names* include those files, in addition to software, that are included with software application packages, for example, sample documents for office application software. Any file that matches a name in the list of Excluded File Names is excluded from further processing.

An initial review of some of the Bush hard drives found files associated with the following operating system and office application software:

dBase III and IV
DOS 4 & 5
Harvard Graphics (HG)
Lotus 1-2-3
Lotus Symphony
OASIS (Office of Administration)
Quattro Pro
REFLECT
Word Perfect 5.0 & 5.1

The Excluded File Names list was constructed from files identified in the manuals for these software products and from the READ.ME files associated with the directories containing the software products.

The dialog box for the Document File Filter is shown in Fig. 1.

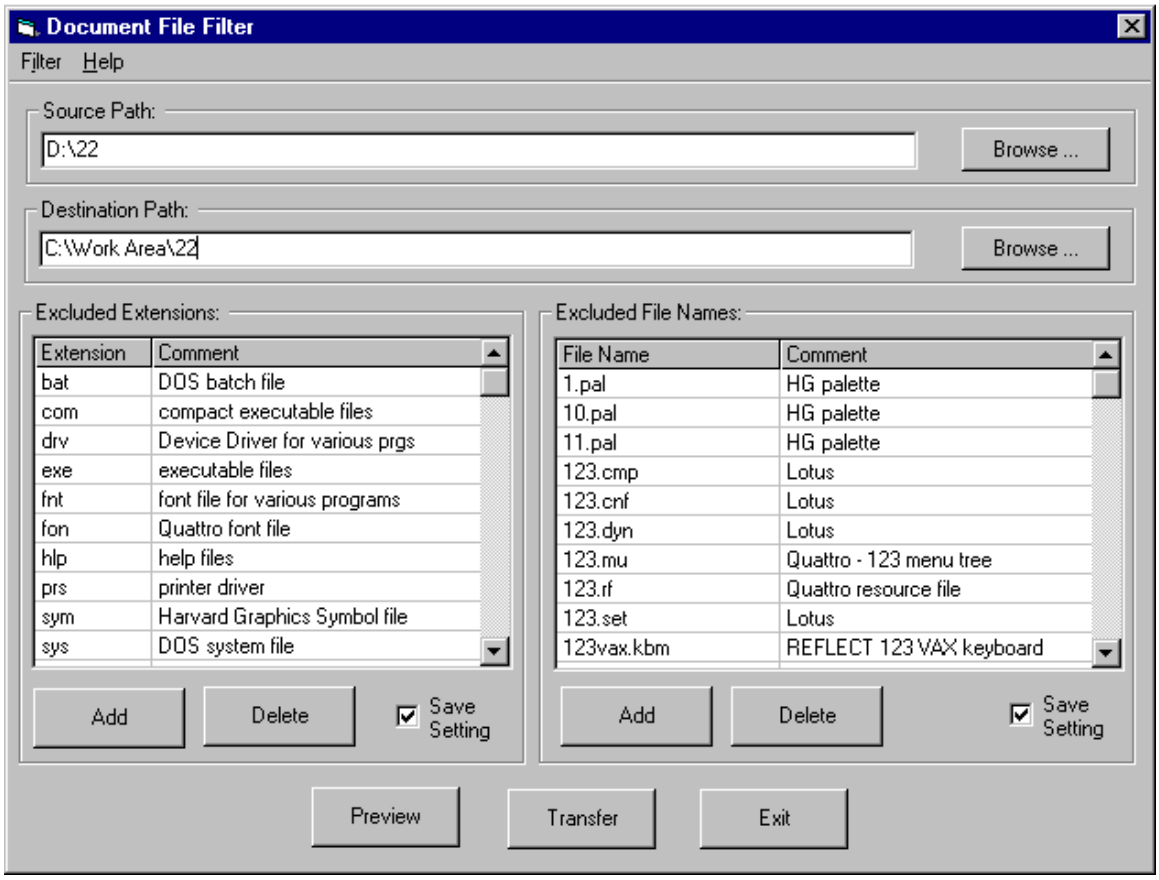

Figure 1. Document File Filter.

An archivist needs to have confidence that the files being excluded are system or software application files, not document files created by a user. To engender this confidence, the dialogue box indicates the operating system or software application name and function in the comment fields associated with the excluded extensions and excluded filenames. The entries in the two columns can be viewed in alphabetical order by pointing at the column heading with the mouse and left-clicking the mouse button.

The location of files to be processed through the filter is identified in the *Source Path* box. Document files that pass through the filter will be saved in the location indicated in the *Destination Path* field. To indicate the drive containing a disk and the directory corresponding to a label of a Bush hard disk to be filtered, one selects *Source Path* and uses the *Browse* button. The device and directory into which document files and their paths should be copied must also be indicated in the *Destination Path* field.

When the *Preview* button is selected, the excluded file extensions and excluded file names are matched against all file names in the directory structure indicated by the *Source Path*. The tool creates two hierarchical directories. One contains the directory structure and filenames of files that have the excluded file extensions and file names. The other contains the directory structure and file names of document files, that is, files without the excluded file extension and excluded file names. A Preview window displays

the two hierarchical directories. The numbers of non-document files and document files are also shown in the preview window. The directories and files are shown in their original order, usually the date they were created or last modified. Examples of these two directory structures are shown in Figure 2.

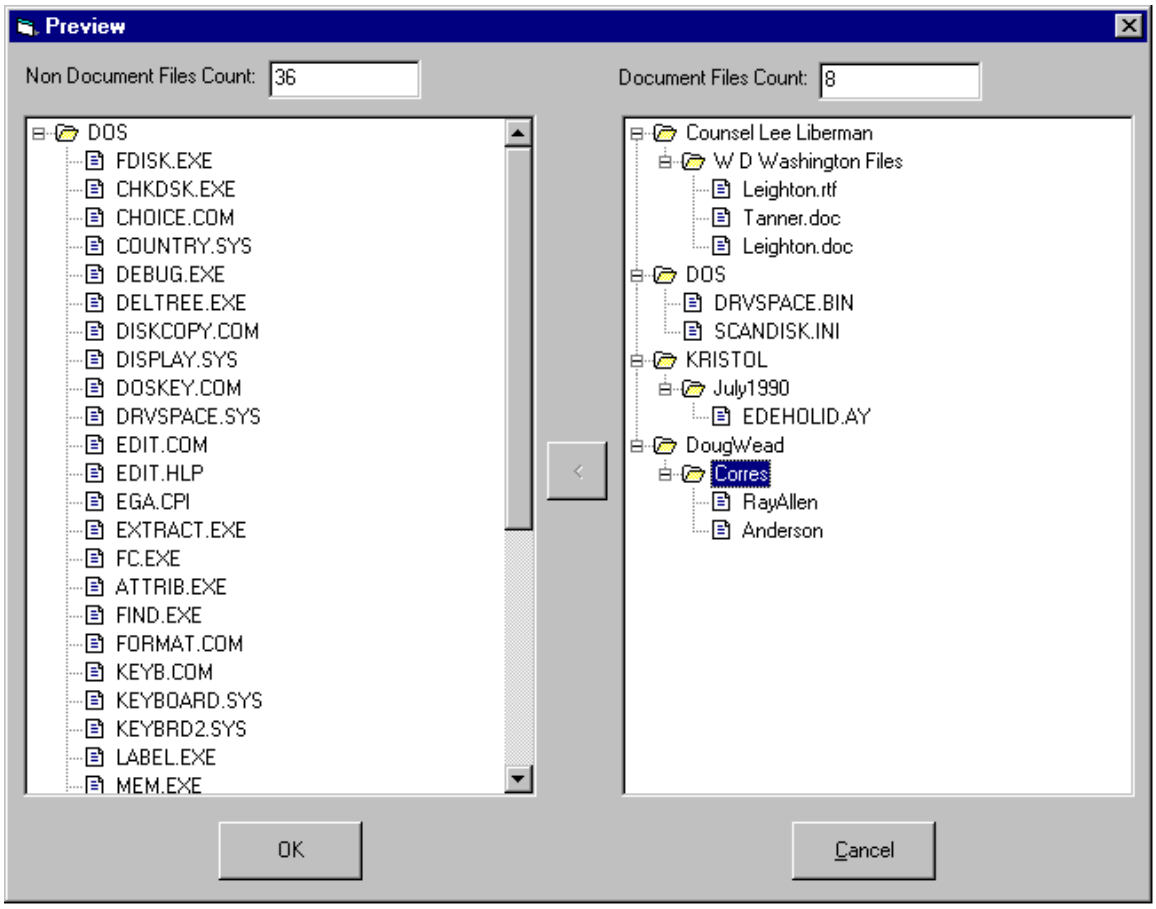

Figure 2. Directories of System Files and Directories of Document Files.

If a directory includes both system files and document files, for example, the DOS directory in Fig. 2, then the directory will appear in the directory structure for system files and document files. Clicking the mouse on a highlighted directory (folder) opens the directory to show other directory and filenames.

The files that pass through the filter probably consist of files saved by the user. The experimental tool enables the archivist to determine what these files actually contain. It does this using Quick View Plus software from INSO. Quick View Plus version 5.1 recognizes over 200 native file formats and displays the contents of a file in the same way the native software would. Double clicking on a filename in the Preview window results either in the file being displayed using Quick View Plus or a dialog box enabling the archivist to select a viewer not included in Quick View Plus. Figure 3 shows an electronic document displayed with Quick View Plus.[5] The name of the office software application

that was used to create the document (WordPerfect 5.1/5.2) is shown below the document.

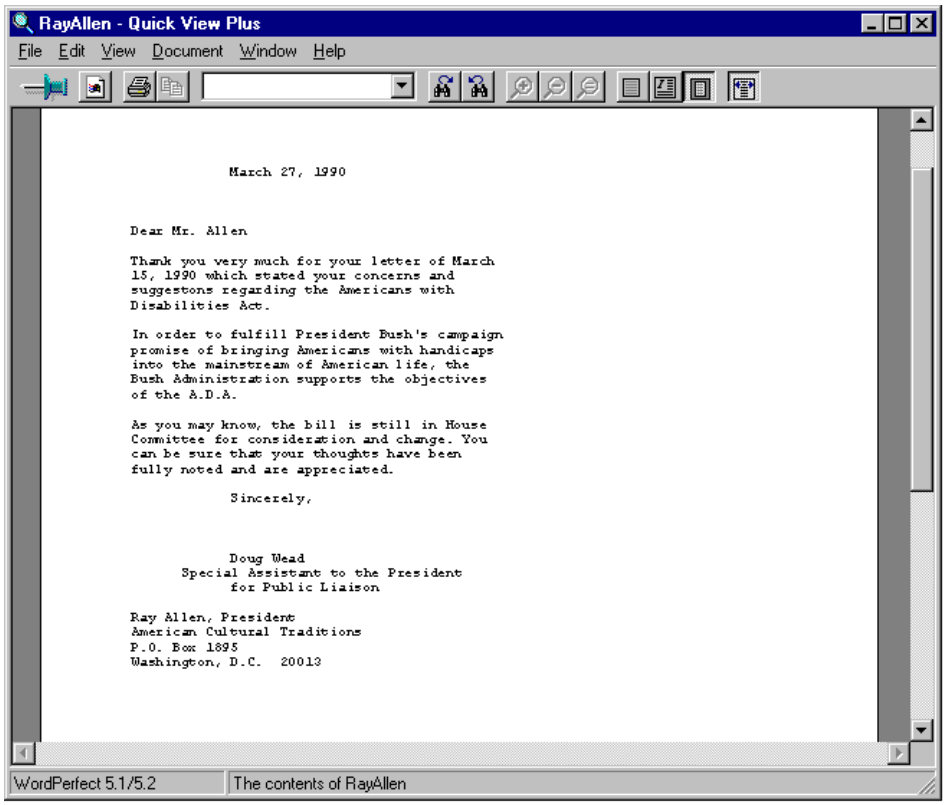

Figure 3. A Document File Displayed with Quick View Plus.

The same capability can be used to attempt to display the contents of a file in the non-document file hierarchy in the Preview window.

Quick View Plus also allows for program integration. For example, Quick View Plus is available as an Active X document server. This means that Quick View Plus can be used to view documents within Microsoft Internet Explorer or Netscape Navigator, and within the file filter.

If a file identified as a document file is actually a non-document file, its file name can be highlighted, and the button in the center of the dialog box in Fig. 2 selected to move it to the non-document files, and vice versa. If the file is in a directory that was not included in the non-document files, the directory name is copied as well. When satisfied that all document files are listed on the right-hand side, one selects *OK* and selects *Transfer* from the previous screen. The document files and their directory paths will then be transferred to the *Destination Path*.

The Document File Filter was applied to the contents of fifteen of the Bush hard drives. Files were excluded based on 15 file extensions and 651 filenames for system and application program files. The results are shown in Table 1.

A	B	C	D
Bush Hard Drive No.	Number of Non-Document Files	Number of Document Files	Processing Time (sec.)
1	579	2809	37
22	1044	1174	29
0105	401	373	7
0101	441	466	6
0115	509	1481	20
0308	618	1017	13
0429	445	31	4
0581	409	661	8
0004A	410	1289	18
0312	426	13	3
0362	595	174	5
0445	524	241	6
0549	408	35	4
0608	487	11	3
0613	451	685	9
Total	7,747	10,460	172
Average	523.77	751.08	12.31

Table 1. Results of an Experiment Using the Document File Filter

The label of the Bush hard drive processed in the experiment is shown in column A. The count of the number of files identified as nondocument (system) files is shown in column B. The count of the number of document files identified by the filter is shown in Column C. The processing time (Column D) includes the time to apply rules 1 and 2 to each of the files in the source path, but not to view the individual files.

The filtering tool sorted over 18,000 files into the document/non-document categories in less than three minutes, or 106 files per second. This processing is immensely faster than what would be required for an archivist to review the files in the *source paths* manually. The filtering was performed on a 125-Megahertz computer. Standard PC's today operate at 400 Megahertz. Thus, the speed of processing would be increased more than three times using state of the art PCs.

On the basis of this experiment and an accessioning experiment described later, the number of document files on 52 of 520 Bush hard drives was determined. It is estimated with 95% confidence that the number of document files per hard disk is 343 ± 139. Thus,

based on this sample, the total number of document files on the Bush hard drives is estimated to be 178,360 ± 72,280,

2.2.1 Exceptions to the Filtering Rules

The 7,747 filenames and extensions of files identified as nondocument (system) files were manually reviewed. If the filenames and extensions were in directories corresponding to those typically prescribed for system and software applications, then they were judged to be correct. If they were in subdirectories that seemed to be user created, they were viewed using the Quick View Plus viewers. Three of the 7,747 files identified as nondocument (system) files were actually found to be document files (see Table 2).

Bush Hard Drive No.	Number of Document Files Identified as Non-Document Files in Column B	Number of Non-Document Files Identified as Document Files in Column C
1	1[6]	15
22	2[7]	230[8]
0105	0	25
0101	0	141[9]
0115	0	29
0308	0	72[10]
0429	0	25
0581	0	17
0004A	0	10
0312	0	11
0362	0	20+[11]
0445	0	34[12]
0549	0	35
0608	0	10
0613	0	280[13]

Table 2. Exceptions to File Categorization Using the Filtering Rules.

Of the three files identified as non-document files that actually were document files, one had a file extension *com* and two had the file extension *prs*. Since these extensions were in the list of file extensions of non-document (system or application software) files, these document files were identified as non-document files. The three files were created using WordPerfect. The WordPerfect application did not enforce a standard file extension for the filenames of saved documents. The creators of these documents used the eight characters allowed for DOS filenames and the three characters allowed for extensions to create mnemonic names for the files.

These three files were easy to spot in the list of non-document files because they were stored in subdirectories with names that one would expect to contain user-created documents, rather than in directories containing system files. For instance, the file SPEECH.COM had the path D\WP5FILES\MAR90\SPEECH.COM. The tool enables the user to move any file from the non-document list to the document list and vice versa. This is accomplished by highlighting the file and clicking on the arrow button between the Non-Document and Document Preview trees. The document is then moved to the proper pathname in the other Preview tree

The filenames of the 10,460 files identified as document files that may be appropriate for preservation as historically valuable records or personal materials were manually reviewed. If the files identified as document files were in user-created subdirectories and had filenames indicative of business activities, they were judged to be document files. The filenames of probable document files that appeared in a directory with a directory name usually associated with system or application software were viewed using Quick View Plus to determine whether they were a system file or a document file.

As indicated in the third column of Table 2, the filter did not filter out all non-document files. Some system files were identified as document files on all 15 hard drives tested. Almost 9% of the 10,460 files categorized as document files proved to be system files. In the cases where there were only a few non-document files included in the group of document files, it was usually because the filename could not be included in the list of filenames of non-document files, because a user might pick that name for a document file, for example TEMP. In the cases that a large number of non-document files remained in those indicated as document files, it was because the filenames of the non-document files had not yet been included in the list of file names for non-document files. An archivist can view the files that are identified as non-document files, and move any that are actually document files to those identified as document files, and vice versa.

The exceptions ocurred when the file extension alone was used to determine whether the file was a system file, and it turned out that the user-created WordPerfect document file used the system file extension. One approach to eliminating this exception to the rule is to use the INSO Quick View Plus package to validate the file format of any files that have a system file extension to see whether they are system files or WordPerfect files. Rule 1(a) includes this refinement and should replace Rule 1.

