PERPOS II:
Scientific and Technical Report

June 19, 2003 – June 18, 2004

William Underwood
Robert Simpson
Elizabeth Whitaker
Dariush Molavi
Marlit Hayslett-Keck
Sandra Laib
Matthew Underwood

PERPOS Technical Report ITTL/CSITD 04-8
August 2004

Information Technology and Telecommunications Laboratory
Georgia Tech Research Institute
Georgia Institute of Technology
Atlanta, Georgia

The Army Research Laboratory (ARL) and the National Archives and Records Administration (NARA) sponsor this research under Army Research Office Cooperative Agreement DAAD19-03-2-0018. The findings in this paper should not be construed as an official ARL or NARA position unless so indicated by other authorized documentation.

Abstract

Archivists must respond to Freedom of Information Act (FOIA) requests beginning five years after the end of a Presidential administration. They do not get around to systematic processing (arrangement, preservation, review and description) until a couple of decades later. However, to support retrieval of records relevant to FOIA requests, archivists at Presidential Libraries create lists of all the folder titles in their collections of Presidential records. Review of records for Presidential Record Act (PRA) restrictions and FOIA exceptions is an intellectually demanding task, requiring page-by-page review. This report describes progress in applying advanced information technology to support these tasks.

Technologies and tools for automatic extraction of information from textual documents are described. This includes recognition of person's names, job titles, dates, locations, organization names, and addresses. This information can be used to recognize document types such as letters, memos, itineraries, and resumes. The recognition of document types supports automated titling of directories and summarization of record series in personal computer filing systems.

A prototype Access Restriction Checker is being constructed that uses content extraction technology to distinguish Presidential Records from Personal Records. Most importantly, by formally representing some of the knowledge and experience that archivists use to decide whether FOIA exemptions or PRA restrictions apply to a document, one is able to automatically recognize probable access restrictions. Such restrictions on release include private information such as social security numbers, marital status, and medical information. With additional semantic and pragmatic knowledge, one is able to recognize PRA restrictions, such as restrictions on release of confidential advice between the President and his staff.

Electronic records stored in digital repositories are vulnerable to system failure, human error, or malicious actions. NARA is particularly concerned with the secure transfer of sensitive files and the security of remote access to such a repository. GTRI has constructed a network configuration that includes a Web portal for Internet access, and an isolated subnetwork behind a firewall containing an archival repository and archival services. The Army Research Laboratory is evaluating firewall and intrusion detection system technologies and products for protecting the subnet. They are also evaluating Virtual Private Network products for remote access and technologies for secure transfer of records.

Experiments are being conducted at the Bush Presidential Library and Archives II to evaluate the models, technologies and tools developed. Archivists at the Bush Library have begun pilot testing of archival processing tools that support accession, arrangement, preservation, review, and description of electronic records.

Keywords: information extraction, content extraction, summarization, knowledge representation, natural language processing, information assurance, E-FOIA.

Table of Contents

1. INTRODUCTION 1

1.1 BACKGROUND 1
1.2 PURPOSE 1
1.3 SCOPE 2

2. INFORMATION AND CONTENT EXTRACTION TO SUPPORT DOCUMENT TYPE IDENTIFICATION AND SUMMARIZATION 2

2.1 INFORMATION AND CONTENT EXTRACTION TECHNOLOGIES AND TOOLS 3
2.2 FILE READERS 7
2.3 NAMED ENTITY RECOGNITION 8
2.4 DOCUMENT TYPE IDENTIFICATION 10
2.5 EXTENSION OF FOLDER/DIRECTORY TITLES 12
2.6 SUMMARIZATION OF RECORD SERIES CONTENT 13

3. FOIA AND PRA REVIEW OF PRESIDENTIAL RECORDS 14

3.1 FOIA AND PRA REVIEW KNOWLEDGE 14
3.3.1 Distinguishing Personal/Political Records from Presidential Records 15
3.3.2 Recognizing Presidential Records Previously Released to the Public 17
3.2.3 Recognizing Records Subject to a(6) b(6): Personal Privacy 17
3.2.4 Recognizing Records Subject to PRA Restriction a(5): Confidential Advice 18
3.2 A CORPUS OF PERSONAL, PRESIDENTIAL AND FEDERAL RECORDS 19
3.3 ARCHITECTURE FOR AN ACCESS RESTRICTION CHECKER 19

4. INFORMATION ASSURANCE 21

4.1 INFORMATION ASSURANCE TEST BED 21
4.2 INFORMATION SECURITY POLICY FOR THE TEST BED AND ITS IMPLEMENTATION 24
4.3 EVALUATION OF NIAP CERTIFIED INFORMATION ASSURANCE PRODUCTS 28
4.3.1 The Common Criteria and NIAP Certification 28
4.3.2 Evaluation of Security with NIAP Certified Firewalls 29
4.3.3 Secure Remote Access and Encrypted Transfer of Files 32

5. PILOT STUDY OF THE USE OF THE ARCHIVAL PROCESSING TOOL AND ITS REFINEMENT 32

6. SUMMARY 38

REFERENCES 41

List of Figures

Figure 1. Default ANNIE Configuration. 6
Figure 2. A Correctly Marked Up Presidential Record in the GATE GUI..... 9
Figure 3. XML Tagged File of a White House Letter. 11
Figure 4. Scope and Content Note for a Record Series and Some of its Folder Titles..... 13
Figure 5. Copy of Record Subject to PRA Restriction a(5)..... 18
Figure 6. Architecture of the Access Restriction Checker..... 20
Figure 7. PERPOS2 Network Architecture. 22
Figure 8. Systematic Archival Processing Supported by the APT. 33
Figure 9. Interface for Accessioning Containers and Viewing Files in a File System. ... 37

1. Introduction

1.1 Background

The National Archives and Records Administration (NARA) faces critical challenges in preserving and delivering authentic Federal and Presidential digital records to future generations. By the time that NARA accessions electronic records, extraordinary efforts are required to gain intellectual and physical control of them, because they were created on systems that are now technologically obsolete.

Archivists at Presidential Libraries must respond to FOIA requests for Presidential Records beginning five years after the end of an administration. Archivists do not get around to systematic processing (arrangement, preservation, review and description) until a couple of decades after the end of an administration. Considerable human effort is needed to construct folder title lists of the contents of thousands of accessioned containers in order to respond to FOIA requests.

Review of Presidential electronic records for access restrictions is an intellectually demanding task that requires page-by-page review of Presidential records. Due to the increasing volume of Presidential electronic records, the need to review these records, and the cost of the limited human resources that can be applied to the review process, the review process is an archival processing bottleneck.

Electronic records stored in digital repositories are vulnerable to system failure, human error or malicious actions. NARA must find ways to leverage advances in information technology in order to address these risks, and improve performance and customer service.

1.2 Purpose

The research objectives for the first year of the PERPOS II project were:

(1) To support improved archival processing through the development and prototyping of advanced technologies to automatically extract information from digital text files; to automatically identify document types; and to summarize folder contents and describe record series;
(2) To determine the kinds of knowledge that archivists use to review Presidential Records for Presidential Record Act (PRA) restrictions and Freedom of Information Act (FOIA) exceptions and investigate the use this knowledge and knowledge-based system technology to support archivist's decisions in reviewing Presidential Records;
(3) In collaboration with U.S. Army Research Laboratory scientists, demonstrate advanced technologies potentially assuring the availability, integrity, authentication, and confidentiality of Presidential records that are preserved

managed, and accessed through distributed, heterogeneous electronic record repositories; and

(4) To pilot test at the Bush Presidential Library the archival processing tools developed during prior research and to refine the tools based on that testing.

This report describes results and progress toward these research objectives.

1.3 Scope

In the next section, progress and results in applying advanced information and content extraction technologies to document type identification and description of folder and record series contents is described. In the third section, results in identifying and representing the knowledge needed for PRA and FOIA review is discussed and an architecture for an Access Restriction Checker is described. In section 4, the PERPOS2 network and security policy is described. It is being used by ARL as a test bed for evaluating NIAP certified security products (firewalls, intrusion detection systems, virtual private networks, and encryption). The fifth section describes the results of pilot testing of Archival Processing Tools at the Bush Presidential Library. Finally, the significant results are summarized, and the planned second year activities are outlined.

2. Information and Content Extraction to Support Document Type Identification and Summarization

In describing record series, archivists often extend or create folder titles, and summarize the contents of a record series in scope and content notes. To do this they must also recognize document types, such as correspondence, memos and press releases. The workload in accomplishing these and other archival processing tasks is substantial and it may be years after accession before these tasks are accomplished. The primary archival activity in newly established Presidential Libraries is FOIA processing rather than systematic processing. Response to FOIA requests could be improved if it were possible to automatically create or extend directory names (folder titles) and describe the contents of unprocessed Presidential Records.

This research task investigates whether information automatically extracted from digital text documents can be used to support identification of document types, description of directory (folder titles), and summarization of record series. The technologies developed can also be applied to the task described in section 3, decision support of FOIA and PRA review.

2.1 Information and Content Extraction Technologies and Tools

Information extraction (IE) is a procedure that selects, extracts and combines data from text in order to produce structured information such as marked-up text, templates or database tables. This structured information will allow automated reasoning in support of business processes, in this case archival processes such as review and description. Facts that are extracted can be used with knowledge to infer additional facts that contribute to text understanding.

Early research in computational linguistics and natural language processing focused on automatic identification of parts of speech and noun phrases. The *part of speech recognition task* is to automatically identify the syntactic category (e.g., noun, verb, adjective) of a term in a document. It is facilitated by an extensive lexicon of terms, but may require partial parsing of sentences to identify the probable syntactic category of a term not in the lexicon, e.g., proper nouns. Another technology used for this task is the Brill Tagger [1]. The *noun-phrase tagging task* is to identify the noun phrases (e.g. "the unemployment rate", "civilian workers") in text. One technology used for this task is Augmented Transition Network (ATN) parsers that identify noun phrases based on having recognized nouns, adjectives, determiners, and prepositions.

The Message Understanding Conferences (MUC) are the more recent driving force for developing this technology [2]. The MUC specifications for various IE tasks have become the de facto standards in the IE research community. MUC divides IE into distinct tasks, namely, Named Entity, Template Element, Template Relation, Co-reference, and Scenario Template Tasks [3].

The *named entity task* is to recognize all named persons, organizations, locations, dates, times, numeric monetary amounts and percentages in text [4]. In the Bush Presidential electronic records, this capability is essential for determining the persons involved in a communication, what organizations they are associated with, the date of the communication, etc. In addition, it will be necessary to recognize addresses, telephone numbers, social security numbers and other named entities.

The *template element task* is to analyze text to fill in a template about entities such as persons and organizations and events (or actions) such as nominations to Federal offices, appointments to Federal offices, and confidential advice. For instance, for a speech act such as advice, one might need to know the date and time it was given, by whom, to whom, what the advice was, and whether the advice was confidential.

The *template relation task* is to identify relations between previously extracted template elements, for example, to relate a person's name in a nomination event to the person's name in their biography. The *scenario template task* is to identify and relate events such as asking for a recommendation for an appointment to the Supreme Court, receiving a

recommendation for that position, deliberation on candidates, nomination of a person for the position, and appointment to the position.

The *co-reference task* is to automatically identify, tag and link co-referring noun phrases and pronouns in text. For instance, in the following text segment, the task is to automatically identify, tag and link *Mr. Andrews* with *chairman of the BNC* and *He*, and to link *Ms Torretta* with *chairman of the BNC*.

> After a long boardroom struggle, Mr. Andrews stepped down as chairman of BNC Holdings Inc. He was succeeded by Ms. Torretta.

Techniques for information extraction can be categorized as *statistical* or *knowledge-based*. Statistical techniques depend on text corpus analysis to discover regularities and learn patterns. They do not require large knowledge-engineering efforts and do not make use of domain knowledge. They depend on the processing of large volumes of text for training of the system. When well trained, these systems have performed relatively well. Knowledge-based techniques make use of the domain knowledge that a person would use in identifying kinds of information. The building of domain-specific rules is labor intensive, but often pays off in performance. Hybrid systems provide the base-level performance of statistical approaches with added improvements of domain-specific knowledge.

