PERPOS II:
Annual Technical Status Report
July 1, 2004 – June 30, 2005

William Underwood
Robert Simpson
Elizabeth Whitaker
Sandra Laib
Brian Harris
Matthew Underwood
Demetrius Campbell
Jason Kau

PERPOS Technical Report TR ITTL/CSITD 05-04
August 18, 2005

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

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. Review of records for Presidential Record Act (PRA) restrictions and FOIA exceptions is an intellectually demanding task, requiring page-by-page review. This report describes progress in applying advanced information technology to support these tasks.

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

A prototype Access Restriction Checker 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 containing confidential advice between the President and his staff.

Electronic records stored in digital repositories are vulnerable to system failure, human error, or malicious actions. GTRI has constructed a Web portal for Internet access and an isolated subnetwork behind a firewall containing an archival repository and archival services. This year, GTRI collaborated with the Army Research Laboratory in evaluating firewall and vulnerability assessment technologies for protecting these resources.

Experiments are being conducted at the Bush Presidential Library and Archives II to evaluate the models, technologies and tools developed. Archivists at the Bush Library have begun pilot testing of archival processing tools that support 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, summarization, knowledge representation, natural language processing, information assurance, E-FOIA.

Table of Contents

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 for the second year of the PERPOS II project 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 document types; and to summarize folder contents and describe record series;

(2) Investigate the use of XQuery to search large collections of e-records that have been annotated using the information and content extraction technologies developed in the first and fourth tasks;

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

(4) 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;

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

(6) 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. In the next section, progress and results on each of the research tasks is described. The final section summarizes research progress.

2. Research Tasks

2.1 Information and Content Extraction to Support Document Type Identification and Summarization

2.1.1 Information Extraction Improvements

During the previous year, the ANNIE information extraction technology was used to recognize named entities in a corpus of 50 Presidential electronic records. These named entities include person, location and organization names, temporal expressions (date, time) and numeric quantities (money, percent). The overall average precision was 78.6% and the overall average recall was 71.3%, The F-measure was calculated as the half weighted mean of precision and recall, and was 74.9% [Underwood 2004].

This year, improvements were made to address the:

- Layout Segmentation Problem, which is that of not recognizing the visually distinct regions & fields of text other than paragraphs consisting of sentences.
- Recognition problems due to inadequate knowledge of organization, location and person names associated with a Presidential Administration,
- Title Plus Caps, Name Format, and the Title+Title=Person problems.

2.1.1.1 Layout Segmentation

In the information extraction experiment, there were a total of 390 person names in the corpus [Underwood 2004]. ANNIE correctly recognized 61% of these, partially found the correct names in 20% of the cases and did not recognize 19% of the person's names in the corpus. In addition, ANNIE misidentified 128 terms as person's names (false positives). It happens that 41% of the false positives and 54% of the partially correct <person> entities occurred in four documents. These documents are attendance lists for meetings in the White House. The items in the lists consisted of a person's name, their title, and organization. Items are separated by a <CR><LF>, but ANNIE did not recognize the item segmentation. We believe that most of the false positives and partially correct person's names will be solved by our solution to the segmentation problem. We believe that the failures to recognize person's names will be corrected by adding person's names to the gazetteers.

Prior to our research, ANNIE had been primarily applied to newswires, newspaper articles and transcriptions of broadcast news. ANNIE's primary processing resources for segmenting these textual documents are the Tokenizer, which recognizes word tokens, and the Sentence Splitter. The sentence splitter relies on punctuation to determine sentence boundaries and ignores carriage return,<CR>, line feed <LF> indicating the end of a line. Document forms such as attendance lists, memoranda, and correspondence contain elements other than sentences, and <CR><LF> may be used to indicate other than paragraph boundaries. A layout segmentation processing resource is being developed to identify document elements other than paragraphs and sentences [Harris 2005]. The algorithm is to:

1. [Create bitmap] Convert the document into a bitmap changing characters to 1's and blanks to 0's. Ignore control characters and any prior markup.
2. [Grow regions]Any 0's in a row that are bordered by 1's are changed to 1's. Repeat this step.
3. [Determine rectangular boundary of region] Determine the upper left coordinates (xu, yu) and lower right coordinates (xl, yl) of a bounding rectangle enclosing each region of 1's.
4. [Mark rectangular regions] Markup the regions in the original document with

The sentence splitter then applies to regions alone and does not cross region boundaries. When sentences are not found, the JAPE rules will apply to lines (terminated with <CR><<LF>) within segments.