Rule 1(a):

> If the filename of a file on a source device has an extension in a list of system or a list of application program extensions, and the file does not have a WordPerfect format,
> then copy the file's filename, extension and path to a directory of system files.

An alternative would have been to check that the file with a system file extension had the file format associated with that file extension. However, while Quick View Plus has viewers for very few system file formats.

Another approach is based on the observation that files (filenames) associated with office application software packages are typically installed in prescribed directories. The list of filenames and extensions of system files could be extended to include the directory name the file is usually included in, the number of bytes in the file and the date the file was created. Rule 2 could be refined to the check that files with filenames in the list of filenames and extensions also are in the right directories, have the expected number of bytes, and the expected date of creation.

To minimize the risk that a file judged to be a system file is actually a document file, it is also possible to save one copy of each unique file judged to be a system file. Another reason for preserving one copy of each system file is that these files include font files and color palettes. If Quick View Plus does not have the precise font used in a file, it displays the document using a character font judged to be most similar to the original font. Quick View Plus also displays legacy graphics files using default color palettes, which might not have been the color palette used in displaying or printing the original digital document file.

In summary, Rules 1(a) and 2 will produce an effective filter. Saving one copy of each file judged to be a system file will minimize the risk of losing a user created file and ensure that system files that may be necessary to render the user files as they were originally created are also preserved.

2.2.2 File Formats and Viewers

Document files with the file formats shown in Fig. 4 were found on the Bush hard drives.

dBase III & IV (DBF, NDX and PRG file extensions)
DisplayWrite
Harvard Graphics for DOS (2. & 3.)
Lotus 1-2-3 for DOS Spreadsheets (WKS file extension)
Lotus Symphony
Plain Text (ASCII and DOS Extended ASCII)
Quattro Pro for DOS Spreadsheets and Charts (WQ1 & WQ2 extensions)
WordPerfect 4.2 for DOS documents
WordPerfect 5.X for DOS documents
WordPerfect Graphics (WPG extension)

PKARC archived and compressed file (ARC extension)
Books Cards and Labels (TXT and DAT extensions)
WordPerfect Office 3.0, WordPerfect Library 2.0 Calendar and Notebook
Harvard Graphics slideshow (SHW)

Figure 4. Types of Software Applications and File Formats Found on the Bush Hard Drives.

INSO's Quick View Plus collection of viewers provides support for all of the file formats in the first group of software applications in Figure 4.[14] It determines the proper viewer by examining the file format. If Quick View Plus does not contain the font file for the display and/or print font indicated in the file format, it uses a default display and print

font. The default font is also used for display and printing of data base and spreadsheet files. If it is important that a document is displayed or printed in the original font, a copy of the font files on the Bush hard drives can be saved.

If Quick View Plus cannot determine the file format, it reports a file's format as *unknown*, and allows the user to view the file in text or hexadecimal format. If Quick View Plus is selected to view a document by clicking the right mouse button, and cannot determine the file format, an archivist is given the option of associating the file with another viewer, for example, WordPerfect Calendar or Notebook files. *WinZip* can be used to open files archived and compressed with PKARC. The file formats for *WordPerfect Office 3.0* Calendar and Notebook files are known and viewers can easily be constructed.[15] The *Books Cards and Labels* application was used for maintaining the addresses and phone numbers of individuals and organizations and printing mailing labels or Rolodex cards. Its file formats are TXT and DAT files that can be displayed as ASCII text files.

In section 4 of this report, information technologies are described that interpret text in order to extract, categorize, index and generalize information contained in documents. Readers are needed to provide information to the text interpretation programs. Readers provide the information in a document without the formatting information that is needed for displaying the document.

The Document File Filter passes four types of dBase III and IV files. The dBase data files have the extension DBF, Memo files associated with a DBF have a DTF extension, the index files have a NDX extension, and programs written in the dBase programming language have a PRG file extension.

PRG files may be user-created programs. It may be necessary to run such programs to display or print the data in order to understand how the data was used. It may be determined that the program output should be preserved, either in addition to or instead of the database file. PRG files may themselves be records. In one White House office, for example, an employee had written several dBase programs for correspondence tracking. Those programs may provide significant evidence of how business activities were carried out. Thus, the document filtering tool should pass the program and indexes, as well as the data files.

Thus, it is possible to use Quick View Plus supplemented by viewers for file formats it does not include for initial preservation of all document files from the Bush Hard Drives. Another function that the Document File Filter could perform is to use the Quick View Plus Package plus additional viewers needed to create a list of the types of native file formats associated with each hard drive. It could also extract the names and copies of font files used with document files. The color palettes used for display of WordPerfect graphics files, Lotus 1-2-3 charts, and Harvard Graphics charts could also be extracted. This information could be preserved as metadata characterizing the types of file viewers, fonts, and palettes needed to view the document files from the hard drive.

The Quick View Plus software used in these experiments has viewers for documents produced by most current office application software and many viewers for legacy software. Documents in more than 200 formats can be viewed. However, with changes in computer technology, it becomes necessary to reprogram the viewers for new computers, new operating systems, and new hardware. The adoption of archival standards for file formats for word-processing, graphics, spreadsheet and database document files is one approach to alleviating this problem. Research is needed to explore and recommend alternatives for standard file formats.

2.3 An Accessioning Experiment

Accessioning is the procedure by which a Presidential Library (or the National Archives) takes physical and legal custody of records or papers and establishes initial intellectual control over the records. The National Archives has physical and legal custody of the electronic records from the Bush hard drives. They do not yet have intellectual control over the material.

In a Presidential Library, the outputs of the accessioning activity are an Accession Register describing the collections of records in the custody of the Library, and the accessioned records are stored in the stacks. Since the Accession Register records the name of the creator, it provides initial intellectual control of the provenance of the records, the office and/or staff member who created the records. The Bush Presidential Library uses a Location Register and Box Description to record the stack location of accessions. Other libraries may record the stack locations in the Accession Register itself. The Location Register records the archival status of unprocessed and processed storage containers to facilitate access to either.

To accession records, archivists must know the creator of the records, that is, the office or the name and title of the person who created the records. The EOP inventory of hard disks usually indicates the last name and initials of the person to whom the PC was assigned and the building and office number, but does not indicate the White House Office or job title of the person. Using the EOP inventory of Bush hard drives and the White House Telephone Directory, an archivist from the Bush Library identified the White House Office that the personal computer was located in and the full names and titles of staff members assigned the hard disks. These were entered in a Microsoft Access data table.

An archivist from the Bush Presidential Library experimentally accessioned the contents of 40 of the Bush hard drives. He used the Access database table described above to identify the records creator. For this experiment, the document files on each drive were treated as a record series. A Microsoft Access database was created containing fields for the data captured in the Bush Presidential Library Accession Register. The archivist entered information relevant to the accession in the databases using a Microsoft Access. Figure 5 shows an example of the accession form.[16] To develop information needed to

describe the records, the archivist used Quick View Plus and other software viewers to review the document files passed by the filtering tool.

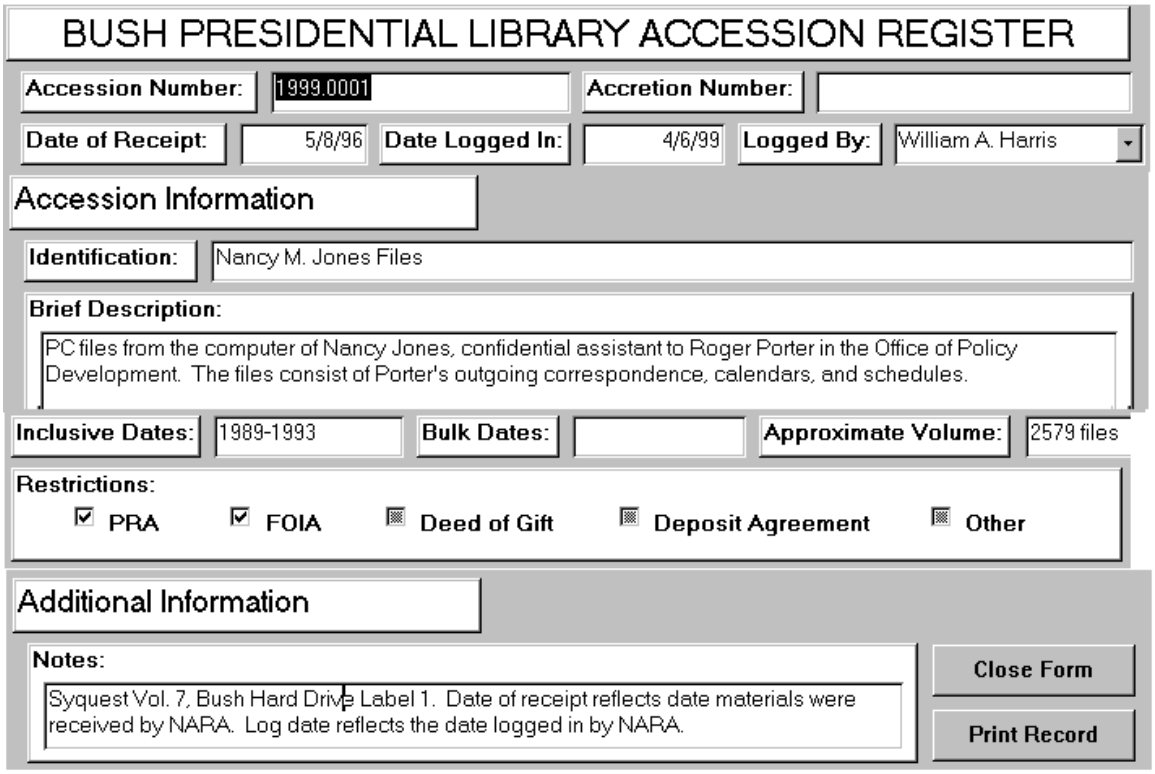

Figure 5. Sample Accession Record for Bush Hard Drives

In accessioning the record series, the archivist found that in addition to Presidential Records, there were also Federal Records and Personal Papers on the Bush hard drives. The Federal Records included electronic records created by the Office of Management and Budget (OMB). The Personal Papers included resumes, recipes and Christmas card mailing lists.

This experiment demonstrates the use of automated tools to process an accession of electronic records following a procedure analogous to the processing of paper records in a Presidential Library. The tools used were the Document File Filter, the Quick View Plus viewers and an Accession Register database. These tools enable the archivist to separate out files that do not need to be preserved as records; they facilitate the identification and review of files that may be records; and they enable the archivist to establish initial administrative and intellectual control of record series.

2.4 Arrangement

Arrangement is the proper ordering of materials within a collection and the placement of materials in an archival storage area. *Proper ordering* consists of structuring a collection in accordance with the filing system that the record creator used to organize its records. Some of the staff member's and office's records on the hard drives were subject to a

recordkeeping plan. For instance, a directory named CORRES was found in which correspondence files were saved in subdirectories whose name was an abbreviation for the month and the last two digits of the year. However, many of the document files from the Bush hard drives are not systematically arranged (ordered) in the DOS hierarchical filing system. For instance, user-created files were found in the root directory along with system files and in the same directory as those containing the WordPerfect software application files.

Filenames appear in a DOS directory in order of the date and time the file with that name was first created or saved, except that when a new directory entry is needed, DOS uses the first available entry, which may be an erased file's old entry. If a file is modified, but saved with the same file name, it remains in the same position in the directory. We will refer to this as *DOS directory order*.

A DOS hierarchical filing system is composed of two types of entities: files and directories. The DOS directory and file names are up to eight characters plus up to a three-character extension. The highest level of a DOS directory hierarchy is called the root directory. The root directory may contain files and directories under it. Any directory may contain other directories under it—which is what makes it hierarchical. Names of nested directory are separated by a backslash (\) character. The pathname of a file indicates its location in the filing system. Two files may have the same filename as long as their pathnames are different. A pathname in DOS may be up to 64 characters in length. To access a file, you must specify the path to the file.

Absent any filing in accordance with a records management plan, the DOS directory order of the document files is the original order of PC files. Directory names are also in DOS directory order. Archivists must decide whether this is the order to be preserved, or whether there is a better logical order such as the last date-time save, or the date of the document in the text of the document file.

The DOS directory names of document files from the Bush Hard Drives are cryptic, usually eight characters or less. Archivists customarily extend folder titles to provide a better description of folder contents. Conventions may need to be established for extending the DOS directory names to provide better descriptions of directory contents.

There is also an issue of long-term preservation of DOS directory names and file structure. DOS is after all a Disk Operating System. It stores directory names and structures in directory files. These directory files are usually accessible to a user only through the DOS *DIR* command. But DOS directory name, file name and path conventions are not the same as the NTFS, Unix and Macintosh filing system conventions. Some filing system independent conventions may be needed to ensure that the filing arrangement remains accessible on operating systems of the future.

There are several approaches to resolving this issue. One is to archive the directories and files of a PC record series using a TAR standard format and provide an UNTAR

capability for any operating system on which the files and file arrangement must be loaded. Another possibility is to emulate DOS commands on any operating system on which the files must be loaded. Thirdly, one might convert all file structures to the filing system structure of a standard operating system such as POSIX. A fourth alternative is to replace the DOS directories with an XML description of the DOS filing structure.

The relationship of records in a data file and the method used to access records in a data file are issues of file format and access, not of file arrangement. However, if a data file is indexed on one or more fields, there may be index files that need to be preserved with the file, e.g., dBase III or IV NDX files. Dbase III and IV DBF files were related in dBase program (PRG) files It may be necessary to determine the relations of the files by interpreting the PRG files and translating the program description of the relations into a SQL description of the relations.

2.5 Analysis of the Contents of the Di Genova Hard Drives

After the conclusion of Independent Counsel di Genova's investigation, three personal computers/textbases used by di Genova's staff were transferred to NARA. Raytheon E-Systems, a NARA contractor, transferred the contents of the hard drives on these computers to nine removable SyQuest drives. The three computers were labeled:

Zyindex 01 Active Files 167 Drives
Zyindex 02 White House Legal
Zyindex 03 Large Files and Erase

The Document File Filter was applied to four SyQuest drives containing information from the first di Genova personal computer. It identified the directories in Fig. 6 as containing operating system and software application files.

\Z1\DOS\	\Z1\ACTIVE
\Z1\LL5\	\Z1\BIN
\Z1\MOUSE	\WINA20.386
\Z1\DELL	\COMMAND.COM
\Z1\BACKUP	\SCANDISK.LOG
\Z1\UTILITY	\CONFIG.SYS
\Z1\TEMP	\AUTOEXEC.BAT

Figure 6. Non-document Files from the First di Genova Hard Drive.

The Document File Filter identified the files in the directory \Z1\DATA\ as containing document files, and it was correct. The other three SyQuest drives from the first di Genova personal computer also contained the directory \Z1\DATA\ and document files. Only the files in the Z1\DATA directory need to be preserved. They are the document files extracted from the Bush hard drives by the FBI for OIC di Genova.

In their inventory of files copied from the Bush hard drives for review by the Office of Independent Counsel, the FBI did not preserve the original disk labeling convention. The label numbers in their reports range from 1-743. They did not preserve the leading zeros that would have distinguished disks labeled 0001-0075 from those labeled 1-75. Instead, the FBI suffixed some of the labels in the sequence 1-75 with the letter A and sometimes suffixed the labels in the sequence 0001-0075 with the letter A. For instance, FBI labels 1 and 1A correspond to labels 1 and 0001 in the original inventory, respectively. FBI labels 25 and 25A correspond to labels 0025 and 25 in the original inventory, respectively.

To determine which persons and offices corresponded to the NARA inventory of hard disks, it was necessary to correlate the original EOP and FBI inventories. Using Microsoft Access, data tables were created for the original EOP inventory and the FBI report of hard disks copied for the Office of Independent Counsel.

The FBI used the ZyIndex application software by ZyLabs for indexing the text of the document files. The index consists of terms that occur in the documents and pointers to the disk locations of files containing that term. The indexing technique used by the ZyIndex software application is hardware dependent. The index pointers are pointing to physical sectors and a relative physical location on the hard drive. Such indexing techniques were common for that particular generation of PC and disk drive technology. The index files and text-retrieval software from the di Genova hard drives are obsolete. They should not be preserved. Any of a number of current hardware-independent, content-based, document retrieval tools such as described in section 4 of this report can be used in place of the original hardware-dependent Zyindex software.[17]

The two SyQuest drives containing the files from the second di Geneva personal computer were processed using the filter. There were only two directories containing document files—\Z2\DATA and \Z2\LEGERASE. The files in Z2\DATA are from 16 additional Bush hard disks. The files in Z2\LEGERASE\ seem to be *erased* files from the same 16 hard disks whose document files are in Z2\DATA.

The di Genova staff seems to have believed that files erased from the Bush hard drives would still be available and might contain information relevant to their investigation. When a file is erased using the DOS delete command, the first byte of the filename in the directory is set to hex E5 and all the File Allocation Table (FAT) entries for a file's space allocation chain are marked as available (set to 0). All other directory information about the file is retained, including the rest of its name, its size, and even its starting cluster number. The actual file date in the data space is not changed. If one made a mistake in erasing a file, and discovers this before the directory entry and the file space is reused, the file can often be recovered using a utility such as the Unerase program in Norton Utilities.[18]

OIC di Genova's staff asked the FBI to make copies of erased files on certain of the disks. The contents of these files have been reviewed. In almost all cases, they consisted of file fragments and null values and only in a few cases might they contain the digital

representation of a complete document. This is because the space occupied by the original file had been reused.

