
William Underwood
Robert Simpson
Elizabeth Whitaker
Lucja Iwanska
Sandra Laib
Brian Harris
Matthew Underwood
Shiela Isbell
Marlit Hayslett
Demetrius Campbell
Dariush Molavi
Jason Kau

PERPOS Technical Report ITTL/CSITD 06-06
December 2006

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

The Army Research Laboratory (ARL) and the National Archives and Records Administration (NARA) sponsored 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

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 record types such as letters, memos, itineraries, and resumes. The recognition of record types supports metadata extraction, automated titling of directories, and summarization of record series.

Archivists must respond to Freedom of Information Act (FOIA) requests beginning five years after the end of a Presidential administration. To respond to these requests, they must be able to search collections of e-records for records that are relevant to the requests. A natural language-based Boolean query language is described that can support FOIA Search and e-discovery in very large record collections. A knowledge acquisition technology is described that has the potential to improve recall of relevant records.

Review of records for Presidential Record Act (PRA) restrictions and FOIA exceptions is an intellectually demanding task, requiring page-by-page review. A prototype Access Restriction Checker has been constructed that uses content extraction and rule-based reasoning technologies to distinguish Presidential Records from Personal Records. 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 e-records relating to Presidential appointments to Federal Office or containing confidential advice between the President and his staff.

The Presidential Electronic Records Pilot System (PERPOS) supports systematic processing of Presidential Electronic Records. It consists of two software components. The Archival Repository Tool (ART) supports accession, cataloging, repository management, and creation of reference copies of e-records that are available for public access. The Archival Processing Tool (APT) supports arrangement, preservation, and review of electronic records in the repository.

This report describes additional functions developed for the PERPOS System that support FOIA Processing of Presidential e-records. The functionality added to PERPOS includes:

- Indexing the accessioned electronic records,
- Creating a FOIA case,
- Searching the indexed records for records relevant to a FOIA request,
- Automatic estimation of the number of pages of e-records relevant to a request,
- Reviewing records for a FOIA case,
- Creating the Scope and Content Note for a FOIA case, and
- Automatically creating a finding aid and a container for a FOIA Collection.
Electronic records stored in digital repositories are vulnerable to system failure, human error, and malicious actions. A Web portal was created for Internet access with an isolated subnetwork behind a firewall containing an archival repository and archival services. The PERPOS Project Staff collaborated with the Army Research Laboratory in evaluating firewall and vulnerability assessment technologies for protecting these resources.

Experiments have been 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 tools that support FOIA processing as well as systematic processing (accession, arrangement, preservation, review, and description) of electronic records.

Paper records can be scanned to produce digital images of the records. These images can be converted to machine-readable records using OCR technology. Thus, the technologies being developed during this project can be applied to processing machine-readable copies of paper records as well as records originally created digital.

**Keywords:** information extraction, content extraction, record types, summarization, query languages, knowledge acquisition, natural language processing, information assurance, E-FOIA, FOIA processing
# Table of Contents

1. INTRODUCTION ............................................................................................................................... 1
   1.1 BACKGROUND ............................................................................................................................. 1
   1.2 PURPOSE .................................................................................................................................... 1

2. INFORMATION EXTRACTION TO SUPPORT RECORD TYPE RECOGNITION AND SUMMARIZATION ............................................................................................................................... 2
   2.1 INFORMATION EXTRACTION EXPERIMENTS .............................................................................. 2
       2.1.1 Enhancements to the Fact Base and Rules ........................................................................... 6
       2.1.2 Experimental Evaluation of the Enhancements ................................................................. 11
   2.2 RECORD TYPE LEARNING AND RECOGNITION ..................................................................... 14
   2.3 CREATION OF FOLDER TITLES AND SCOPE AND CONTENT NOTES ..................................... 18
       2.3.1 Method for File Unit Description ......................................................................................... 19
       2.3.2 Summarization of Record Series Content ............................................................................ 20

3. TECHNOLOGIES FOR RETRIEVAL OF PRESIDENTIAL E-RECORDS ........................................... 22
   3.1 NATURAL LANGUAGE-BASED RECORD RETRIEVAL ................................................................. 24
   3.2 ACQUISITION OF KNOWLEDGE TO SUPPORT RECORD RETRIEVAL ....................................... 27

4. DECISION SUPPORT FOR FOIA AND PRA REVIEW ........................................................................ 30
   4.1 SPEECH ACTS ............................................................................................................................... 31
   4.2 ACCESS RESTRICTION CHECKER PROCEDURE ......................................................................... 33
   4.3 THE USER INTERFACE .................................................................................................................. 34
   4.4 DECISION RULES FOR RECOGNIZING PERSONAL RECORDS .................................................. 37
   4.5 DECISION RULES FOR RECOGNIZING PRA RESTRICTION A(5) ............................................ 38

5. FOIA PROCESSING ............................................................................................................................. 41
   5.1 FOIA PROCESSING DATAFLOW ................................................................................................. 41
   5.2 INDEX HOLDINGS ......................................................................................................................... 43
   5.3 FOIA CASE MANAGEMENT .......................................................................................................... 44
   5.4 PERFORMING A SEARCH ............................................................................................................. 46
   5.5 ESTIMATING THE NUMBER OF PAGES TO BE REVIEWED ....................................................... 51
   5.6 REVIEWING RECORDS FOR FOIA CASES ................................................................................. 52
   5.7 FOIA CASE DESCRIPTION ............................................................................................................ 60
   5.8 ARRANGEMENT AND FINDING AID OF A FOIA REFERENCE CONTAINER ................................ 61
   5.9 VIEW FOIA COLLECTIONS .......................................................................................................... 62
   5.10 PILOT EVALUATION .................................................................................................................... 65

6. EVALUATION OF ADVANCED TECHNOLOGIES FOR INFORMATION ASSURANCE ...... 65
   6.1 INFORMATION ASSURANCE TEST BED .................................................................................... 66
   6.2 INFORMATION SECURITY POLICY FOR THE TEST BED AND ITS IMPLEMENTATION ............. 68
   6.3 EVALUATION OF NIAP CERTIFIED INFORMATION ASSURANCE PRODUCTS ........................... 72
       6.3.1 The Common Criteria and NIAP Certification ................................................................... 72
       6.3.2 Evaluation of Security with NIAP Certified Firewalls ....................................................... 73

7. SUMMARY OF RESULTS .................................................................................................................. 76

REFERENCES ........................................................................................................................................... 78
LIST OF FIGURES

FIGURE 1. THE ANNIE APPLICATION PIPELINE IN DEFAULT CONFIGURATION................................. 4
FIGURE 2. A JAPE RULE FOR ANNOTATING A JOB TITLE FOLLOWED BY PERSON’S NAME................ 5
FIGURE 3. ANNIE PERFORMANCE ON SAMPLE CORPUS 1 .......................................................... 6
FIGURE 4. OUR APPROACH TO RECORD TYPE LEARNING............................................................ 14
FIGURE 5. XML MARKUP OF NAMED ENTITIES IN SAMPLE CORRESPONDENCE. .................... 15
FIGURE 6. INTELLECTUAL FORMS OF SAMPLE WHITE HOUSE CORRESPONDENCE .................. 16
FIGURE 7. INDUCED STOCHASTIC CFG FOR CORRESPONDENCE RECORD TYPE .................... 17
FIGURE 8. APPROACH TO IDENTIFYING RECORD TYPE............................................................... 17
FIGURE 9. PARSE OF INTELLECTUAL ELEMENTS INDICATING A CORRESPONDENCE RECORD TYPE........ 18
FIGURE 10. LEXICAL CONSTRUCTS THAT EXPRESS TAXONOMIC KNOWLEDGE ....................... 28
FIGURE 11. SOME EXAMPLES OF SPEECH ACTS ......................................................................... 32
FIGURE 12. PERPOS2 NETWORK ARCHITECTURE......................................................................... 66
1. Introduction

1.1 Background

Due to the need to respond to FOIA requests, archivists at the Presidential Libraries do not get around to systematic processing (arrangement, preservation, review and description) of records until a couple of decades after the end of an administration. With the increasing volume of Presidential electronic records being accessioned, there is a need for technologies to support automatic description of file units and record series, so that archivists have better intellectual control of accessioned records.

Archivists at Presidential Libraries must respond to FOIA requests for Presidential Records beginning five years after the end of an administration. With the increasing number of Presidential e-records being accessioned, there is a need to support search of e-record collections for records that are relevant to FOIA requests.

Presidential Library Archivists who must search very large e-record collections face the problem that the larger the collection to be searched, the lower the precision (relevance) of the retrieved documents to the request. The result is that a large number of documents reviewed end up not being relevant to the request. NARA also faces the challenge of searching large collections of electronic records related to federal court litigation involving federal and presidential records in NARA's custody. There is a need for improving precision in e-discovery in very large heterogeneous e-record collections.

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 e-records, the need to review these records, and the cost of the limited human resources that can be applied to the review process, there is a need for automated support of archival review decisions.

With the increasing accessions of Presidential e-records, there is a need to support FOIA case management for e-records, E-FOIA review, and automatic creation of finding aids for e-FOIA collections.

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 assurance technology to address these risks.

1.2 Purpose

The research tasks of the Presidential Electronic Records PilOt System (PERPOS) Project - Phase II were to:
(1) Support improved archival processing through the development and prototyping of advanced technologies to automatically extract information from digital text files; to automatically identify record types; and to summarize folder contents and record series;

(2) Extend natural language-based document search and retrieval technology to support FOIA Search and e-discovery in very large record collections;

(3) Represent 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;

(4) Refine the archival processing tools developed during prior research to support FOIA processing; and

(5) In collaboration with U.S. Army Research Laboratory scientists, evaluate advanced technologies assuring the availability, integrity, authentication, and confidentiality of Presidential records that are preserved, managed, and accessed through distributed, heterogeneous electronic record repositories.

The purpose of this report is to describe results and progress toward these research objectives. The following sections describe progress and results on each of the research tasks. The final section summarizes the research results.

2. Information Extraction to Support Record Type Recognition and Summarization

For experiments in information extraction, archival description and FOIA review, a corpus of 150 Presidential records from the Bush Administration was created [Underwood and Hayslett-Keck 2004].

2.1 Information Extraction Experiments

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 facilitates automated reasoning in support of archival processes such as review and description. Extracted information can be used to recognize record types such as letters, memos, itineraries, and resumes. The recognition of record types supports metadata extraction, automated titling of directories, and summarization of record series. Extracted content, supplemented with knowledge of access restrictions, also facilitates recognition of documents or passages in documents
that might be subject to Freedom of Information Act exemptions and Presidential Record Act restrictions.

Technologies and tools for automatic extraction of information from textual documents were described [Underwood 2004]. In particular, ANNIE (A Nearly-New Information Extraction System) was described as an example of a lexical resource and rule-based approach to information extraction. A similar system had been used in the Message Understanding Conferences and achieved performance comparable to that of humans performing the same task.

The Seventh Message Understanding Conference (MUC-7) had as one of its task the automatic recognition of the following named entities.

<table>
<thead>
<tr>
<th>MUC-7 Entity Name</th>
<th>Named Entity Description</th>
<th>Named Entity Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Person</strong></td>
<td>Named person or family</td>
<td>Mr. President, Mr. and Mrs. Tom Walker</td>
</tr>
<tr>
<td><strong>Location</strong></td>
<td>Name of a politically or geographically defined location (cities, provinces, countries, international regions, bodies of water, mountains, etc.)</td>
<td>Peru, Montana, Red Sea, 15 Elm Street NW</td>
</tr>
<tr>
<td><strong>Organization</strong></td>
<td>Named corporate, governmental, or other organization entity</td>
<td>Navy, Red Cross, US Congress, Republican Party</td>
</tr>
<tr>
<td><strong>Time</strong></td>
<td>Complete or partial expression of time of day</td>
<td>2 pm, 1400 EST, 09:11:34 EST</td>
</tr>
<tr>
<td><strong>Date</strong></td>
<td>Complete or partial date expression</td>
<td>tomorrow, last year, 9/07/1946</td>
</tr>
<tr>
<td><strong>Percent</strong></td>
<td>percentage</td>
<td>10-15%, 25 to 35%, 9.75 percent</td>
</tr>
<tr>
<td><strong>Monetary</strong></td>
<td>monetary expression</td>
<td>18.00 dollars, $18.00</td>
</tr>
</tbody>
</table>

The MUC-7 evaluations of named entity recognition systems showed that they had achieved performance levels comparable to human annotators. In particular, the LaSIE system, from which ANNIE was developed, achieved an F-measure of .86-.90.

Figure 1 shows the user interface to the ANNIE application pipeline.
Document Reset removes previous annotations and any undesired markup in preparation for a clean run of the IE system. The English Tokenizer segments a document into words, numbers, symbols, punctuation, and space tokens. The ANNIE Gazetteer processing resource compilation of word lists into finite state machines for fast lookup. The Sentence Splitter segments documents at the sentence level. The Parts of Speech Tagger uses a default lexicon and rule set to classify tokens as nouns (singular, plural, proper), verbs (base form, past tense), adjectives, determiners, etc. The Named Entity Transducer (aka. Semantic Tagger) is based on the JAPE (Java Annotations Pattern Engine) language. ANNIE’s JAPE rules recognize and annotate the named entities in text. The left-hand-side (LHS) of the rules consists of an annotation pattern that may contain regular expression operators (e.g. *, ?, +). The right-hand-side (RHS) consists of annotation manipulation statements [Cunningham et al 2006]. The Orthomatcher processing resource correlates Persons, Locations, Organizations, and Date annotations as being similar enough to each other to create a “match”.