2.1.1.2 Domain Knowledge Needed for Information Extraction

ANNIE correctly recognized 50% of the organization names and partially recognized an additional 34% of the organization names in the corpus. ANNIE failed to recognize 16% of the organization names and incorrectly identified 78 terms as organization names. The partially correct cases and failures to recognize organization names can be overcome by addition of organization names to the gazetteers and extension of the JAPE rules for recognizing organization names.

Factual, domain, knowledge is essential to extraction of information from electronic records in order to determine document types, summarize documents, and summarize record series. It is also essential in extracting content from documents in order to determine whether documents are personal/political records or might have FOIA or PRA restrictions on their release to the public.

The kinds of factual knowledge that are needed include knowledge of

- George H. W. Bush Family Members
- President Bush's Friends
- Campaign Staff
- RNC Staff
- Presidential Nominations and Appointments to Federal Office
- White House Staff Members, Titles and Offices
- Bush Administration Senior Officials (Cabinet Secretaries and Undersecretaries)
- Presidential Advisors
- Members of 101st and 102nd Congresses
- Foreign Heads of State

Each of these kinds of knowledge has been acquired and represented in a form that can be used by the ANNIE information extraction system and the Template Filling rules of the Access Restriction Checker [Harris and Underwood 2004]. A significant result of this work includes the capability to automatically create the list who was nominated for appointment to federal office and when they were nominated from appendices in the Bush Public papers. Another accomplishment is the capability to automatically create the list of foreign heads of state from the CIA World Fact Book. This is significant because it reduces the effort required to acquire the knowledge.

2.1.1.3 Pattern-Action Rules for Recognizing Named Entities

The *Title+Caps problem* occurs when ANNIE finds a <Title> or <JobTitle> in the gazetteer lists, and uses a rule to match the title with a capitalized word immediately following the title to create a Person annotation, but the capitalized word is not a person's name. The *Name Format problem* is the failure to correctly annotate a person's name because the name is in an unexpected format. The *Title+Title=Person problem* occurs when a sequence of title+jobtitle, which often refers to a person, does not refer to a

person, but to a job title. We are refining the corresponding JAPE rules to solve these problems.

We are also incorporating pattern-action rules to recognize U. S. postal addresses, global political entities (locations of type government) and facilities (buildings or structures). Many entities, such as museums and schools, have characteristics of both facilities and organizations. In some cases, whether entities should be typed as facilities or as organizations can be indicated in the gazetteer. For other ambiguous entities, which are not explicitly listed in the gazetteer, e.g., a hospital or a hotel, the pattern-action rules are used to classify the entity as a facility or organization.

We are testing these additions and modifications on the corpus of 50 documents used in the original information extraction experiment [Underwood 2004]. When the modified ANNIE is able to recognize the named entities that it missed, correctly recognizes those that were only partially correct, and no longer miscategorizes named entities, we will conduct a second experiment on a different set of 50 documents, to evaluate whether the modifications generalize to other samples. When there is a high degree of precision and recall on the sample documents, an experiment will be conducted on the larger collection of Bush PC records at the Bush Presidential Library.

2.1.2 Document Type Learning and Recognition

The name associated with the form of a document (e.g., memorandum, correspondence, agenda, minutes) is an important element of archival description. The form of a document also aids in identifying the administrative context (author, addressee(s), date) and a description of the action or matter of the document.

Diplomatics defines *documentary form* as the complex of rules of representation used to convey a message, that is, as the characteristics of a document that can be separated from the determination of the particular subjects, or places it concerns. Documentary form is both physical and intellectual. The term *physical form* refers to the external make-up of the document, while the term *intellectual form* refers to its internal articulation. Therefore, the elements of the former are defined by diplomatists as external or *extrinsic*, while the elements of the latter are defined as internal or *intrinsic*. [Duranti 1998, p. 134]. In XML, the intellectual form of a document would be defined by a Document Type Definition (DTD), while the physical form of a document would be defined by an XSL Style Sheet.