Finally, the three SyQuest drives from the third di Genova computer were processed using the filter. The first drive contained only system and application program files. The second contained system files and one directory—LARGEFIL—of very large document files, on the order of one megabyte. The Zyindex software could not index files that were this large. The FBI broke these files into smaller files and included the parts on the first di Genova personal computer where they were indexed and could be searched. The large files on the third computer were not indexed.

The third SyQuest drive from the third di Genova personal computer contained a few software application files and files in a directory named ERASED. These files have file names that indicate they were possibly constructed from the areas marked for deletion on the hard drives whose document files were copied to the first di Genova computer.

3. An Object-Oriented Analysis of Archival Processing

To understand the current procedures and information used by the Bush Presidential Library to accession, process and provide citizen access to paper textual records, the processing and FOIA review manuals of the Bush Library were read and analyzed. An AS-IS activity model of the accession, processing and reference service activities was constructed using the IDEF0 activity modeling methodology. An AS-IS IDEF1X relational data model of the types of information created and used during these activities was also constructed.[19]

Use case analysis is an object-oriented analysis technique that clarifies and defines the functions of a system from the user's viewpoint. A use case is a case of a user using the system. A use case diagram can be used to specify or characterize the functionality and behavior of a system interacting with users.[20] Figure 7 is a use case diagram for some of the functions that could be performed to support an archivist in accessioning records, archival processing, and responding to FOIA requests.

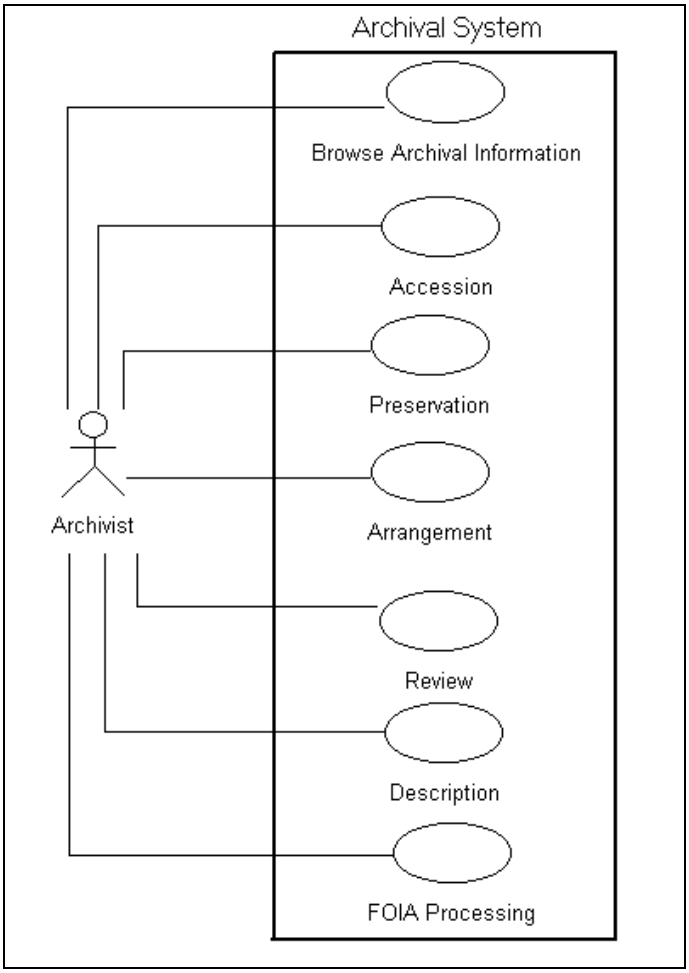

Figure 7. Use Case Diagram for Archival System.

The use case diagram is represented in the Unified Modeling Language (UML) notation.[21] The use case diagram and other models shown in this report were constructed using the Rational Rose CASE tool.[22] A stick figure depicts an actor (a type of user or other system). A labeled ellipse represents a use case. A labeled solid line represents an association relationship. Each use case can be decomposed into subcases.

For each use case, use case scenarios are written that describe how an actor interacts with the system. For instance, to browse archival information, an archivist might interact with the system according to the following scenario.

Use Case 0: Browse Archival Information

1) Archivist requests display of Accessions, Administrative Control Groups, Series, File Units, Items, Storage Containers, or Archival Storage Locations.
2) System displays requested information.
3) Archivist retrieves digital storage module from storage location and loads on a drive.
4) Archivist requests display of file structure of loaded storage module.
5) System displays file structure.
6) Archivist requests to view digital file.
7) System displays requested digital file.
8) Archivist requests system to print or copy a digital file and then close the file.
9) System performs the requested file operation(s).

An archivist can interact in any order with the use cases to browse, preserve, arrange, review or describe records. For instance, if record series are processed systematically, preservation and arrangement are performed before review and description. However, the volume of FOIA requests received by the Bush Library is so large that there is little or no time for systematic archival processing. First, the archivists partially describe the record series by creating a finding aid (folder title list) for the series. They can then respond to FOIA requests, review relevant records, and while doing so arrange and preserve them. Furthermore, as is the case for the Bush hard drives, it may be necessary to preprocess the record series prior to accessioning by interacting with a preservation use case to filter out nonrecords. This is a very flexible framework for archival processing.[23]

An (archival) system consists of types of archival objects, attributes of and operations on these objects, interface objects and control objects. An object-oriented model of the types of objects manipulated by an archival system is shown in Figure 8.[24] Types (classes) of objects are depicted as a two-part box, with the class name in the top part and a list of attributes in the middle part. A more detailed version of this model showing the operators associated with a class is described in Appendix A. The notation used in the model is also explained in that appendix. The glossary in Appendix B defines the classes in the model.

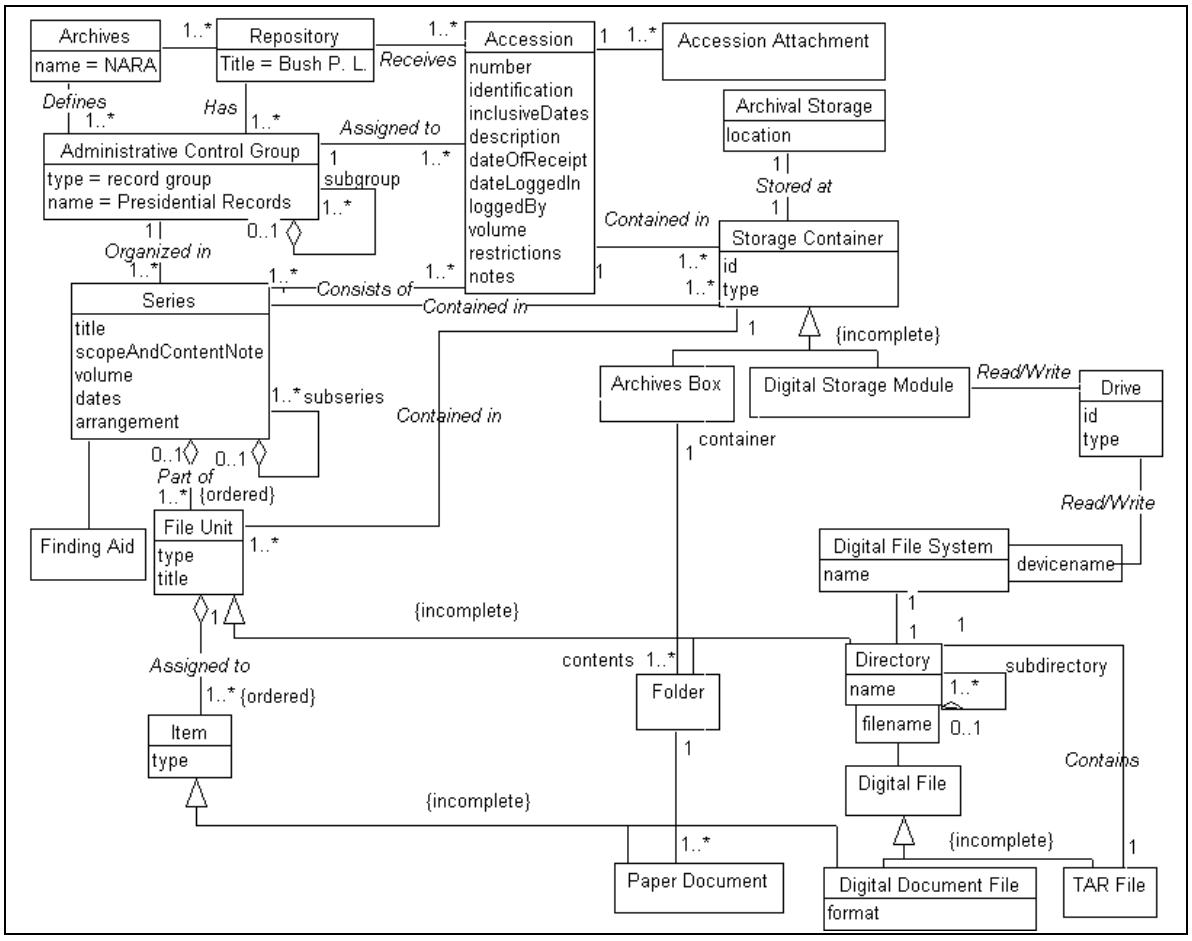

Figure 8. Class Diagram for Types of Archival Objects.

Beginning at the upper left of the diagram, the interpretation of the diagram is that an archival agency, such as NARA, has one or more repositories, such as the Bush Presidential Library. A repository has one or more administrative control groups, e.g., agency fonds, Bush Presidential Records, and record groups or collections of personal papers. NARA defines these administrative control groups.

A repository receives one or more accessions. These accessions have one or more attachments such as a transmittal document, deed of gift, or finding aid. An accession is contained in one or more storage containers. One or more accessions are assigned to an administrative control group. An administrative control group may have subgroups, for instance, Bush Presidential Records have White House offices as subgroups.

An administrative control group is organized in one or more series. A series may have subseries as parts. An accession is assigned to one or more series, and a series may consist of one or more accessions (accruals or accretions). A series (or subseries) is contained in one or more storage containers and a storage container contains one or more series.

A series (or subseries) may have one or more file units as parts.[25] File units are ordered. A storage container contains one or more file units.[26] A file unit contains one or more items. Items are also ordered.

A storage container is stored at an archival storage location. Kinds of storage containers include archives boxes and digital storage modules, e.g., a magnetic tape cartridge.[27] An archives box contains one or more folders, which are a kind of file unit.[28] A folder contains one or more paper documents, which are a kind of item.[29]

A digital storage module may be loaded on a drive, and the drive reads or writes a digital storage module. A digital filing system reads or writes to a drive specified by a (logical) device name.[30] There is a one-to-one relationship between a file system and a (root) directory. A directory is a list of filenames that point to either digital files or to other sub(directories). A directory name plus a filename specifies a unique digital file. The kinds of digital files include digital document files and TAR files. This specialization of digital files is incomplete.[31] Digital document files are a kind of item. A directory may have subdirectories, which are themselves directories. A directory is a kind of file unit. A TAR file contains a directory, which in turn contains digital files that are digital document files or other sub(directories).

Instances of the File Unit type contain the title of folders and directories. In other words, the instances of File Unit are information objects describing the folders and directories. Similarly, the Item type can contain information about paper documents and digital document files. These instances of File Unit and Item can be used as finding aids for folders and directories, paper documents and digital document files. On the other hand, drives and storage modules are types of physical objects.

In section 3.4 the Digital File System, Directory and Digital Document types will be refined to model the arrangement of shadow folders and mirror files needed for segregating files temporarily closed to researchers.

Interaction analysis can be used to create sequence diagrams from the scenarios used to describe use cases. Sequence diagrams describe the sequence of interactions between an archivist, interfaces to the archival system and objects in the archival system, e.g., a specific series, file unit, or item. During interaction analysis, classes of objects, their attributes and relationships are generalized from the specific objects. An example of a sequence diagram used to create this model is shown and explained in Appendix C.

3.1 Accession

To accession a record series, an archivist should be able to interact with an archival system according to a scenario such as the following.

Use Case 1: Accession: Extract information and register acquisition

1) Archivist reviews transfer document and attachments.
2) Archivist requests that system create entry in accession register.
3) System displays accession form.
4) Archivist requests that system display administrative control groups.
5) System displays administrative control groups.
6) Archivists indicates accession is for a particular record group/fonds, e.g., Presidential Records.
7) Archivist indicates drive and pathname of an accessioned digital storage module.
8) Archivist requests display of directory structure.
9) System displays directory structure.
10) Archivist requests extraction of information.
11) System reads digital document files and
    Counts number of user-created files,
    Accumulates the storage volume of the user-created files,
    Determines inclusive dates of files,
    Identifies record types, e.g., letters, memos, reports, calendars, schedules,
    Extracts names and titles of authors /writers of files,
    Generalizes information extracted from files to determine primary subjects.
    Infers the primary activities associated with documents, and
    Displays the results.
12) Archivist may view individual files in directories to extract other information.
13) Archivist uses extracted information to compose a brief description of the accession and enters into Accession Form
14) Archivist indicates type and id of storage module, e.g., SyQuest Drive 19, and the name of a directory or TAR file containing the accession.
15) System associates the storage module with the accession.
16) Archivist chooses an archival storage location for the storage container.
17) System associates archival storage location with storage container.
18) System saves contents of the accession form and associates attachments with accession.

3.2 Preservation

During the analysis of the files on the Bush and di Genova hard drives, the need was identified to support archivists in removing files that are not user-created documents. This requirement is considered a preservation requirement because it discriminates digital document files that need to be preserved from those that do not. Information filtering was demonstrated as a technology that can provide automated support for this function. The following scenario defines the requirement.

Use Case 2.1: Remove nonrecords and extract collection metadata.

1) Archivist indicates location of record series to be filtered.
2) System identifies nonrecords, e.g., system and application software files.
3) System counts the number of digital system and digital document files
4) System determines native formats of digital document files.
5) Archivist reequests to view particular files.
6) System displays the indicated files.
7) Archivist decides which files to preserve.
8) System saves indicated directories and digital document files as a preservation copy.

A digital preservation strategy is needed to preserve an authentic, available, accessible, understandable copy of original records.[32] Migration is a digital preservation strategy that seeks to achieve this goal by periodically transferring digital records from one hardware/software configuration to another or from one generation of computer technology to another. During the analysis of the Bush hard drives, the need was identified to support archivists in converting obsolete digital file formats to standard or current file formats. The following scenario, patterned after the use case for detecting and removing nonrecords, suggests a method to meet this need.

Use Case 2.2: Convert digital document files in obsolete, nonstandard file formats to current, standard file formats.

1) Archivist enters into the system a list of standard/current file formats for which there are file viewers.
2) Archivist enters into the system the obsolete and nonstandard file formats for which there are converters.
3) System retains this information
4) Archivist indicates source location of digital document files to be checked to determine whether they need to be converted and the destination location for the converted digital document files.
5) System determines each digital document's file format, and shows those that need to be converted, and whether there is a converter to a current/standard format.
6) Archivist reviews the files that need to be converted. If additional converters are needed, archivist seeks to obtain them. If all digital document files needing to be converted have converters, the archivist requests the system to perform the conversion.
7) System converts digital document files needing conversion to current or standard formats, replicates those that do not need conversion, and writes the digital files to the indicated destination.

Realization of this use case will require establishing current/standard file formats for word-processing, spreadsheet, database, image and many other kinds of digital files. It will also require a large number of file format conversion methods, and validation that there is no loss of information in the conversion, or at least identification of the kind of

information that is lost. Methods will also be needed for decoding attachments to e-mail that are MIME, BinHex, BASE64, or uuencoded.

3.3 Arrangement

Arrangement is "the intellectual and physical processes and results of organizing documents in accordance with accepted archival principles, particularly provenance, at as many as necessary of the following levels: repository, collection, record group or Fonds, subgroups, series, subseries, file unit, and item."[33] Intellectual/administrative arrangement by provenance provides information about records creators. Intellectual arrangement by filing structure provides information about records. The physical arrangement of files refers to arranging documents within folders and folders within boxes.[34] For electronic files, it refers to arranging files within directories in a digital file system.

An archivist needs system assistance in arranging the PC document files within the types of objects in a digital archive. He may also need to rearrange files that are in DOS directory order into a more logical order, e.g., alphabetic by file name or correspondent name, or by document date, rather than the date in the DOS directory. DOS and other operating systems such as Unix and Windows NT do not provide adequate support for rearranging files into a different order.

Assume than an archivist uses the Browse Archives use case to locate an accession to be arranged.

Use Case 3.0 Arrange Record Series and Files within Archives.

1) Archivist requests display of administrative control group and subgroups associated with accession.
2) System displays requested administrative control group(s).
3) Archivist may select or create a subgroup, e.g., a White House Office.
4) Archivist may request display of existing series within administrative control group or may request the system to create a new series, in which case the archivist will give it a title.
5) Archivist requests the system assign the accession to the series.
6) If subseries are needed, the archivist requests the system to create the subseries and enters a title for each subseries.
7) Archivist requests the system to create file units associated with a (sub)series and enters a title for each (sub)series.
8) Archivist determines the logical order of file units in the (sub)series.
9) If the logical order should be different than the original order of directories, the archivist indicates the order in which to arrange the file units.
10) System arranges the file units in the indicated order, and puts the name of the order in the arrangement attribute of series.

11) Archivist may perform steps 7-10 for items and request the system to associate them with a file unit.
12) If there was a finding aid attached to the accession of this series, the archivist asks the system to associate that finding aid with the series.