The MUC evaluation metrics are precision (P), recall (R) and F-measure. In simple terms, recall (R) is the number of relevant features extracted divided by the total number of relevant features in the corpus. Precision (P) is the number of relevant featured extracted divided by the number of features extracted. In simplest terms, the F-measure[1] is

$$F= RP/.5\ (R + P)$$

The Automatic Content Extraction (ACE) program began in 1999 with the objective of developing natural language processing technology to support automatic understanding of textual data. The ACE program requires the development of technologies that automatically detect the meaning conveyed by the text. The ACE tasks are focused on semantic analysis of document sets whereas the MUC tasks were based upon syntactic analysis.

The ACE program is administered by NIST. ACE uses a greater breadth of document input from newswire texts, scanned newspapers (OCR), and from the automated speech recognition (ASR) of broadcast news. Performance evaluations are conducted bi-annually, but the evaluation results are closed. Performance results are sent only to the participants of the conference.

The ACE program uses a list of document-level mentions and entity types that differ from the original MUC named entity set. For instance, the entity types include Geographical Political Entities (GPE) that are governed Locations such as cities,

[1] The MUC recall, precision, and F measures also consider partially correct annotations, and are actually more complex. See References 4 or 11 for details

countries, and states. Facilities are Locations that have the name of Organizations but are buildings.

The ACE program has Entity Detection and Tracking (EDT) and Relation Detection and Characterization (RDC) tasks. These tasks are analogous to the MUC Named Entity and Template Element tasks, albeit with added complexities. The ACE tasks evaluate the recognition of entities, their attribute types, the relation of entity mentions, mention roles, and mention extents. For example, when a mention of "The White House" is made within a document it could be considered as either an Organization entity or a Facility entity dependent on the document context. "The White House" is therefore linked to two co-reference chains that describe both possible entity type interpretations. This is useful during subsequent processing stages to resolve ambiguity.

There is a large academic research community providing advances in basic and applied information and content extraction technologies and approaches. It was decided to use the GATE/ANNIE information extraction toolset developed at the University of Sheffield. ANNIE is a toolset derived from LASSIE-II, which performed superbly on the named entity task in MUC-7 [5], F-measure .90. GATE/ANNIE embodied as MUSE has also been used in the subsequent Automatic Content Extraction (ACE) conferences [6]. GATE/ANNIE is open source, has good documentation, and is widely used.

GATE/ANNIE

The GATE (General Architecture for Text Engineering) is a framework for associating visual, language and processing resources used in text analysis [7, 8]. GATE is an open source (GPL) distribution from the University of Sheffield (UK) and is used on a global scale by researchers in the analysis, development, and furthering of information extraction and information retrieval technologies. Written in the Java Programming Language and with a development API, GATE is platform independent (Windows, Unix, Linux, etc.) and extensible through the modification, replacement, or creation of additional processing resources.

ANNIE (A Nearly New Information Extraction) is an application containing processing resources [8]. These resources are configured, ordered, and then assigned to language resources for processing. Annotations are assigned to regions of text within the documents. Using its default setup parameters, ANNIE provides the following processing resources:

Document Reset	(Clears all or part of previous annotation sets)
English Tokenizer	(Document atomization into constituent parts)
Gazetteer	(Finite State-based Lookup)
Sentence Splitter	(Separation of Document into Sentences)
PartOfSpeech Tagger	(According to the Penn Treebank POS Tagset)
Named Entity Transducer	(JAPE Rules as cascaded Finite State Transducers)
OrthoMatcher	(Orthographic Co-reference)

The ANNIE processing pipeline is shown in Figure 1.

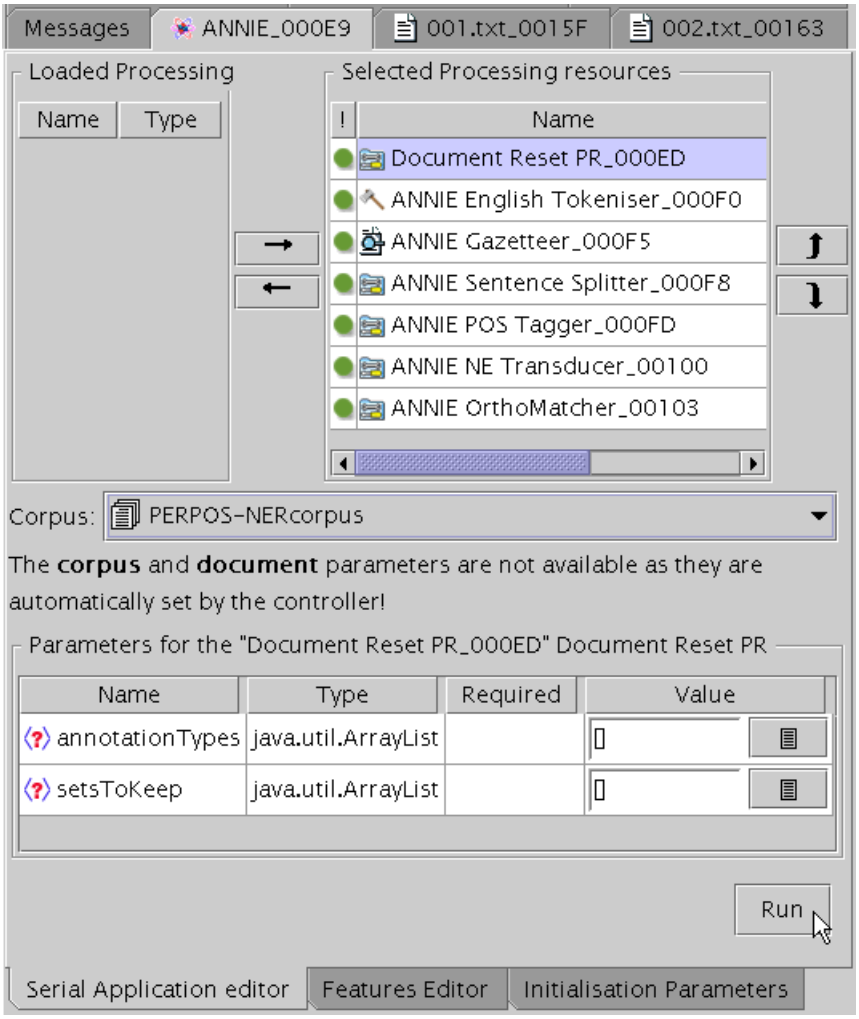

Figure 1. Default ANNIE Configuration.

The authors of ANNIE use the term gazetteer to refer to lists of location names, person's names, organization names, and time expressions. While the term gazetteer is appropriate for location names, because they are geographical, one of the terms lexicon, terminology, or fact base would probably be more appropriate for characterizing the other lists.

The Java Annotations Pattern Engine (JAPE) used to recognize named entities contains rules that have regular expressions on the left-hand side of the rule and an action, such as markup a term, on the right-hand side. The regular expressions represent a pattern that is matched against text in a document.

2.2 File Readers

To apply information extraction techniques to personal computer files, it is necessary to have plain ASCII text or HTML versions of the files. A *file reader* is a program that converts text in a proprietary or standard file format to an ASCII text or HTML file. The Stellent Filters provided with the Oracle Text are file readers. They convert about 200 legacy and current file formats to ASCII text or HTML [9].

The Bush Administration's personal computer files were created using a variety of legacy word processors, for example,

AMI Professional (SAMNA)
DCA-RFT
IBM DisplayWrite 2&3
IBM DisplayWrite 4&5
Lotus Manuscript
MSWord for DOS
MSWord for Windows 2.0
MultiMate Advantage 2
Windows Write
WordPerfect 4.2
WordPerfect 5.0
WordPerfect 5.1/5.2
WordPerfect Notebook 2.0

There are other types of legacy digital objects than text documents that occur in the Bush Administration's personal computer files that also contain textual information, for example,

Windows 3.1 Calendar
WordPerfect Calendar 2-3
dBase II database
dBase III database
dBase IV database
Advanced Revelation database
Borland Reflex 2.0 database
Paradox 4.0 database
Lotus 123 1.0 and 2.0 Worksheets
Microsoft Excel 2.0 Worksheet
PlanPerfect 5.1 Worksheet
QuatroPro for DOS Worksheet
QuatroPro for Windows 3.x Workbook
Harvard Graphics 2.0 Chart
Harvard Graphics 3.0 Chart

There are no Stellent Filters for WordPerfect Notebook, WordPerfect Calendar, Windows 3.1 Calendar, Advanced Revelation database, Borland Reflex database, dBase II database, and PlanPerfect Worksheet. File readers will be created for these formats, or file converters will be used to transform them to file formats for which there are Stellent Filters.

An experiment was conducted in which files in twenty-one formats for which there are Stellent filters were passed through the filters (file readers) and converted to ASCII and to HTML [10]. ASCII Files read from all twenty-one file formats were judged to be adequate as input to GATE/ANNIE. The transformation to HTML had a few glitches, but the HTML preserves the appearance of the document (e.g., bold, italics, paragraphs) and GATE accepts the HTML markup and preserves it as original annotations.

2.3 Named Entity Recognition

An experiment was conducted to empirically evaluate the performance of the named entity recognition techniques applied to the files from the Bush Administrations PC filing systems [11]. A fifty-document corpus was selected from a corpus of Personal and Presidential electronic records that was constructed from records that have been opened to the public, or that are fictitious because they simulate records that have FOIA exemptions or PRA restrictions [12]. With the aid of GATE, a version of the fifty-document corpus was manually marked up that contains the "correct" markup of named entities. The file is called the "key" file. A human-annotated "key file" inside the GATE GUI is shown in Figure 2.

ANNIE then extracted the named entities from same files and put marked up copy in "response files." The performance of ANNIE was then measured by comparing the key file to the corresponding response file.

The effectiveness of ANNIE in recognizing named entities was measured using the MUC evaluation metrics. The precision measure was 0.7857, the recall measure was .7132, and the F-measure was .7477. This performance is significantly less the performance of ANNIE on the MUC-7 corpus (F-measure ≈ .90)

Analysis of the results indicates that while many of the named entities were correctly identified, many were only partially correct, and others incorrect. This was in large part due to the fact that our corpus of documents contained document types, such as letters, memos, and agenda, that did not occur in the corpora of the MUC conferences, which were limited to newswire and broadcast news document types. The performance in recognizing organization names could be improved by adding to the gazetteer

> White House Office Names
> Federal Offices to which the President appoints or nominates individuals
> that are subject to Senate approval, and those that are not.[2]

[2] *Federal Yellow Book*, *Judicial Yellow Book*, Leadership Directories, Inc.