In the following JAPE rule, the Lookup annotation refers to the wordlist (gazetter) for all words associated with the category of jobtitle. If the rule matches a jobtitle followed by the TempPerson annotation, the jobtitle will be annotated with jobtitle and TempPerson will be annotated with person name. This rule would annotate "President George Bush" as <jobtitle>President</jobtitle><person>George Bush</person>.

Rule: PersonJobTitle
Priority: 20

Figure 1. The ANNIE Application Pipeline in Default Configuration
Is the rule-based information extraction system technology as evaluated at the MUC conferences extendable to the domain of Presidential electronic records? In particular, can the ANNIE processing resources as initially configured achieve a comparable level of performance in recognizing named entities in Presidential correspondence, memos, press releases, etc? To answer these questions, an experiment was conducted that is similar to the MUC evaluations. A corpus of 50 representative Presidential records was selected. The ANNIE system in default configuration was applied to this corpus. A human annotator marked up the named entities in the same 50 documents for the corpus key. The “key” files are the versions of the corpus that have been marked-up by a person as being correct. Of the fifty documents used to evaluate ANNIE’s Named Entity Recognition performance, the majority of them were OCR'd text files from scanned photocopies of records from the George Bush Presidential Library and Museum [Underwood 2004].

GATE’s Corpus Benchmark Tool was used to calculate performance measures. It calculated precision, recall and F-measure for each entity type.

\[
\begin{align*}
\text{Precision} &= \frac{\text{Correct} + \frac{1}{2}\text{Partial}}{\text{Correct} + \text{Spurious} + \text{Partial}} \\
\text{Recall} &= \frac{\text{Correct} + \frac{1}{2}\text{Partial}}{\text{Correct} + \text{Missing} + \text{Partial}} \\
\text{F-measure} &= \frac{(\beta^2 + 1)\text{Precision} \times \text{Recall}}{(\beta^2 \text{Precision}) + \text{Recall}}
\end{align*}
\]

The partially correct annotations count off of ANNIE’s precision and recall scores only half as much as the missing and False positive responses do. Setting $\beta=1$ results in recall and precision being weighed equally in the F-measure.

The Corpus Benchmark tool also computed the average precision, recall and F-measure for all entity types. Figure 3 shows the results.

<table>
<thead>
<tr>
<th>Annotation Type</th>
<th>Correct</th>
<th>Partially Correct</th>
<th>Missing</th>
<th>Spurious</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person</td>
<td>242</td>
<td>94</td>
<td>75</td>
<td>107</td>
<td>0.65237020316</td>
<td>0.70316301703</td>
<td>0.67681498829</td>
</tr>
<tr>
<td>Location</td>
<td>249</td>
<td>14</td>
<td>111</td>
<td>9</td>
<td>0.94117647058</td>
<td>0.68449197860</td>
<td>0.79256965944</td>
</tr>
<tr>
<td>Organization</td>
<td>146</td>
<td>105</td>
<td>59</td>
<td>65</td>
<td>0.62816455696</td>
<td>0.64032258064</td>
<td>0.63418530351</td>
</tr>
<tr>
<td>Date</td>
<td>162</td>
<td>3</td>
<td>30</td>
<td>6</td>
<td>0.95614035087</td>
<td>0.83846153846</td>
<td>0.89344262295</td>
</tr>
<tr>
<td>Money</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>Percent</td>
<td>9</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>1.0</td>
<td>0.69230769230</td>
<td>0.81818181818</td>
</tr>
</tbody>
</table>
The performance of the ANNIE system in recognizing named entities in Presidential e-records was fair—overall average precision .786, overall average recall was .713, and F-measure .749. The precision for Location, Date, Money and Percent annotations was comparable to results achieved during the MUC conferences on other corpora (> .94). However, the precision for Person (.65) and Organization (.62) annotations was poor, as reflected in the high number of partially correct and spurious annotations. With the exception of Date and Money annotations, the recall was poor (.64-.70) as indicated by the relatively large number of missing and only partially correct annotations. The system missed a number of named entities, was only partially correct in a number of others, and incorrectly identified some terms. An analysis of the types of errors and their sources was described [Underwood 2004].

2.1.1 Enhancements to the Fact Base and Rules

Factual knowledge needed for information extraction was identified [Harris and Underwood 2004]. The analysis of the results of the first information extraction experiment [Underwood 2004] was reviewed to identify potential areas for improving the performance on the original corpus. Improvements to the IE system are in three areas. First, 69 wordlists containing about 15000 terms were added. Second, modifications were made to the default JAPE rules and about 50 rules were added [Isbell et al 2006].

Additions to the wordlists to improve the recognition of person's names include:

- Last names of Appointees and nominees to Federal Office during the Bush Administration
- Ambiguous first names and last names
- Most frequent US female first names
- Most Frequent US male first names
- Last names of Senators and Representatives in the 101st and 102nd Congress
- White House staff members last names

Wordlists added to improve recognition of organization names include:

- Organizational prefixes, such as National and Federal
- Names and abbreviations of Departments and Offices of the Executive Branch of the US Government
- Names of the White House Offices during the Bush Administration
- Names for U.S. Federal Judicial Organizations, e.g., US District Court and Supreme Court
• Names of Offices and Committees of the Legislative Branch of the US Government
• Names of US Public Policy Groups, industry trade associations, charities, patriotic, fraternal and civic organizations

Wordlists added to improve the recognition of job titles include:

• Job titles of President Bush's Nominees and Appointees to Federal Office
• Job titles of the White House Staff during the Bush Administration
• Initial terms in job titles, such as Deputy, Assistant, Chief, and Vice
• US Military ranks and their abbreviations

Wordlists added to improve recognition of dates and times include:

• Federal US Holidays, for example, New Years, Independence Day and Thanksgiving
• Names for parts of the day, e.g., the witching hour, dawn, dusk
• Spans of time, e.g., seconds, minutes, hours

Additions to the wordlists to improve the recognition of addresses and locations include:

• Secondary unit designations such as Suite, Apartment, and Building
• Pre- and Post-directionals, e.g., North, Northeast, and their abbreviations
• Names for common street suffixes, e.g., Avenue, Boulevard, and their abbreviations
• Major US city names, US state names and their postal abbreviations

The following list shows some of the problem areas in recognizing person's names.

• Misrecognized entity, e.g., B.A. Degree taken as Person Initials + Last Name
• Middle names, e.g., J Boyden Gray, and merged names, e.g., Mr. and Mrs. Lightfoot, were not recognized
• Many organization names recognized as Person names, required modifications to both Person and Organization JAPE rules
• Ambiguous names, e.g., Virginia as a state name and as a first name of a female person.
• Titles and names not merged: <title>President</title> <person>George Bush</person> should be <person><title>President</title> <person>George Bush</person></person>
• Made changes to lastname, firstname JAPE rule to recognize person name.

To eliminate these problems, nine JAPE rules were added to the name phase, and others were modified. The following is an example of a rule added to recognize person's names that are preceded by a title word, and optionally followed by a social title such as Jr. or Sr.
Rule: PersonTitle
Priority: 35.
// Mr. Jones
// Mr. Fred Jones
// Mr. Fred Howard Jones Jr.
// The Honorable George Bush
// Ms. Marlit Hayslett-Keck

(!{Token.category == DT} | //Determiner as in "the".
!{Token.category == PRP} | // Personal Pronoun
!{Token.category == RB} | // Adverb)

(!TITLE)+  //1 or more "title" words: Honorable, Mr., etc.
((FIRSTNAME | FIRSTNAMEAMBIG) //zero or more names.
)*
(PREFIX)* //LAST NAME PREFIX, i.e. De from De Gallo.
(UPPER) //uppercase word or hyphenated word (Last Name)
(UPPER)? //optional uppercase word.
(PERSONENDING)? // Jr., Sr., III

:person ➔ //goes to person entity.

The default JAPE rules recognize the following types of addresses.

Address Types:
Street – 101 Broadway
Phone – (202) 212-5555
URL – http://www.yourdomain.com
Email – username@domain.org
IP – 250.0.0.1

However, the Default JAPE rules failed to recognize many US Postal Addresses. Twelve
JAPE rules were added to the postal address recognition phase. They perform better than
the JAPE rules provided with Default ANNIE at recognizing street addresses and Post
Office Box numbers. They recognize military addresses, rural routes and highway
contract addresses. For instance, the following JAPE rule was added to check for the last
line format: city, state zip.

// LAST LINE Rules
// us mod
Rule: LastLineAddress
// atlanta, ga 33333
// atlanta, ga 33333-1234

( !{Lookup.minorType == city}
//!
!{Token.orth == upperInitial} |
!{Token.orth == allCaps}
)+
//(Token.string == ",")?

8
The information extractor needs to recognize the names of organizations that do not appear in the wordlists, for example, Allen Law Firm, Mayo Clinic and Underwood Trust. The following rule was added to help achieve this. A proper noun singular followed by a term from the organization_noun.lst gazetteer is an organization name.

Rule: NNPOrg
Priority: 25
// Tech University
// Allen Law Firm

{(Token.category == NNP)+ // Proper Noun Singular, e.g., Allen
{Lookup.majorType == organization_nouns} //
}
:orgName -->
:orgName.TempOrganization = {kind = "org", rule = "NNPOrg"}

Additional rules are needed to recognized organizations that are not in the gazetteers. For instance, the following was added to recognize the names of departments, e.g., Department of Defense, Department of Mathematics and Computer Science.

Rule: OrgDept
Priority: 25
// Department of Pure Mathematics and Physics

{Token.string == "Department"}
{Token.string == "of"}

{Token.orth == upperInitial}+ // Pure Mathematics

{Token.string == "and"}
{Token.orth == upperInitial}+ // One or more initially capitalized words, e.g., Physics
?
:orgName -->
:orgName.TempOrganization = {kind = "department", rule = "OrgDept"}

While expressions such as 4.7% and 5.2 Percent were correctly annotated, ANNIE missed expressions such as "0.3 percentage point." This is an easy fix. One simple adds "percentage" to the PERCENT Macro.

Macro: PERCENT
{{Token.string == "]"} | {Token.string == "percent"}| {Token.string == "percentage"}) | (Token.string == "per")}
A position, post, office, official position or function is a job in an organization. Following are a few of the job titles appearing in the Bush PC records.

President
Vice President
Ambassador to the U.N.
Governor
Secretary of Commerce
Secretary of State
Secretary
Deputy Secretary
Senator
Representative
Chief Lobbyist
Director
Assistant Director
Chief of Staff
Staff Assistant for Policy
Deputy Assistant to the President
Assistant to the Vice President for Domestic Policy

A title, title of respect, or form of address is an appellation signifying status or function, e.g., Mr. or General or Professor. An appellation is an identifying word (or words) by which someone or something is called and classified or distinguished from others.

Some job titles can be used as titles, that is, as a form of address. For example

President Bush
Vice President Quayle
Governor Sununu
Senator Kennedy

Hence, some terms for job titles are ambiguous as to whether they are job titles or titles. A job title is usually considered a title when it precedes a person's full or last name.

In addition to the additions to the wordlists for job titles it was necessary to add JAPE rules to recognize and markup job titles such as the following

<jobtitle>Deputy Assistant</jobtitle> to the <jobtitle>President</jobtitle> for <organization>Policy Planning</organization>

This is accomplished by adding the following JAPE Rule:

Rule: Jobtitle1
The interpretation of this rule is that a job title or a title followed by "to" and "the" followed by a job title or a title followed by "for" and then an organization name should be marked up as a job title.

### 2.1.2 Experimental Evaluation of the Enhancements

The results of the Default ANNIE gazetteers and JAPE rules applied to the original 50 record corpus are shown in the table below.

#### Previous Default ANNIE:

<table>
<thead>
<tr>
<th>Annotation Type</th>
<th>Correct</th>
<th>Partially Correct</th>
<th>Missing</th>
<th>Spurious</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person</td>
<td>242</td>
<td>94</td>
<td>75</td>
<td>107</td>
<td>0.65237020316</td>
<td>0.70316301703</td>
<td>0.67681498829</td>
</tr>
<tr>
<td>Location</td>
<td>249</td>
<td>14</td>
<td>111</td>
<td>9</td>
<td>0.94117647058</td>
<td>0.68449197860</td>
<td>0.79256965944</td>
</tr>
<tr>
<td>Organization</td>
<td>146</td>
<td>105</td>
<td>59</td>
<td>65</td>
<td>0.62816455696</td>
<td>0.64032258064</td>
<td>0.63418530351</td>
</tr>
<tr>
<td>Date</td>
<td>162</td>
<td>3</td>
<td>30</td>
<td>6</td>
<td>0.95614035087</td>
<td>0.83846153846</td>
<td>0.89344262295</td>
</tr>
<tr>
<td>Money</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>Percent</td>
<td>9</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>1.0</td>
<td>0.69230769230</td>
<td>0.81818181818</td>
</tr>
</tbody>
</table>

Overall Average Precision: 78.6 %
Overall Average Recall: 71.3 %

A test was conducted using these modifications to the ANNIE wordlists and JAPE rules. The results are shown in the following table.