Interaction diagrams for this use case have been created for arranging the following subgroups and series:

- Digital document files from a Bush Hard Drive created by a staff member in a White House Office
- Paper files from a White House Office
- Paper files for a Staff Member in that office.
- WHORM Subject Files including case file folders.
- Alphabetic Name File
- Personal Papers of William J. Bennett.
- Digital Storage Module containing only email messages.

This analysis demonstrates that the class diagram has the requisite classes, attributes, operators and relations for arranging the common kinds of paper and electronic records in the Bush Presidential Library. An example of an interaction (sequence) diagram for arranging digital document files from a Bush hard drive is described in Appendix C.

3.4 Review

Review is the process of identifying and segregating materials that are to be temporarily closed to researchers due to restrictions of the Presidential Records Act (PRA) or exemptions of the Freedom of Information Act (FOIA), or that should be removed from the collection because they are Personal Papers. An archivist must review Presidential Records page-by-page for PRA restrictions or FOIA exemptions.

Spelling checkers and style checkers are common adjuncts to word-processing software. They essentially use a lexicon, grammatical rules and pattern-matching technology to discover possible spelling or stylistic errors. The possibility of applying this technology to FOIA and PRA review suggests the following scenario.

An archivist could use the Browse function (use case 0) to locate the digital storage module containing the digital files to be reviewed. Assuming that the storage container had been loaded on a drive, the archivist could use the system as follows.

Use Case 4.1: Check Record for FOIA Exemptions, PRA Restrictions, or both.

1) Archivist indicates digital document file(s) to check.
2) System checks document for FOIA exemptions, PRA restrictions, or both.

3) System highlights suspect portions of document's text, indicates the exemption and/or restriction that may apply, and explains the line of reasoning that led to that conclusion.
4) If Archivist decides that the document does not contains restricted or exempt information, then Archivist opens the document. If the document contains restricted or exempt material, but there is a significant enough amount of material that is not restricted, and opening that material would not compromise the restricted material, then Archivist enters *Redaction Scenario 4.3*, otherwise the Archivist enters *Withdrawal Scenario 4.4.*

The knowledge-based system, natural language understanding, and finite-state technologies described in section 4 of this report provide an opportunity to realize this decision support capability. As will be shown in section 4, these information technologies can also support the identification of Personal Papers, so the following scenario is technically feasible.

Use Case 4.2: Check whether document(s) is a Personal Paper.

1) Archivist indicates location of document files to be checked.
2) System checks whether each document file is Personal Paper.
3) System shows those document file(s) that may be Personal Papers, and for each explains why.
4) Archivist decides whether each document files is a Personal Paper and withdraws those document files judged to be Personal Papers.

Shadow folders and mirror files are created to segregate paper textual records that are temporarily closed to researchers. A shadow folder is an exact copy of folder in the open files that is used to store documents closed under access restrictions of the PRA or exempt from disclosure under the FOIA, withdrawn security classified documents, and withdrawn personal records. The shadow folders are stored in archive boxes in the order in which they occur in the open files. Thus, they shadow, mirror, or parallel the original arrangement.

If a document should be withdrawn, a shadow folder is created that contains the title of the original folder. A withdrawal sheet is included in the shadow folder along with the withdrawn document. A withdrawal marker is put in the original folder in place of the withdrawn document. If all materials in a folder are withdrawn, the original empty folder includes only the withdrawal markers. If a document containing text that is subject to PRA restrictions or FOIA exemptions is redacted, a redaction marker is created to indicate that the document has been redacted, the original withdrawn and replaced by the redaction marker and the redacted document.

These concepts can be extended to digital file structures as shown in the class diagram in Figure 9. In the figure, there is an existence dependency between a Shadow Directory and a corresponding Directory in the open files. That is to say, an instance of a shadow

directory would not exist without a corresponding directory object. There is also an existence dependency between a redacted digital file and a withdrawn, closed, unredacted file, and between a withdrawal marker and a closed digital file.

If a redacted digital file is somehow or other marked as *redacted*, a redaction marker is not needed. If the attributes associated with the withdrawal sheet are associated with the Closed Digital Files, the Shadow Directories and attributes of the Closed Digital Files are equivalent to the Withdrawal sheet.

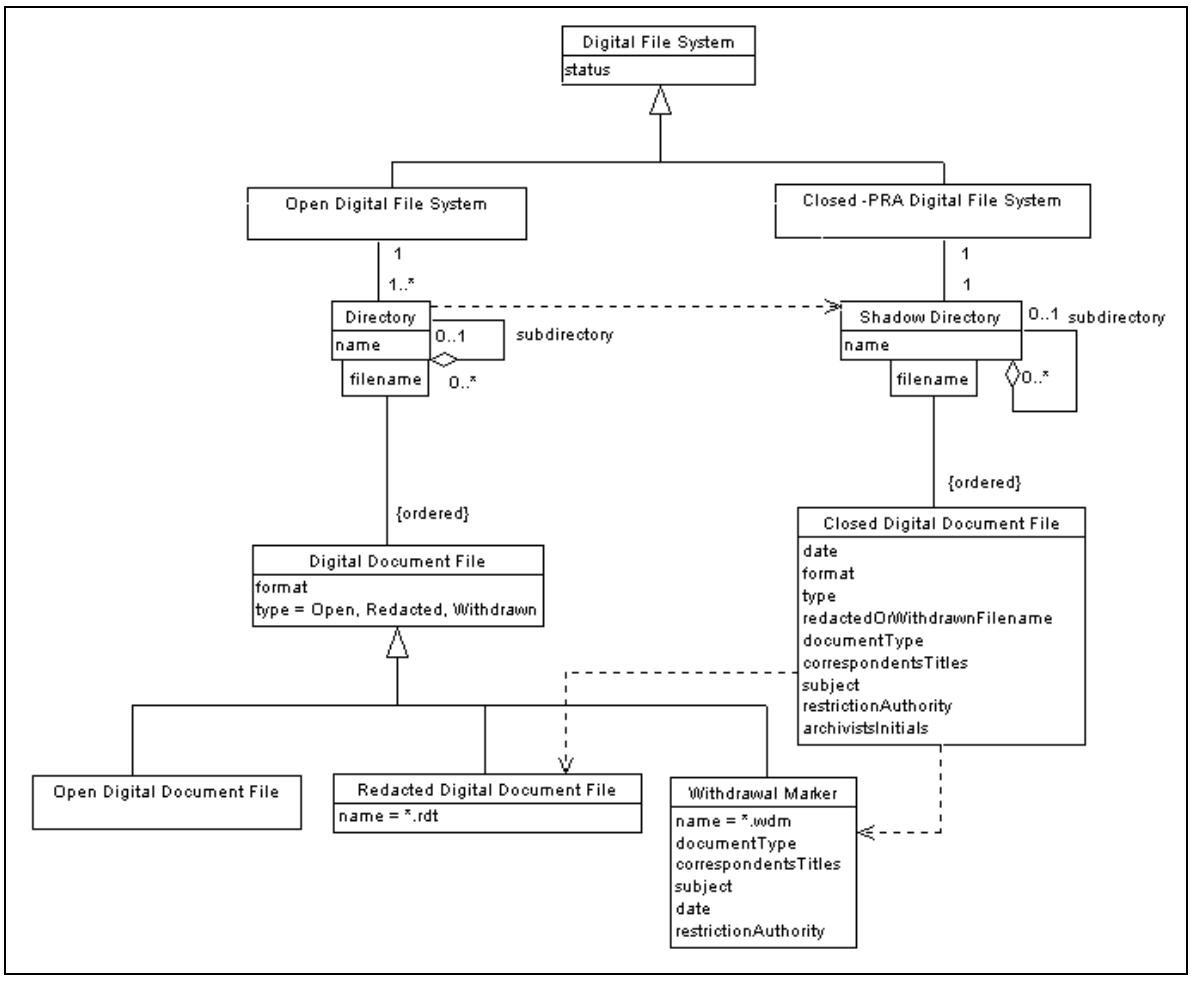

Figure 9. Class Diagram for Shadow Folders and Mirror Files.

The class diagram in Figure 9 must be extended either to include an *In Process* digital file system or to allow digital files in the open digital file system to be of type *in process* in addition to open, redacted or withdrawn.

The following scenarios describe the system functions required to support an archivist in redacting and withdrawing documents.

Use Case 4.3: Redact Document

1) Archivist indicates document to be redacted.
2) System makes a copy of the document.
3) Archivist highlights the portion of text to be redacted and indicates the relevant PRA restriction or FOIA exemption.
4) System replaces the highlighted text with symbol corresponding exactly to the length of the highlighted text, and shows the reason for redaction with FOIA or PRA codes.
5) Archivist requests withdrawal of original and saving redacted copy in open directory.
6) System moves withdrawn digital document file to a shadow directory and writes it as a closed digital document file. System also saves the filename of the original digital document file as an attribute of the closed digital document file.
7) System saves the redacted copy in the open directory in the same logical sequence as the original. The redacted copy has the file name extension *rdt*.
8) System extracts correspondents, subject and date from original file and displays them to the archivist.
9) Archivist edits or approves the information.
10) System saves the information as attributes of the closed digital file.

Use Case 4.4: Withdraw Document.

1) Archivist indicates the displayed digital document is to be withdrawn and the restriction authority.
2) System moves the digital document to a closed PRA directory as a closed digital document file.
3) System reads the withdrawn document and extracts information for the Withdrawal Marker.
4) System displays a Withdrawal Marker containing the extracted information.
5) Archivist edits/approves the contents of the Withdrawal Marker.
6) System copies information from Withdrawal Marker to attributes of closed digital file.
7) System places the Withdrawal Marker in the Open Directory in the same logical location originally occupied by the withdrawn digital document.

3.5 Description

Description is "the process of analyzing, organizing, and recording information that serves to identify, manage, locate and explain the holdings of Archives and Manuscript repositories and record systems from which those holdings were selected."[35] Thus, description provides intellectual control over the Library's holdings and facilitates reference services. For paper textual document collections at the Bush Library, it involves

creation of a finding aid consisting of a scope and content note, biographical sketch when applicable, series description, and folder title list.

The creation of File Unit (folder, directory) titles was described in Use Case 3: Arrangement. The creation of File Unit titles could be performed during description instead of during arrangement. Folder title lists that are used as finding aids are immediately derivable from File Unit titles associated with a series (subseries, etc.).

Assume that an archivist is browsing a record series for the purpose of describing it.

Use Case 5: Description

1) Archivist indicates the kinds of information that should be extracted: primary subject, activities, record types, arrangement, time frame.
2) System extracts requested information from digital document files and displays it.
3) Archivist uses the information to create a series description, scope and content note, and requests that the system saves this information.
4) System saves the series description and scope and content note.

Archivists do not customarily describe series at the item level. However, as described in use cases 4.3 and 4.4 , when withdrawing documents they do describe the correspondents, their titles, the subject, and record type of documents withdrawn. Archivists inherit filing systems in which documents are described at the item level, for instance, correspondence-tracking systems such as C-Track used in the White House Office of Records Management. In the future, archivists will be accessioning records in electronic recordkeeping systems that are also described at the item level. Furthermore, access to large files of electronic mail that are not aggregated into folders would also be facilitated by descriptions at the item level. Use case 5 can be extended to extract information from each digital document file that can be retained either as attributes of an item or of a digital document file.

3.6 FOIA Processing

The Bush Presidential Library responds to a FOIA request by using finding aids for its holdings. To support quick response to FOIA requests satisfied by paper textual documents that have not been processed, the Bush Library created a folder title list of every file folder of paper records in the Staff Member and Office Files. As mentioned earlier, the directory names (folder titles) of electronic document files from the Bush Hard Drives are cryptic, usually eight characters or less. So creating a list of the subdirectory names is not likely to be of much help in supporting quick response to FOIA requests unless an archivist reviews the contents of the document files in a directory and extends the directory name. Search for digital document files relevant to a FOIA request can be performed using text-based search, rather than by a search of folder titles. The content of

the digital document files themselves can be used to determine relevance to the search criteria.

Use Case 6: Search for Digital Documents Relevant to a FOIA Request.

1) Archivist formulates a query corresponding to FOIA request and requests system search for relevant documents.
2) System searches text of document files or document surrogates (attributes of Items, XML attribute values, indexes) for relevant documents.
3) System ranks search results and displays most relevant series, most relevant directories within series, and most relevant document files within directories.
4) Archivists reviews search results and determines relevant series and directories.
5) System indicates the number of digital document files in relevant series and directories.
6) Archivist requests that the system saves the results of the search and associates them with the FOIA request.

Records responsive to a FOIA request should be identified with high precision and recall. In the next section, document retrieval technologies are reviewed and a technology that meets this requirement is recommended.

4. Information Technologies to Support Archival Processing

Information filtering, viewer/reader, and object-oriented analysis technologies have been described in prior sections. Other information technologies are needed to realize the archival support functions specified in the use case scenarios of the preceding section. These include natural language processing, knowledge-based system, and document retrieval technologies.

4.1 Natural Language Processing

Natural Language Processing (NLP) technology is of importance to archival processing because it supports automatic text interpretation. Text interpretation is of importance because it enables information extraction, generalization, categorization and indexing of documents that are needed to support archival functions.

Linguists characterize the implicit linguistic knowledge of a human language user by specifying levels of representation and then specifying the complex mapping between them. Lexical knowledge, knowledge about the individual words in a language, is represented in a lexicon. Terminological knowledge, knowledge of the terms of specialized domains, e.g., foreign policy or economic policy, is represented in ontologies. Syntactic knowledge, knowledge about the relationship that words bear to each other in a sentence, is specified as a set of pattern-matching rules, that can be applied either to

produce the sentences in a language or else to recognize whether a given string belongs to the language. Semantic knowledge, knowledge of the relationship of words to meaning, is represented with a set of interpretation rules. Discourse knowledge, knowledge of the meaning of sequences of sentences, paragraphs, or documents, is represented in terms of references, writer's intentions, and communication or speech acts.

NLP is not a single type of computation, but a set of related computations, lexical analysis, ontological analysis, syntactic analysis, semantic analysis and discourse analysis. Furthermore, NLP has proven difficult because text interpretation is so dependent on real-world knowledge outside of the information contained in the text.

4.1.1 Lexicons and Ontologies

The WordNet Project at Princeton has created a large English language lexical database that is a key resource for natural language processing.[36] Figure 10 shows the WordNet senses for the noun *transfer*.

The **noun** "transfer" has 6 senses in WordNet.

1. transportation, **transfer**, transferral, conveyance -- (the act of transporting something from one location to another)
2. **transfer** -- (someone who transfers or is transferred from one position to another; "the best student was a transfer from LSU")
3. **transfer**, transference -- (the act of transfering something from one form to another; "the transfer of the music from record to tape suppressed much of the background noise")
4. **transfer** -- (a ticket that allows a passenger to change conveyances)
5. **transfer**, transfer of training, carry-over, generalization -- (generalization of a skill learned in one situation to a different but similar situation)
6. **transfer**, transference -- (transferring ownership)

Figure 10. Senses of the Noun Transfer in WordNet.

WordNet includes a set of functions for determining from the lexical database the parts of speech and senses of a word. It also includes functions to determine a word's synonyms and antonyms, hypernyms (transfer is a kind of…), hyponyms (… is a kind of transfer), meronyms (parts of transfer), and troponyms (particular ways to transfer), preconditions for transferring, and effects of transferring. Perhaps most importantly, WordNet contains sentence frames for verbs that aid in determining the sense of a word appearing in text. The semantic relationships between concepts are sometimes represented as a semantic network.

In addition to generic lexical knowledge, natural language processing systems usually need knowledge of facts and terminology specific to the contexts in which documents are created, e.g., legislative or diplomatic terminology. These sources of knowledge are represented as terminology databases or ontologies. An *ontology*, for natural language processing (NLP) purposes, is a body of knowledge about a domain (a particular subject field) that contains primitive symbols used to represent meaning, organizes these symbols in a hierarchy, and further connects these symbols with other semantic relations among the concepts. Figure 11 shows an open knowledge base connectivity (OKBC) representation of the concepts of diplomatic representation and chancery extracted from the CIA World Fact Book.[37]

```
(DEFINE-OKBC-FRAME DIPLOMATIC-REPRESENTATION
  :SUBCLASS-OF (LEGAL-GOVERNMENT-ORGANIZATION)
  :INSTANCE-OF (CLASS)
  :OWN-SLOT-SPECS ((MY-SOURCE WORLD-FACT-BOOK-1995))
  :TEMPLATE-SLOT-SPECS ((CHIEF-OF-MISSION)
(MISSION-IN)
(COUNTRY-REPRESENTED))
)

(DEFINE-OKBC-FRAME HAS-CHANCERY-IN-US
  :OWN-SLOT-SPECS ((RANGE CHANCERY))
)
```

Figure 11. An Example of Domain Knowledge from the World Fact Book.

A natural language processing system can use these representations to interpret domain facts that are not in a document. It can also use them to interpret documents that involve diplomatic concepts and facts.

4.1.2 Syntax

Grammars for English that have been developed by linguists have thousands of rules and constraints. Parsers that are used with these grammars automatically create excellent descriptions of the syntactic structure of English sentences.

Feature logic has been developed to represent syntactic features and constraints. These feature structures are manipulated by the operation of unification, hence the term unification-based grammars. These tools have a close connection with similar developments in knowledge representation and programming, in particular with constraint-based programming. Furthermore, parsing using these syntactic rules can be viewed as a deductive process, which allows us to cast the syntactic analysis problem as a problem of knowledge-based reasoning.[38]

Text interpretation systems based on linguistic theory still lack the coverage, robustness and efficiency needed for realistic applications. Most practical applications use statistical methods and performance-oriented finite-state methods.