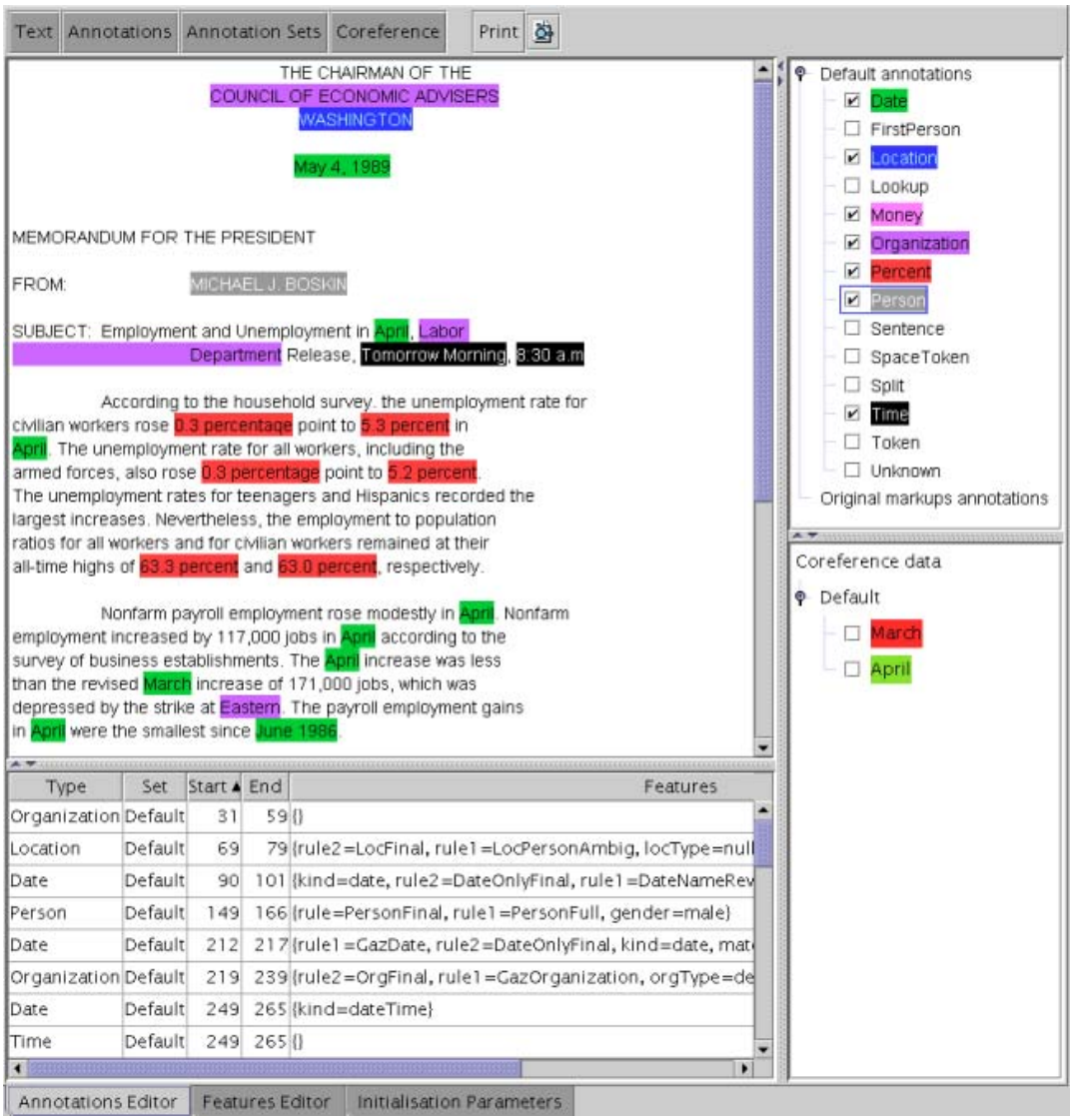

Figure 2. A Correctly Marked Up Presidential Record in the GATE GUI.

The performance in recognizing person's names can be improved by adding to ANNIE's gazetteer the names of

White House Staff Members[3]
Members of the 101st (1989-90) and 102nd (1991-92) Congresses
George H. W. Bush Family Members
President's Friends
Campaign Staff
RNC Staff Register
Names of Members of Cabinet and other Bush Administration Senior Officials
Names of Foreign Heads of State (1989-1992)[4]

[3] White House Telephone Directory

The analysis also showed that there are other annotations that are needed for document type identification such as addresses, greetings and formulas of respect, for example, Sincerely, or Yours Truly. JAPE rules can be used to recognize these entity types and integrated with ANNIE's other processing resources. This is being done and after testing, a second experiment will be conducted at the Bush Presidential Library on a larger sample of Presidential Records.

2.4 Document Type Identification

Archival descriptions include the names of the types of documents that occur in record series, file units (directories or folders) and items. The form or type of a document derives from the business activity and procedure used to create it. The kinds of documents observed to occur in the Bush PC files include, but are not limited to:

Agenda
Attendee List
Bar Chart
Biography
Briefing (Presentation)
Briefing Memo
Decision Memo
Diary
Executive order
Information Memo
Job Application
Letter
List of Candidates
Mailing List
Memo
Memo from President
Minutes
National Security Directive (NSD)
Newsletter
Nomination to Federal Office
Notes
Presidential Statement
Press Pool Report
Press Release
Recipe
Referral Memo
Resume
Schedule
Signature Memo
Situation Report
Summary
Transcript of Speech
Staff Register
Telephone Call Recommendation
Transcript of News Conference

These document types are being analyzed to identify those kinds of annotations that are necessary to characterize the intellectual form of the document type. For instance, Figure 3 shows a manually marked up (XML) copy of a piece of correspondence from the office of a Bush Administration staff member.[5] The letter was printed on letterhead and signed by the author. The digital copies of such letters do not include the letterhead or the signature.

The XML tags are indicative of the kinds of information that need be extracted from the PC files in order to determine whether they are letters. They include annotations that were not in the named entity recognition task, for instance, greetings, formulas of respect,

---

[4] 1990, 1991, 1992, 1993 CIA World Fact Books http://manybooks.net/titles/usciaetext93world192.html
[5] Doug Wead's Files, [Alpha Correspondence File 2-90 to 6-90 'A' 'B'], Bush Presidential Library. The original document has been reviewed and opened to the public.

job titles, postal addresses (or street addresses), and zip codes. JAPE rules are being constructed to extend ANNIE's existing annotation capabilities to recognize those entities (or features) necessary to characterize the 35 document types.

```
<date>March 27, 1990</date>
<greeting>Dear</greeting><person>Mr. Allen</person>
<p>Thank you very much for your letter of <date>March
15, 1990</date> which stated your concerns and
suggestions regarding the Americans with
Disabilities Act.</p>
<p>In order to fulfill <person>President Bush's</name> campaign
promise of bringing Americans with handicaps
into the mainstream of American life, the
Bush Administration supports the objectives
of the A.D.A.</p>
<p>As you may know, the bill is still in <organization>House
Committee</organization> for consideration and change. You
can be sure that your thoughts have been
fully noted and are appreciated.</p>
<formula of respect>Sincerely,</formula of respect>
<person>Doug Wead</person>
<job title>Special Assistant to the President
for Public Liaison</job title>
<person>Ray Allen</person>, <job title>President</job title>
<organization>American Cultural Traditions</organization>
<postal address>P.O. Box 1895</postal address>
<location>Washington, D.C.</location> <zipcode>20013</zipcode>
```

Figure 3. XML Tagged File of a White House Letter.

Two approaches are being investigated to recognizing document types. The first is to create JAPE rules that that can recognize each document type. The JAPE rules will then be tested on a representative sample of document types. Finally, an experiment will be conducted at the Bush Library to determine its performance on Presidential electronic records.

The second approach involves the use of GATE's Machine Learning Module or the Hidden Markov Model Module. In training mode, these modules learn the essential features of a marked-up sample of documents of the same type. Then in recognition mode, they can use what was learned to classify marked-up documents.

A "Test Corpus" has been constructed of 125 files of 26 different document types. A larger sample is needed because many of the document types have just one or two examples in the corpus. A sample of several hundred documents representing a large number of document types will be automatically marked up with ANNIE's extended capabilities to recognize and annotate a document with features necessary to characterize document types. Then the Machine Learning module will be used in a supervised mode to learn the document types of those documents in the sample. Finally, an experiment will be conducted at the Bush Library to determine the performance of the Document Type Identifier in recognition mode on Presidential electronic records.

2.5 Extension of Folder/Directory Titles

Archivists at the Bush Presidential Library use a folder title list as an aid in responding to FOIA requests. The operating systems used on personal computers in the White House Offices during the Presidential Administration of George H. W. Bush were PCDOS, MSDOS and Windows 3.1. The directory (folder) names could be no longer than 11 characters (8 plus a 3 character extension). The directory names of the files on the Bush hard drives are cryptic, e.g., CORR for Correspondence, or WORK for whatever a particular person's tasks might be. The user-created files may be stored in a directory containing the name of the application used to create them, e.g., WP51 (for WordPerfect 5.1), or they may even be stored in the root directory. The feasibility of using the information extracted from files in a directory to automatically create more descriptive directory (folder) titles for the contents of the Bush hard drives is being explored.

The terminology that archivists and information technologists use to name the elements of filing systems differs. The following table shows the correspondence between archival terminology and computer terminology for filing systems.

Archival (Filing System) Concepts	**Instance of Archival Concept**	**Computer Filing System Concepts**	**Instance of Computer Filing System Concepts**
Series (or File)	Subject File	Directory (a Folder Icon)	C:\SUBJ
File Unit (Folder)	Adoption	Subdirectory (a Folder Icon)	C:\SUBJ\AGRICULT
Item	Document or Record	Filename (File)	Displayed file CROPRPT.WP5

A *series* is "file units or documents arranged in accordance with a filing system or maintained as a unit because they result from the same accumulation or 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*." [6]

Series of electronic records (computer files) are contained in the directories of computer filing systems. File Units or Folders of electronic records are subdirectories. Items are files (or the displayed files) in a subdirectory.

The archivists at the Bush Presidential Library have conventions for titling or extending the titles of file folders. For instance, a folder or directory title CORR, if it contained correspondence, might be extended using square brackets as follows: CORR[ESPONDENCE]. If it was titled WORK and contained correspondence, its title might be extended as follow: WORK[CORRESPONDENCE].

[6] L. J. Bellardo and L. L. Bellardo. *A Glossary for Archivists, Manuscript Curators, and Records Managers*. Chicago: The Society of American Archivists, 1992.

For directories that are characterized by the kind of documents in the directory, the Document Type Identifier will be used to identify the types of documents in the directory, and a set of rules will be used to check for the type(s) of documents found in a directory, check the current directory name, and when appropriate suggest an extension to the directory (folder) title.

A test corpus of 125 files of 26 different document types has been organized into directories (or folders). This corpus is being used to test and refine the Folder Titler. An experiment will also be conducted at the Bush Presidential Library to determine the Folder Titler's performance on file directories from the filing systems of White House Offices and Staff Members' personal computers.

It is not unusual for a single directory of a file system from a White House personal computer to include letters, memos, agendas, applications for a post administration job, resumes and other document types. In this case, it may not be appropriate to extend the folder title, but to leave it as it is and to summarize the contents of the directory. This issue is discussed further in the following section.

2.6 Summarization of Record Series Content

The contents of most of the Bush Administration's PC filing systems will not have complete descriptions until archivists find the time to systematically review the contents of the file systems. The feasibility of using the results of the two prior research tasks, plus additional content extraction capabilities, to automatically create descriptions of the contents of the Bush Administration's PC filing systems is being investigated.

NARA and the Presidential Libraries have established conventions for describing record series, file units and items. The scope and contents notes and folder titles that archivists have created for paper records in the holdings of the Presidential Libraries are being studied. Figure 4 shows an example of a scope and contents note and some folder titles for Marvin Fitzwater Files, in the White House Press Office.

Series: Subject File 1989-1983
The Subject File contains material related to a wide range of issues and topics. Much of this consists of press briefings, talking points, transcripts, itineraries, publications, invitation lists, fact sheets, photographs, statistics, reports, press pool reports, press clippings and press releases prepared for the press. It also contains correspondence, notes and memoranda reflecting the daily functions of the press Secretary and his interactions with the President, the White House Staff, federal agencies and the public. The file is arranged alphabetically.
Folder Titles
Box 1
ABC TV - *Prime Time Live*
Abortion
Administrative Appointments
Adoption

Figure 4. Scope and Content Note for a Record Series and Some of its Folder Titles.

A series description begins with an introductory word or phrase, "Series consists of …" or "Series contains…" The scope is the period of time. The content of a series describes the specific activity or activities generating the records, and information is given about the internal structure of the series, including the arrangement and documentary forms of the records.

A Record Series Summarizer is being constructed that uses the information and content extraction capabilities described in section 2.5, and the Folder Titler described in section 2.6 to describe the contents of an accessioned file system. The "Test Corpus" described in the previous section will be used to test the Record Series Summarizer, and then experiments will be conducted at the Bush Library to determine its performance on file systems containing Presidential Records.

3. FOIA and PRA Review of Presidential Records

Review of Presidential electronic records for access restrictions is an intellectually demanding task that requires page-by-page review of Presidential records. Due to the increasing volume of Presidential electronic records, the need to review these records, and the cost of the limited human resources that can be applied to the review process, the review process is an archival processing bottleneck. The purpose of this research task is to determine the kinds of knowledge that archivists use to review Presidential Records for Presidential Record Act (PRA) restrictions and Freedom of Information Act (FOIA) exceptions, to use this knowledge to develop an automated Access Restriction Checker to support an archivist's decisions in reviewing Presidential Records, and to experimentally evaluate its performance.

3.1 FOIA and PRA Review Knowledge

The first step in our research was to become acquainted with the FOIA and PRA statutes and understand how they were used in review of Federal and Presidential records. A Freedom of Information Act and Privacy Act Workshop conducted at the USDA Graduate School was attended. The "Justice Department Freedom of Information Act Guide" was studied to understand the interpretation and application of FOIA exemptions.[7] Archivists described the terms that occurred in reviewed records that indicate that the record or a portion of it might be subject to a specific FOIA exemption or PRA restriction.