#### Modified Information Extractor:

<table>
<thead>
<tr>
<th>Annotation Type</th>
<th>Correct</th>
<th>Partially Correct</th>
<th>Missing</th>
<th>Spurious</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person</td>
<td>311</td>
<td>52</td>
<td>23</td>
<td>40</td>
<td>0.8362</td>
<td>0.8731</td>
<td>0.8542</td>
</tr>
<tr>
<td>Location</td>
<td>305</td>
<td>16</td>
<td>48</td>
<td>16</td>
<td>0.9288</td>
<td>0.8482</td>
<td>0.8867</td>
</tr>
<tr>
<td>Organization</td>
<td>292</td>
<td>29</td>
<td>10</td>
<td>35</td>
<td>0.861</td>
<td>0.926</td>
<td>0.8923</td>
</tr>
</tbody>
</table>
Improvements in overall average precision were from .786 to .904 and in overall average recall from .713 to .907.

Fifty additional Presidential records from the Bush Presidential Library were scanned and OCRed. First, the corpus was manually marked up with the correct named entities, creating the Key files. ANNIE in its default configuration (no improvements) was applied to the corpus. The results are as shown below [Isbell et al 2006].

<table>
<thead>
<tr>
<th>Annotation Type</th>
<th>Correct</th>
<th>Partially Correct</th>
<th>Missing</th>
<th>Spurious</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person</td>
<td>294</td>
<td>75</td>
<td>94</td>
<td>286</td>
<td>0.5061</td>
<td>0.716</td>
<td>0.593</td>
</tr>
<tr>
<td>Location</td>
<td>216</td>
<td>12</td>
<td>149</td>
<td>14</td>
<td>0.9174</td>
<td>0.5889</td>
<td>0.7173</td>
</tr>
<tr>
<td>Organization</td>
<td>218</td>
<td>64</td>
<td>167</td>
<td>91</td>
<td>0.6702</td>
<td>0.5568</td>
<td>0.6083</td>
</tr>
<tr>
<td>Date</td>
<td>222</td>
<td>6</td>
<td>53</td>
<td>21</td>
<td>0.9036</td>
<td>0.8007</td>
<td>0.8491</td>
</tr>
<tr>
<td>Money</td>
<td>19</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0.975</td>
<td>0.8864</td>
<td>0.9286</td>
</tr>
<tr>
<td>Percent</td>
<td>22</td>
<td>1</td>
<td>4</td>
<td>0</td>
<td>0.9783</td>
<td>0.8333</td>
<td>0.9</td>
</tr>
</tbody>
</table>

Overall average precision: 0.7868893138665035 Overall average recall: 0.6920527077032016

The performance of ANNIE on this new corpus is very similar to its performance on the first corpus of 50 records.

<table>
<thead>
<tr>
<th></th>
<th>First Corpus</th>
<th>Second Corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Recall</td>
<td>0.713</td>
<td>0.692</td>
</tr>
<tr>
<td>Average Precision</td>
<td>0.786</td>
<td>0.787</td>
</tr>
</tbody>
</table>

The new wordlists and JAPE rules were applied to the second corpus. The results are shown below.

<table>
<thead>
<tr>
<th>Annotation Type</th>
<th>Correct</th>
<th>Partially Correct</th>
<th>Missing</th>
<th>Spurious</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person</td>
<td>333</td>
<td>66</td>
<td>64</td>
<td>145</td>
<td>0.6728</td>
<td>0.7905</td>
<td>0.7269</td>
</tr>
<tr>
<td>Location</td>
<td>249</td>
<td>13</td>
<td>115</td>
<td>6</td>
<td>0.9534</td>
<td>0.6777</td>
<td>0.7922</td>
</tr>
<tr>
<td>Organization</td>
<td>279</td>
<td>46</td>
<td>124</td>
<td>67</td>
<td>0.7704</td>
<td>0.6726</td>
<td>0.7182</td>
</tr>
<tr>
<td>Date</td>
<td>268</td>
<td>1</td>
<td>12</td>
<td>17</td>
<td>0.9388</td>
<td>0.9555</td>
<td>0.9471</td>
</tr>
<tr>
<td>Money</td>
<td>19</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0.975</td>
<td>0.8864</td>
<td>0.9286</td>
</tr>
<tr>
<td>Percent</td>
<td>24</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0.9107</td>
<td>0.9444</td>
<td>0.9273</td>
</tr>
</tbody>
</table>

Overall average precision: 0.858151383920109 Overall average recall: 0.8066458715367074
Average recall is .1 better and average precision is .07 better than the results for the default wordlists and JAPE rules. However, the performance of the modified ANNIE on the second corpus dropped from its performance on the first corpus.

<table>
<thead>
<tr>
<th></th>
<th>First corpus</th>
<th>Second Corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Recall</td>
<td>.907</td>
<td>0.807</td>
</tr>
<tr>
<td>Average Precision</td>
<td>.904</td>
<td>0.858</td>
</tr>
</tbody>
</table>

This is largely because new record types occurred in the second corpus that did not occur in the first corpus. This shows that the additions to the gazetteers and JAPE rules to improve the performance on the first corpus of Bush e-records are general enough to apply to other Bush e-records.

There are still a number of improvements that need to be made before this technology can be practically applied to the Bush PC records. Additional entity types such as telephone numbers and social security numbers need to be recognized in order to support recognition of access restrictions on release of personal information.

Locations, such as the United States, need to be distinguished from Global Political Entities, such as the United States Government or the Russian Federation. Sources of such knowledge include the CIA World Factbooks. Phrases, such as "The White House," need to be disambiguated as to whether they are a facility (the building) or an organization the President and the White House Staff.

Relative temporal expressions, such as "Day after tomorrow" need to be recognized in order to sequence events mentioned in the text. This is needed in order to understand the context of specific actions and events mentioned in series of records.

For determining the subject of records and summarizing series of records, semantic categories such as Congressional Bills (e.g., House Resolutions) and Statues or Acts (e.g., Americans with Disabilities Act) need to be automatically identified in text and annotated.

To cost effectively extend this technology to larger and more diverse corpora, we will investigate grammatical induction to create rules characterizing mailing addresses. We will also investigate the machine learning of Hidden Markov Models (HMMs) to characterize person's names, organization names, or geographic place names.

The layout segmentation problem is the improper segmentation of text resulting in an extensive number of improperly conjoined entities across segment boundaries in documents that have headers or footers, double spacing, multiple columns, tables and lists. This problem has been solved in part by modifying the sentence splitter rules. We will continue to refine a text segmenter that distinguishes paragraphs of sentences from blocks of text that contain addresses, tables, and lists.
2.2 Record Type Learning and Recognition

Information extraction and grammatical induction technology are being applied to the problem of learning the documentary form of a variety of Presidential electronic records [Underwood and Harris 2006]. Given a sample of records of a particular type, such as correspondence, information extraction technology is used to identify and markup semantic categories, such as person's names, organization names, location names, job titles, dates, and postal addresses. The content is then removed from the annotated records leaving the intellectual elements of documentary form. From the intellectual elements of the sample records, a stochastic context-free grammar is automatically induced that defines the documentary form of that particular record type. Grammars learned for a variety of record types can then be used with a parser to recognize documentary forms of records of unknown record type.

The significance of this research is that record types have a role in archival description and review. Archival descriptions include the names of the types of records that occur in a record series. In addition, knowing a record's type aids in understanding the action communicated by a document. Knowing record type can aid in discriminating personal records from Presidential records. It can also aid in determining access restrictions. The ability to recognize record type can also contribute to searching for and retrieving records based on the intellectual elements of their documentary form.

Figure 4 shows our approach to record type learning.

Figure 4. Our Approach to Record Type Learning

A set of documents of a particular type is selected and converted to text or html format. The document shown below was one such document in a sample of White House Correspondence.
The information extractor is run against this sample to produce documents that describe the named entities in the document using XML markup.

```
<paragraph><Date>Feb 7, 1989</Date></paragraph>

<paragraph><Address><Person><Title>Miss</Title> Ashley Walker Bush</Person>
c/o <Title>Mr.</Title> and <Title>Mrs.</Title> Neil M. Bush</Address></paragraph>

<paragraph>Dear <Person> Ashley</Person>, </paragraph>

<paragraph>On this the first day of your life, your old grandfather sends you his love. Today was the day after my Savings and Loan proposal, the day of my visit to Capitol Hill to see a lot of Congress members, 2 days before my speech to the nation—but on this day of your birth, I'm thinking of you. You have 2 great parents, an older sister who will teach you and brother who will protect you. You have grandparents who love you a lot already. Welcome, welcome to this big loving family—I am a happy Gampy because you're here. </paragraph>

<paragraph>Devotedly,  
George Bush</paragraph>
```

Figure 5. XML Markup of Named Entities in Sample Correspondence.

A content removal algorithm is used to remove content from the marked up documents leaving a list of the intellectual elements of documentary form for each document in the sample. The figure below shows the intellectual elements of the documentary form of
sixteen pieces of correspondence. The third list in the figure corresponds to the intellectual elements the document in Figs 4 and 5.

Figure 6. Intellectual Forms of Sample White House Correspondence.

To induce a stochastic context-free grammar for the intellectual form of White House correspondence, Stolke's algorithm for grammatical induction is applied to this sample [Stolke and Omohundro 1994]. This algorithm begins with a grammar constructed by

1. Creating a nonterminal symbol producing each terminal (e.g., DATE, Dear, PERSON) in the lists.
2. Creating rules with the symbol S on the left-hand side the string of nonterminals replacing each terminal in a list on the right hand side.

Then two operators are applied to produce alternative candidate grammars:

1. Chunking -- substituting a nonterminal for a sequence of two or more nonterminals that occur on the right-hand-side of two or more production.
2. Merging -- the recognition that two nonterminals generate the same string and merging the two nonterminals into one nonterminal that generates the string.
The alternative grammars are evaluated to determine the best alternative so far. Then the two operators are applied to the selected grammar and the process is continued until there is no improvement in applying the operators. The result for the sample intellectual forms for correspondence shown in Figure 6 is the following grammar.

```
| NT219     | (NT2109 NT2109) | 1.000 | 43.0000 |
| NT2219    | (NT2219 NT2219 NT2219) | 1.000 | 12.0000 |
| NT2392    | (NT2392 NT2392) | 0.305 | 7.000000 |
| (NDATET-7896-2081 NT2084 NT2084) | 0.675 | 16.0000 |
| NT2709    | (NT2109 NT2084 NT2084) | 0.349 | 13.0000 |
| (NT2109 NT2709) | 0.120 | 5.000000 |
| (NT1211 NT2084) | 9.14e-2; 4.000000 |
| (NT2239 NT2709) | 0.205 | 8.000000 |
| (NT2709 NT2087) | 0.234 | 9.000000 |

S   (NT2682 NT1211 NT2082) | 1.518e-2; 1.000000 |
| (NT2682 NT2239 NT2085 NT2084 NT2087 NT2082) | 0.171 | 4.000000 |
| (NT2682 NT2709 NT2082) | 0.434 | 9.000000 |
| (NDATET-7896-2081 NT2082 NT2083 NT2084 NT2103 NT2709) | 0.362 | 8.000000 |

NT2083  address  | 1.000 | 22.0000 |
NT2083  dear    | 1.000 | 22.0000 |
NT2084  person  | 1.000 | 43.0000 |
NT2086  sincerely | 1.000 | 17.0000 |
NT2087  jobtitle | 1.000 | 13.0000 |
NT2109  best    | 3.899e-2; 5.000000 |
| paragraph | 0.683 | 97.000000 |
| regards  | 3.899e-2; 5.000000 |
| with    | 3.899e-2; 5.000000 |
NT2121  devotedly | 0.375 | 2.000000 |
| warmly  | 0.025 | 0.000000 |
| NTDATE-7896-2081 | date | 1.000 | 22.0000 |
```

Figure 7. Induced Stochastic CFG for Correspondence Record Type.

Our approach to record type recognition is illustrated in Figure 8.

Figure 8. Approach to Identifying Record Type.

An individual text document of unknown record type is converted to text or html file format. The information extractor identifies named entities and uses XML markup in to indicate these entities. The content removal program produces a list of intellectual
elements of the document. Finally a context-free grammar parser uses the induced grammars to parse the list of intellectual elements. The parse tree indicates the record type.

Shown below is a list of intellectual elements were derived from a document using the above method.

(date address dear person paragraph paragraph paragraph paragraph paragraph paragraph paragraph sincerely person jobtitle)

Using the grammars induced for correspondence and memoranda, the parser for context-free grammars produces the following parse, which indicates that the original document is of record type correspondence.

```
0.0023696290970279974
{s (ntdate-7896-2081 . date) (nt2082 . address) (nt2083 . dear)
 (nt2084 . person) (nt2109 . paragraph)
 (nt2709
 (nt2239 (nt2219 (nt2109 . paragraph) (nt2109 . paragraph)))
 (nt2219 (nt2109 . paragraph) (nt2109 . paragraph))
 (nt2219 (nt2109 . paragraph) (nt2109 . paragraph)))
 (nt2709 (nt2109 . paragraph) (nt2086 . sincerely) (nt2084 .
 person)))
 (nt2087 . jobtitle)))
```

Figure 9. Parse of Intellectual Elements Indicating a Correspondence Record type.

The Record type learning method described in this section has been applied to small samples of White House Correspondence and White House Memoranda to infer grammars describing the documentary forms of these record types. The grammars have been automatically combined and used with a parser to automatically identify and discriminate the record types from the combined samples and of other e-records not included in the samples [Underwood and Harris 2006].