Probabilistic parsing is an example of applying statistical methods to syntax. For instance, syntax can be thought of as a constraint on word order. One way to represent this is to construct a list of allowable n-grams, sequences of n words, in a language. Each n-gram can be associated with the probability of a word given its n-1 probabilities. The probabilities can be estimated by determining the relative frequencies from a large corpus of texts. This reduces syntax to a Markov chain that can be used to correct an error-ridden OCR text by finding the sequence of n-grams of maximum probability.

Languages have been categorized with regard to the complexity of the patterns that their sentences must satisfy. For instance, the patterns of a regular language can be specified by a *regular expression*, a formula that indicates the order in which symbols can be concatenated, whether there are alternative possibilities at each position, and whether substrings can be arbitrarily repeated. Patterns can be defined as regular expressions in a language such as Perl.

The syntax of a language can also be specified using an automaton, a machine that operates either to produce or to recognize the sentences of a language. For instance, a finite-state machine consists of a finite number of states and a function that determines transitions from one state to another as symbols are read from an input string. The machine starts at a distinguished initial state and is positioned to read the first character of the input string. The machine transitions from state to state as it reads the input string, eventually coming to the end of the string. At that point, if the machine is in one of a set of final states, the machine has accepted the string; that is, the string belongs to the language that the machine characterizes. Finite-state machines accept the same class of languages as are defined by regular expressions.

Finite-state recognition of sentence syntax is a key element of most current text interpretation methods. Finite-state parsers have been constructed that approximate the descriptive power of context-free and unification grammars but out-perform them in parsing coverage and efficiency. Furthermore, the approximation is usually good enough for practical purposes. For instance, the information extraction problem requires that documents and passages be identified that are likely to contain information relevant to a particular user's needs. Syntactic and semantic analyses of documents using formal linguistic models of competence would certainly help to solve this problem. However, such analyzes usually provide much more information than the task actually requires. Finite-state solutions can be constructed for the information extraction process that are very efficient.[39]

Connectionist methods (artificial neural networks (ANNs)) have been explored for a number of language processing tasks. Research in applying ANNs to natural language

processing involves simulating finite-state machines. For instance it has been shown that a recursive ANN feeding back output activations to the previous layer can recognize the same strings recognized by a finite-state acceptor or generated by a context-free grammar. Connectionist models have thus far played a relatively small role in large or difficult language processing tasks, except for the information categorization task. This application of ANNs is discussed in section 4.3 on document retrieval technologies.

4.1.3 Semantic Interpretation

Two common tenets of theories of semantics are that the meaning of a sentence is a function of the meaning of its parts, and consists of the inferences that can be drawn from the sentence. The interpretation of a sentence must be represented so that it can be combined with other knowledge to determine the meaning of the sentence. Predicate logic has been used for this purpose. Knowledge representation languages such as frame languages and conceptual networks have also been used to represent meaning. Logical deduction is used for inferring meaning, but non-monotonic reasoning, abductive inference, and case-based reasoning are also needed.

As with syntax, methods of sentence interpretation based on statistical and finite-state methods have proven more efficient than those based on linguistic theory and effective enough for many applications. Word sense disambiguation is an example of a statistical linguistic method applied to semantics. The meaning of a word is dependent on the words appearing just before it or after it in a sentence. By determining the possible word sequences associated with the meaning of a particular word, we can maximize the joint probability of word sequence and word sense to improve the accuracy of interpretation.

Augmented Transition Networks (ATNs) are an example of a finite-state method of semantic interpretation. Figure 12 shows an ATN for the context-free grammar:

S→ NP VP | VP
VP → verb [NP] [PP]*

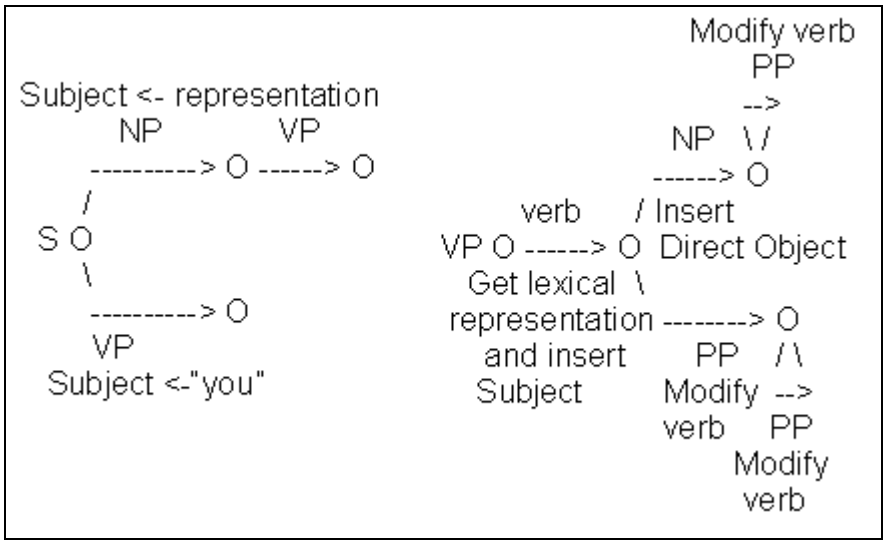

Figure 12. Sample ATN for Interpreting Text.

An ATN is a finite-state machine that parses a sentence according to a grammar. The arcs labeled VP on the left ATN invoke the ATN on the right to parse a verb phrase (VP). Some of the arcs have procedures attached to them, e.g., the arc labeled verb has the procedure "Get lexical representation and insert Subject." These procedures are executed when the arc is traversed. The lexical representation is the sentence frame from the lexicon that represents the sense of the verb. The result is a semantic representation of the meaning of a sentence.

4.1.4 Discourse Analysis

The methods of discourse analysis support the interpretation of sequences of sentences, sequences of paragraphs, and sequences of documents. There are two linguistic approaches to discourse analysis, inference-based and intentional. According to the inference-based approach, the coherence of the text derives from the semantic relationships between the information conveyed by successive sentences. As a result, the major computational tools used are inference and abduction on representations of the propositional content of sentences. This approach is particularly useful for resolving the problem of anaphora, that is, phrases that refer to objects that have been mentioned previously in the text. It also helps in resolving references to objects that are not mentioned in the text, that is, resolved using common sense knowledge.

According to the intentional approach, the coherence of discourse derives from the intentions of writers, and understanding depends on recognition of those intentions. As a result, the major computational tools used are based on the theory of speech acts and Artificial Intelligence planning techniques.

Based on the observation that questions are generally followed by answers, and proposals by acceptances or rejections, rules can be formulated that state constraints on acceptable dialogues. The primitive predicates of these rules typically check to determine whether a sentence, paragraph, or document represents an illocutionary act, such as a request, reply, offer, question, answer, proposal, acceptance, rejection, warning, suggestion, confirmation, etc.

Linguistic theories of discourse, e.g., rhetorical structure theory,[40] have great explanatory power, but computational models are usually based on finite-state methods.[41] A speech act becomes a state transition label in an augmented transition network. When a speech act is recognized in the text, the ATN makes the appropriate transition. The outgoing arcs from the resultant state allow the system to predict the next type of speech act to expect. For instance, if one document represents a proposal, the system would expect another document in a case file to correspond to an acceptance or rejection of the proposal.

Document structure or form might be the subject of discourse analysis, but linguists also study document structure in pragmatics, which is concerned with the proper use of language depending on the context of the dialogue or discourse. The content and form of linguistic expression used in a situation depends on the social position of the people involved, the matter at hand, and the intentions of those involved. This behavior is determined mostly by social factors.

The science of Diplomatics studies the genesis, inner constitution and transmission of documents, of their relationship with the facts represented in them and with their creators.[42] The Diplomatic theory of the intellectual form of documents has explanatory power, but computational models for use in text interpretation have not been developed from this theory.

Practical methods for interpretation of document structure are based on rules for parsing document structure and finite-state methods.[43] Practical methods for representation of document structure are based on markup language technology, that is, SGML, HTML and XML. SGML provides a method for defining document structure called a Document Type Definition (DTD). A DTD is a sort of formal grammar specifying the structural relations in a particular type of document. In the next section, a software technology will be described that uses grammatical rules to recognize digital document types such as letters, memos, schedules and agenda. Of course, if the document has been created with an SGML DTD, the document type is determined from the tags in the document.

4.1.5 Applying NLP Technologies to Archival Processing Functions

The technologies described in section 4 can support the system functions specified in the use case scenarios of section 3. Text interpretation and information extraction provide the information needed in use case 1, Extract Information and Register Accession. Text interpretation, summarization, generalization, and scripts support use case 5, Description.

This section gives an example of software that uses NLP technologies to support information extraction, generalization, categorization and indexing of documents. The example is for electronic records at the time of creation rather than for electronic records after they are transferred to an archive for long-term preservation, but the technology is applicable to the latter as well.

The Records Management Team is a set of software tools that uses knowledge-based text interpretation technology to support requirements for managing electronic records.[44] The Filing Assistant, a member of the Team, uses text interpretation technology to support office personnel in filing and finding records according to the office file plan. To interpret text it uses a lexicon of words, parts of speech, senses of words, synonyms and antonyms. It uses domain specific knowledge of an enterprise's terminology, forms, abbreviations and acronyms, an office's business functions and office positions. It uses a finite-state parser for interpreting the document text and descriptions of record categories. It uses knowledge of document structure and content to identify document types and to extract information from documents such as names, dates and subjects. It uses the indicated subject of a document, or if there is not one, abstracts the document to determine possible subjects. Figure 13 illustrates the user interface for extracting information to profile a document and for categorizing the document according to the enterprise/office record classification scheme.

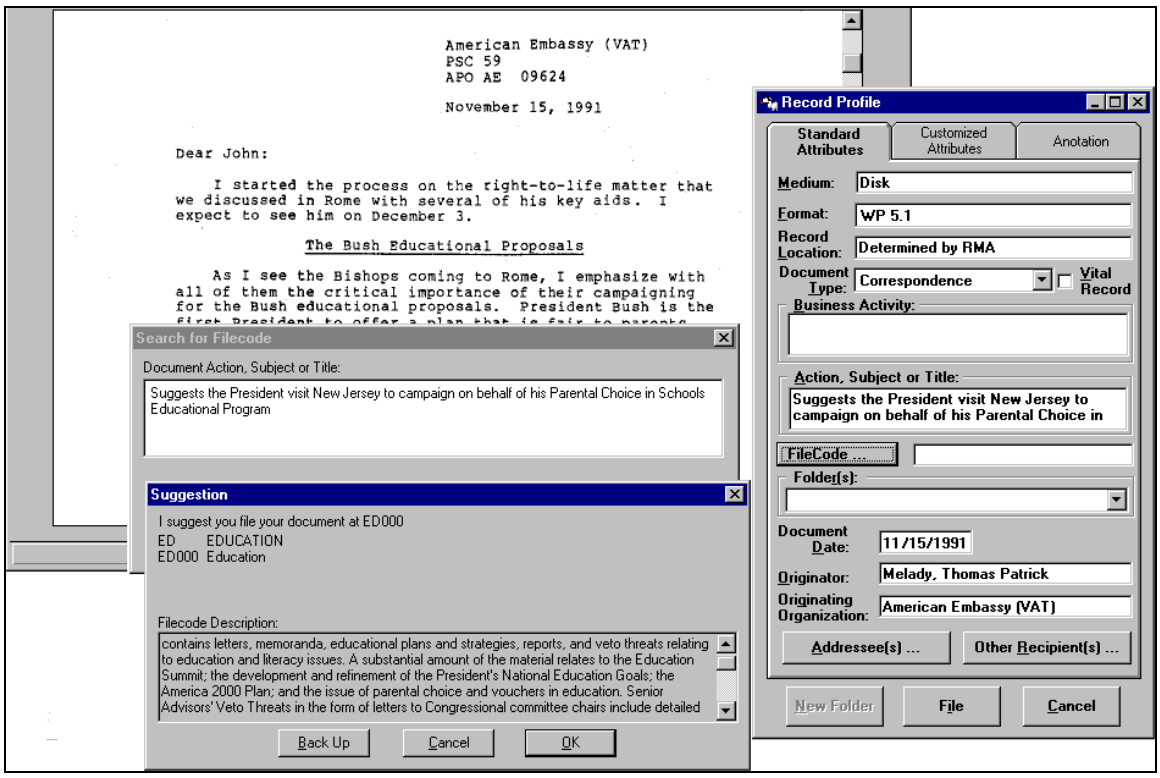

Figure 13. Illustration of Text Interpretation to Extract Information and Categorize Records.

The document shown in Fig. 13 is an open paper document from the WHORM Subject Files at the Bush Library. It was scanned, OCRed and saved in WordPerfect 5.1 format. The Filing Assistant was asked to aid in filing the document. The Filing Assistant determined the document type and extracted information from the document needed in the Record Profile dialog box. The information in the Action, Subject or Title Field was entered. This is the subject that will be saved in the record profile. The File Code button was selected. The Search for Filecode dialog box was displayed. The Subject, Action or Title could be edited in that box, but that description is not the one that will be saved in the record profile. When the user selects the OK button on the Search for File Code dialog box, the Filing Assistant categorized the document according to the WHORM subject classification scheme. To do so, it interpreted the description given in the Subject, Action and Title field using the lexicon that was previously described and an ATN interpreter. It used the representation of the meaning of the description to search through the meanings of WHORM subject category descriptions, which had been previously loaded and interpreted. It recommended the WHORM category *ED Education,* which was the subject code and category that had been originally assigned to the document. If the OK button is selected, that file code will be inserted in the field for file code in the Record Profile.

4.2 Knowledge-Based System Technology

.

Knowledge-based system technology can be used to control the text interpretation process. This section describes the application of this technology to checking documents for FOIA exemptions and PRA restrictions and for distinguishing Personal Papers from Presidential and Federal Records.

The knowledge-based system (KBS) framework is illustrated in Fig. 14. It distinguishes the declarative knowledge needed to solve a problem from the control and inference component which uses this knowledge. The problems to be solved are stored on an agenda. As the system *reasons* about the problem, it creates subproblems that it also stores on the agenda. Intermediate results are stored in a working memory. Through a user interface, the KBS framework can be used to support user decisions.

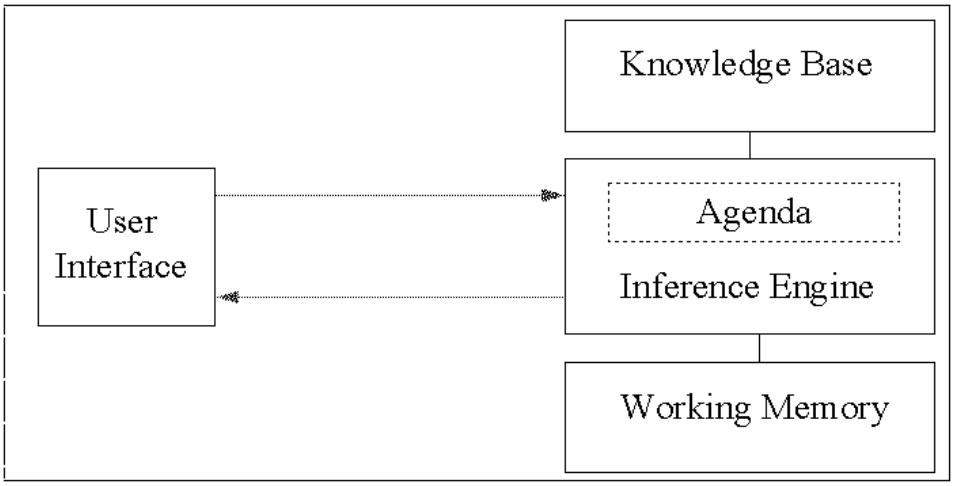

Figure 14. Knowledge-Based System Framework.

Within this general framework there are many technological alternatives that depend on choice of knowledge representation and inference mechanisms. For instance, rule-based (expert) systems technology represents knowledge as if-then rules and uses forward or backward chaining as the inference mechanism. The rules may include the use of confidence factors to address the need to represent knowledge and conclusions that are not definite. Logic programming is an alternative implementation of this framework that represents knowledge as rules and facts and uses resolution as a deductive inference mechanism.

4.2.1 PRA and FOIA Review

In section 3, the kinds of knowledge needed to interpret text were identified. These include lexical knowledge, commonsense knowledge, domain specific knowledge, syntactic knowledge, and semantic interpretation knowledge. To this must be added task specific knowledge, for instance, knowledge of PRA restrictions and FOIA exemptions.

A knowledge-based system for text interpretation begins with a model of concepts relevant to its goals. For FOIA or PRA review, these are the concepts that an archivist would believe to be related to limitations on release of information. These concepts are used to guide the information extraction process. The KBS first expands lexical and domain knowledge structures related to the concepts. Then it uses these knowledge structures to process the text of a document. Information is extracted from the document that matches the conceptual representations it is looking for. The result is a representation of the meaning of the document. The KBS then checks the meaning against PRA/FOIA rules and prunes away those that do not satisfy the conditions of the rules. Then the terms, sentences and paragraphs in the document that are most relevant to the limitations are highlighted on the user interface.

Fig. 15 shows some if-then rules for PRA restrictions.