We are applying information extraction and grammatical induction technology to learn the documentary form of a wide variety of Presidential electronic records [Underwood and Harris 2005]. The approach is to apply information extraction technology to identify and markup the intrinsic elements of a document's form (person's names, organization's names, job titles, dates, postal addresses, greetings (Dear, Hi) and salutations, e.g., Sincerely, Yours Truly. Given an annotated sample of documents of a particular type, e.g., decision memo, a stochastic context-free grammar is automatically induced that

defines the intellectual form. This is equivalent to constructing an XML DTD, as a DTD is defined using an extended context-free grammar.

Hong [2003] describes a heuristic search method for inferring a stochastic context-free grammar describing the structure of html documents. It involves four components:

- an evaluation function on hypotheses,
- an initial hypothesis,
- a set of successor operators in the hypothesis space, and
- a hill-climbing search method.

Evaluation Function

The hypothesis space is the set of stochastic (probabilistic) context free grammars. The evaluation function is based on the grammar complexity function defined below, which is designed to prefer grammars that have fewer rules and simpler productions.

$$C(G) = \sum_{i=1}^{N} \sum_{j=1}^{m_n} -\log p_{ij} + c(w_{ij})$$

The w_{ij} are right-hand sides of productions with left-hand nonterminal X_i and the p_{ij} are the probabilities that the production $X_i \rightarrow w_{ij}$ will be chosen to expand X_i .

The complexity of a string w is:

$$c(w) = (K + 1) \log(K + 1) - \sum_{i=1}^{r} k_i \log k_i$$

where w has length K and contains r distinct symbols each occurring k_i times, respectively.

The evaluation function is

$$E(G) = -C(G)$$

So the search tries to find the least complex grammar.

Initial Hypothesis

The initial grammar is:

$$A \rightarrow w_1 \mid w_2 \mid \ldots \mid w_m \quad [p_1, p_2, \ldots, p_m]$$

where w_i are annotated documents of the same document form (or type) and p_i are their relative frequencies. If all the strings are different, then all the p_i will be equal to 1/m.

Successor Operators

Successor operators take the current hypothesis and generate new hypotheses. The operators used are summarized in the following table.

Operator	From	To
Substitution	A → awb B → cwd	A → aXb B → cXd X → w
Substitution (variant)	A → awb X → w	A → aXb X → w
Disjunction	A → awb B → avb	A → aXb B → aXb X → w | v
Disjunction (variant)	A → awb B → avb X → w	A → aXb B → aXb X → w | v
Recursion	A → aXX…Xb X → w	A → aXb X → wX | $\in$
Expansion	A → aXb X → w_1 | … | w_m	A → aw_1b | … | aw_mb
Normalization	A → w | w | A | B B → v C → u	A → w | v

For instance, the interpretation of the substitution operator is that "if a substring w occurs multiple times in different productions, create a new production X → w and replace all occurrences of w with X's.

The normalization operator is applied immediately after using one of the other operators. It merges any redundant productions, expands nonterminals that are only referenced once, and drops productions that are inaccessible, result in some nonterminal that has no productions, or is of the form X → X.

Search Strategy

A hill-climbing strategy is used to search the hypothesis space.

1. From the initial hypothesis, perform every possible successor transformation to generate a new set of candidates.
2. Score the candidates according to the evaluation function, and the least complex candidate is chosen as the next hypothesis.
3. If no candidates can be found that score better than the current hypothesis, then stop, else chose the least complex candidate as the next hypothesis and repeat the process.

Example: White House Correspondence

Suppose that the initial hypothesis is:

> A → <date> </date> <greeting> <greeting> <person> </person> text <salutation> </salutation> <person> </person> <jobtitle> </jobtitle> <address> </address>
>
> A → <date> </date> <greeting> <greeting> <person> </person> text   text <salutation> </salutation> <person> </person> <jobtitle> </jobtitle> <address> </address>

These two production rules consist of the annotations to two documents that are White House correspondence and differ in dates, greeting, etc. and have one and two paragraphs respectively. The initial complexity of the grammar is 45.3.

After applying substitution and recursion operators the following grammar would be produced.

> A → <date> </date> <greeting> <greeting> B C <salutation> </salutation> B <jobtitle> </jobtitle> <address> </address>
> B → <person> </person>
> C → text C | $\in$

The complexity of this grammar is approximately 26.57.