### 2.3 Creation of Folder Titles and Scope and Content Notes

The directory names of personal computer files are often abbreviated and sometimes uninformative. A method is being developed for extending these directory names. It is based on using the record type identifier to recognize the record types in a directory. A method is also being developed for creating "Scope and Content Notes" for record series [Underwood 2005]. It is based on formulating a list of questions that can be asked about the record series, and that can be answered by using background knowledge of the context of the series, natural language understanding technology to understand the content of the series, and question answering technology to obtain the information that is needed in "Scope and Content Notes."
The significance of this task is that a number of metadata extraction and description tools have been and are being developed that automatically provide some of the data elements needed in archival description or provide archivists with the information that they need to provide these descriptive elements. Furthermore, in the cases that there are not enough archivists to immediately describe the large volume of accessioned record series, these tools can provide a surrogate description until archivists have time to construct the description.

### 2.3.1 Method for File Unit Description

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 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[RESPONDENCE]. If it was titled WORK and contained correspondence, its title might be extended as follow: WORK[CORRESPONDENCE].

For directories that are characterized by the kind of documents in the directory, the Record Type Identifier is used to identify the types of documents in the directory, and a set of rules is 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 record types was organized into directories (or folders). This corpus was 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 record 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.
2.3.2 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 information extraction and record type recognition, 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 content notes [NARA] and the scope and contents notes and folder titles that archivists have created for paper records in the holdings of the Presidential Libraries were studied. The figure below 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

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.

NARA's *Lifecycle Data Requirements Guide* (LCDRG) [NARA] mandates the creation of Function and Use and Scope and Content data elements for record series. Function and Use is mandatory for all new descriptions of organizational records, but not for personal papers or previously describing archival materials. Legacy archival descriptions of record series usually included function and use in the Scope and Content note.

"Function and Use is the description of why the archival materials were created. This element differs from Scope and Content Note, which describes the significant information contained within the records. Function and Use is *about the activities* that resulted in the creation of the archival materials. Scope and Content Note is *what is in* the archival materials."
NARA uses a *Program Area Thesaurus* as an authority list for program areas or functions of organizations. NARA created the Program Area Thesaurus based upon the Federal Register Thesaurus of Indexing Terms, and the Center for Army Lessons Learned (CALL) Thesaurus.

A Scope and Content Note is "The description of the breadth and depth of the record group, collection, series, file unit or item."

In the guidance with regard to creating Scope and Content Notes, the LCDRG states:

"Write a note that provides answers to basic questions that users might ask about the record group, collection, series, file unit, or item described. Explain any significant or heavily-represented topics, people, organizations, geographic places, or languages represented in the record group, collection, series, file unit, or item, as well as the types of materials present."

During accession using the PERPOS System, each file system is placed in a container (Jar file) and is associated with a Library collection, an office and is given a series title. The container is stored in Archival holdings. This contextual information is available for use in creating the series Function and Use and Scope and Content Note.

The publication *A Guide to the White House Office* prepared by the archival staff at the Bush Presidential Library lists the Offices, Officers, and many of the responsibilities and activities of the White House Offices. As part of the research activities to determine the feasibility of identifying probable access restrictions on release of records to the public (See section 4 of this report), a capability is being developed to determine the communication act of a record [Underwood and Harris 2005].

The form of description for Function is:

These records were created by *<Person name>* in performing the *<list of activities>* activities of the *<office name>*.

The activities are those of the office or those represented by the communication actions of the records.

The approach to creating a Scope and Content Note for a record series is based on answering the questions such as the following.

What are the record types in the series?
Who created the records in the series?
Who are the records about, that is, to what person or organization do they relate?
Who contributed to the production or authorship of the records and what is their relationship to the activities documented?
What are the records generally about?
What are the main topics or subjects mentioned?
What unusual or historically significant topics are mentioned?
Where does the action or events take place?
What specific geographic places are mentioned?

The template for presenting the answers to the questions is (in part):

"This series consists of <list of record types from most numerous to least numerous or in order of arrangement>. "The material relates to<list of subjects or topics>; <list of organizations>"

If the record types are database files, then the form of description is:

This series consists of a database tables: (1) <name of table> which includes <attributes of the table> (2) …(n) <name of table> which includes <attributes of the table>

When the performance of the information extraction system and record type identifier are substantially improved, experiments will be conducted at Archives II and the Bush Presidential Library to evaluate the performance of the Folder Titler and the Series Summarizer.

3. Technologies for Retrieval of Presidential E-Records

Archivists need to respond to Freedom of Information Act (FOIA) requests for electronic records that have not yet been systematically processed. In responding to FOIA requests and conducting e-discovery, NARA currently uses commercial-off-the-shelf document retrieval systems to search its increasingly large collections of e-records. Results of the Information Retrieval Track of the Text Retrieval Evaluation Conferences (TREC) indicate that current Information Search and Retrieval technologies do not scale well. The larger the collection of documents, the lower the degree of recall and precision of retrieved e-records relevant to the queries. This means that, first, NARA archivists must review many, many more documents than are actually relevant to FOIA requests and that during e-discovery, its attorneys must review many more documents than are actually relevant to a litigant's case. Second, researchers submitting FOIA requests experience serious delays in receiving records relevant to their requests, the likelihood of having to read many more records than were relevant to their request, and the likelihood of not having all that were actually relevant to their request. Third, it takes archivists much longer to process FOIA requests because there are many e-records to be reviewed that are not actually relevant to the request. Fourth, in e-discovery, attorneys who agree to using a limited set of search terms with a text retrieval system receive case-relevant e-records, but have to review many e-records, that while relevant to the query, are not relevant to the case.
In previous research, a document retrieval experiment was conducted with systems representing three document retrieval technologies. WebGlimpse was used as an example of a Boolean retrieval technology without relevance ranking; Oracle Text with word queries as an example of Boolean search technology with relevance ranking; and Sun's NOVA precision content passage retrieval system as an example of natural language-based search technology. The NOVA precision content retrieval system is particularly interesting because in addition to retrieving relevant passages/documents, it provides a conceptual index of the entire collection that allows the searcher to navigate the conceptual space around the conceptual areas related to the documents retrieved, thus supporting interactive search and retrieval, and potentially increasing precision and recall. Queries used in the experiments were derived from actual FOIA requests submitted to the Bush Presidential Library. The experiments were conducted using the Bush Public Papers as a sample collection [Underwood and Underwood 2002].

Recall is a measure of the ability of a system to present all relevant documents. Precision is a measure of the ability of a system to present only relevant documents. For response to FOIA requests, a document retrieval system must have high recall. To reduce the number of documents that have to be reviewed the retrieval system should have high precision, without sacrificing recall. Average precision is a good measure of the utility of a document retrieval system. Average precision combines precision, relevance ranking and overall recall. Average precision is the sum of the precision at each relevant hit in the hit list divided by the total number of relevant documents in the collection. In the experiment, the average precision of Oracle Text with word queries was .7620, NOVA .6165, and WebGlimpse .5436.

The results of the experiments were analyzed to explain the difference in performance for different topics. Oracle Text with word queries had the best performance with regard to average precision, and especially for broad general queries with many alternatives. NOVA's Precision Content Retrieval, while not performing as well overall, out performed Oracle Text on topics where the request was for specific information, and the query involved just a few words. NOVA's performance would have been better if the user interface allowed a larger number of passages to be retrieved and relevancy feedback had been used to refine the NOVA queries. WebGlimpse, using a Boolean search technology without relevance ranking, did not perform as well as the other search technologies.

The average precision of the systems evaluated was significantly greater than the average precision of the document retrieval systems evaluated in the Ad Hoc Query Track of the Eighth Text Retrieval Conference (TREC-8). Precision is dependent on the size of the document set searched and is typically lower for larger document sets. The experiment was conducted on a document set of about 5000 documents. There were more than 500,000 documents in the TREC-8 document set. Hence, the results of our experiment are not conclusive.

We want to conduct a similar experiment using a larger corpus, namely the Bush PC e-record collection, which contains approximately 150,000 e-records in about 50 different file formats. This collection is 30 times the size of the corpus used in the information
retrieval experiment conducted in the previous phase of research, which also contained files of just one format.

### 3.1 Natural Language-Based Record Retrieval

In responding to FOIA requests and conducting e-discovery, NARA currently uses commercial-off-the-shelf document retrieval systems to search its very large collections of e-records. For example, in the PERPOS prototype, we use Oracle Word Search to support Presidential Library archivists in retrieving e-records responsive to FOIA requests. Results of the Information Retrieval Track of the Text Retrieval Evaluation Conferences (TREC) indicate that current Information Search and Retrieval technologies do not scale well. The larger the collection of documents, the lower degree of recall and precision of retrieved e-records relevant to the queries. This means that, first, NARA archivists must review many, many more documents than are actually relevant to FOIA requests and that during e-discovery, its attorneys must review many more documents than are actually relevant to the case. Second, researchers submitting FOIA requests experience serious delays in receiving records relevant to their requests, the likelihood of having to read many more records than were relevant to their request, and the likelihood of not having all that were actually relevant to their request. Third, it takes archivists much longer to process a FOIA requests because there are many e-records to be reviewed than are actually relevant to the request. Fourth, in e-discovery, attorneys who agree to using a limited set of search terms with a text retrieval system getting fewer case-relevant e-records, but having to review many e-records, that while relevant to the query, are not relevant to the case.

The Natural Language Boolean Query Interface to NOVA that we have designed includes AND, OR and NOT operators as well as parentheses for grouping expressions. The difference from other Boolean query languages lies in the semantics for interpretation of the query. The following paragraphs discuss some of the issues addressed [Underwood and Iwanska 2006a].

The natural language queries are Boolean expressions of noun phrases, and NOVA's conceptual index of a collection is organized in terms of noun phrases that subsume other noun phrases [Woods 1997]. The results set for a query on a single noun phrase is often not just the passages (documents) containing the noun phrase but passages (documents) containing all the noun phrases that are be subsumed by a noun phrase in the query.

Consider the following natural language text retrieval query: “I’m interested in articles on NLP but not semantics and parsing since 1995." It is possible to interpret a query like this in many different ways. It has been found that the higher the number of disjunctions, conjunctions and prepositions used in a statement with a negation, the higher the ambiguity of the statement. Negation is difficult, due to the ambiguities as to which components are negated and which aren’t.

Suppose that one wanted to retrieve records related to "tax deductions but not charitable tax deductions." There is currently no way to express this in the NOVA query language,
though it could be accomplished by finding all records related to "tax deductions" and those related to "charitable tax deductions" and removing the later from the former, unless the former also included "tax deductions" that were not "charitable tax deductions." There is a need for a Boolean NOT operator in NOVA queries, so the query could be expressed "tax deductions AND NOT charitable tax deductions."

The negation of the noun phrase "charitable tax deductions," describes the set that is the set complement of "charitable tax deduction" and includes three sets: (1) documents that are not about "deductions," documents that are about deductions that are not tax deductions, (3) documents about tax deductions that are not charitable. The later is what is meant in the example from the preceding paragraph, and might be represented as "NOT (charitable) tax deductions." The Boolean NOT operator will need to be constrained so that using it does not result in an ambiguous expression.

A query interface is being developed for NOVA in which Boolean natural language queries can be expressed. This does require any modifications to NOVA itself. The natural language Boolean query will not have the same meaning as the Boolean query on terms used in WebGlimpse or Oracle Word Search, because NOVA is using a conceptual map of the concepts in the documents that are related by lexical subsumption.

To address the problem of negation in natural language, we are developing a capability to parse a negated noun phrase and present to the user the possible interpretations to determine what is actually intended. From these, the user selects the intended meaning. The resulting negated set is not just the documents containing the negated noun phrase, but all documents containing the negated noun phases subsumed by that noun phrase.

The results set is documents, not just passages in documents. The penalty scoring method used by NOVA applies to each of the phrases in the Boolean query. The ranked results of each subquery need to be merged. The negated expressions should now appear in the document, not just in passages.

**Syntax**

Our definition of phrases that can be part of a Boolean NL query is dictated by the noun and verb phrases that NOVA currently recognizes and indexes. A noun phrase consists of a noun with any associated modifiers including adjectives, adjective phrases, and nouns in the possessive case. A verb phrase consists of a verb and a noun phrase. We define a PHRASE of a Boolean query as follows.

```
PHRASE → NP2
PHRASE → VP
NP → DET NP2
NP → NP2
NP → NP PP
NP2 → NOMINAL
NP2 → ADJP NOMINAL
NOMINAL → N
```
Only one prepositional phrase can occur in a noun phrase. A noun phrase might also include adjective clauses (relative clauses), for example, "The coat which is red." Adjective clauses are not identified and indexed by NOVA. Nor are conjunctions of noun phrases and prepositional phrases identified. Predicate adjectives that modify a noun subject (The coal mines are dark and dank) are also not identified. Nor are adverb modifiers of adjectives identified. Adjective phrases that modify nouns and that are formed with participles are also not included in phrases (We picked up the records broken in the scuffle). Hence, we do not include any of the above in the definition of phrases that can be part of a Boolean NL query.

The syntax of a Boolean NL query is defined as follows.

\[
Q \rightarrow \text{PHRASE} \\
Q \rightarrow (Q \text{ AND } Q) \\
Q \rightarrow (Q \text{ OR } Q) \\
Q \rightarrow \text{NOT } Q
\]

Semantics

NOVA returns passages from documents that include the phrase in a query, phrases subsumed by phrase, and other passages containing terms in the phrase ranked in relaxation order. The following paragraphs describe the semantics of the Boolean queries.

**PHRASE**

The NL Boolean query for a phrase returns the document ids of the documents containing passages that contain exactly the phrase, those phrases subsumed by the phrase or whose relaxation ranking is greater than .90. Ranking is by penalty scoring of the highest valued phrase(s) occurring in a document.

\[(Q_1 \text{ AND } Q_2)\]

This query returns the document ids of the documents containing both query $Q_1$ and $Q_2$. Ranking is by the highest scores for $Q_1$ and $Q_2$.

\[(Q_1 \text{ OR } Q_2)\]

This query returns the document ids of the documents containing phrases $Q_1$ or $Q_2$. Ranking of documents is by the highest penalty score for $Q_1$ or $Q_2$. 
NOT Q

This query returns the document ids of documents that do not include the documents that are the result of query Q.