IF: There is evidence of someone being appointed to Federal Office at
text location (p, l),
THEN: There is evidence of Appointment to Federal Office (PRA (a (2)),
Create next possible restriction in document at text location (p, l).

IF: Addressee of document is the President, and
Author of document is an advisor to the President, and
There is evidence of advice to the President at text location (p, l),
THEN: There is evidence of Confidential Advice (PRA a(5)),
Create next possible restriction in document at text location (p, l).

IF: Document has been checked for PRA restrictions, and
There is evidence of PRA restriction in document,
THEN: Highlight text locations of possible restrictions.

IF: User selects highlighted text in document
THEN: Display PRA Checker dialog box.

Figure 15. Sample If-Then Rules for PRA Restrictions.

The PRA checker uses the lexicon to expand the concept of advice to include concepts such as counseling, consulting, recommending, urging, persuading, dissuading, warning, discouraging, admonishing, requesting, submitting, advocating, proposing, approving disapproving, suggesting, advancing, discussing, commending, evaluating, assessing, appraising, and judging. It uses knowledge of the persons who are advisors to the President, and kinds of Federal Offices to which the President may make appointments. It extracts information from the document that matches the conceptual representations that it is seeking. It then checks these against the conditions of the PRA rules. Figure 16 shows the results of a prototype PRA checker used to demonstrate these concepts. It was applied to a copy of a Presidential Record that has been opened by the Presidential Library.

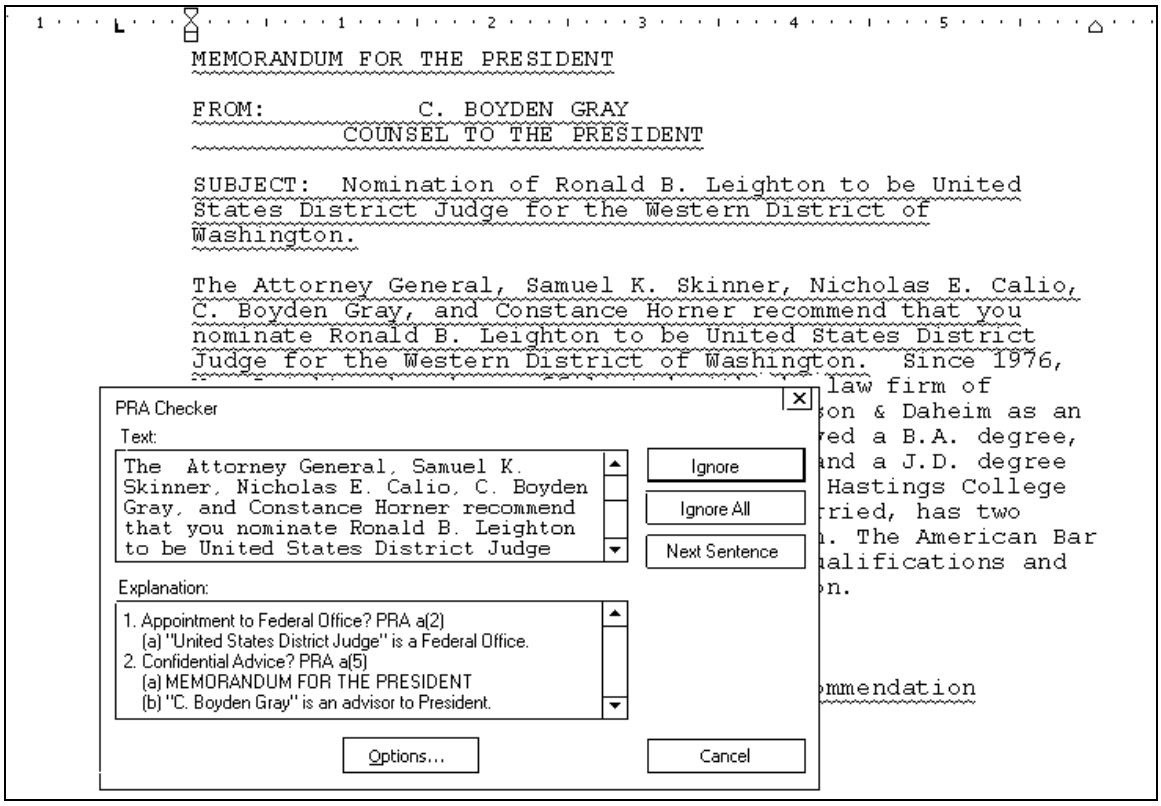

Figure 16. User Interface to PRA Checker.

The PRA checker concludes that there is evidence that the document is a candidate for withdrawal under the restrictions of paragraphs a (2) and a (5) of the Presidential Records Act.[45]

Much of the knowledge that archivists use to judge whether parts of a document are related to a PRA restriction or FOIA exemption seems to be based on their prior experience and in particular on prior cases. This seems to call for case-based knowledge and case-based reasoning.[46] Case-based interpretation is used to form a judgment about or to classify a new situation by comparing and contrasting it with cases that have already been classified.

To construct case-based knowledge for a PRA/FOIA checker, a corpus of digital documents is needed that archivists have already reviewed and classified as to restriction or exemption. A method such as cluster analysis can be used to identify features of the cases common to a restriction that distinguish them from other cases. Cluster analysis can also be used to identify features that distinguish subcases within a restriction category. To determine relevant cases, features of a new case are compared against the features of cases in the stored knowledge base. The performance of a system based on case-based reasoning can be self-improving because it can automatically incorporate the results of new PRA/FOIA decisions into the case knowledge base.

4.2.2 Distinguishing Personal Papers from Presidential and Federal Records.

When reviewing unprocessed documents, Archivists at the Bush Presidential Library must distinguish Presidential Records and Federal Records from Personal Papers. Personal Papers are not owned by the Library unless they are donated. The text interpretation and knowledge-based system technology described for FOIA/PRA review can also be applied to distinguishing Personal Papers.

Personal Papers, in the context of Presidential Records, are defined as "all documentary materials, or any reasonable segregable portion thereof, of a purely private or nonpublic character which do not relate to or have any effect upon the carrying out of the constitutional, statutory, or other official or ceremonial duties of the President. Such term includes –

(A) diaries, journal, or other personal notes serving as the functional equivalent of a diary or journal which are not prepared or utilized for, or circulated or communicated in the course of, transacting Government business;
(B) materials relating to private political associations, and having no relation to or direct effect upon the carrying out of constitutional duties of the President, and
(C) materials relating exclusively to the President's own election to the office of the Presidency; and materials directly relating to the election of a particular individual or individuals to Federal, State or local office which have no relation to or direct effect upon the carrying out of constitutional, statutory, or other official or ceremonial duties of the President."

This is not an operational definition that would enable a person or system to determine whether a document was a Personal Paper. An operational definition can be constructed from examples of Personal Papers, for instance,

IF: Document is a personal diary, or
Document is a personal journal, or
Document is personal notes, or
Document is a recipe, or
Document is a Christmas card mailing list, or
Document is a grocery list, or
Document is a letter and letter is from a parent to children,
THEN: There is evidence that the document is a Personal Paper.

In section 4.4, a method was described for defining document types. This method can be used to define and identify the document types in the preceding rule. The conclusions drawn from such a rule are not definite conclusions, for instance, a resume that is part of a job application to work at the White House is not a Personal Paper, but may contain information subject to Privacy Act restrictions.

4.3 Document Retrieval

Document retrieval technology can support response to FOIA requests and Citizen access to opened Presidential Library holdings. There are two traditional approaches to text-(content-) based document retrieval—Boolean and statistical retrieval models. The Boolean retrieval model uses a query specification expressed as words or phrases combined using the operators of Boolean logic. All documents containing the exact combination of terms specified in the query are retrieved.

A major limitation of this retrieval method is that it provides no means of ranking documents by relevance to the query. Another limitation is that it excludes some documents that do not precisely meet the query specification even though they may be relevant.

Statistical retrieval models are an improvement over the Boolean retrieval model in that they do not require an exact match and they rank relevant documents. The vector space model treats queries and texts as vectors in a multidimensional space. The dimensions of the space are the terms occurring in the texts. The terms of the query are weighted to account for their relevance to the user, and the terms in the texts are weighted to account for their statistical distribution. Best matches of queries and texts are determined by comparing their vectors.

Probabilistic retrieval models compute the probability that a user's information need is satisfied given the evidence in a particular document, and thus to produce a ranked list of documents. The source of such evidence is typically the statistical distribution of terms in all the documents and in relevant and non-relevant texts. For instance, Automony's knowledge server is a COTS product that uses a proprietary method known as Adaptive Probabilistic Concept Modeling to classify and retrieve documents. The method combines Bayesian probability with an artificial neural network.[47]

Many content-based document retrieval systems are implemented using index files of the most frequent document terms or of document surrogates (attributes or metadata extracted from the text). These surrogates are constructed using information extraction techniques such as those described in section 4.1.

Content-based methods of document retrieval that rely solely on keywords to identify relevant texts often fail because they involve no understanding of the text of the documents. For instance, keyword matching often fails due to polysemy (more than meaning for a single word) and synonymy (many ways of referring to the same concept).

Another method of document retrieval is so-called conceptual search and retrieval. A text interpreter can use a lexicon (or a semantic net) to distinguish the specific meaning of a term in a document from other possible meanings. A user's search request can also be used to determine the meaning of the terms in the request. Furthermore, if the meaning of the term in the search request is synonymous with the meaning of a different term, the

search request can be extended to search for terms that have the same meaning. This document retrieval technology can substantially reduce the consideration of irrelevant documents while increasing the retrieval of relevant documents.

Excalibur Technologies' RetrievalWare[48] is a COTS system that supports Boolean, pattern and concept search. RetrievalWare's conceptual search uses lexical knowledge represented as a semantic network. Figure 17 shows the results of a search of News Wire data from September and October 1996 using a conceptual query "Impact of low petrol prices."

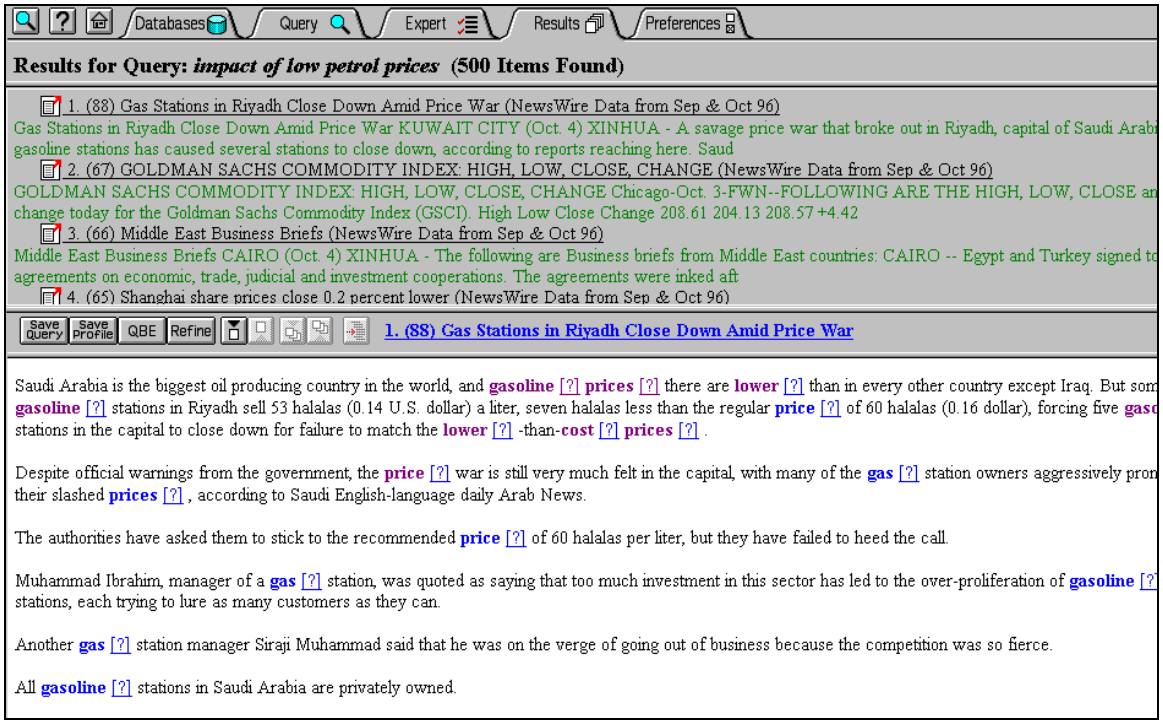

Figure 17. Sample Screen from RetrievalWare's Concept Search.

The results are ranked by relevance to the query. Words synonymous with petrol such as gas, gasolene, and gasoline were used in the search. The search results were of higher precision and recall than those that would be obtained with a Boolean search of the original search terms.[49]

The results of a cluster analysis of the concepts in a collection of documents can be displayed as a topographic map. Dots on the map represent documents. Peaks represent concentrations of similar documents. The major concepts or subjects of documents in an area of the map can be displayed. A Boolean search on topics results in the display of the map locations of documents meeting the Boolean conditions. The documents at those locations also can be displayed. Fig. 18 shows a sample screen from ThemeScape, a COTS product that exhibits these features.[50]

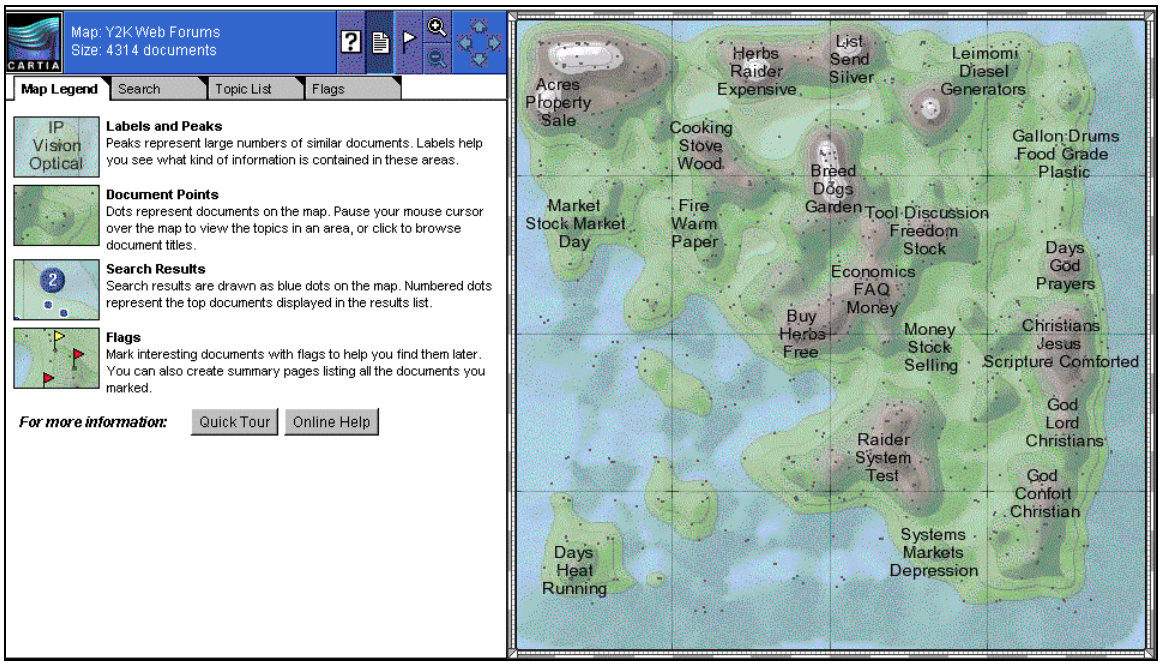

Figure 18. A Sample Screen from ThemeScape.

The screen shows an information map for 4314 documents from Y2K web forums. One of the features of this type of interface is its support of relevance feedback to the searcher.

Acquaintance is a system developed by NSA/DoD that clusters documents using n-grams.[51] It also provides an interface similar to that of Themescape for interactive retrieval of similar documents. The Army Research Lab has used this tool for categorizing large volumes of e-mail messages that had not been filed according to a records classification scheme.[52]

In summary, search of document surrogates or conceptual search and retrieval with relevance ranking of search results are the preferred document retrieval methods to support FOIA response, because they provide the best precision and recall. However, interactive search of maps that cluster documents with similar features, can also enhance FOIA response, by providing visual relevance feedback.

4.4 Technologies to Support Redaction

To comply with the 1996 Electronic FOIA (E-FOIA) Amendments, Public Law No. 104-231, redaction software must support redaction of words, sentence and pages, indicate the length of material redacted, and show the reason for redaction with FOIA codes. There are two technologies that meet these requirements: (1) redaction of character encoded documents in a proprietary format, and (2) redaction of images of documents or images embedded in character encoded documents.

The *Word DOT Redactor* is freeware developed by the Department of Veteran Affairs.[53] It is a document template (DOT) for Microsoft Word 7.0 and 8.0. Assume, for the purpose of example, that the phrase "married, has two children and resides in Tacoma, Washington" in the document shown in Fig.19 needs to be redacted because it contains information that if provided to the public would constitute an unwarranted invasion Leighton's personal privacy. Text to be redacted is highlighted. The reason for redaction is selected from a set of buttons corresponding to FOIA exemptions.