From the inferred Probabilistic context-free grammars (PCFGs) for a variety of document types, we create a document type identifier and apply it to a corpus of documents that were not used in inducing the grammars. Experiments are being conducted to determine the performance of the document form induction and identification algorithms.

The significance of this research is that these tools are applicable to any document types of textual form including not only such traditional forms as correspondence and memos, but computer source programs (C-programs, Visual Basic Programs), batch command files, tables, databases, and spreadsheets. Thus, they can be used as tools to learn and identify the document (record) types occurring in any collection of textual e-records, including the holdings of future Presidential Libraries. Furthermore, these two functions, document type learning and identification, are not limited to the PERPOS prototype, but have Java wrappers that enable their use on other platforms and in other archival systems.

2.1.3 Creation of Folder Titles and Scope and Content Notes

During the current year of research, methods were developed for extracting information (data elements, metadata) about e-records, for example, record (document) types, time span (beginning and ending chronological dates), origin/function of the records,

arrangement, volume (number of bytes and pages), and file format type. We also developed tools for presenting this information in an archival catalog of Presidential Library Holdings (Archival Repository Tool), and in a manifest of metadata elements included in the containers of a record series. Metadata captured, maintained and displayed includes restrictions on access determined by Presidential Library Archivists. We also studied the standards and rules to be followed in presenting this information and creating these tools, e.g., NARA's Lifecycle Data Requirements Guide (LCDRG). 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."

We have developed a method for creating "Scope and Content Notes" for record series, file units and items. 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 a "Scope and Content Note."

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.

2.2 Experiments to Evaluate Advanced 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.

2.2.1 Boolean Query with Relevance Ranking Document Retrieval Technology

The Oracle DBMS has been has been integrated with the Archival Repository Tool (ART) installed on a PERPOS workstation at the Bush Presidential Library, and will be installed at Archives I and II. ART now supports indexing and search of the Bush PC e-record collection. (See section 2.5 of this report).

We want to evaluate the performance of Oracle word text on this larger Presidential e-record collection as compared to (1) XQuery on an XML Database (DB) of annotated copies of the Presidential e-records and (2) an enhanced NOVA natural language-based retrieval system.

2.2.2 Document Retrieval using an XML Database of annotated E-records and XQuery

We will use Oracle 10g XML DB and Oracle Xquery. The Presidential e-records will be annotated using the enhanced Gate/Annie information extraction system described in sections 2.1.1 - 2.1.2 of this report. This will enable searches constrained by document type, subject, person's names, job titles, organization names, global political entities, locations, facilities, and dates. The Communication Act Identifier described in section 2.4 of this report can also produce annotations as to author, addressee, communication act, purpose, and propositional content that might be used to constrain the search.

2.2.3 Natural Language-Based Record Retrieval

In our initial experiment, the user interface to NOVA required one to select 20, 50, or 100 relevant hits (passages). However, if there were more documents relevant to a query that the number of hits selected, the number retrieved were stopped at the number selected. For instance, on query 10, "Operation Desert Shield Storm Persian Gulf War IRAQ Kuwait", *Search for 100 hits* was selected and 100 passages were retrieved. 66 passages were relevant from 59 documents. However, there were 150 documents judged as relevant to this query. If it had been possible to set the cutoff to a higher number, say 500

hits, the average precision (and recall) on this topic would probably have been higher. There were at least ten NOVA queries in which the recall and average precision would have been significantly higher had the retrieved passages not been cutoff too soon. We conclude that the user interface should allow a larger number of hits, e.g., 500 or even 1000, before cutoff.

Another conclusion of this analysis was that NOVA would have performed significantly better if it had the capability to express it natural language queries in a Boolean query language. For example, the eighth FOIA request was for "Materials pertaining to Human Immunosuppressant Virus or HIV, and Acquired Immune Deficiency Syndrome or AIDS. [Note: HIV is more often the abbreviation of Human Immunodeficiency Virus]." Query 8 to NOVA was "HIV Human immunodeficiency virus AIDS." While NOVA retrieved passages corresponding to 29 of 55 relevant documents in the collection, its performance would have been better had separate queries been issued for "Acquired Immune Deficiency Syndrome," "AIDS," "Human Immunosuppressant Virus", and "HIV," and the results combined. Even better would be to allow the form of the NOVA query to be "Human Immunosuppressant Virus OR HIV OR Acquired Immune Deficiency Syndrome OR AIDS," and have the results automatically combined.