### 3.2 Acquisition of Knowledge to Support Record Retrieval

Background knowledge of the persons, policies and events of a Presidential Administration guides Reference Archivists search for relevant documents and determination of their relevancy to a request. Similarly, such background knowledge is essential to interpreting e-records during review to determine whether an e-record, or passages of an e-record are subject to Presidential Record Act restrictions or FOIA exemptions.

One of the basic kinds of knowledge needed to understand documents is the knowledge of whether a phrase is more general or specific than another. People acquire this kind of knowledge from dictionaries, experience, education, or reasoning using what one already knows. Automatic interpretation of text requires similar knowledge, and it can be acquired by sources created by lexicographers or knowledge engineers, machine learning techniques, or by inference using knowledge already acquired.

One source of such knowledge is WordNet [Fellbaum 1998]. WordNet 2.1 records basic relationships of over 155,000 words with over 207,00 meanings. Information in WordNet is organized around logical groupings called synsets. Each synset consists of a list of synonymous words or collocations. A collocation in WordNet is a string of two or more words, connected by spaces or hyphens. Nouns and verbs are organized into hierarchies based on the hypernym/hyponym relation between synsets. A hypernym is a word that is more general than a given word. A hyponym is a word that is more specific than a given word. Formally, the noun X is a hyponym of the noun Y, if X is a (kind of) Y. For example, "milk" is a hyponym of "food" because milk is a kind of food.

Natural language lexical constructs have been investigated as a means for acquiring taxonomic knowledge [Hearst 1992]. Figure 10 shows some lexical constructs for identifying hyponyms with examples identified in the Bush Public Papers.

<table>
<thead>
<tr>
<th>Construct</th>
<th>Example from Bush Public Papers</th>
</tr>
</thead>
<tbody>
<tr>
<td>NP₀ such as {NP₁} * (and</td>
<td>or) NP</td>
</tr>
</tbody>
</table>
domestic problem) hyponym(need to revitalize American industry, domestic problem)

NP \{, NP\}^{*} \{;\} or other NP … antidiscrimination or other subjects of enormous concern => hyponym (antidiscrimination, subject of enormous concern)

NP \{, NP\}^{*} \{;\} and other NP … Federal judges and other high officials in the Government. => hyponym (Federal judge, high official in the Government)

NP \{,\} including \{NP ,\}^{*} \{or | and\} NP … beautiful natural resources, including our oceans, estuaries, our beaches, and our shoreline => hyponym (oceans, beautiful natural resources) hyponym (estuaries, beautiful natural resources) hyponym (beaches, beautiful natural resources) hyponym (shoreline, beautiful natural resources)

NP is \{a | an\}NP The shamrock is a symbol of unity => hyponym(shamrock, symbol of unity)

NP, which are NP … housing and urban policy, which are among the most important and challenging issues in America today => hyponym(housing, challenging issue) hyponym(urban policy, challenging issue).

NP like NP basic needs like shelter => hyponym(shelter, basic needs)

<table>
<thead>
<tr>
<th>Natural language construct</th>
<th>Construct occurs in n documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>NP_0 such as {NP,}^{*} (and</td>
<td>or)} NP</td>
</tr>
<tr>
<td>NP {, NP}^{*} {;} or other NP</td>
<td>88</td>
</tr>
<tr>
<td>NP {, NP}^{*} {;} and other NP</td>
<td>783</td>
</tr>
<tr>
<td>NP {,} including {NP ,}^{*} {or</td>
<td>and} NP</td>
</tr>
<tr>
<td>NP is {a</td>
<td>an}NP</td>
</tr>
<tr>
<td>NP which {are</td>
<td>is} NP</td>
</tr>
<tr>
<td>NP like NP</td>
<td>1595</td>
</tr>
</tbody>
</table>

Automated methods were developed to acquire such knowledge from text and were applied to the 5, 173 Bush Public Papers [Underwood and Iwanska 2006 b].

For instance, the method to recognize the "such as" construct is
The *such as* algorithm applied to the Bush Public Papers had the following results.

<table>
<thead>
<tr>
<th>Natural language construct</th>
<th>No of documents in which construct occurs</th>
<th>No of occurrences of construct</th>
<th>Number of useful occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>X such as Y, ..., and Z</td>
<td>369</td>
<td>473</td>
<td>197</td>
</tr>
</tbody>
</table>

The following table shows some of the useful concepts acquired from the Bush Public Papers using the *such as* algorithm.

<table>
<thead>
<tr>
<th>Concept</th>
<th>Definition or Subtypes</th>
</tr>
</thead>
<tbody>
<tr>
<td>tax on consumerism</td>
<td>value-added tax</td>
</tr>
<tr>
<td>international organizations</td>
<td>World Meteorological Organization, United Nations Environment Programme, International Council of Scientific Unions</td>
</tr>
<tr>
<td>domestic problems of our country</td>
<td>affordable housing, a crumbling public infrastructure, a need to revitalize American industry</td>
</tr>
<tr>
<td>State-administered programs</td>
<td>Medicaid</td>
</tr>
<tr>
<td>vehicles</td>
<td>passenger cars, trucks, buses</td>
</tr>
<tr>
<td>Soviet-bloc weapons</td>
<td>AK - 47's</td>
</tr>
<tr>
<td>nongovernmental entities</td>
<td>the American Red Cross, Pan American Union</td>
</tr>
<tr>
<td>historically black colleges</td>
<td>Alcorn</td>
</tr>
<tr>
<td>major structural barriers to imports</td>
<td>rigidity in the distribution system</td>
</tr>
<tr>
<td>hardliners believed responsible for the massacre</td>
<td>Deng Xiaoping [Chairman of the Central Military Committee] and Premier Li Peng</td>
</tr>
<tr>
<td>transnational problems</td>
<td>need to protect the environment</td>
</tr>
<tr>
<td>murders committed in violation of other Federal statutes</td>
<td>murder-for-hire, murder of a kidnap victim or a Federal prison guard</td>
</tr>
<tr>
<td>professionals</td>
<td>scientists and surgeons</td>
</tr>
<tr>
<td>education programs</td>
<td>Head Start, merit schools, adult literacy</td>
</tr>
<tr>
<td>discretionary programs</td>
<td>Veterans Medical Care; refugee programs; U.N. Peacekeeping activities</td>
</tr>
</tbody>
</table>

There are approximately 5000 documents (29 Mb) in the Bush Public Papers. The Bush PC files consist of about 150,000 documents (1.5 GB). Using the *such as* algorithm, we would expect to acquire many more types and subtypes from the Bush PC files. Experiments will be conducted using these methods to acquire taxonomic knowledge from the Bush PC Files.
The significance of these results is that the knowledge acquired from the Bush Public Papers cannot be found in lexicons such as WordNet. The knowledge is specific to the Administration. This kind of knowledge is needed by Natural language-based document retrieval systems such as NOVA and is essential when applying natural language technology to recognizing restrictions on disclosure of information in Presidential records. Furthermore, it is not necessary for a knowledge engineer to handcraft the taxonomies.

4. Decision Support for FOIA and PRA Review

Review of Presidential electronic records is an intellectually demanding task that requires page-by-page review of Presidential Library accessions. Due to the increasing volume of electronic records from all branches of government, the need to review these records, and the limited number of archivists performing the task, the review task is an archival processing bottleneck. The purpose of this investigation 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 and to use this information to develop an automated review assistant to support archivist's decisions in reviewing Presidential Records.

There are many potential benefits to such a tool, including:

1) reducing the risk of opening a document or passage of a record whose access should be restricted,
2) a tutoring tool during training of review archivists.
3) a tool that novice reviewers could use to check their work.
4) provision of additional evidence in case a reviewer's judgment was uncertain, or point out uncertainties, where the reviewer thought the decision was certain.
5) support estimation of FOIA review workload in terms of the number of restrictions and types of restrictions likely to apply.
6) support review of Federal Records for FOIA exemptions.

Rules were defined in Jess for identifying Personal Record Misfiles and Press releases. Rules were also represented for identifying restrictions on release of information concerning appointments to Federal Office, confidential advice, and Personal Privacy [Underwood and Harris 2005].

Machine Intelligence for understanding of Presidential e-records requires background knowledge. This knowledge needs to be automatically acquired in order to make it economically feasible to employ natural language understanding technology. Knowledge of the names of persons actually appointed or nominated to Federal Office was automatically acquired from the Bush Public Papers and imported into the Access Restriction Checker.

The user interface of the Access Restriction Checker now supports definition and display of rule-based knowledge and import and display of factual knowledge. The integrated
environment now supports not only demonstration of the application of the technology but includes tools for acquiring and defining the knowledge, and for testing and refining the knowledge.

We have a collection of 150 Presidential e-records that represent Personal Record Misfiles, appointments to Federal Office, confidential advice, personal privacy information, and Press Releases. We are testing and refining the Access Restriction Checker using this corpus.

### 4.1 Speech Acts

A speech act is the use of language to perform some act. Speech acts are to be contrasted with other human actions in which something is done as opposed to said, for example, walking, eating, gardening, etc. Figure 11 shows some examples of speech acts. The speech acts include resignation, appointment, nomination, advice, recommendation, requesting, briefing, reporting and many other human actions that are carried out in presidential records.

Among the participants in a speech act, linguists distinguish a *speaker*, who is the utterer of a message and an *addressee* who is any of the immediate intended recipients of the speaker's communication. They also distinguish the propositional content of a message and its illocutionary force. A *proposition* is that part of the meaning of a clause or sentence that is constant, despite changes in such things as the voice or illocutionary force of the clause.

*Illocutionary force* is the combination of the illocutionary point of an utterance, and particular presuppositions and attitudes that must accompany that point. An *illocutionary point* is the basic purpose of a speaker in making an utterance. According to certain analyses, there are five kinds of illocutionary points:

- An **assertive** illocutionary point is an illocutionary point in which the speaker purposes to present that the state of affairs described by the propositional content of the message is actual. "Alberto Gonzales currently holds the office of US Attorney General."
- A **commissive** illocutionary point is the illocutionary point of a speaker committing to bring about the state of affairs described in the propositional content of the message, for example, "I will prepare for you an analysis of the War Powers Act."
- A **directive** illocutionary point is an illocutionary point in which the speaker attempts to get someone to bring about the state of affairs described by the propositional content of the message, for example, to "I want an analysis of the war powers act."
- A **declarative** illocutionary point is an illocutionary point in which, by making an utterance, a speaker brings into existence the state of affairs described in the propositional content of the message, for example, I nominate Alberto Gonzales for the position of US Attorney General."
- An **expressive** illocutionary point is an illocutionary point that communicates an attitude or emotion about the state of affairs described in the propositional content of the message. "I approve of the nomination of Alberto Gonzales to the position of US Attorney General."
o congratulation - the speech act of acknowledging that someone has an occasion for celebration.

o approval - the speech act of expressing a favorable opinion
  o recommendation - the speech act of commending a person as worthy or desirable

o proposal - the speech act of making a proposal
  o presentation - the speech act of presenting a proposal
  o advice - the speech act of advising as to an appropriate course of action.
    ▪ recommendation - the speech act of recommending something as advisable.

o command - the speech act of authoritatively directing or instructing that someone do something.
  o order - the speech act of a superior giving a command that must be obeyed.

o agreement - the speech act of agreeing.
  o subscription - the speech act of agreement expressed by signing your name.
  o ratification, confirmation - the speech act of making something valid by formally ratifying or confirming it.

o request - the speech act of requesting
  o invitation - the speech act of requesting someone participate or be present or take part in something.
  o questioning, inquiring - the speech act of requesting information.
  o interrogation, examination, interrogatory - the speech act of formal systematic questioning.
    ▪ deposition - (law) the speech act of a pretrial interrogation of a witness usually conducted in a lawyer's office.
    ▪ interview - the speech act of questioning a person (or a conversation in which the information is elicited); often conducted by a journalist.
      • job interview, employment interview - the speech act of interviewing a person to determine whether an applicant is suitable for a position of employment.

o reply, response - the speech act of continuing a conversational exchange.
  o answer - the speech act of replying to a question.

o description - the speech act of describing something.

o affirmation, assertion, statement - the speech act of affirming or asserting or stating something.

o complaint - the speech act of expressing a grievance or resentment.

o informing - a speech act that conveys information.

o briefing - the speech act of providing detailed instructions, as for a military operation.

o report - the speech act of informing by report.
  o summarization, - the speech act of preparing a summary, stating briefly and succinctly.

o promise - a speech act by one person committing to another agreeing to do (or not to do) something in the future.

o address, speech - the speech act of delivering a formal spoken communication to an audience.

o resignation - the speech act of giving up a claim or office or possession.

o appointment - the speech act of putting a person into a non-elective position.
  o nomination - the speech act of officially naming a candidate.

Figure 11. Some examples of Speech Acts
We represent the speech (communication) act represented by a record in a template with slots indicating the elements of the communication act. In written documents the speaker is referred to as the author.

(deftemplate communication_act
  (slot documentID)
  (slot act)
  (slot purpose)
  (slot author)
  (slot addressee)
  (slot date)
  (slot content))

Below are shown some rules for filling in communication act templates.