The kinds of knowledge identified include the knowledge that is needed to interpret the content of a record, the knowledge (concepts) characterizing a FOIA exemption or PRA restriction, and the knowledge acquired from experience in reviewing Presidential, Federal and Personal records.

[7] http://www.usdoj.gov/oip/foi-act.htm

This knowledge needs to be represented in such a form that automated methods can be used to analyze and interpret a digital record and conclude whether an access restriction applies. Two ways are described in which this knowledge can be represented, as rules and as cases. The knowledge acquisition problem and the need to investigate methods for machine learning of new rules from cases in which archivists have made FOIA or PRA access restriction decisions were also discussed.

A report was prepared that described the results of the study and analysis to identify the kinds of knowledge and reasoning that archivists use to judge whether records are Federal or Presidential records or whether they are personal records. Some of the kinds of knowledge used to determine whether passages of a record or an entire record is subject to access restrictions, or can be opened for public access were identified. Archivists at the Bush Library reviewed the draft report, and suggested improvements. Their suggestions were incorporated in the report [13].

The following sections describe some of the kinds of knowledge that are needed to distinguish Personal from Presidential Records; to recognize records that have no restrictions because they have already been released to the public; and to recognize records that are subject to two PRA restrictions. See Reference 13 for more examples. The Jess (Java Expert System Shell) scripting language [14, 15] is being used to represent factual knowledge, such as names of Bush family members and names of advisors to the President, to represent FOIA and PRA review knowledge and for rule-based inference.

3.3.1 Distinguishing Personal/Political Records from Presidential Records

The Presidential Records Act defines personal records as follows:

“The term "personal records" means all documentary materials, or any reasonable segregatable 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, journals, 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."

In addition to a diary, journal or personal notes, personal records of the President or First Lady can include:

- Correspondence with friends expressing sentiments, e.g., "Congratulations," “Happy Birthday,” “Thanks for the Gift.”
- Correspondence with a family member that does not involve a constitutional duty.
- First Lady's records - personal correspondence.
- A record concerning the President's plans for attending a funeral of a personal friend, rather being a state occasion.
- Material pertaining to a personal appointment, rather than an appointment for Presidential business.

Material of a Personal/Political nature can include:

- Bush Presidential Campaign Material or correspondence
- Document relating to an individual's campaign for local, state, or federal office.
- Letter to or from the Chairman of the Republican National Committee (RNC).
- Document sent to the RNC.
- Document that contains phrases such as "I was saying to [Chairman of RNC], …" "I sent so and so to [Chairman of RNC]."
- RNC staff register
- Poll on political issues by private pollsters.
- Correspondence relating to the election of a particular individual or individuals to Federal, State or local office.

Personal Records of White House Staff Members can include:

- A staff resume for a post-administration job
- A job application for a position outside the White House Offices
- A recipe
- A Christmas card mailing list
- A grocery list

Some of these types of personal records can be recognized by document type alone. Assuming a Document Type Identifier can recognize recipes, Christmas card mailing lists, and grocery lists, and that ID is a variable whose value is a specific file identifier, the following rule will handle these simple cases.

```
If      Document_type(ID) = Recipe or
        Document_type(ID) = Grocery List or
        Document_type(ID) = Christmas Card Mailing List
Then    Personal_Record(ID).
```

Assume that the Document Type Identifier can recognize letters and that templates have been filled in indicating whom the letter is from, and whom it is to. A rule such as the

following could be used to infer that a communication from the President to the Chairman of the RNC is a Personal/Political Record.

```
If      Document_type(ID) = Letter and
        From(ID, X) and
        Job_Title(X) = President and
        To(ID, Y) and
        Job_Title(Y) = Chairman of RNC
Then    Personal_Political_Record(ID).
```

Similar rules are needed personal records of the President and First Lady, for other Personal/Political records, and for personal records of White House Staff Members.

3.3.2 Recognizing Presidential Records Previously Released to the Public

There are some copies of presidential records that can be opened for public access because they have been previously released to the press or the public. These include White House press releases, press pool reports, newswires, and Presidential (Vice Presidential, First Lady) trip itineraries.

Assuming that a Document Type Identifier can recognize White House Press Releases, the following simple rule might suffice for such press releases.

```
If      Document_Type(ID) = White_House_Press_Release
Then    Access_Restriction(ID) = None and
        Reason_for_Opening(ID) = "White House Press Release"
```

Similar rules have been formulated for press pool reports, newswires and Presidential trip itineraries.

3.2.3 Recognizing Records Subject to a(6) b(6): Personal Privacy

This restriction on release concerns records that contain "personal information" whose disclosure would "constitute a clearly unwarranted invasions of personal privacy." To recognize information that is subject to this restriction, annotation capabilities will be added to ANNIE to recognize the following:

- Social Security Numbers
- Home addresses,
- Home telephone numbers
- Home email addresses.
- Religious affiliation
- Marital status (separated, divorced)

Rules are being formulated that will enable the Access Restriction Checker to infer that a part of the text contains "personal information" and should be redacted.

3.2.4 Recognizing Records Subject to PRA Restriction a(5): Confidential Advice

This restriction applies to "confidential communications requesting or submitting advice, between the President and his advisers, or between such advisers." Figure 5 shows a copy of a memo from the President to one of his advisors requesting confidential advice.[8]

December 5, 1990

MEMORANDUM FOR BOYDEN GRAY
THROUGH: BRENT SCOWCROFT
FROM: THE PRESIDENT
Boyden —

Please prepare for me a short analysis of the War Powers Resolution. . . . With out recognizing the constitutional validity of the War Powers Resolution, is there a way for the President to fulfill all his responsibilities to Congress by saying, a few days before any fighting was to begin, "hostilities are imminent—period!!

I am several thousand miles south, but these questions stay on my mind:

1. How do we fully involve Congress?
2. If we have to attack from a cold start how does the latest UN Resolution impact on congress?
   1. Is there something short of "declaring" war that satisfies Congress yet doesn't risk tying the President's hands?
   2. As the clock on the UN resolution keeps running toward the time when force has international authority, what possible official requests can/should a President make of Congress?

If you reply to this memorandum before I return, please hand carry your reply to Brent for "Eyes Only" transmission to me.

Please share a copy of this memorandum with John Sununu and Brent only.

Warm regards.

Figure 5. Copy of Record Subject to PRA Restriction a(5).

For an Access Restriction Checker to recognize that this is a request by the President for confidential advice, the capabilities to extract person's names and to identify the record as a memorandum are necessary. From this it can be inferred that the memo is to Boyden Gray and from the President. Factual knowledge of the names of the President's advisors is then needed, and would include that Boyden Gray, Counsel to the President, was an advisor.

Requests for action, questions, and for that matter, advising someone are examples of speech acts. They are sentences that not so much say something as they are sentences that

[8] George Bush. *All the Best: My Life in Letters and Other Writings*, New York: Scribner, 1999, pp. 491-492. Access to this document would have been restricted for a period of twelve years after the end of the Bush Administration had former President Bush not waived the restriction on this particular document by making it public in his book. The ellipsis at the end of the first sentence may indicate that part of the document is still restricted.

do something. The statement "Please prepare for me a short analysis of the War Powers Resolution" is a request for an action. It is followed by a number of questions. The implication is that the President would like Boyden Gray to answer these questions. Similarly, there are many ways of advising that never explicitly state, "I advise that …" For instance, Boyden Gray's reply might simply provide an analysis of the War Powers Resolution, and provide answers to the questions. To recognize it as confidential advice, one might need the context of the President's memo. A large sample of Presidential records requesting confidential advice and providing confidential advice needs to be analyzed in order to construct a set of rules that can be used to recognize various means of requesting and providing advice.

The President indicates that the reply should be via "Eyes Only" transmission. While not conclusive, Administrative markings such as "Eyes Only," "Administratively Confidential," and "Personal and Confidential," may indicate that the requested advice is confidential.

3.2 A Corpus of Personal, Presidential and Federal Records

A corpus of about 125 documents has been created that represent kinds of documents encountered by archivists who review Presidential records [12]. These include personal/political records, Federal records and Presidential records corresponding to records with no restrictions and with each kind of PRA restriction or FOIA exception. Most of those with restrictions are "simulated" records with fictitious information that capture the essence of a restricted document without actually being a restricted document. Fifty of these documents have been used in the Information Extraction experiments. These documents are also being used in testing the document type identifier. They will also be analyzed to determine criteria for distinguishing personal from Federal and Presidential records and for identifying passages or documents that have access restrictions. This corpus is being extended to include an additional 125-150 sample records.

3.3 Architecture for an Access Restriction Checker

An intelligent agent architecture for an Access Restriction Checker has been developed. Figure 6 shows an interface software agent that communicates with the archivist about the archivist's task and sends information to a set of specialized software agents that have knowledge about elements of access restriction. These software agents share information about partial solutions to the archivist's task in a shared working memory called a "blackboard." All of the software agents have access to specialized "domain knowledge" including the specific names associated with a particular Presidential administration and other concepts necessary for aiding the archivist. These domain knowledge repositories are represented as ovals in the following diagram.

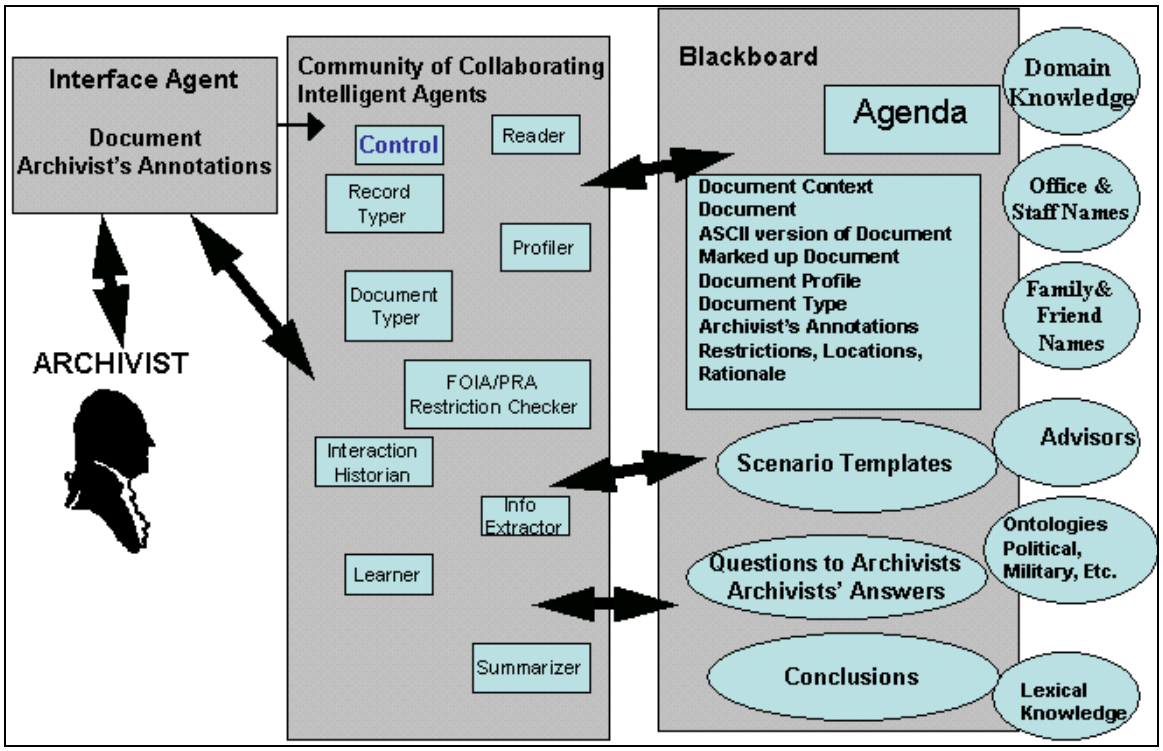

Figure 6. Architecture of the Access Restriction Checker.

When an archivist who is reviewing a document selects the tool, Check Access Restriction, an interface agent will place on the Blackboard Agenda the tasks that need to be accomplished. This includes: *Determine document context*, i.e. Collection, office, series, file unit; *Read the document*, i.e. convert it to ASCII text; *Extract information from document*; *Determine document type*, *Determine record type* (personal, Federal, Presidential); *Profile the document*, i.e., determine the author, addressee(s), titles, subject, date; and for Presidential records, *Check PRA/FOIA restrictions*, i.e., check for a(2) restrictions, check for a(5) restrictions, etc.