The forty-ninth FOIA request was for:

> Records relating to Bilaterals (Ministerial Meetings)
>
> 3/2-3/90 Bush - Kaifu Ministerial Meetings (Palm Springs, CA)
> 4/4/91 Bush - Kaifu Ministerial Meetings (Newport Beach, CA)
> 7/11/91 Bush - Kaifu Ministerial Meeting (Kennebunkport, Maine)
> 9/1/89 Bush - Kaifu Ministerial Meetings (Washington)

The NOVA query for FOIA request 49 was "Kaifu Bush meeting." The query response would be improved if one could ensure that a particular term such as "Kaifu" occurred. The query results would have been better had the query simply be "Kaifu." This can be accomplished by using Boolean AND, "Kaifu AND President Bush AND (meeting OR discussion)."

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 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 remove 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, 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.

We are formulating a Boolean query language for NOVA based on a Boolean semantics for natural language [Iwanska 1992]. A query interface will be developed for NOVA in which Boolean natural language queries can be expressed. We do not believe that this will 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.

An experiment will be conducted in which Oracle word Search, Xquery with annotated e-records in an XML DB, and Boolean query interface to NOVA are used to perform 50 FOIA searches of the Bush PC e-records.

2.3 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 reviews of Federal Records for FOIA exemptions.

Progress on this task was in the following areas:

- Automatic interpretation of a document to determine the communication act that it conveys.
- Representation of decision rules for access restrictions
- Acquisition of domain knowledge
- Refinement of the User Interface
- Testing of the Access Restriction Checker on a sample corpus

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.

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.

2.3.1 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 1 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.

- congratulation - the speech act of acknowledging that someone has an occasion for celebration.
- approval - the speech act of expressing a favorable opinion
  - recommendation - the speech act of commending a person as worthy or desirable
- proposal - the speech act of making a proposal
  - presentation - the speech act of presenting a proposal
  - advice - the speech act of advising as to an appropriate course of action.
    - recommendation - the speech act of recommending something as advisable.
- command - the speech act of authoritatively directing or instructing that someone do something.
  - order- the speech act of a superior giving a command that must be obeyed.
- agreement - the speech act of agreeing.
  - subscription - the speech act of agreement expressed by signing your name.
  - ratification, confirmation - the speech act of making something valid by formally ratifying or confirming it.
- request - the speech act of requesting
  - invitation - the speech act of requesting someone participate or be present or take part in something.
  - questioning, inquiring - the speech act of requesting information.
  - 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.
- reply, response - the speech act of continuing a conversational exchange.
  - answer - the speech act of replying to a question.
- description - the speech act of describing something.
- affirmation, assertion, statement - the speech act of affirming or asserting or stating something.
- complaint - the speech act of expressing a grievance or resentment.
- informing - a speech act that conveys information.
- briefing - the speech act of providing detailed instructions, as for a military operation.
- report - the speech act of informing by report.
  - summarization, - the speech act of preparing a summary, stating briefly and succinctly.
- promise- a speech act by one person committing to another agreeing to do (or not to do) something in the future.
- address, speech - the speech act of delivering a formal spoken communication to an audience.
- resignation - the speech act of giving up a claim or office or possession.
- appointment - the speech act of putting a person into a non-elective position.
  - nomination - the speech act of officially naming a candidate.

Figure 1. Some examples of Speech Acts

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." [This is where requests for actions and requests for information goes.]

- 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 which 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."