If sentence is imperative,
  and object of sentences is ?z,
then assert (act "request")
  assert (content ?z)

If document is memorandum,
  and "From <person> ?x </person>"
  and "To <person> ?y </person>"
then assert (author ?x), assert (addressee ?y)

Below is shown is an example of a communication act template for a specific document.

(communication_act
  (document Doc-0014)
  (act request)
  (purpose directive)
  (author "The President")
  (addressee "Boyden Gray")
  (date "December 5, 1999")
  (content "analysis of War Powers Resolution")
)

4.2 Access Restriction Checker Procedure

The following sketches the procedure used by the Access Restriction Checker. When requested to check for possible access restrictions in a document [Harris et al 2005]:

1. Convert the record from its original format into an html version of the document.
2. Use factual knowledge and information extraction rules to identify person's names, job titles, organization names, addresses, dates and other relevant information and markup the html version of the record.
3. Identify the record type of the record.
4. Use factual knowledge and template filling rules to fill in templates indicating the kind of communication action the record conveys, the purpose of the action, the author, addressee and its content.
5. Use personal/political record decision rules and access restriction decision rules and subsumption-based reasoning to infer from the filled in template(s) whether there is an access restriction.
6. Display the results to the archivist in the user interface.

4.3 The User Interface

The figure below shows the user interface to the access restriction checker. You must first load a document to annotate. You can do this by selecting File → Open. This will bring up a file dialog that will allow you to select a file; the file must either be an html, xml, or plain text document.

After selecting your file it should load in the Original Document into a text window as illustrated in the figure below.
Then to identify and annotate the document with named entities click on the *Annotate Original Document* button. The Post Processed Document area will load the annotated version of the document as shown in the next figure.
Different entity types are identified by different color text. The colors are defined in the properties file that is loaded but they can be changed in the edit annotations box on the left of the application. The figure below shows a color change in progress.

To check for restrictions, click on the Check For Restrictions button. The tool will identify any restrictions that satisfy its restriction rules. If there are restrictions identified, another window will pop up that will allow you to cycle through the recommended restrictions on the document. An example of the restriction window is shown in the figure below.
The dialog box indicates

- The sentence where the restriction is found.
- The rule within Jess that is activated
- Evidence for the restriction

4.4 Decision Rules for Recognizing Personal Records

The Presidential Records Act defines personal records as follows:

"The term "personal records" means all documentary materials, or any reasonable segregable portion thereof, of a purely private or nonpublic character which do not relate to or have any effect upon the carrying out of the constitutional, statutory, or other official or ceremonial duties of the President. Such term includes

a) Diaries, 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."

Examples of Personal Record Misfiles (PRMs) have been analyzed to determine criteria for distinguishing them from Presidential Records [Underwood and Harris 2005]. The following rules express such criteria.

If the record is addressed to the President or the First Lady, and is from a person who is a member of the Republican National Committee (RNC), or the record is addressed to a person who is a member of the RNC and is from the President or First Lady, then the record is a communication between the President or First Lady and the RNC.

If the record is a communication between the President or First Lady and the RNC, and is about political issues, then the document is a PRM because it is personal/political.

These rules are expressed in Jess as follows.
(defrule prmr1a
 (communication_act
  ((author ?person&: (=?person ?rnc_staff_member))
   (addressee ?person&: ((=?person ?presidentID) | (=?person ?firstLadyId)))
  )
 )
 =>
 (assert communication_between_president_or_first_lady_and_rnc)
 (printout t "Communication between President or First Lady and RNC")
)

(defun prmr1b
 (communication_act
   (addressee ?person&: (=?person ?rnc_staff_member)))
  )
 )
 =>
 (assert communication_between_president_or_first_lady_and_rnc)
 (printout t "Communication between President or First Lady and RNC")
)

(defun prmr2
 (communication_between_president_or_first_lady_and_rnc
  (communication_act
   (content political_issue)
   )
 )
 =>
 (assert review_class (type PRM) (rule prmr1))
 (printout t "review_class type PRM PRMR1")
)

4.5 Decision Rules for Recognizing PRA Restriction a(5)

PRA Restriction a(5) "Confidential Advice" applies to "confidential communications requesting or submitting advice, between the President and his advisers, or between such advisers." This includes, but is not limited to, policy or legal advice. It includes all documentary forms containing or requesting advice including final memoranda, draft memoranda, notes from meetings, letters, etc.

The President's advisors include counselors and assistants to the President, Deputy Assistants, Special Assistants to the President, and the Director of Media Affairs. It could include a Senator or Congressman who writes to the President as a personal friend and trusted adviser, rather than in his or her official capacity. It could also include anyone in the Executive Branch providing advice, including interagency groups and committees generating options or advice. PRA restriction a(5) applies for twelve years after the expiration of the President's term in office.

Twenty-five documents that represent "confidential communications requesting or submitting advice, between President Bush and his advisors or between such advisors"
were analyzed to determine the features that would enable one to conclude that they were subject to PRA restriction a(5). Some of these were not actually restricted under a(5) because President Bush waived his restriction rights under PRA a(5) to certain subclasses of them. Twenty-four documents that reflect communications between the President and his staff or between staff members that were not confidential were analyzed to determine those features that would enable one to determine that these were not subject to restriction a(5) [Underwood and Harris 2005].

Following are examples of some of the decision rules for recognizing a PRA restriction a(5), Confidential Advice.

If the author of the record is the President and the addressee is a presidential advisor, or the author of record is a presidential advisor and the addressee is the President, then the record is a communication between the President and an advisor.

If the author of the record is a presidential advisor and the addressee is a presidential advisor, then the record is a communication between presidential advisors.

If record is a communication between the President and a presidential advisor, or the record is a communication between presidential advisors, and the purpose of the communication is a request (for action, information) or an order, and the content involves Domestic Economic Policy issues, then access is restricted under PRA a(5).

Domestic Economic Policy addresses economic growth and tax revenues. Fiscal and Monetary policy is a part of Domestic Economic policy and addresses the budget, especially taxation and borrowing. This knowledge is not represented as decision rules, but as rules such as the following:

domestic_economic_policy_issue(X), if equal(X, "economic growth") or subsumes("economic growth", X) or equal(X, "tax revenues"), or subsumes("tax revenues", X), or fiscal_and_monetary_policy_issue(X).

fiscal_and_monetary_policy_issue(X), if equal(X, "federal budget"), or subsumes("federal budget", X), or equal(X, "taxation"), or subsumes("taxation", X), or equal(X, "federal borrowing"), or subsumes("federal borrowing", X).

The decision rules are represented in Jess as follows.
(defrule p5r1
  (communication_act
   (author ?person &:(= ?person ?presidentID))
   (addressee ?person_id &:(presidential_advisor ?to_person_id)) )
=>
  (assert communication_between_president_and_advisor)
  (printout t " communication between president and advisor")
)

(defrule p5r1a
  (communication_act
   (author ?person &:(=?person ?presidential_advisor))
   (addressee ?person &:(=?person ?presidentID))
  )
=>
  (assert communication_between_president_and_advisor)
  (printout t "communication between president and advisor")
)

(defrule p5r2
  (communication_act
   (author ?person &:(= ?person ?presidential_advisor))
   (addressee ?person_id &:(=?presidential_advisor ?to_person_id))
  )
=>
  (assert communication_between_presidential_advisors)
  (printout t "communication between presidential advisors")
)

(defrule p5r3
  "Confidential advice on domestic economic policy issues"
  (communication_between_president_and_advisors)
  (communication_act
   (purpose "directive")
   (content domestic_economic_policy_issue)
  )
=>
  (assert (review_class (type P5) (rule p5r3) (waived (is_waived P5r3))))
  (printout t " review_class type p5r3")
)

(defrule p5r3a
  "Confidential advice on domestic economic policy issue"
  (communication_between_presidential_advisors)
  (communication_act
   (purpose "directive")
  )
5. FOIA Processing


During the current project, additional functions were developed for PERPOS that support FOIA Processing of Presidential e-records [Laib and Underwood 2004, 2005]. FOIA Processing is initiated by a citizen's request for records under the Freedom of Information Act (FOIA). A search is made of accessioned records (usually unprocessed) to determine the records related to the request. The requestor is notified of the volume of records (in pages) that are potentially relevant and an estimate is made of the time needed to process them. An archivist will then review just those records that might be relevant, not considering a series in its entirety, but just relevant records in several series.

5.1 FOIA Processing Dataflow

The diagram below illustrates the dataflow of FOIA Processing activities that are supported by the PERPOS tools. The numbered, labeled circles are activities. The two parallel lines represent data stores. The rectangles represent entities external to the PERPOS system, and the labeled, directed arrows are data flows. Stepping through the diagram in the numerical sequence of the activities, one sees the dataflow.
Holdings consist of accessioned containers of e-records. The PERPOS Tools first support FOIA processing by providing the capability to Index Holdings (Activity 1). This function creates an index of all the terms in the textual records in Holdings.

The Bush Presidential Library Database is a Microsoft Access database of tables, forms, reports, and queries. The database includes an accession register, location register, folder title list, Reference Request Form and a Reference Search Form. The Presidential Library Database is external to the PERPOS Tools. To perform a search of the Staff Member and Office electronic records in Holdings, a FOIA case is created (Activity 2) that has the same FOIA Case Number as that on the Reference Request Form. An archivist translates the FOIA request into a FOIA Query, which is submitted to FOIA Search (Activity 3) that returns a result set of pointers to records that are relevant to the FOIA query. The result set is associated with the FOIA case. The requestor is notified of the volume of records (in pages) that is relevant and an estimate is made of the time needed to process them (not shown in the data flow). When an archivist is ready to begin work on the FOIA case, they check out one of the containers associated with the case (Activity 4) to an archivist’s work area. The archivist then uses the review activity of the Archival Processing Tool (APT) to review just those records in the container that are relevant to the query, not considering an entire record series or container (Activity 5). Review actions include opening a record for public access; withdrawing or redacting a record because of access restrictions; marking a record as a Personal Record Misfile or marking a file for transfer to the Library because it is a nonrecord. When the records that are in the result set for a container are reviewed, the archivist checks the container back into
Holdings (Activity 6). Once all containers with relevant e-records are reviewed, the Archival Repository Tool is used to make a FOIA Reference Collection (Activity 7) and Finding Aid (Activity 8). Records Marked for Transfer are transferred to the Library (Activity 9). The Public Access System has not yet been developed.

The following sections describe each of these support functions in detail and show the user interface.

5.2 Index Holdings

Before one can search for electronic records relevant to a FOIA case, one must create an index of e-records in the Repository (Holdings). Only containers in Holdings that have been filtered will be indexed. Filtering is the process of removing operating system files, office software application files, system or software documentation, or sample application files from accessed records. These files are not Presidential records and indexing them would result in an archivist having to consider non-records during the review process. Filtering of containers of record series is discussed in section 2.4 of this Reference Manual. Archive files that have not been expanded, and password-protected files, image, and audio files will not be indexed.

To index the contents of all accessioned and filtered containers, select FOIA Case Mgmt from the Activity drop-down menu. Next, the archivist selects Index from the Tools dropdown menu.

The message “Indexing Containers …” will appear in the status bar at the bottom of the screen. Depending on the number of containers and number of files in each container, this process will take from a few minutes up to several hours. If unfiltered containers are encountered while indexing, a dialog box will appear that lists the OAID of the containers that are not indexed because they are not marked as filtered.
The archivist can get a printout of these container ids by selecting the Print button at the bottom of the dialog box. Select the OK button at the bottom of the form to complete the process.

It is necessary to re-index the containers in the holdings area in the following cases.

1. Since the index was last created, new containers have been accessioned and/or filtered.
2. Since the index was last created, non-record files have been transferred from containers to the Library during the arrangement, preservation, or review process.
3. Removal of Personal Record Misfiles (PRMs).

In the first case, records relevant to a FOIA search may have been added to Holdings, but will not appear in a result set because they are not indexed. In the second case, a FOIA search may return a pointer to a record that is no longer in its original container. In the third case, a FOIA search may return a pointer to a record that is no longer in its original container because upon review it is found to be a PRM and is subsequently transferred out of the container. It is not necessary to re-index holdings before each search of holdings, if these conditions have not occurred.

### 5.3 FOIA Case Management

The Archival Repository Tool is used to manage the processing of electronic records relevant to a FOIA Case. To do this, select FOIA Case Mgmt from the Activity drop-down menu shown below.
The archivist then selects *Add FOIA Case* from the *Edit* drop-down menu.

This causes the Add FOIA Case dialog box to appear.

The Archival Repository Tool automatically fills in the Archivist field with the user name and the *Date* field with the system date. In the dialog box, the archivist enters the FOIA Case number from the Bush Presidential Library Reference Search Form and selects *OK*. There is a check to be sure that there is not a collision with an existing FOIA case
number. The *Scope and Content Note* should not be filled out until the FOIA Case is complete and it is time to create a FOIA Reference copy.

After *OK* is selected, the archivist is returned to the *FOIA Case Mngmt* activity with the FOIA Case Number highlighted in the left pane and the FOIA Case properties in the right pane.

### 5.4 Performing a Search

To search for records relevant to a FOIA request, the *FOIA Case Mngmt* Activity must be the current activity. To begin a FOIA Search, the archivist must first select a FOIA Case. When a FOIA Case is highlighted, the archivist selects *Search* from the *Tools* drop-down menu.

This starts the FOIASearch application. The FOIASearch’s main screen appears with the FOIA Case No in the screen title.
The archivist translates the FOIA request from the Reference Request Form into an Oracle Text with word query. For instance, if the request had been for "Any materials related to Iraq and Kuwait," the Oracle Text query might be "Iraq & Kuwait." The archivist enters the search criteria into the Query box.