To accomplish these tasks, the PRA/FOIA checking agents will use their knowledge of PRA/FOIA restrictions and apply them to the marked-up document on the blackboard. They will also need lexical and domain knowledge. The domain knowledge includes knowledge of office and staff names and titles, the names of Bush family and friends, the names of advisors, and knowledge of political, military, and foreign policy terms.

In some cases, the PRA/FOIA agents may need information that cannot be determined from the document, its context, or the lexical and domain knowledge. In this case they will formulate a question to the archivist. This question will be passed to the interface agent, who will ask the question, obtain the answer and pass it back to the blackboard for use by the agent that needed the answer.

The results of these tasks will be put on the blackboard. A Summarizer Agent will gather the results and pass them to the Interface Agent who will highlight those terms or

passages in a document that might have PRA restrictions or FOIA exemptions, and provide the rationale for the agent's conclusions

If an archivist accepts the recommendation of the Access Restriction Checker, he will open or withdraw the document or redact the relevant associated term or passage and annotate it with the proper codes for the PRA/FOIA restriction. Alternatively, the archivist can reject or modify the recommendation. Finally, the archivist may identify a restriction that was not detected by the Access Restriction Checker.

An Interaction Historian agent will record the decisions of the archivist and save them with the contents of the blackboard in an Access Restriction archive. At a future phase of the project, it is intended to develop a Learning agent that will attempt to improve the performance of the Access Restriction Checker using the information in the Access Restriction archive to modify the knowledge used by the PRA/FOIA agents.

4. Information Assurance

A distributed, heterogeneous computing environment for digital archives will be vulnerable to equipment failure, human error and physical or cyber attacks. If records are transferred between these repositories, their integrity and confidentiality must be protected.

Information technology (IT) security is the protection of the confidentiality and integrity of information and the assurance of its availability by countering threats to that information arising from human or systems-generated activities, malicious or otherwise. *Confidentiality* is the prevention of the unauthorized disclosure of information. *Integrity* is the prevention of the unauthorized amendment or deletion of information. *Availability* is the prevention of the unauthorized withholding of information from authorized users.

GTRI has constructed a network configuration that includes a Web portal for Internet access, and an isolated subnetwork behind a firewall containing an archival repository and archival services. The Army Research Laboratory is evaluating firewall and intrusion detection system technologies and products for protecting the subnet. They are also evaluating Virtual Private Network products for remote access and technologies for secure transfer of records.

4.1 Information Assurance Test bed

An Information Assurance Test bed has been created that consists of the network architecture shown in the following Figure [16].

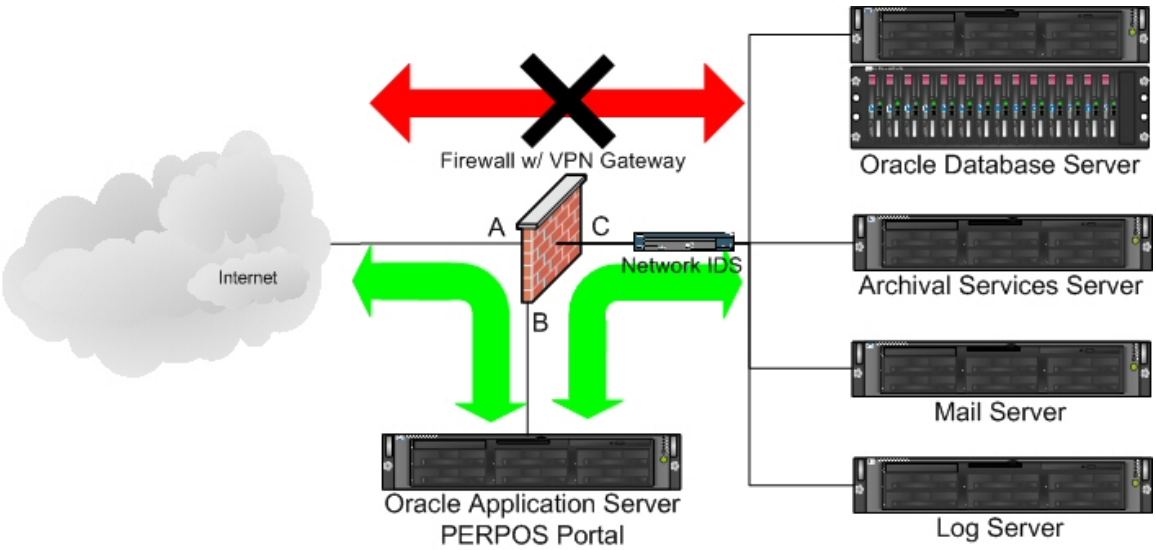

Figure 7. PERPOS2 Network Architecture.

Communication Ports

Every computer has the capability to communicate through 65536 communication ports numbered 0-65535. A computer can send or receive information on any of these ports, but cannot send and receive at the same time on the same port. Ports 0-1023 are assigned to specific services.[9] For example, http services (used to connect to a website) are assigned to port 80. Secure Shell (SSH) is assigned to port 22. Secure Socket Layer (SSL) is assigned to port 443. Specific communication protocols are associated with each service. A decision will be made as to what services to make available and only expose those ports.

Firewall

A firewall is a device that has a set of rules specifying what communications traffic it will allow or deny. For instance, for the configuration shown in Fig. 6, the rules might be:

Disallow A -> C
Allow A -> B only on port 80
Allow B->A
Allow B <-> C
Allow C -> A only if C initiates the connection.

Firewalls are implemented using a variety of technologies, including static packet filters, stateful firewalls, and proxy firewalls. These will be further discussed in sections 4.2 and 4.3 of this report.

[9] http://www.iana.org

Virtual Private Network (VPN) Gateway

A VPN is a secure channel established for a network session formed across unprotected channels, such as the Internet. A VPN allows an outside user to communicate with the internal network as if connected directly to it. The secure channel is formed using client encryption/decryption software and a device at the firewall that enables encryption/decryption.

Intrusion Detection System (IDS)

An IDS is an alarm system that is used to detect and alert the Network Manager of malicious events. Two types of IDS exist: Host-based (HIDS) and network-based (NIDS). HIDS sensors reside on an individual server. NIDS sensors are devices that monitor network traffic. They might reside between the firewall and the Oracle application server or between the firewall and the four protected servers. When IDS sensors detect a suspicious event, they can send an alert by using the mail server or by logging the occurrence.

Oracle Application Server

The Oracle Application Server will store no critical data. Apache web server and Oracle Application Server currently run on that machine. The web server will only have port 80 opened for public access. It will pass requests through a firewall, one of the protection mechanisms, to the Database or Archival Services Server.

The Secure PERPOS Portal accessible via Internet II contains the Gentoo Linux operating system, the Oracle Application Server, the Apache web server, and Secure Shell (SSH) to provide secure remote access.

The Demilitarized Zone (DMZ) and Screened Subnets

The term Demilitarized Zone, when applied to networks, refers to the area outside of the firewall that is accessible by the public. The Oracle Application Server makes up the DMZ of this network. The screened subnet is an isolated network behind the firewall. It consists of the Oracle Database Server, the Archival Services Server, the mail server and the Log Server.

Oracle Database Server

The Oracle database server is an HP ProLiant server with dual Intel Processors and a HP StorageWorks SCSI based enclosure for archival storage. Two hard disks are internal to the server, for system drive and system drive mirroring. Five high capacity hard disks (146.8GB) are plugged into the HP StorageWorks providing 734GB of archival storage capacity. The StorageWorks is capable of housing 14 hard disks, allowing for over 2

terabytes of data using similarly sized disk drives. The Oracle database resides on the StorageWorks hard disks.

The rack-mounted devices will reside in a climate-controlled Data Center with additional redundancy in their power and cooling systems. An Advanced Intelligent Management System by Hewlett Packard (Integrated Lights-Out, SmartStart, & Insight Manager) also protects the configuration. The operating system on the server is Gentoo Linux 1.4, with the 2.4.24 Linux kernel sources. The Oracle Database Server is installed on a subnet inside of the GTRI firewall and will be protected with a NIAP certified firewall.

Archival Services Server

Initially, the Archival Services Server contains the PERPOS tools (Archival Processing Tool and Archival Repository Tool) [see section 4 of this report]. It will also include information extraction modules, the document type identifier, and the record series summarizer.

Mail Server

A mail server is needed for such functions as notification of illicit login or of suspicious behavior on the network detected by an IDS. The software is a Postfix mail server.

Log Server

All transactions with the Oracle Database server, the Archival Services Server, and mail Server are logged. The log server is SYSLOGD operating on Linux.

4.2 Information Security Policy for the Test bed and Its Implementation

A Security Policy has been developed for the PERPOS Network Architecture described in the previous section [17, 18]. The purpose of this policy is to provide general guidelines and specific recommendations for the protection of information stored on the Presidential Electronic Records PilOt System (PERPOS) computer network. The information security measures outlined in this document are designed to protect and preserve PERPOS records and metadata and as well as any data related to authorized users of the system. It also restricts access to Archival Services.

This policy is relevant only to the network configuration that is described Fig. 7, and only in the environment in which the initial configuration was deployed. If the network is deployed in another environment, this policy must be updated to reflect any security policies already in place. All creation, processing, storage, communication, distribution, and disposal of PERPOS information in digital form are covered by this policy.

Physical Security

Access to the hardware configuration will be restricted to those needing access. A guard at the door, and keycard access to the ITTL facilities and keycard access to the ITTL Data Center accomplish this.

Operating System Security

Operating system kernel functionality that is not needed will not be installed. Gentoo Linux allows one to disable capabilities that are not needed. Gentoo Linux was chosen, in part, because of its ease of configuration.

To prevent denial of service attacks and minimize the impact of having to reinstall the operating system, the file system is partitioned into three separate hard drives for /root (system drive), /opt (oracle), and /usr /var (users). This prevents users or attackers from filling drives with log entries. User spaces are partitioned, thus establishing quotas. This prevents a user from filling the drive space. If the operating system has to be reinstalled, the installation of Oracle is not lost.

Failed passwords will be logged and logs inspected on a daily or weekly basis. To ensure OS password security, OS passwords are not to be shared. They must have at least a minimum number of characters and include numbers and special characters.

Users shall only have access to the files they need. The Network Administrator will make sure directory/file permissions/ ownership are properly set.

The operating system (Gentoo Linux) security advisory will be monitored for immediate notification of new vulnerabilities. Patches will be applied immediately.

Application Security

Golden Rule: Do not install any application that is not needed. The only applications that will be installed are: Apache (web server), Oracle Application Server, Secure Shell (SSH), Postfix (mail server), Oracle Database, and modules of the Archival Repository Tool and Archival Processing Tool.

CERTS and other security lists for each of the applications will be monitored for updates. Patches will be applied immediately.

Known operating system and application vulnerabilities can be identified by searching the CERT Advisories, Incident Notes, and Vulnerability Notes at the CERT Coordination Center.[10] For instance, a search for "Oracle 9i Application Server" vulnerabilities yielded the following vulnerabilities, among others.

[10] www.cert.org

- Vulnerability Note VU#977251: Oracle 9iAS XSQL Servlet ignores file permissions allowing arbitrary users to view sensitive configuration files

- Vulnerability Note VU#611776: Oracle9i Application Server PL/SQL Gateway web administration interface uses null authentication by default

- Vulnerability Note VU#278971: Oracle 9i Application Server does not adequately handle requests for nonexistent JSP files thereby disclosing web folder path information

- Vulnerability Note VU#476619: Oracle 9iAS default configuration allows arbitrary users to view sensitive configuration files

- Vulnerability Note VU#936507: Oracle 9iAS allows access to CGI script source code within CGI-BIN directory

The Secure Sockets Layer (SSL), i.e., https, of the Apache web server, is used to encrypt communications back and forth between the client and the server. This includes certificates and 128-bit encryption.

The Secure-Shell (SSH) is used to secure remote access to the system. This provides a secure console that the network administrator can control with up to 4096 bits of encryption.

A mail server is needed for such functions as notification of illicit login. Only those capabilities of Postfix, the mail server, that are needed will be enabled. Remote relay will not be enabled, as this could allow a remote user to use the email server to mail out spam. Only SSL connections will be enabled.

The database resides on the storage array. It is separated in hardware from other aspects of the system to minimize the effects of errors in other parts of the system (user error, developer error, administrator error, hacker error, hardware error) on database integrity/security. This separation, in particular, keeps any possible circumventions of operating system security isolated to the compromised system.