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")
)
```

2.3.3 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 document type of the record.

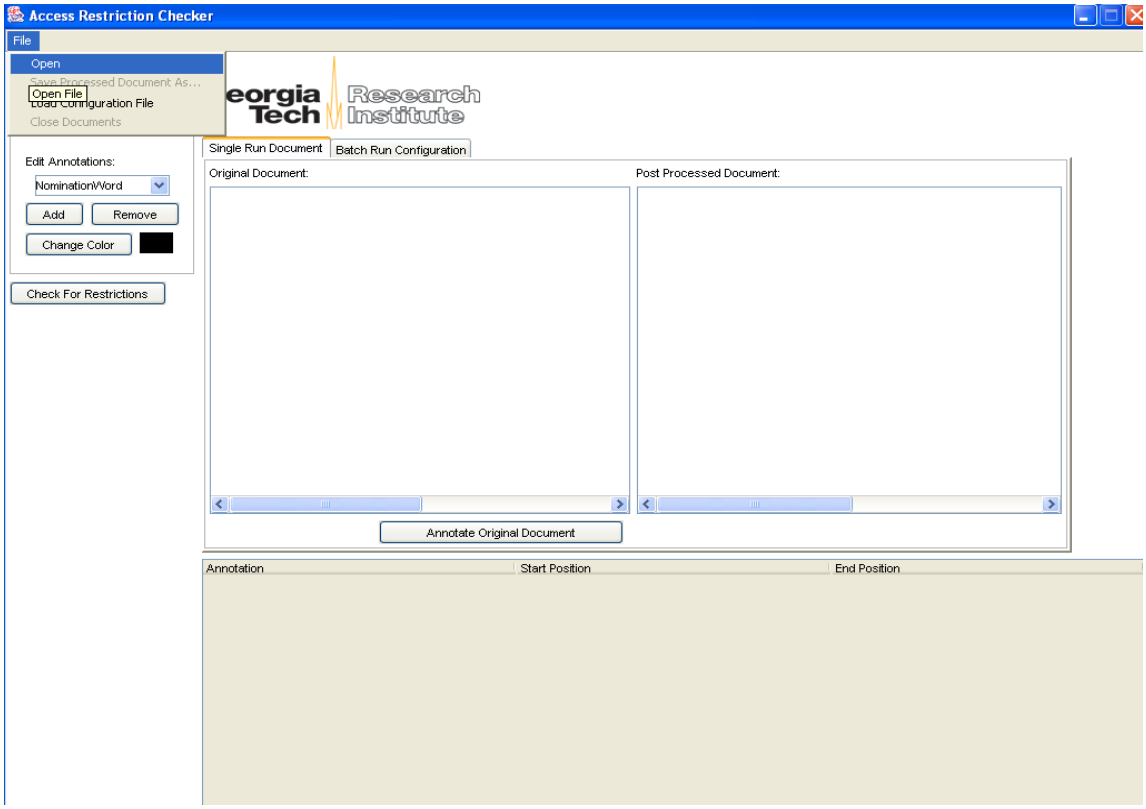
4. Use factual knowledge and template filling rules to fill in templates indicating the kind of communication action the record conveys, the purpose of the action, the author, addressee and its content.
5. Use personal/political record decision rules and access restriction decision rules and subsumption-based 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.