Oracle Text uses the Boolean operators AND (&), OR (|) and NOT (~). Parentheses are used for grouping expressions. For example '(George | Barbara) & Bush' is for all records containing the words either 'Barbara' or 'George' and the word 'Bush'.
A root word prefixed with a dollar sign ($), e.g., $broadcast, will find all documents containing its root word (stem) or derivatives, e.g., broadcasts, broadcasting, or broadcaster. The EQUIV operator (=) is used to indicate that two or more words are equivalent, for instance (91=1991).

Using the ACCUM(ulate) (,) and weight (*) operators, one can increase the score for documents that match a query by weighting terms differently. For instance, in searching for documents related to the *Clarence Thomas nomination to the Supreme Court*, the expression

\[(\text{justice, judge, Supreme Court}^5, \text{Clarence Thomas}^10)\]

will increase the score of the term *Supreme Court* by 5 times and the term *Clarence Thomas* by 10 times. This signifies that documents related to *Clarence Thomas* and *Supreme Court* are most relevant to the query. The ACCUM operator gives the highest scores to documents that contain the terms within the scope of the operator; for instance, ACCUM (dog, pet, Millie) will give the highest score to documents that contain all three terms.

One can search for terms that are in close proximity with the NEAR operator. For example, to find all documents where Soviet is within six words of Revolution, the following query would be issued.

\[\text{NEAR}((\text{Soviet, revolution}), 6)\]

The default and maximum value for the NEAR operator is to search for terms separated by no more than 100 words. In conjunction with Boolean operators, the NEAR operator constrains the scope of a query. Used with the section-searching operator WITHIN, the NEAR operator can constrain the search to predefined zones (sentence, paragraph, HTML sections). Examples of FOIA requests and the corresponding Oracle Text queries are shown in Appendix C.

Once the query is entered, select the *Search* button.
Once a search has been performed, the View Document, Limit Results, and Save Results buttons are enabled. The result set is displayed as a table with the column headings Rank, OAID, and FilePath. The Rank is the relevance value assigned by Oracle Text Search. The OAID is the unique identifier the archivist assigned to the container when it was accessioned. The FilePath is the path within the container to the file (record).

To determine whether a record is relevant to the FOIA request, an archivist performing the FOIA search is able to view individual records with the search terms highlighted. The View Document screen also displays the OAID along with the path within the container in the title of the screen. To do this, the archivist selects the row in the table containing the path to the file he/she wishes to view and selects the View Document button.

The archivist can limit the results of the query to any result with a rank greater than one of their choosing. To do this, the archivist selects the Limit Results button.
However, in this case, it is not necessary to limit the result since the second record is relevant. The archivist can remove all results that do not have rank greater the seven by selecting the OK button. The archivist can also choose to change the search criteria by simply changing the query and selecting the search button. The old results will be discarded and the results of the new search will be displayed.

When the archivist selects the Save Results button at the bottom of the Search screen, the query is returned to the Archival Repository Tool along with the query results and is associated with the appropriate FOIA Case for further processing. At any time, the archivist can leave the FOIA Search application by selecting the Cancel button at the bottom of the dialog box. If this button is selected, neither the query nor the result list will be returned to the Archival Repository Tool and associated with the FOIA case.

If the FOIA results search is saved, both the query and the results will be displayed in the right windowpane of the Archival Repository Tool when the FOIA Case is highlighted. The containers that contain the relevant records will be listed under the FOIA Case. The results will be displayed in a list with the headings ‘Status’, ‘OAID’, and ‘Path’. The status of a record can be ‘s’, ‘f’, or blank. The status will be ‘s’ for a record that has been systematically processed, so will not need to be reviewed. It will be ‘f’ for a record that has been processed for a previous FOIA case, so will not need to be reviewed. It will be blank for a record that is unprocessed, so will need to be reviewed. Under the result list, the properties Processed Files, Unprocessed Files, Processed Pages, and Unprocessed Pages are listed. The Archival Repository Tool calculates the Processed Files and Unprocessed Files properties. These property values are updated each time a container is returned to holdings, even if the container was checked out for another FOIA case or systematic processing.
5.5 *Estimating the Number of Pages to be reviewed*

An archivist responding to a FOIA request must enter into the Reference Search Form the estimated number of pages to be reviewed (unprocessed) and those that are already reviewed (processed). Files of the same size but different file format can correspond to very different numbers of pages. For example, a Microsoft Excel file may take up a relatively small amount of file space, but generally converts to a large number of pages. Whereas, an image file may have a large number of bytes but correspond to a single page. Another factor is the size of the header that some file formats have. An experimental method has been developed for estimating the number of pages of processed and unprocessed records relevant to a FOIA case.

To use this method, highlight the FOIA Case number, and then select *Estimate Pages* from the *Tools* drop-down menu.

When the Archival Repository Tool has estimated the number of pages in the FOIA Case, it updates the currently selected FOIA Case and displays the estimates in the right pane.
5.6 Reviewing Records for FOIA Cases

To begin the review of records relevant to a FOIA case, an archivist opens the Archival Repository Tool. The archivist then selects *FOIA Case Mgmt* from the *Activity* drop-down menu. The archivist selects the FOIA case to be processed. The FOIA case's properties appear in the right panel. If FOIA search is performed, the query criteria and the result list will appear in the property list. Lists of containers in which there are records relevant to the request are displayed beneath the FOIA Case in the left pane. Click on the ‘+’ symbol beside the FOIA case, to view the list of containers. The archivist then selects a container to review. If the container is already checked out for processing by another archivist, the *Archivist, Processing Type,* and *Case No* properties will appear in the container’s property list in the right panel. Containers can only be checked out to one archivist at a time.

If the container is not already checked out to another archivist, the archivist can check it out by selecting *Checkout Container* from the *File* drop-down menu. The archivist who is checking out the container must be the archivist who created the FOIA case.

<table>
<thead>
<tr>
<th>Archival Repository Tool</th>
</tr>
</thead>
<tbody>
<tr>
<td>File</td>
</tr>
<tr>
<td>Open Container</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accession Number</td>
<td>2005,0001</td>
</tr>
<tr>
<td>UAIID</td>
<td>15004</td>
</tr>
<tr>
<td>Collection</td>
<td>Bush Presidential Records: Staff</td>
</tr>
<tr>
<td>Office Name</td>
<td>President, Office of the</td>
</tr>
<tr>
<td>Series Title</td>
<td>George H. W. Bush's Files</td>
</tr>
<tr>
<td>Container Type</td>
<td>Tar</td>
</tr>
<tr>
<td>Original Container Name</td>
<td>0004</td>
</tr>
<tr>
<td>Status</td>
<td>Inprocess</td>
</tr>
</tbody>
</table>

A dialog box will appear to allow the archivist to indicate where the checked out container should be placed. Recommended practice is to put it in the Work Area directory in a folder titled with the initials or name of the archivist.
When the container is checked out, the container’s manifest is modified to allow the Archival Processing Tool to know which records in the container belong to the FOIA case and which archivist has checked it out. Then, a copy of the container with the modified manifest is placed in the location indicated in the dialog box. The original container with its original manifest is kept in the repository as a backup.

When save is selected, the archivist is returned to the *FOIA Case Mgmt* screen. The *Archivist* who checked out the container, the *Processing Type* under which the container was checked out, and the *Case No* of the case under which it was checked out are
displayed in the right pane of the Archival Repository Tool. Containers can only be checked out to one archivist at a time.

When a container is checked out, only the archivist who checked out the container or a user with administrator privileges can open the container for processing. Anyone else who opens the container will be opening a read-only copy that remains in the Holdings area. The archivist opens the container using the Archival Processing Tool. The archivist selects Review from the Activity drop-down menu, and selects Open from the File drop-down menu.

From the archivist's folder in the Work Area, the container can be opened.
The filenames of the records associated with the FOIA case and the folders that contain the records appear in **boldface**.

To review a record, the archivist opens the file whose file name is boldfaced and it is displayed in a viewer. The archivist reads the document to determine whether it is a non-record, a Personal Record Misfile (PRM), or a Presidential record. Review of Presidential records, whether as a part of FOIA Processing or Systematic Processing, is to determine whether there are PRA restrictions or FOIA exemptions that apply.

The record shown below is a memorandum from President Bush to Brent Scowcroft, his National Security Advisor. PRA restriction a (5), Confidential Advice, would have applied to this record, had the President not waived that restriction for this record which he included in his book.\(^1\)

---

1  *All the Best, George Bush: My life in letters and other writings*, 1999, pp 490-491.

---
If a record is closed, the archivist must indicate the reason for withdrawal, in this case PRA a (5). There are no FOIA exemptions.

The archivist must also enter withdrawal information—record type, correspondent's name(s), subject or title, and chronological date. Some of this information can be copied from the record and pasted into the withdrawal form. The Archivist's username is automatically captured, as is the date of withdrawal.
A box next to the filename of the record is color coded to indicate the type of access—green for open, red for closed, blue for redacted. Yellow indicates a PRM and grey indicates marked for transfer. The access type, access restrictions, and withdrawal information for a record are displayed in the right panel of the window and are stored in a manifest file in the container.

A check in a grey box next to a folder (or container) symbol indicates that some of its records have been reviewed. When all of the records in a folder (or container) have been reviewed, the check will appear in a white box.
If a record is redacted, a document image is created for the record, and the archivist can block out text in the image, select the reason for withdrawal and stamp an area near the text with the reason for withdrawal.

The review support capabilities are essentially the same as were developed for systematic processing, except that only the records relevant to the FOIA query need to be reviewed. The support capabilities for review of records are described detail in section 2.7 of this Manual.

When an archivist has finished reviewing all the records in a container that are part of a FOIA Case, he checks the partially reviewed container back into the archival repository (including both reviewed and un-reviewed records). This is accomplished in the Archival Repository Tool by selecting Checkin Container from the Files drop-down menu.
When a container is checked back into the repository, the FOIA case under which the container was checked out and all other FOIA cases that contain one or more of the same records are updated. Their Processed Files and Unprocessed Files properties are updated. The Case’s Processed Pages and Unprocessed Pages properties are not automatically updated. It is the responsibility of the archivist to update these properties by selecting the Estimate Pages tool. The Status field, which displays Results, is also updated for this case and any other case that contains the reviewed records. Any additional records, created during redaction or conversion, are added and any path properties that were changed due to changes in arrangement are modified. Temporary information (Archivist's name and results list for this container) is removed from the manifest and the reviewed container and manifest replaces the container and manifest in holdings.

When a record is removed from a FOIA case, all other records with the same GroupID are removed from that case, but no records are removed from any of the other cases. The records removed from one FOIA case are still relevant to other FOIA cases to which they belong.

Once all pertinent updates have been made, the temporary information that was added to the container’s manifest is removed and the container replaces the original container that had been kept as a backup in the repository.
When an archivist has processed a container, he can proceed to the next container that is not checked out to another archivist. All the relevant records in some of the containers may have already been processed. If this is the case, the container does not need to be checked out. The archivist can tell if a container has any unprocessed records associated with the FOIA case by looking at the FOIA case's result list. The status of a record is listed beside the filename of the record. They can stop review at any time and resume review later by selecting the relevant FOIA case number from the Archival Repository Tool.

5.7 FOIA Case Description

When an archivist has reviewed the records in all the containers and folders that are associated with a FOIA Case, the FOIA Case should be described. To describe the FOIA Case an entry is made in the FOIA Case's Scope and Content Note property. To accomplish this, select the FOIA Case in the FOIA CASE Mngmnt Activity, and then select Properties from the Edit drop-down menu.
Once the Scope and Content Note has been entered, select the OK button. The archivist will be returned to the FOIA Case Mgmt Activity screen. The Scope and Content Note will be displayed with the other properties in the right pane. After a description has been entered into the Scope and Content Note field, the Make FOIA Ref menu option on the File drop-down menu will be enabled.

5.8 Arrangement and Finding Aid of a FOIA Reference Container

The last step in FOIA Processing is creating the FOIA Collection Reference Container and Finding Aid. The Archival Repository Tool creates them at the same time. The arrangement of the information in the Finding Aid is the same as the arrangement of the records in the Reference Container.

The contents of the FOIA Case Reference container and the Finding Aid are arranged following the Bush Presidential Library guidelines for arranging a FOIA Collection. The FOIA Collection Reference container is arranged by Collection with the Bush Presidential Records: Staff Member and Office Files being the first collection. Under each collection, the offices are arranged alphabetically. Under each office, the series are arranged alphabetically by the staff member's last name. Under each series, the containers are arranged numerically by OAID number (smallest to largest). Within OAID number, the directories and records are in the order in which they appeared in the original container, unless there was some rearrangement. The collection of Quayle Vice Presidential Records are arranged next in alphabetical order by office, then alphabetically by staff member name, and then in OAID number order.
If no staff person is identifiable for a series, the series are arranged alphabetically after the last staff person for an office. If Federal records are processed in response to a FOIA request, they are arranged at the very end of the FOIA Collection by Record Group.

To create a FOIA Reference container select the FOIA Case in the FOIA Case Mgmt Activity, then select Make FOIA Ref from the File drop-down menu. If the Make FOIA Ref menu option is not enabled, then not all files in the result set have been reviewed or the Scope and Content Note for the FOIA Case has not been created.