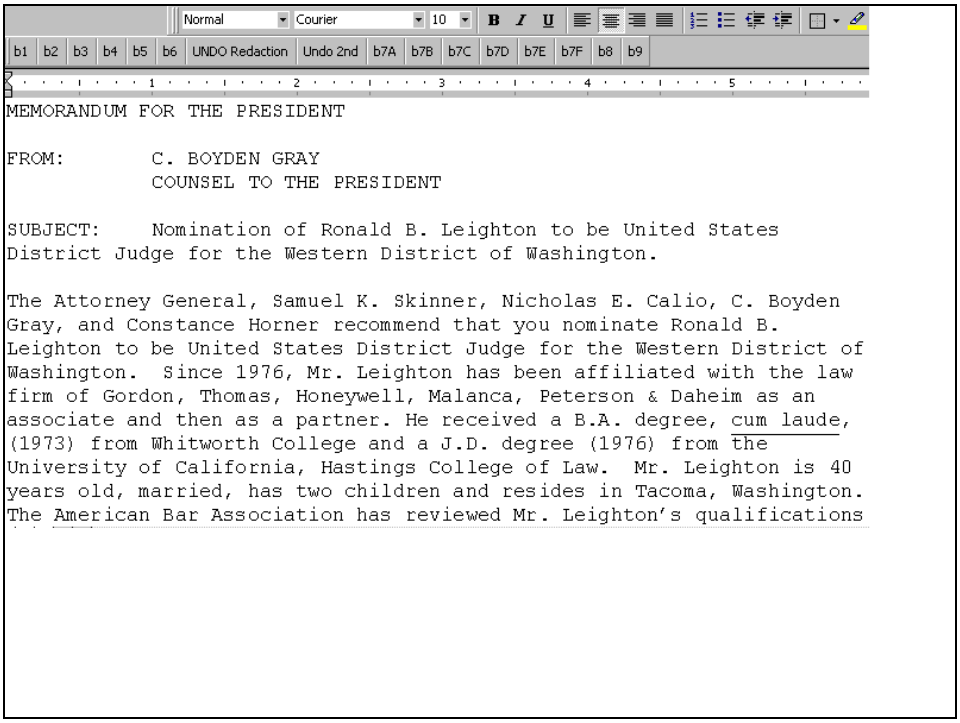

Figure 19. Sample Document Needing Redaction.

A macro associated with each button replaces the highlighted text with the FOIA exemption code and dots as shown in Fig. 20.

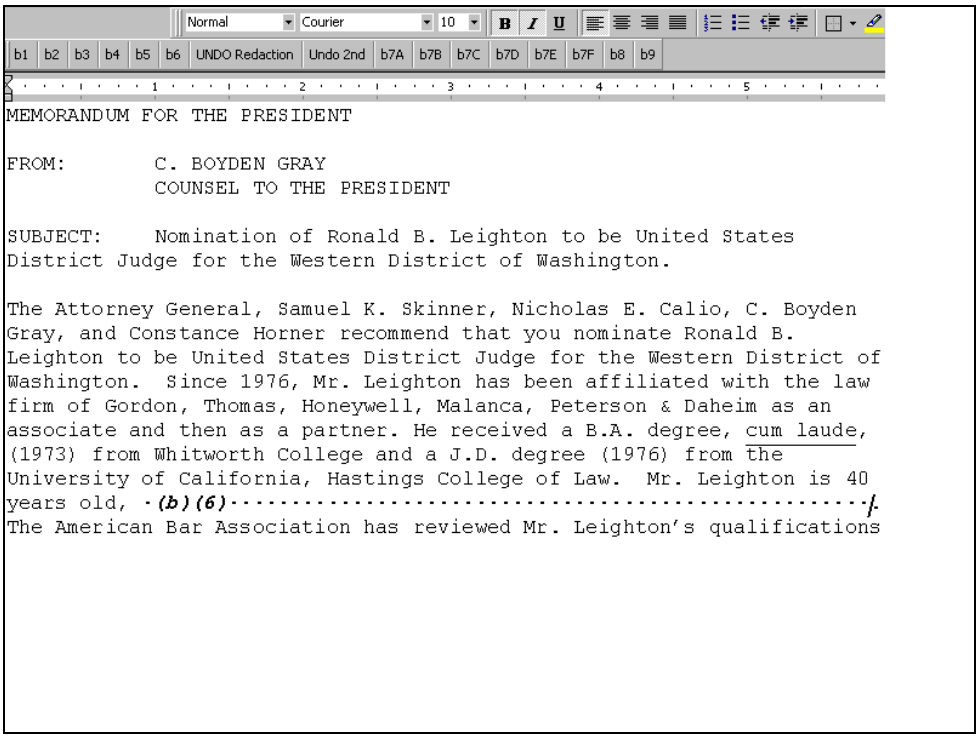

Figure 20. Redacted Phrase in a Sample Text Document.

The macros are easily modified to correspond to PRA restriction codes. The DOT redactor works only on text, not on embedded images.

The Standard Edition of *Imaging for Windows* from Eastman software is free.[54] It can import document images in tif, awd, bmp, jpg, pcs, and xif formats. It outputs tif, or awd file types. It redacts with a solid box and has a customizable rubber stamp. Rubber stamps can be used to indicate the reason for the exemption and to identify the redactor and the date.

The Professional Edition of *Imaging for Windows* reads document images in awd, bmp, dcx, gif, jpeg, pcx, tif, wif and xif formats. Optically recognized images can be saved as text. It can save data as Hypertext Markup Language, bmp, awd and tif files. It can straighten crooked images. It can annotate with a highlighter, sticky note, solid and transparent boxes, or a customizable rubber stamp.

Redax is a plug-in for Adobe Acrobat Exchange that supports redaction of text and images from PDF (portable document format) documents.[55] Redacted text is replaced with the reason for exemption and a dashed line between brackets. You can also redact all or part of an image in a PDF document. Redacted images are replaced with black pixels. You can search for specified text and overlay it with specified exemption codes.

5. Summary and Recommendations

The contents of the Bush hard drives were analyzed to identify issues that need to be addressed to gain intellectual and physical control of the document files. There are on average 579±109 files per hard disk that are system files and 342±139 user created files per hard disk. There is a need to support archivists in distinguishing user-created documents from operating system and office application software files. A prototype document file filter was developed that interactively supports these decisions. An experiment was conducted to test the filtering rules. The experiment indicated that to preserve all user-created files, a rule was needed to check file formats as well as file extensions. To minimize the risk that a file judged to be a system file is actually a document file, it is also possible to save one copy of each unique file judged to be a system file.

An archivist from the Bush Presidential Library experimentally accessioned document files from 40 of the Bush hard drives. To do so, it was also necessary to establish the White House Office and staff member associated with the Bush hard drive. It was discovered that in addition to Presidential Records, there are Federal Records and Personal Papers on the drives.

User-created files created in fourteen different proprietary file formats were discovered on the Bush hard drives. The Quick View Plus viewers can be used to view most of the user-created files from these drives. Viewers for the remaining file formats are easily constructed. File readers are also needed, if text interpretation technologies are used to support archival tasks.

Preservation of user-created files in proprietary file formats requires that viewers for these file formats be migrated when there are changes in computer and software technology. Research is needed to identify alternatives to migration of legacy viewers for proprietary file formats.

The di Genova hard drives were analyzed with the aid of the document file filter. Some of the hard drives contain the ZyIndex application software for indexing the text of the document files and its associated indexes. The indexing technique used by ZyIndex is hardware dependent. The ZyIndex software and indexes to the documents are obolete and should not be preserved. Any of a number of current hardware-independent, content-based, document retrieval tools could be used to support Boolean search of the document files from the di Genova hard drives.

An object-oriented use case model was developed to characterize the kinds of support needed by archivists in processing of electronic records created on personal computers. Archival support functions were specified for accession, preservation, arrangement, review, description, and response to FOIA requests.

An archivist must arrange accessioned collections within the current holdings and may need to rearrange a record series into a coherent logical order. A data model is needed to represent the types of archival objects in a digital archive and to relate archival metadata to digital document files stored in digital file systems. An object-oriented model of archival objects was described that represents the arrangement of archival objects in a digital archive. Software tools to support archival tasks will use operators associated with the classes of objects in the model. The model also identifies the kind of information needed to segregate open from closed files. Object-oriented software and database technology offers the National Archives the most cost-effective path to implementation and maintenance of archival software due to its support of software reuse.

Information technologies that are needed to realize these capabilities include natural language processing, knowledge-based systems, and document retrieval. These technologies were described and specific technology alternatives evaluated as to their utility in supporting archival processing requirements.

Natural Language Processing (NLP) technology is of importance to archival processing because it supports automatic text interpretation. Text interpretation is of importance because it enables information extraction, generalization, categorization and indexing of documents that are needed to support archival functions such as accession, review and description. Lexicons such as WordNet are used to support text interpretation. Domain specific knowledge such as White House Offices and Staff Members names, advisors to the President, and lists of Federal Offices appointed by the President, are also required for interpreting presidential electronic records. Finite-state methods can be used for parsing, interpreting, and extracting information from PC files. Such methods can also be used for determining document types such as memos, letters, schedules and agenda.

To respond to FOIA requests, a tool is needed to support FOIA search of unprocessed electronic records. Document retrieval technologies that can support this function were described. Search of document surrogates and conceptual search with ranked search results are recommended as the best document retrieval methods.

PRA and FOIA review of Presidential and Federal records is the major bottleneck in providing citizen access to archived collections. Automated tools are needed to support this intellectually demanding task. Text interpretation technology combined with rule-based representation of PRA restrictions and FOIA exemptions was recommended as an approach to supporting the review function. The feasibility of this approach was demonstrated with a prototype PRA checker.

To meet the requirements of electronic FOIA, redaction tools must be able to support the redaction of character-encoded digital documents and of document images. The technology to support the redaction of document images is well developed, but additional work is needed to produce a redactor for character-encoded digital documents that are not based on a proprietary file format.

Some of the information technologies described are available to Federal Agencies at low or no cost, because they are products of federally sponsored research. Some needed technologies are available commercially. However, to satisfy the archival processing requirements identified in section 3, additional software development is needed.

During the process of accessioning and processing the PC records from the Bush hard drives, it is recommended that the information technologies and resources identified in this report be acquired and used to develop software to support each of the archival functions defined in section 3. The result of such an effort would be an archival toolkit that could be applied to future accessions of Presidential and Federal Records. Such a toolkit would support gaining archival control of electronic records, support increased productivity, and reduce the time from accession to opening of Presidential and Federal records for Citizen access.

Appendix A: Notation Used in the Class Diagrams

This appendix defines the notation used the class diagrams in section 3. It also shows some of the operators that have been identified for the class diagram of an archival system.

-

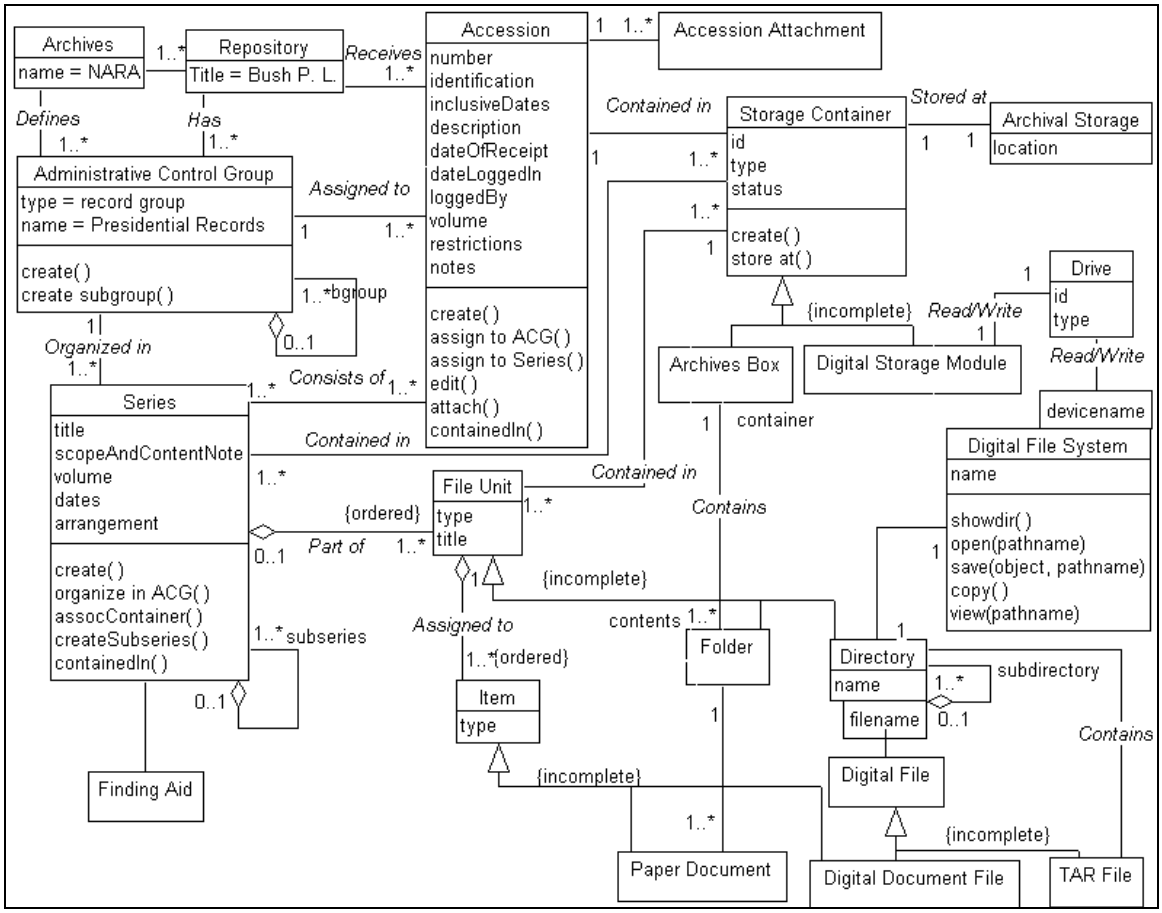

Figure 21. Class Diagram Showing Some Operators for Types of Archival Objects.

Types (classes) of objects are depicted as a 3-part box, with the class name in the top part, a list of attributes in the middle part, and a list of operations in the bottom part. For instance, create() is an operator in the class Administrative Control Group that creates an object that is an instance of an Administrative Control Group. In the same class is an operator that creates a subgroup of an Administrative Control Group.

A solid line between classes depicts an association relationship. Association names are italicized and are verbs or verb phrases. For instance, an Archives Box contains one or more folders.

The cardinality of associations is indicated by 1, 0..1 (zero or one), 1..* (one or more), 0..* (0 or more). For instance, an Administrative Control Group is assigned to one or more Accessions. If cardinality is not indicated, the association relationship is one-to-one.

A *qualified association* relates two object classes and a qualifier. The qualifier is a special attribute that reduces the multiplicity of an association. The qualifier distinguishes among the set of objects at the many end of an association. For example, in Fig. 21, *filename* is a qualifier. A digital document file corresponds to a directory and a filename.

A *role name* is a name that uniquely identifies one end of an association. Each role name on a binary association identifies an object or set of objects associated with a class of objects at the other end. Role names are nouns, possibly modified by an adjective, and often followed by the preposition *of*. For instance, an Archives Box is a container of one or more folders. A folder is the content of an Archives Box.

A solid line with a diamond at one end depicts an aggregate relationship. The diamond end designates the client class, which is sometimes called the aggregate class. An instance of the aggregate class is an aggregate object. The class at the other end of the relationship is called the supplier class. It is the part whose instances are contained by the aggregate object. If the diamond is hollow it indicates that the client class has a pointer or reference within it to an instance of the other class, that is, it indicates logical containment. For instance, a File Unit is part of a Series.

A generalize relationship between classes shows that a subclass shares the structure or behavior defined in one or more superclasses. A generalize relationship shows an “is-a” relationship between classes. A generalize relationship is depicted with a solid line with an arrowhead pointing to the superclass. For instance, a Folder is a File Unit.

A constraint is an expression of a semantic condition that must be preserved while the system is in a steady state. A constraint can apply to an association as a whole or to a particular role. Constraints on relationships are displayed surrounded by braces. For instance, the constraint {ordered} indicates that the file units that are part of a series are ordered. It does not state what the order is because the order varies from series to series.

Appendix B: Glossary of Class Names in Class Diagrams

Sources:

ARC Glossary of NARA Archival Data Model
SAA SAA Glossary
NARA A Federal Records management Glossary, Second Edition, 1993.

Accession

1. An addition to the holdings of an archives, whether by transfer under an established and legally based procedure, by deposit, purchase, gift, or bequest. (compare SAA accession)
2. An accession so recorded (compare SAA Acquisition and Accession List/Register)

Accession Attachment

Records transmittals, finding aids, software documentation or any other materials that accompany an archival accession.

Administrative Control Group

Identifiers used to control archival materials, such as Record Group numbers and titles; Collection identifiers and titles. (ARC)

Archival Storage

1. The storage areas in a Records center, archives, or manuscript repository for shelved material.
2. The shelving units in such a storage area. (Compare SAA Stacks)

Archives

The organization or agency responsible for appraising, accessioning, preserving and making available permanent records. Also called archival agency. In the U. S. Government, the National Archives and Records Administration (NARA). (NARA)

Archives Box

A storage container, variable in terms of composition, construction, and dimensions, intended to protect and facilitate the handling of archival materials. Archives boxes are also called manuscript boxes. (SAA)

Cartridge

A closed container of film or of tape, designed for loading and unloading in a reader, projector, recorder, or computer tape drive, without prior threading or rewinding. (SAA)

Closed File

1. A File unit or series containing documents on which action has been completed and to which additional documents are not likely to be added. See also Open File (1). (SAA)

2. A file unit or series to which access is restricted or denied. See also Open File (2). (SAA)

Closed File Unit
See Closed File

Collection
A grouping of records/archives created by private individuals and organizations. (SAA)

Digital File
1. Data in a computer system that is saved on a digital storage module and capable of being manipulated as an entity. A file must have a unique name within a directory. Also called file.
2. A related collection of records. Also called a data set.