The PERPOS Portal, implemented with Oracle Application Server, uses Single Sign-on (SSO) to authenticate the user in one location and then provides that user with whatever it is that she is slated to see based upon her profile or permissions. While SSO is less secure than having multiple password-protected sites, fewer passwords reduces the likelihood that the password will be hacked.

Oracle Application Server provides the Administrator with the capability to limit the Content Area / Page Access / Component control provided to the user so that they only see what they have authority to see.

Data will only be encrypted when it is sent, and decrypted when it is received. Data in the database will not be encrypted. This prevents sniffing and spoof threats. Integrity checks will be implemented on transmitted and received data. This is what is meant by End-to-End encryption. The Advanced Security Option (ASO) of the Enterprise edition of Oracle provides this capability. It can use private key (symmetric algorithms), private / public key pairs (asymmetric algorithms), and it supports Public Key Infrastructures (PKI). ASO implements message digests (MD5 and SHA-1) for integrity checks. ASO also allows integration with smart cards, token cards, and biometrics.

The ability of users to see or not to see aspects of the database at a high granularity is restricted. This is the meaning of “Row-level security. " This enables security closer to the bone than just permitting or denying a user to see a database table. Oracle's ASO Label Security satisfies this requirement. Additional information can be created which sticks with the data to determine how sensitive it is and all the data in the database can be viewed hierarchically. Categories such as a SSN (Social Security Number), or a sensitive name are only viewable by those users with the correct authentication and access privileges. It is a NARA requirement that no security classified information ever get on the server. However, many unclassified Presidential and Federal records are still sensitive and may have restrictions on access.

Oracle Label Security has an Evaluation Assurance Level (EAL) 4 rating (out of a possible EAL1-7. That is not actually very high. It basically says that the software does what it says it can do. However, Oracle is the only database provider to have such a rating.

Backup and recovery of data will be provided, so that if there is a hardware fault, data can be recovered. Database access will be logged and the logs checked daily or weekly.

Passwords of sufficient length will be used to restrict access to the database. Passwords must differ from previously used passwords. A mixture of alphabetic and numerical characters will be required, and passwords that are a single dictionary word disallowed. Passwords must be changed every sixty days, and users will be locked out if they do not change their password within their grace period. Users will be notified when their password must be changed, preferably automatically and at an adequate time before expiration, so that they have time to change their password. A user's account will be locked out after a certain number of username and/or password failures. If a user has not used the account for 60 days, they will be locked out. Oracle provides all of these capabilities.

Network Security

Documents uploaded to the system may contain viruses or. Hence, every uploaded document will be scanned for viruses. Detected viruses will be quarantined.

Initially, intrusions will be detected via an open source IDS such as Port Sentry or Snort.[11] A policy has been developed for incident response [18]. ARL is evaluating COTS, NIAP certified NIDS for protecting the network (see next section).

Web traffic is limited to Port 80 of the web server. The network perimeter will be protected with a firewall. Initially, the subnet will be protected with IP Tables from Net Builder, an open source packet filter firewall, or an Application Layer Proxy firewall, such as Squid.[12] ARL is evaluating COTS, NIAP certified firewalls for protecting the network (see next section).

ISS Internet Scanner will be used to evaluate the implementation of the Security Policy. Internet Scanner comes with preprogrammed templates for testing Oracle. One can add or remove tests, and Internet Scanner will test for all known exploits. Since Internet Scanner is quickly updated when new vulnerabilities are announced, it is also a tool to determine whether your system has been updated.

4.3 Evaluation of NIAP Certified Information Assurance Products

GTRI is collaborating with ARL researchers who are conducting experiments to evaluate the performance of the secure portal operating in unsecured and secured mode. Also to observe and measure the overhead associated with deployed cryptographic products use to transfer electronic records from the PERPOS portal to ARL computers.

4.3.1 The Common Criteria and NIAP Certification

The Common Criteria for Information Technology Security Evaluation (CCv.1) [19] unifies the national IT security evaluation specifications of Canada, the European Commission and the U.S. Essentially, they consist of protection profiles (PPs) and security targets (STs) to help nonexperts understand complex IT security evaluations.

PPs detail the security requirements for a type of product. Product developers have to match their products' functions with a list of STs. By specifying STs, vendors can get their products evaluated according to the CCv.1. Ratings range from Evaluation Assurance Levels 1 through 7.

The seven Evaluation Assurance Levels are as follows:

EAL1 - functionally tested
EAL2 - structurally tested
EAL3 - methodically tested and checked
EAL4 - methodically designed, tested and reviewed

[11] http://www.linux.ie/articles/portsentryandsnortcompared.php http://linux.rice.edu/help/tips-sentry.html http://www.snort.org/ http://www.grennan.com/Firewall-HOWTO.html
[12] http://squid.nlanr.net/

EAL5 - semi-formally designed and tested
EAL6 - semi-formally verified design and tested
EAL7 - formally verified design and tested

The National Institute of Standards and Technology (NIST) and the National Security Agency (NSA) have established a program under the National Information Assurance Partnership (NIAP) to evaluate IT product conformance to international standards. The program is known as the *NIAP Common Criteria Evaluation and Validation Scheme for IT Security*. Its purpose is to help consumers select commercial off-the-shelf IT products that meet their security requirements and to help manufacturers of those products gain acceptance in the marketplace [20].

NIAP evaluates the security features of the following kinds of security products.[13]

- Anti-virus
- Firewalls
- Intrusion Detection Systems (IDS)
- Network Management
- Operating Systems
- Public Key Infrastructure (PKI)
- Trusted DBMS
- VPN

4.3.2 Evaluation of Security with NIAP Certified Firewalls

There are three types of firewalls: Packet Filters, Stateful Inspection/Filtering and Application Layer Proxy. A Packet Filter firewall inspects the packet header and forms an action. Stateful inspection/Filtering tracks the state of the network connection. For instance, when a request is sent to a web page, if the initial request is allowed through, the subsequent requests are allowed through, but tracked, and when the requests are finished the access is sealed up. Cisco Secure PIX Firewall and VPN-1/Firewall-1 Next Generation are commercial products that implement this type of firewall.

Application Layer Proxy is a per-service firewall. It would permit only outgoing traffic that was associated with port 80 incoming traffic. Sidewinder, from Secure Computing Corp., is an example of a commercial firewall of this type.

ARL has surveyed commercial firewall products that are National Information Assurance Partnership (NIAP) certified [21]. These products include

- Sidewinder G2 Firewall, Secure Computing Corporation
- Symantec Enterprise Firewall with VPN 7.0, Symantec Corporation
- Gauntlet Firewall V6.0, developed by Network Associates Inc., owned by Secure Computing Corporation

[13] List of Validated Products (by Technology Type) http://niap.nist.gov/cc-scheme/vpl/vpl_type.html

- StoneGate Firewall, Version 2.0.5, Stonesoft Corporation
- VPN-1/Firewall-1 Next Generation, Check Point Software Technologies Inc.
- Netscreen Appliance Models, Netscreen Technologies, Inc.
- Cisco Secure PIX Firewall Models, Version 6.2(2), Cisco Systems, Inc.
- Cyberguard Firewall for UnixWare / Premium Appliance Firewall, Release 4.3,
- Cyberguard Corporation

Open source firewalls such as IPTables and Squid are not submitted to NIAP because communities rather than companies develop them, and the communities don't have the funds to support certification.

ARL has chosen two of these for evaluation. Both have been certified at EAL4. As a matter of fact, this is the maximum EAL for which a firewall protection profile has been written [22].

Symantec Enterprise Firewall with VPN 7.0

The Symantec Firewall provides two basis functionalities: controlling the information traveling through it and protecting the information sent from site to site and from a remote client to site using virtual private network (VPN) capabilities. Here are some basic features of the product:

- Application-Layer Proxy: The firewall provides full application level inspection in addition to circuit-layer protection and packet filtering as in conventional firewalls. This inspection allows complete check of all levels of the protocol stack to detect and prevent attacks inserted in every level.
- Built-In support for popular protocols: Covers most popular protocols.
- High Speed Performance: Firewall throughput exceeds common network connection such as ATM OC-3 (155 Mbps), and slower networks such as Fast Ethernet, etc.
- Integrated Virtual Private Network (VPN) Support: Provides secure, high-speed connection between site to site and site to client using IPSec security protocol, with AES, DES, Triple DES encryption to protect data, and IKE key management for user authentication and key exchange.
- Operating System Hardening: Built-in detection mechanism to protect itself from intrusion.
- Port Blocking: Automatically blocks all unauthorized TCP and UDP ports to protect services from the firewall.
- Anti-Spamming and Anti-Spoofing: Protect email servers from spamming and prevent unauthorized access to internal systems.
- Centralized, Remote Management: Equipped with Symantec Raptor™ Management Console, a graphical user interface, to help create and enforce security policy, receive automatic alerts for specified log events, and generate detailed reports.

- Platform Requirements:
    - Solaris: Single processor, 400 Mhz, Solaris 7/8 UltraSparc I/II, 256 MB RAM, 8 GB disk space, CD-ROM drive, at least 2 NICs.
    - Windows NT/2000: Intel Pentium III, 400 Mhz, 256 MB RAM, 8 GB disk space, CD-ROM drive, and at least 2 NICs.

Check Point Firewall-1/VPN-1, Nokia IP350

The Check Point Firewall-1/VPN-1, Nokia IP350, full-featured designed for small and medium enterprises, operates in two modes: supervising all traffics passing between networks connected to the firewall by inspecting packets, blocking all unwanted access attempts, and protecting communication channel over the Internet (public network) between two Check Point Firewalls or a Check Point Firewall and a SecureClient.

- Stateful Inspection Technology: The firewall enforces its security policy and desktop security policy by taking action one of the following operations: either accept the IP packet flow between the source and destination, or reject with notifying the source, or drop without notifying the source. Inspects traffic from data to application layer.
- Internet Protocols: Covers most popular Internet Protocols.
- High Speed Performance: The firewall throughput for large packets is up to 350 Mbps, with VPN throughput for large packets is up to 80 Mbps, 3DES, AES.
- Integrated VPN Support: enables secure connectivity between sites, remote offices and users.
- Management and remote supervision: Equipped with management tools such as Nokia Horizon Manager, Network Voyager, etc. to simplify installation, configuration, management, and maintenance.
- Anti-Spoofing: Administrator can create a filter with particular sets of network addresses either to allow, reject or drop packets which each conforms to the allowed set of networks for particular interfaces and for the direction of movement.
- Data Filtering: capable of having FTP, HTTP and SMTP based connections diverted to an interface for packet content analysis, as a precondition for accepting.
- IP Security platform for the small enterprise: combined with the Nokia IPSO™ secure operating system which is the industry-proven hardened Nokia operating system with web-based element management interface and Command Line Interface.
- Audit: Capable of generating audit records, logs, and alerts corresponding to audit events.

Internet Security Systems' *Internet Scanner* will be used for vulnerability assessment. *Internet Scanner* performs probes of network communication services, operating systems, routers, email, web servers, firewall and applications to identify weaknesses that could be exploited by intruders to gain access to the server.

4.3.3 Secure Remote Access and Encrypted Transfer of Files

The National Archives is particularly concerned with the secure transfer of sensitive files and the security of remote access to archival systems. The PERPOS2 Network Architecture provides a Test bed for evaluating technologies for securing such a system.

The acquired firewalls also have Virtual Private Network Capability. GTRI will collaborate with ARL in assessing the capability of a VPN based on these products to protect the confidentiality of electronic records traveling across the Internet. ARL will observe and to measure the overhead associated with deployed cryptographic products used to transfer electronic records from the PERPOS portal to ARL computers. Experiments can also be conducted to assess the effectiveness of the Secure Sockets Layer (SSL) i.e., the https of the Apache web server in encrypting communications back and forth between the client and server.

The Advanced Security Option (ASO) of the Enterprise edition of Oracle provides the capability for end-to-end encryption with integrity checks on transmitted and received data. It can use private key (symmetric algorithms) private / public key pairs (asymmetric algorithms), and it supports Public Key Infrastructure (PKI). ASO implements message digests (MD5 and SHA-1) for integrity checks. ASO also allows integration with smart cards, token cards and biometrics. The effectiveness of this technology may also be evaluated.