<table>
<thead>
<tr>
<th>Archival Repository Tool</th>
</tr>
</thead>
<tbody>
<tr>
<td>File</td>
</tr>
<tr>
<td>Open Container</td>
</tr>
<tr>
<td>Checkout Container</td>
</tr>
<tr>
<td>Make FOIA Ref</td>
</tr>
<tr>
<td>Exit</td>
</tr>
</tbody>
</table>

Make FOIA Ref

<table>
<thead>
<tr>
<th>Ingmt Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>property</td>
</tr>
<tr>
<td>Case No:</td>
</tr>
<tr>
<td>Date:</td>
</tr>
<tr>
<td>Archivist:</td>
</tr>
<tr>
<td>Scope and Content Note:</td>
</tr>
<tr>
<td>Criteria:</td>
</tr>
<tr>
<td>Processed Files:</td>
</tr>
<tr>
<td>Unprocessed Files:</td>
</tr>
<tr>
<td>Processed Pages:</td>
</tr>
<tr>
<td>Unprocessed Pages:</td>
</tr>
<tr>
<td>Results:</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

The FOIA Reference Container will contain opened records, redacted records, FOIA markers for relevant records that are in another FOIA Collection or in a Reference container for a systematically processed record series, withdrawal sheets and transfer sheets. It will not contain closed records, originals of redacted records, PRMs, or transferred records.

When the FOIA Case is selected, the case containers are shown under the FOIA Case and the contents of the Scope and Content Note along with query and the result list are displayed in the right pane. The only difference is that the Open FOIA Ref Copy and the Open FOIA Finding Aid option of the File drop-down menu are enabled when a FOIA Case is highlighted that has a Reference Container.

### 5.9 View FOIA Collections

This view includes a catalog of FOIA collections that have been opened for public access and includes the capability to browse the catalog and open FOIA collection reference containers. The catalog is similar to the left panel of the FOIA Case Management Activity, except it includes just those FOIA cases whose review is complete and for
which a reference copy and finding aid have been created. When a FOIA Case id is highlighted, the finding aid for that FOIA collection is displayed in the right panel. The archivist can browse the finding aid seeing the folder titles and filenames. If they want to view the records, they highlight the name of that container in the left panel and open it. The container is opened in a version of the APT that only allows them to explore the contents of the container. While this view is not yet implemented, the following two sections describe how it is currently possible to open a FOIA collection Finding Aid and FOIA Reference Container.

To open the Finding Aid of a FOIA Case, first select the FOIA Case. Then select Open Finding Aid from the File drop-down menu.

The Finding Aid for the FOIA Case is created as an HTML document so that it can be published on the Bush Presidential Library Web Site.

To open the FOIA Reference container, select the FOIA Case. Then select Open FOIA Ref Copy from the File drop-down menu. This will open the FOIA Reference Container inside the Archival Processing Tool.
The “wds” extension on the highlighted filename above indicates that it is a “Withdrawal Sheet.” In right pane, the Access property is shown as “Closed.” The grayed out Activity buttons on the tool bar indicate that the container is open as read only. It is still possible to open a record in a read-only viewer. To view the Withdrawal Sheet, select QuickView from the drop-down View menu.
5.10 Pilot Evaluation

Archivists at the Bush Presidential Library are pilot testing PERPOS in processing actual personal computer records from the White House Offices of the Presidential Administration of George H. W. Bush [Spencer et al 2005]. The archivists are evaluating the functionality of the tools for both systematic and FOIA processing.

Of particular interest are the capabilities for converting records from their legacy formats to current or standard formats. The overall preservation strategy has been to leave the files in their original format, so long as they can be viewed. However, there are some file formats for which we cannot find viewers but we can find converters that transform the original into a file format that can be viewed. There are a few file formats for which there is neither a viewer nor a converter. For these records it is necessary to execute the application in MSDOS or Windows 3.1 (16-bit) to view the record and then consider some method such as screen capture and saving the screen contents as an image.

During review of Presidential electronic records, Bush Library archivists discovered that the redacted copies of records were difficult to read. This is due to the fact that the Quick View Plus viewers used to display 75 or so user-created files do not use the same fonts as those used by the software application that created the file. Furthermore, the copy of the record used for redaction is a TIFF image. Another COTS redactor, Redax, was added to the APT that supports redaction of PDF Files. The APT tools include file format converters that can convert 50 of the 75 or so legacy file formats that occur on the Bush Hard Drives to PDF, html, and other standard or current file formats. The readability of the documents produced is better, and the documents produced by the converters more exactly represent the physical form of the original document.

Another benefit of being able to convert legacy file formats to html and PDF formats that it is possible to have corresponding to a redacted copy of a document, an unredacted copy not released to the public, but that is marked up showing the content of what was redacted. When it is necessary to re-review a record, the reviewing archivist can read this copy. An additional benefit is that redacted records in html or PDF format can be indexed and searched, whereas this is more difficult for redacted Tiff images.

6. Evaluation of Advanced Technologies for 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.

### 6.1 Information Assurance Test bed

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

![Figure 12. PERPOS2 Network Architecture.](image)

**Figure 12. 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.\(^2\) 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

\(^2\) [http://www.iana.org](http://www.iana.org)
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. 12, 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.

**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 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 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). It will also include information extraction modules, the record 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.

6.2 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 [Molavi 2004a, 2004b]. 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. 12, 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. Linux allows one to disable capabilities that are not needed.

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 (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. For instance, a search for "Oracle 9i Application Server" vulnerabilities yielded the following vulnerabilities, among others.

- 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.

3 [www.cert.org](http://www.cert.org)
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.\(^4\) 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.\(^5\) ARL evaluated COTS, NIAP certified firewalls for protecting the network.

ISS Internet Scanner was 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.

6.3 Evaluation of NIAP Certified Information Assurance Products

The PERPOS Project Staff collaborated with ARL researchers who conducted experiments to evaluate the performance of the secure portal operating in unsecured and secured mode [Nguyen and Racine 2005]. ARL also observed and measured the overhead associated with deployed cryptographic products used to transfer electronic records from the PERPOS portal to ARL computers.

6.3.1 The Common Criteria and NIAP Certification

The Common Criteria for Information Technology Security Evaluation (CCv.1) [NIST a] unifies the national IT security evaluation specifications of Canada, the European Commission and the US. 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


\(^5\) [http://squid.nlanr.net/](http://squid.nlanr.net/)
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
- 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 [NIST b].

NIAP evaluates the security features of the following kinds of security products.\(^6\)

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

### 6.3.2 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.

---

\(^6\) List of Validated Products (by Technology Type) [http://niap.nist.gov/cc-scheme/vpl/vpl_type.html](http://niap.nist.gov/cc-scheme/vpl/vpl_type.html)
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 surveyed commercial firewall products that are National Information Assurance Partnership (NIAP) certified [Nguyen et al 2004]. 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
- 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.
- 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 chose 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. GTRI collaborated with the Army Research Laboratory in the evaluation of these two firewall products with regard to their capabilities to control access to the PERPOS repository and archival services [Kau and Nguyen 2005].

- CheckPoint Firewall-1 Next Generation-Application Intelligence (NG-AI) R55 on a Nokia IP350 appliance (256 MB RAM, Pentium 3 700 MHz) running Nokia IPSO 3.8.1BUILD28.
- Symantec Enterprise Firewall 8.0 for Windows on a Dell PowerEdge 1750 (2GB RAM, dual Pentium 4 Xeon 3.06 GHz) running hardened Windows 2000 Server SP4.

The firewall network configuration used for PERPOS places the web server on a dedicated DMZ interface and the database and archive server on a dedicated internal interface as is shown in Figure 12.

This type of network configuration allows for the best security with a single firewall as it allows you to configure the firewall such that Internet systems cannot initiate connections with internal network systems and optionally DMZ systems cannot initiate connections with inside systems (for PERPOS, the DMZ must talk to the database server on the internal network). For example, if a system is compromised in the DMZ, the firewall is still providing some degree of protection to the internal systems. If there was only an
outside (Internet) interface and inside (internal network) interface, a compromised web server would have unrestricted access to the database and archive servers.

Our primary conclusions with regard to firewalls are:

- Firewalls should be classified by the degree to which they do deep packet inspection and on a per protocol basis.
- Firewall appliances should be used instead of firewall plus a general purpose operating system in order to provide increased security, reduced management costs, optimized configurations, and higher performance.
- While NIAP certification of firewall products is a Federal requirement, it is not sufficient to control access to protected systems.

Two vulnerability assessment network scanner products, Nessus and Internet Security Systems (ISS) Internet Scanner, were evaluated on their ability to detect vulnerabilities and the usefulness and depth of their reports. These vulnerability assessment network scanners were also used to provide vulnerability assessment for PERPOS project systems and firewalls. We illustrate why “outside the firewall” vulnerability assessment scanning is necessary in order to verify that firewall rules are configured correctly/working as expected, that inadvertent external access to internal resources has not occurred, and that the firewall is not leaking information about the internal network or the firewall products themselves that could be used by hackers trying to penetrate the firewall. We illustrate why "inside the firewall" vulnerability scanning is necessary in order to identify application system (e.g., Oracle) vulnerabilities, to identify operating system vulnerabilities, to identify unnecessary network servers, and to suggest enhanced security configurations for necessary network servers.

Our primary conclusions regarding vulnerability assessment tools are:

- More than one vulnerability assessment scanner should be used in order to compare results to ensure that one of the scanners is not missing vulnerabilities due to configuration errors, lack of updated signatures, or differences in detection methods.
- Vulnerability assessment scanners can return false positives. Administrator knowledge about the scanned systems, comparison of results with another scanner, and consultation with the vendors of the target systems must be performed in order to distinguish false positives from true positives.
- The results from Nessus and ISS Scanner show that vulnerability assessment scanners are useful in identifying unnecessary network services, suggesting enhanced security configurations for necessary network services, and revealing inadvertent external access to internal resources.
7. Summary of Results

A previous experiment showed that the ANNIE information extraction system performed reasonably well in recognizing person, organization and locations names, as well as dates, times and monetary amounts in a corpus of 50 Presidential e-records. Refinements were made to the wordlists and JAPE rules of the ANNIE information extraction system. A test was conducted using these modifications to the ANNIE Factbase (wordlists) and JAPE rules. Improvements in overall average precision were from .786 to .904 and in overall average recall from .713 to .907.

Fifty additional Presidential records from the Bush Presidential Library were scanned and OCRRed. Their subjects are shown in Appendix A. First, the corpus was manually marked up with the correct named entities, creating the Key files. The additional and modified gazetteers and new JAPE rules were applied to the second corpus. The performance of the modified ANNIE on the second corpus dropped from its performance on the first corpus. The average recall dropped from .907 to .807 and average precision fro .904 to .858. This is largely because new record types occurred in the second corpus that did not occur in the first corpus. However, this shows that the additions to the gazetteers and JAPE rules to improve the performance on the first corpus of Bush e-records are general enough to apply to other Bush e-records.

The significance of this information and content extraction task is that it provides the foundation for automatic metadata extraction for e-records (e.g. chronological date, author, subject), automatic identification of record type, automatic summarization of record series, and understanding the content of electronic records in support of FOIA review.

A grammatical induction method has been developed that uses a sample of annotated e-records of a particular type to learn the documentary form of that record type. The induced grammars are used with a parser to determine the record types arbitrary e-records. Methods have been developed to use the information extracted from the records in a directory to extend the titles of cryptic directory names, to describe the contents of the directory (file unit), and to describe the contents of a record series. Experiments will be conducted this coming year to evaluate the performance of these methods.

A natural language-based Boolean query language has been developed that can be interfaced to Sun's NOVA natural language-based passage retrieval system. This technology has the potential to improve both the precision and recall of searches for Presidential records relevant to FOIA requests.

Methods have been developed for acquiring taxonomic knowledge from text. It has been applied to the Bush Public papers. The kinds of knowledge acquired are specific to the Bush administration and not available in WordNet or other thesauri. Potentially, they can be used with Sun's NOVA natural-language passage retrieval system and the natural language Boolean query language to improve the recall of searches for Presidential records relevant to FOIA requests.
A method has been developed for determining the communication (speech) act conveyed by a record. Decision rules have been developed to distinguish personal records from presidential records. Decision rules have also been developed for recognizing PRA restriction a(2), Appointments to Federal Office, a(5), Confidential Advice, and a(6) b(6), Personal Privacy. A tool for supporting review decisions has been prototyped and is being used with sample personal records and presidential records to test and refine the decision rules for access restrictions. Background, domain knowledge of the Bush Presidential Administration has been acquired to support the natural language processing and rule-based reasoning required.

The capability to support Systematic Processing Case Management and FOIA Processing Case Management has been added to the Archival Repository Tool (ART). This includes the capability to search the Bush PC e-record collection using the Oracle DBMS and a Boolean Query Language with relevance ranking. An estimation of the number of pages of e-records associated with a FOIA case is provided. Review of records relevant to a FOIA request is supported by saving the query results and indicating to the archivist those records in a container that are relevant to the request that have not been reviewed, as well as those that are relevant and have already been reviewed. FOIA collections and Finding Aids are automatically created after completion of the review.

The PERPOS Project Staff collaborated with the Army Research Laboratory in the evaluation of two firewall products with regard to their capabilities to control access to the PERPOS repository and archival services. Two vulnerability assessment network scanner products, Nessus and Internet Security Systems (ISS) Internet Scanner, were also evaluated on their ability to detect vulnerabilities and the usefulness and depth of their reports.
References

PERPOS Technical Reports can be found at http://perpos.gtri.gatech.edu


[Harris et al 2005] B. Harris, E. Whitaker, R. Simpson. Access Restriction Checker, PERPOS TR 05-07, ITTL/CSITD, Georgia Tech Research Institute, August 2005.


Conference and Workshop Presentations