Digital File System
1. A system for organizing directories and files.
2. The collection of directories and files stored on a given digital storage module.

Digital Storage Module
A self-contained digital storage medium that is used in combination with a device for reading or writing files on the medium, for example, a magnetic tape cartridge.

Directory
A kind of digital file that contains a list of digital file names and the locations of the corresponding digital files on a digital storage medium, e.g., a magnetic tape cartridge or hard disk. Depending on the operating system, also called a catalog or folder.

Document
1. Recorded information regardless of medium or characteristics. (SAA)
2. A single item. (SAA)

Drive
An electromechanical device that reads and writes digital information on a storage medium contained in a digital storage module.

File unit
An organized unit (folder, volume, etc) of documents grouped together either for current use or in the process of archival arrangement. (SAA File (1))

Finding Aid
The descriptive tool, published or unpublished, manual or electronic, produced by a creator, records center, archives or manuscript repository to establish physical control and/or intellectual control, over records and/or archival materials. Basic finding aids include local, regional, or national descriptive databases; guides; inventories; registers;

location registers; catalogs; special lists; shelf and container lists; indexes, calendar and, for electronic records, software documentation. (SAA)

Folder
A folded sheet of cardboard or heavy paper serving as a container for a number of documents. (SAA)

Item
The smallest indivisible archival unit, e.g., a letter, a memorandum, report, leaflet, or photograph. (SAA)

NARA
See Archives.

Open File
1. A file to which documents are being added. (SAA)
2. A file with no restrictions as to access as distinct from a closed file. (SAA)

Open File Unit
See Open File.

Record Group
A body of organizationally related records established on the basis of provenance by an archives for control purposes. A record group constitutes the archives of an autonomous recordkeeping corporate body. Collective record groups and general record groups represent modifications of this basic concept for convenience in arrangement, description, and reference service. (SAA)

Redacted Digital File
A digital file in which text has been suppressed because of exemptions from disclosure under the Freedom of Information Act or restrictions under the Presidential Records Act.

Repository
A place where documents are kept. Repository is frequently used synonymously with depository. (SAA)

Series
File units or documents arranged in accordance with a filing system or maintained as a unit because they result from the same accumulation of filing process, the same function, or the same activity; have a particular form; or because of some other relationship arising out of their creation, receipt, or use. A series is also known as a record series. (SAA)

Series Description
A written analysis describing a series, usually including such elements as the series title, scope and content notes, size or volume, inclusive dates and/or Bulk dates of the material, arrangement and subjects dealt with by the series. (SAA)

Shadow Directory
An exact copy of a directory containing open digital files that is used to store digital files closed under restrictions of the Presidential Records Act or exempt from disclosure under FOIA, withdrawn security classified digital files or withdrawn personal digital files.

Storage Container
Something that holds things, especially for storage or transport, for example, an archives box, or a digital storage module.

Storage Medium
The physical material in or on which information may be recorded (e.g., clay tablet, papyrus, paper, parchment, film, magnetic tape. (SAA medium)

Subgroup
A body of related records within a record group or fonds, corresponding to administrative subdivisions in the originating agency or organization or, when that is not possible, to geographical, chronological, functional, or similar groupings of the material itself. When the creating body has a complex hierarchical structure, each subgroup has as many subordinate subgroups as are necessary to reflect the levels of the hierarchical structure of the primary subordinate administrative unit. (SAA)

Subseries
A body of documents within a series readily identified in terms of filing arrangement, type, form, or content. (SAA)

TAR File
TAR is an acronym for *T*ape *AR*chive. It is a method used to group digital files into a single file.

Volume
1. Manuscript or printed sheets bound together in a cover (SAA)
2. The physical space occupied by a group of documents. (SAA)

Withdrawal Marker
A document (or digital file) placed in a folder (directory) in place of a document (digital file) that has been withdrawn and placed in a shadow folder (directory).

Appendix C: Interaction Analysis and Sequence Diagrams

Figure 22 shows a sequence diagram constructed during the analysis of the Arrangement Use case and represents one scenario.

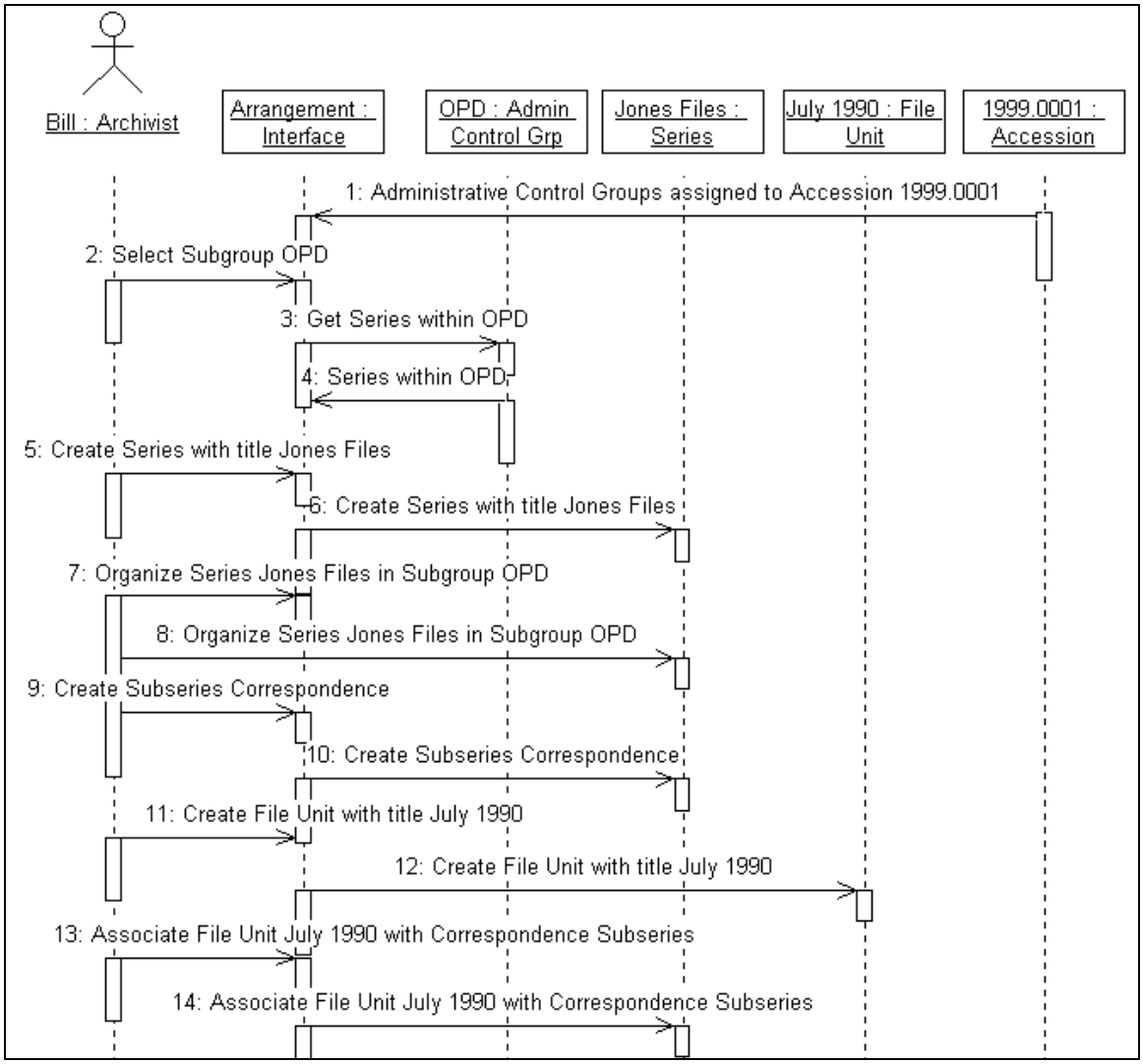

Figure 22. Example of Sequence Diagram for Arrangement Use Case.

The analysis begins with the objects Bill, OPD (Office of Policy and Development) Nancy Jones Files, a Correspondence directory, and a July 1990 subdirectory. Each of these objects is represented in a label in a rectangle at the top of the diagram. The term following the name of an object is the name of the type (or class) of the object.

The scenario begins with the assumption that Bill, an Archivist, has asked for a display of the Administrative Control Group assigned to Accession 1999.0001. The Accession class sends the requested information to the Arrangement (user) interface object which displays

it. Bill selects the subgroup OPD. The user interface sends a message to the Administrative Control Group Class to get series organized in that control group. The Administrative Control Group has an operator that enables it to get these and sends them back to the user interface which displays them.

Objects communicate with each other by sending messages. Classes and objects have operators associated with them that enable them to change their attributes, get the values of attributes, and to establish or change relationships between objects. This interaction analysis technique is used to identify classes of objects, their attributes and the operators associated with a class of objects. The relationships between objects shown in the class diagram of section 3 exist because the objects *know* about and communicate with each other, i.e., interact with each other. Operators associated with the class of the objects are used to establish and determine the relationships. The reader can, if he or she wishes, continue through the sequence of interactions between objects to see the other classes of objects, operators and relationships that were identified from the analysis of this particular scenario.

The object-oriented model (class diagram) created from this kind of analysis is not ad hoc. The interaction analysis technique is empirical and systematic. It abstracts concepts from real objects and data about the objects. The analysis is constrained by the author's analysis and understanding of the archival processing of paper documents by the Bush Presidential Library Staff, but extends that understanding to digital documents. It is also constrained by definitions of archival concepts (see Glossary in Appendix B), but extends those definitions to include classes of objects needed for archival processing of digital documents contained in a digital file system.

Sequence diagrams are being constructed for each of the use cases described in section 3. They can be the basis of user-oriented functional requirements for archival processing. The class diagrams in section 3 and Appendix A show the classes and some of the attributes, operators and relationships that have been identified so far. They can be the basis of the information requirements for a system for processing collections of both paper and digital documents in a Presidential Library.

Notes

[1] K. Thibodeau. Electronic Records Research at the US National Archives, *International Seminar: The Future of Archives, the Archives of the Future*. Associazione Nazionale Archivistica Italiana, Cagliari, Sardinia, Oct. 29-31, 1998.
[2] FBI Facsimile from Special Agent Harvey Barlow, FBI to Attorney Alan Strasser, February 8, 1993.
[3] End of Task Report, Facsimile from Jim Gilbert, Raytheon E-Systems to Bruce Ambacher, NARA.
[4] The files on hard disks from the Office of Management and Budget are Federal Records, not Presidential Records.
[5] This document and others shown in this report are not from the Bush hard drives. The document files on those drives have not been reviewed for PRA restrictions or FOIA exemptions. This and other documents shown are digital copies of paper copies of Presidential Records that have been opened by the Bush Library in response to FOIA requests.
[6] D\WP5FILES\MAR90\SPEECH.COM is a document file that contains an address (speech) to the Commonwealth Club.
[7] D\LABOR\EMPLOY.PRS is a user file concerning employment. D\WP5\INVSALES.PRS is a user file concerning sales.
[8] Includes WP Office 3.0 Calendar & Notebook software application files.
[9] Includes Books office application software files and game files.
[10] Includes game files.
[11] Includes Books office application software files.
[12] Includes Novell network system files.
[13] Includes game files.
[14] Quick View Plus 5.1 provides viewers for over 200 file formats and is particularly strong in providing support for legacy software applications such as those that run under DOS. It retails for $59.
[15] WordPerfect Corporation. *Developer's Toolkit*, 1991.
[16] The figure includes only portions of the actual Bush Library Accession Form. Other fields relevant to donations of Personal Records have been omitted.
[17] The current version of ZyIndex is a case in point. http://www.zylab.com/
[18] P. Norton and R. Jourdain. *The Hard Disk Companion*, New York: Brady Books, 1988.
[19] W. Underwood. AS-IS IDEF Activity and Data Models of the Archival Processing of Presidential Textual Records. Georgia Tech Research Institute, January 1999.
[20] I. Jacobson. *Object-Oriented Software Engineering—A Use Case Driven Approach*. Addison-Wesley, 1992.
[21] *UML Semantics*. Ver 1.1, 1 Sept 1997 (http://www.rational.com/uml/)
[22] http://www.rational.com CASE is an acronym for Computer-AIDED Software Engineering.
[23] An IDEF0 activity model would not show this as clearly as a use case model does, because the IDEF0 diagrams and method do not support specification of user interfaces.
[24] This model reflects the review and suggestions of Ken Thibodeau including the adoption of NARA's archival concept of an Administrative Control Group. However, any oversights or errors remain the responsibility of the author.
[25] Specifically, zero or one series has one or more file units as parts. If a series contains subseries, the series may not be associated with file units, but the subseries are. Furthermore, a simplifying assumption has been made. A record series has at least one file unit. If a record series consists only of items that are digital document files, this file unit is a directory of filenames.
[26] It is from the aggregation of file units within series (or subseries) and their association with storage containers that folder title lists used as finding aids are derived.
[27] The constraint {incomplete} indicates that there are other storage containers, e.g., document cases for oversized documents.
[28] The constraint {incomplete} indicates that there are kinds of file units other than folders and directories, for instance, volumes.
[29] The constraint {incomplete} indicates that there are kinds of items other than paper documents and digital document files, for instance, microfilm documents, audio recordings, and motion picture documentaries.

[30] Instances of a digital filing system could be DOS, Unix, NT, MacIntosh, etc.
[31] For instance, there are software application files, system files, and even directories and subdirectories are digital files.
[32] *Guide for Managing Electronic Records from an Archival Perspective*, International Council on Archives, 1997.
[33] L. J. Bellardo and L. L. Bellardo. *The Glossary of Archivists, Manuscript Curators, and Records Managers* (Chicago: Society of American Archivists) 1992.
[34] Frederic M. Miller. *Arranging and Describing Archives and Manuscripts*. Chicago: The Society of American Archivists) 1991, p. 60.
[35] Bellardo and Bellardo. *The Glossary of Archivists.*
[36] C. Fellbaum (ed.). *WordNet: An Electronic Lexical Database*. MIT Press, 1998.
[37] There are other knowledge representation languages than OKBC for representing ontologies, e.g., the knowledge query manipulation language (KQML) and the Knowledge Interchange Format (KIF), draft proposed American National Standard, NCITS.T2/98-004.
[38] S. M. Shieber. *An Introduction to Unification-based Approaches to Grammar*. CSLI, Stanford, 1988.
[39] D. Appelt, J. Hobbs, J. Bear, D. Israel, M. Kameyama, and M. Tyson. The SRI MUC-5 JV-FASTUS information extraction system. *Proceedings of the Fifth Message Understanding Conference*, Baltimore, Maryland, August 1993. Morgan Kaufmann.
[40] W. C. Mann and S. A. Thompson. Rhetorical structure theory: A theory of text organization. Technical Report ISI/RS-87-190, Information Sciences Institute, University of Southern California, 1987.
[41] N. Dahlback and A. Jonsson. An empirically based computationally tractable dialogue model. In *Proceedings of the 14th Annual Conference of the Cognitive Science Society* (COGSCI-92), Bloomington, Indiana, July 1992.
[42] L. Duranti. Diplomatics: New Uses for an Old Science. *Archivaria* 28, Summer 1989, pp. 7-27.
[43] B. Adelberg. NoDoSE: A Tool for Semi-Automatically Extracting Structured and Semi-Structured Data from Text Documents. *Proceedings of ACM SIGMOD International Conference on Management of Data*. June 2-4, 1998, Seattle, Washington.
[44] W. Underwood and S. Laib. Evaluation of a Restructured MARKS in Filing, Retrieval and Retention of Electronic Records, Technical Report AIAI-TR-97-04, Army Research Laboratory, Computer Software Technology Division, Georgia Institute of Technology, July 1997.
[45] In a January 6, 1997 letter from President Bush to the National Archivist, the President outlined materials that he waived from restriction. Included in his list were "announcements of appointments to office and accompanying press releases" and "routine letters of recommendations for positions." The archivist who reviewed these materials decided that they could be opened since the restriction regarding appointments to Federal office had been waived and the advice to the President regarding this appointment was routine.
[46] J. L. Kolodner. *Case-Based Reasoning*, Morgan Kaufman, 1993.
[47] http://www.autonomy.com
[48] http://www.excalib.com/home2.html
[49] The Intelligent Filing Assistant used to illustrate information extraction and categorization in section 4.1.5 also retrieves documents using lexical knowledge and conceptual search.
[50] http://www.cartia.com
[51] S. Huffman. Acquaintance: Language-Independent Document Categorization by N-Grams. *NIST Special Publication 500-236: The Fourth Text Retrieval Conference (TREC-4)*. pp. 359-372.
[52] B. Bryant "Experiments with Acquaintance in Categorizing E-mail. Army Research Laboratory, Georgia Institute of Technology, Atlanta, August, 1996.
[53] Word DOT Redactor (http://www.va.gov/foia/redactor/)
[54] Imaging for Windows Standard Edition (or Professional Edition) (http://www.eastmansoftware.com/)
[55] Redax (http://www.digapp.com/)