5. Pilot Study of the Use of the Archival Processing Tool and its Refinement

Developed during Phase I research, the Archival Processing Tool (APT) supports the activities of archivists when they systematically process records received from the White House offices at the end of a Presidential administration. Figure 8 illustrates these activities. The labeled circles (bubbles) in the figure represent activities in archival processing. The labeled parallel lines represent the kinds of information that are created and used during the process. The labels on directed edges represent the kinds of information that are stored as a result of activities and subsequently used by other activities.

Archivists at the Presidential Library must accession Presidential Records that the National Archives takes custody of at the end of a Presidential Administration. Record series are loaded from storage devices, such as floppy and compact disk drives or file transfer areas. The record series (file systems) are loaded into a tool called the Archival Processing Tool. An entry is made in an accession register, and the containers are associated with the accession entry. The Archival Processing Tool supports browsing and viewing the records so that the archivist can give a preliminary description to the accessioned series and can even enter directory names or folder titles. Information about the series is stored in the Catalog of Holdings and the containers are stored in Holdings.

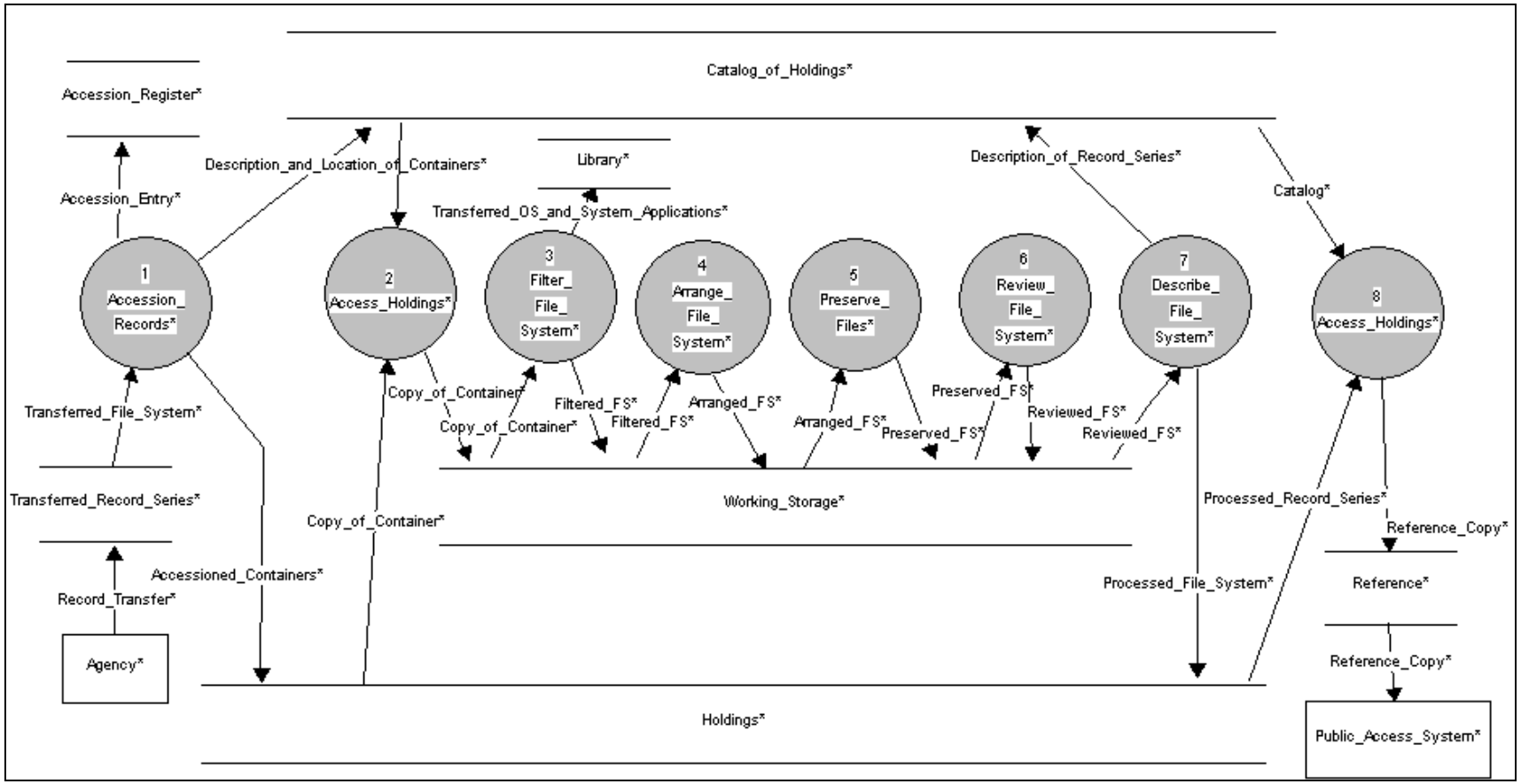

Figure 8. Systematic Archival Processing Supported by the APT.

Archivists start work by using the accession register to identify the accessioned containers they need to process. They copy the containers associated with the assigned work from Holdings into an archivist's subdirectory in a Work Area.

The personal computer records from the Bush Presidential Administration (1989-1993) include the entire file system of personal computers, operating system and application files as well as those files created or received by the White House Office Staff. The APT supports filtering file systems by blocking operating system and applications files and passing through user-created files. This activity involves separating records and non-records. After each step of work, the archivist saves their work back to the work area.

Archivists should attempt to maintain the original order of files in a file system, but some files may not have been saved in the proper directory (folder). For instance, some word processing files that should have been stored in a CORR[espondence] folder may have been stored in the root directory or in the directory including the word processing application. Archivists may need to perfect the arrangement my moving misplaced files into the proper directory. During the arrangement activity, an archivist may also extend or create better directory (folder) titles.

Some PC files may be in obsolete and/or in proprietary formats that can no longer be viewed. Other files may be corrupted due to media deterioration or file transmission errors. Other files may be encrypted, so that there is a need to recover a password and decrypt the file so that it can be viewed. These activities are referred to as archival preservation.

PC records must next be reviewed for Freedom of Information Act (FOIA) exemptions on their release to the public. The must also be reviewed for Presidential Record Act (PRA) restrictions on their release. The Archival Processing Tool supports the viewing, redaction, and withdrawal of records with access restrictions.

When archivists have completed the preceding activities, they must describe the record series. Since this involves defining the record creator (organization, office and/or individual records creator), it has traditionally been described as not only description, but as arrangement of the processed record series. This involves loading the containers containing the record series to view and describe their contents, and moving the containers from the work area to archival storage of processed records. The result is an updated description in the Catalog of Holdings.

Since the master copy that is stored in archival storage (Holdings) may contain records whose access is restricted in whole or in part, it is then necessary to create a reference copy that includes just those records that are open to the public. To do this, an archivist accesses the current holdings and creates a reference copy that is transferred to the Public Access System.

Archivists need to access the holdings of processed record series to review those closed files or originals of redacted documents when access restrictions have expired. They also need to access those holdings for preservation actions such as converting to new file formats when current file formats become obsolete.

Pilot use of the APT (version 2.04) by two archivists began at Archives II in the fall of 2002 and continued through the winter. A report was written that summarizes what are believed to be new requirements for processing records created on legacy personal computer platforms [23]. The requirements are identified based on the results of laboratory experiments and use of the APT prototype by archivists. With each requirement is described the experimental or practical situation in which the requirement was identified, risks that will be encountered if the requirement is not satisfied, and the technological alternatives that have been, or are being, explored to meet the requirement and mitigate the risks. The report also describes advanced technologies that can be used to satisfy previously known requirements for processing electronic records.

The APT User Guide was revised to reflect changes made to the functionality of the tool [24]. These changes included integrating accessioning and description into the APT. The capability to identify file types was extended to over 225 file types.

Version 2.05 of the APT was installed at the Bush Presidential Library and Museum in College Station, Texas. Two members of the archival staff were trained in its use. During the spring, archivists at the Bush Presidential Library used Version 2.05 of the Archival Processing Tool (APT). The contents of three file systems were processed (filtering, arrangement, review). The archivists identified additional functionality and revisions to functionality that was needed. These included, but were not limited to:

- The capability to merge two containers, for example, when system and software applications are discovered after filtering and other archival activities have been performed.
- The capability to move a file that was previously believed to be a non-record, back into a system of records.
- Personal Record Misfile (PRM) is not a reason for closing a file. The action should be "Mark as PRM."
- It must be possible to view the document at the same time one is entering Reasons for Withdrawal.
- It must be possible to transfer Personal Record Misfiles (PRMs) out of a reviewed file system.
- Three items needed to be added to the Withdrawal Information Dialog box: Withdrawal Sheet (WS) No., Case No., and Pages.
- When Transfer is selected during the Filtering activity, a Transfer Information Dialog box is needed that contains the following information:
    - Accession No.
    - Case No.
    - OAID No.
    - Container Id

- Collection
- Unit
- Series
- Folder Title
- Transferred To
- Description
- Transferred by
- Date of Transfer

This information will be used to fill in a Presidential Library Transfer Sheet that will be included with the Transferred files and in the User-created Files. The name of the file will be TRANSFER and it will be stored in the Manifest_Info Directory.

- During review, it must be possible to change closed, opened, redacted files, files marked for transfer, and files marked as PRM back to unreviewed.
- A viewer is needed for PIF (shortcut files). Without such a viewer, QuickView Plus (the set of file viewers) attempts to take the shortcut.
- An operation "Withdraw P1F1 material" will be needed on the File pull-down menu. This is needed for material that an archivist finds is not marked as security classified, but in their judgment is restricted from release as containing National Security Information, for instance, a document that was a letter to a Foreign Head of State. The file should be removed from the file system and written to a CD with a withdrawal sheet placed where it was, and a copy of the withdrawal sheet written to the CD. The withdrawn file will have to be processed on a computer accredited/certified for processing national security materials.
- The Reference Copy of a Record Series will have only open and redacted Files. In place of the closed files and PRMs, it will have a withdrawal sheet created from the withdrawal information stored in the manifest of the master copy. The filename of the withdrawal sheets will be DNnnnnnn.pdf where nnnnnn is the sequence number of the document in the directory containing the withdrawn document. It is not the filename of the withdrawn document. It will also contain transfer sheets in place of records that were PRMs.
- The preservation activity will need to support conversion to formats such as PDF and JPG so that files that are in legacy formats can be made available for display in NARA's archival research catalog (ARC)

Another result of the pilot study is some reviewed PC records that have been opened to public access. Such records are needed for research in the information extraction and FOIA/PRA review research tasks described in sections 2 and 3 of this report.

The Accession, Description, and Access Holdings activities are creating, modifying and using metadata about the Collection, Office, Record Series and Folders of a file system and storing or retrieving containers from Archival Storage. The Filter, Arrangement, Preservation, and Review activities are operating on the files in a file system. However, it is necessary to view the file system during Accession and Description. Consequently, the activities of Accession, Description and Access Holdings have been removed from the Archival Processing Tool (APT) and have been and placed in the Archival Repository

Tool (ART). These functions are described in a new version of the User Manual [25]. Figure 9 illustrates the user interface to ART and the APT when accessioning a record series.

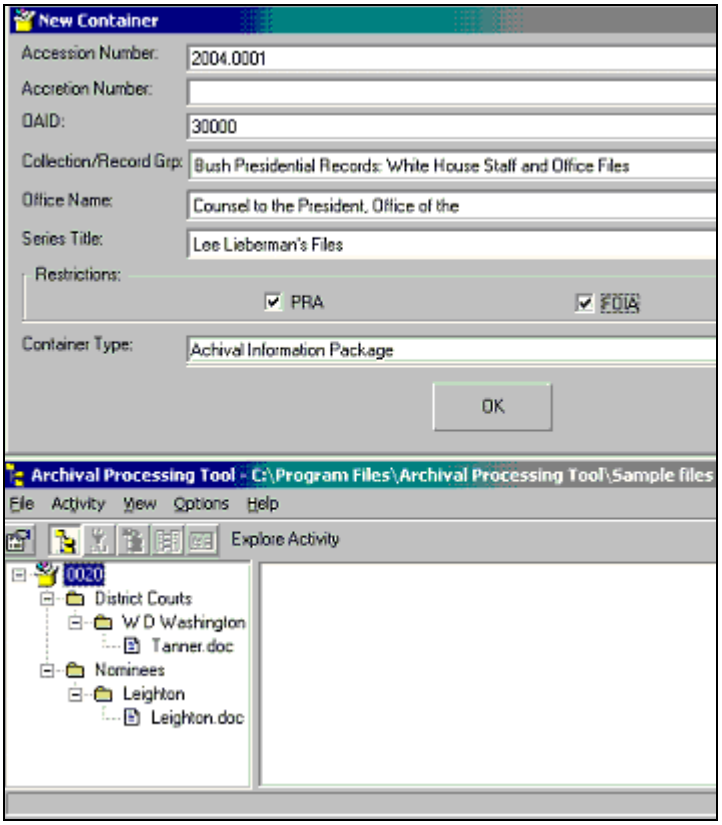

Figure 9. Interface for Accessioning Containers and Viewing Files in a File System.

The information in the upper window is a part of the Archival Repository Tool. It relates to the Accession Register and Catalog of Holdings. The information in the lower window relates to browsing a file system and viewing Files. These are functions of the Archival Processing Tool.

Most of the recommendations from Bush Library Archivists have been implemented in a new version of the APT. During late spring, the new versions of the APT and ART were installed at the Bush Presidential Library so that the Archivists could continue the Pilot use of the tools.

The APT was primarily designed to support systematic processing of records created using personal computers. This means arranging, preserving, reviewing and describing all records in an accessioned record series (or file system). It was learned from working the archivists at the Bush Presidential Library that their work primarily involves response to FOIA requests, rather than systematic processing. This activity, which is referred to as FOIA processing, involves finding all records relevant to a FOIA request, regardless of which record series they are in. This often involves processing just a few folders in a series and not the entire series.

The text-based information retrieval research scheduled for FY 2005 will address in part additional the functionality that will be needed for FOIA response. To support FOIA processing, changes will be needed to the metadata in the Catalog of Holdings to reflect the storage of FOIA collections in addition to record series.

6. Summary

The General Architecture for Text Engineering (GATE) developed at the University of Sheffield is a good framework for integrating advanced lexical and processing resources for textual analysis and understanding. GATE and the ANNIE information extraction processing resources are open source, have good documentation, and are widely used in academic research.

ANNIE has a good collection of processing resources for information extraction. An experiment has shown the performance of those resources in recognizing many of the kinds of named entities that are necessary for document type identification. Analysis of the results indicates that while many of the named entities were correctly identified, many were only partially correct, and others incorrect. This is in large part due to Presidential Records and Personal/Political Records containing document types, such as letters, memos, and agenda that were not used in the MUC and ACE conferences, which were limited to newswire and broadcast news document types. However, these partially correct and incorrect annotations can be corrected by extending the location, person's name, and organization gazetteers, and by modification of the JAPE rules. The experiment also showed that there are other annotations, such as addresses and job titles, that are needed for document type identification. JAPE rules can be used to recognize these entity types and integrated with ANNIE's other processing resources. This is being done and after testing, a second experiment will be conducted at the Bush Presidential Library.

The Bush Presidential electronic records contain textual information in at least 28 legacy file formats. To extract information or content from them it is necessary to read or convert them to ASCII plain text or HTML. An experiment with the Stellent Filters available with Oracle Text demonstrated that Stellent Filters could correctly (with few errors) read 21 of those file formats. File readers for the other 7 file formats will be needed. Readers that convert proprietary file formats to ASCII text are relatively easy to create.

The thirty-five document types that are in our current corpus of Presidential and Personal Records were analyzed to determine the features that characterize each document type. JAPE rules are being constructed for recognizing those annotation types needed for document type recognition, but that are not among the current ANNIE processing resources. Two approaches to document type recognition are being investigated, a rule-based approach based on JAPE rules constructed to recognize each document type, and a machine learning approach is which a Machine Learning Module is trained to recognize

document types in a record sample, and then used to recognize document types in file systems that were not in the training sample.

A Folder Titler is being developed that uses the Document Type Identifier to identify the types of documents in the directories of a PC file system. A set of rules is being constructed that can be used to infer whether a directory contains correspondence, memos, or is a subject file, and suggests an extension to the directory (folder) title.

Archivists at the Bush Presidential Library use a folder title list as an aid in responding to FOIA requests. A directory (folder) title list of the contents of the White House Office and Staff Members' personal computers would aid them in responding to FOIA requests for these electronic records. The contents of the Bush Hard Drives will not have scope and content notes for record series until archivists find time to systematically review the contents of the hard drives. A Folder Titler and a Record Series Summarizer are being developed that use the Document Type Identifier and content extraction capabilities to automatically extend folder titles and create scope and content notes for series of electronic records. This year, experiments will be conducted at the Bush Presidential Library to determine the performance of these tools.

A report was prepared that described the results of a study and analysis to identify the kinds of knowledge and reasoning that archivists use to judge whether records are Federal or Presidential records or whether they are personal records. Some of the kinds of knowledge used to determine whether passages of a record or an entire record is subject to access restrictions, or can be opened for public access were identified.

An Access Restriction Checker is being prototyped that incorporate the information and content extraction technologies described in section 2 of this report. The Jess (Java Expert System Shell) scripting language is being used to represent factual knowledge, such as names of Bush family members and names of advisors to the President, to represent knowledge of FOIA exemptions and PRA restrictions and for rule-based inference. It will be tested on a corpus of about 300 documents that represent the kinds of documents encountered by archivists who review Presidential records. These include personal/political records, Federal records and Presidential records corresponding to records with no restrictions and each kind of restriction typically encountered. After testing, experiments will be conducted at the Bush Presidential Library to evaluate the performance of the tool.

GTRI is collaborating with ARL in experiments to evaluate the performance of secure portal technology and security products protecting the availability, integrity, authentication and confidentiality of records stored, processed, transmitted and received on a server connected to an Internet II environment.

An Information Assurance Test bed has been created that consists of the Oracle Application Server implementing the PERPOS Portal, an Oracle Database Server, an Archival Application Server, providing archival services, an E-mail Server, and a Log Server. A Security Policy was developed for the PERPOS Network Architecture. The

purpose of this policy is to provide general guidelines and specific recommendations for the protection of information stored on the Presidential Electronic Records PilOt System (PERPOS) computer network.

ARL has surveyed commercial firewall products that are National Information Assurance Partnership (NIAP) certified. ARL has chosen two of these for evaluation. Both have been certified at EAL4. Internet Security Systems' *Internet Scanner* will be used for vulnerability assessment. *Internet Scanner* performs probes of network communication services, operating systems, routers, email, web servers, firewall and applications to identify weaknesses that could be exploited by intruders to gain access to the server.

The National Archives is particularly concerned with the secure transfer of sensitive files and the security of remote access to archival systems. The firewalls selected for evaluation also have Virtual Private Network capability. GTRI will collaborate with ARL in assessing the capability of a VPN based on these products to protect the confidentiality of electronic records traveling across the Internet. ARL will observe and to measure the overhead associated with deployed cryptographic products used to transfer electronic records from the PERPOS portal to ARL computers.

Version 2.05 of the Archival Processing Tool (APT) was installed at the Bush Presidential Library and Museum in College Station, Texas. Two members of the archival staff were trained in its use. During the spring, the archivists used the APT to process (filtering, arrangement, review) the contents of three accessioned file systems. They identified additional functionality and revisions to functionality that was needed. Most of the recommendations from Bush Library Archivists have been implemented in a new version of the APT. During late spring, new versions of the APT and Archival Repository Tool (ART) were installed at the Bush Presidential Library so that the archivists could continue the pilot use of the tools.

References

1. E. Brill. A simple rule-based part-of-speech tagger. In *Proceeding of the Third Conference on Applied Natural Language Processing*, pages 152-155, Trento, Italy, 1992.

2. N. Chinchor. Overview of MUC-7/MET-2. *Proceedings of the Seventh Message Understanding Conference*. 1998. http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_proceedings/overview.html

3. N. A. Chinchor and E. Marsh. MUC-7 Information Extraction Task Definition. *Proceedings of the Seventh Message Understanding Conference*. July 1998. HTML

4. N. A. Chinchor and P. Robinson. MUC-7 Named Entity Task Definition. *Proceedings of the Seventh Message Understanding Conference*. September 1997. HTML

5. K. Humphreys, R. Gaizauskas, S. Azzam, C. Huyck, B. Mitchell, H. Cunningham, and Y. Wilks. University of Sheffield: Description of the LASIE-II System as Used for MUC-7. Department of Computer Science, University of Sheffield. 1998. PDF

6. D. Maynard, H. Cunningham, and Y. Wilks. Sheffield ACE Report. Natural Language Processing Group, University of Sheffield, UK. 2002. HTML

7. H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*. Philadelphia, July 2002. PDF

8. H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan, C. Ursu, and M. Dimitrov. Developing Language Processing Components with GATE: A User Guide. Department of Computer Science, University of Sheffield. February 2003. PDF

9. Oracle Text Reference Release 9.2, Appendix B: Supported Document Formats http://www.cise.ufl.edu/help/database/oracle-docs/text.920/a96518/toc.htm

10. M. G. Underwood. File Readers for Legacy File Formats, Working Paper, Information Technology and Telecommunications Laboratory, Georgia Tech Research Institute, June 2004

11. M. G. Underwood. Recognizing Named Entities in Presidential Electronic Records. PERPOS Technical Report 04-4, Information Technology and Telecommunications Laboratory, Georgia Tech Research Institute, June 2004.

12. W. Underwood and M. Hayslett-Keck. A Collection of Presidential, Federal and Personal Records for use in Information Extraction and FOIA/PRA Review Experiments. PERPOS Technical Report 04-5, CSITD/ITTL, GTRI, June 2004.

13. M. Hayslett-Keck and W. Underwood, An Analysis of the Knowledge Required to Perform FOIA and PRA Review, PERPOS Technical Report ITTL/CSITD 04-1, Information Technology and Telecommunications Laboratory, Georgia Tech Research Institute, January 2004, revised March 2004.

14. E. J. Friedman-Hill. Jess, The Rule Engine for the Java Platform, Version 6.1p7. SAND98-8206 (revised), Sandia National Laboratory, Livermore, CA, 7 May 2004. http://herzberg.ca.sandia.gov/jess/docs/61/

15. E. Friedman-Hill: Jess in Action: Rule-based Systems in Java. Manning Publications, July 2003.

16. B. Q. Nguyen. An Information Assurance Architecture for a Web Portal of Presidential Electronic Records, Army Research Laboratory, February 2004.

17. D. Molavi. Information Security Practices Recommended for Use with the PERPOS2 Server, PERPOS Technical Report ITTL/CSITD 04-2, Information Technology and Telecommunications Laboratory, Georgia Tech Research Institute, January 2004

18. D. Molavi. PERPOS Network Use and Security Policy, PERPOS Technical Report 04-6, Information Technology and Telecommunications Laboratory, Georgia Tech Research Institute, June 2004.

19. NIST. *Common Criteria for Information Technology Security Evaluation Version 1* (also ISO 15408:1999) http://csrc.nist.gov/cc/

20. Common Criteria Evaluation and Validation Scheme for IT Security - Organization, Management, and Concept of Operations. http://niap.nist.gov/cc-scheme/policy/ccevs/guidance_docs.html

21. S. Nguyen. Firewall Selection for the PERPOS Network, ARL Report, May 2004.

22. Application-level Firewall Protection Profile for Medium Robustness Environments Version 1.0, Information Assurance Directorate, National Security Agency, October 28, 2003. http://niap.nist.gov/cc-scheme/pp/PP_VID1016-PP.pdf

23. W. E. Underwood. The Presidential Electronic Records PilOt System: Results of Laboratory Experiments and Use by Archivists. PERPOS TR ITTL/CSITD 03-01, Information Technology and Telecommunications Laboratory, Georgia Tech Research Institute, November 2003. http://perpos.gtri.gatech.edu/perpos/publications/index.htm

24. W. Underwood, M. Hayslett-Keck and S. Laib. The Archival Processing Tool: User's Guide, Version 2.06. PERPOS Technical Report ITTL/CSITD 03-2, Information Technology and Telecommunications Laboratory, Georgia Tech Research Institute, Revised December, 2003

25. W. Underwood, M. Hayslett-Keck and S. Laib. The PERPOS Tools: User's Guide, Version 3.0. PERPOS Technical Report ITTL/CSITD 04-2, Information Technology and Telecommunications Laboratory, Georgia Tech Research Institute, March 2004