2.3.4 The User Interface

The figure below shows the user interface to the access restriction checker.

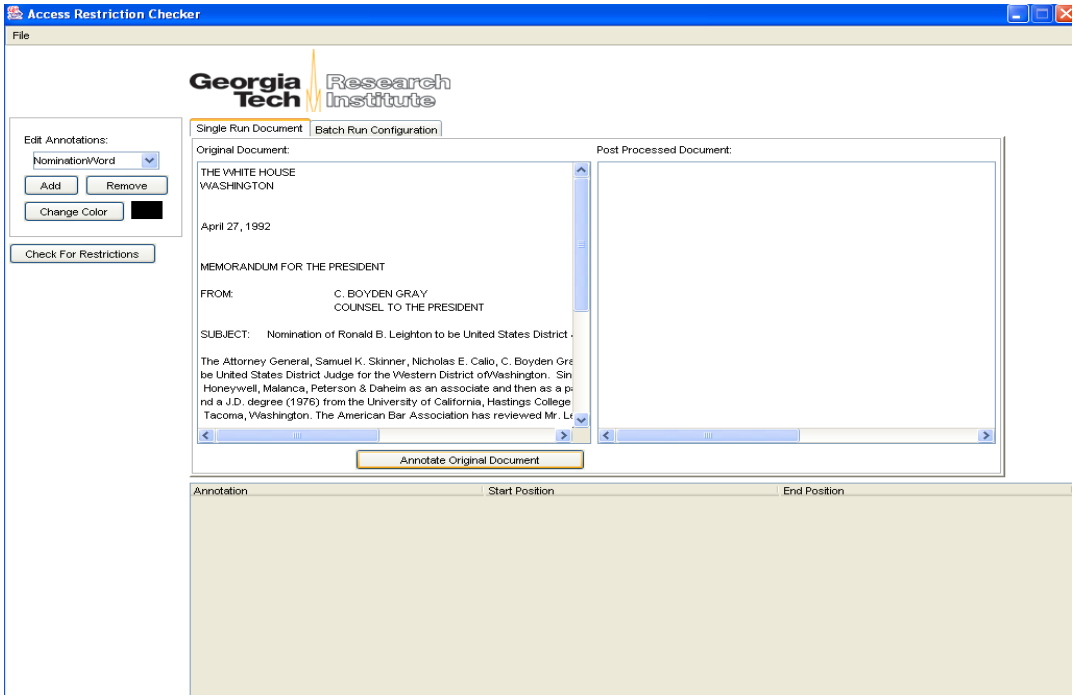
Step One

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.

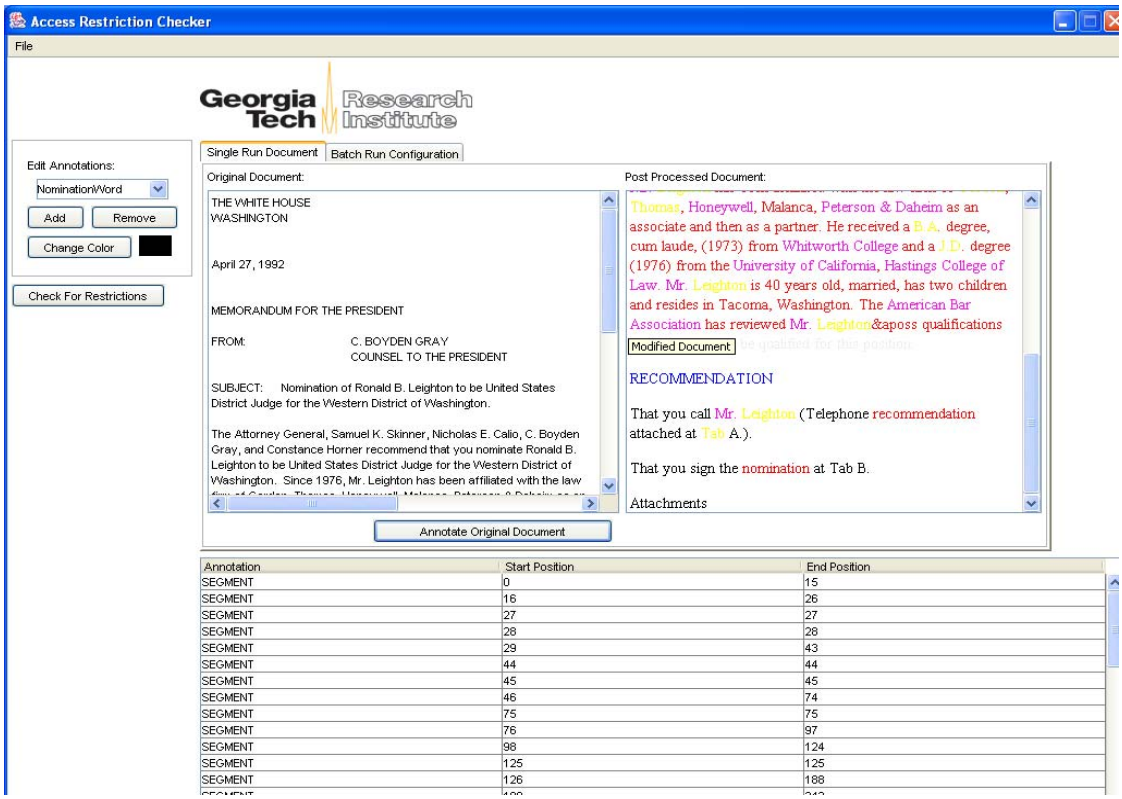


Step Two

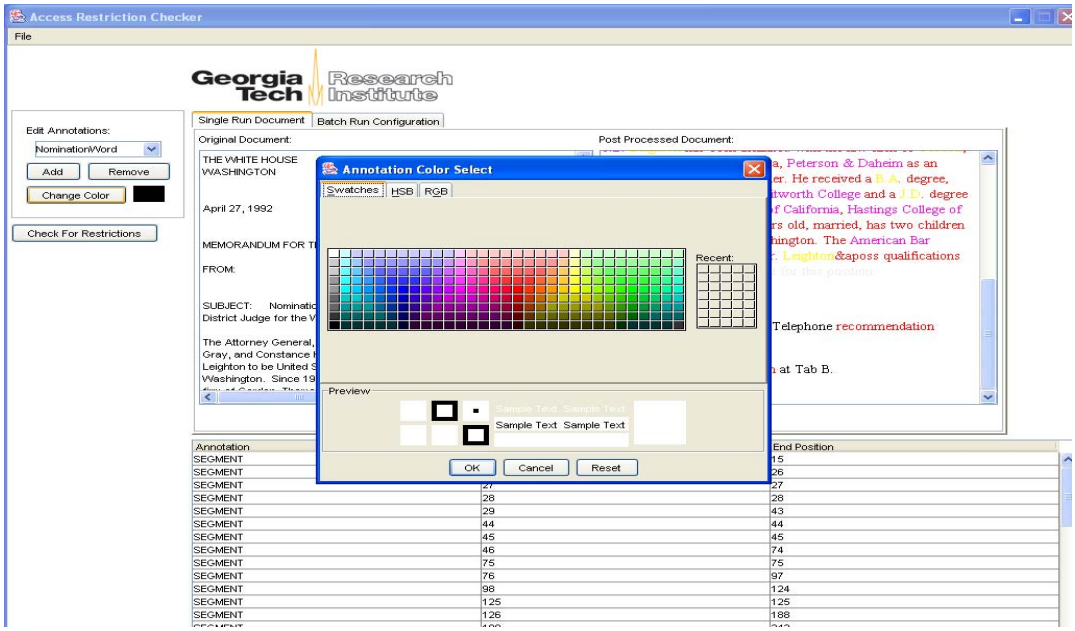
After selecting your file it should load in the Original Document into a text window as illustrated in the figure below.



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.

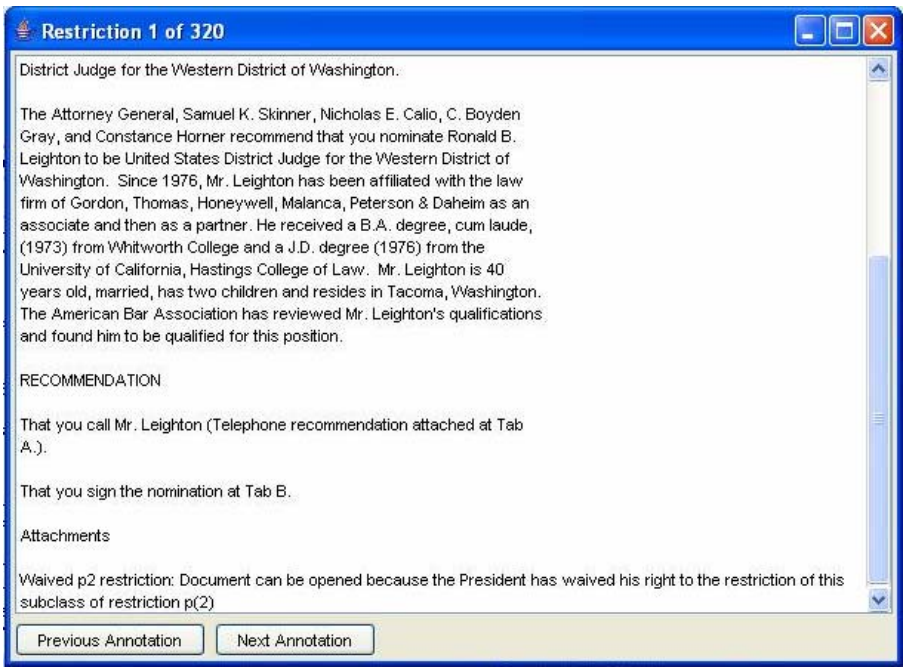


Different entity types are identified by different color text. The colors are defined in the properties file that is loaded but they can be changed in the edit annotations box on the left of the application. The figure below shows a color change in progress.



Step Three

To check for restrictions click on the *Check For Restrictions* button. The tool will identify any restrictions that satisfy it 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

2.3.5 Example of Decision Rules for Recognizing Personal Record Misfiles

The Presidential Records Act defines personal records as follows:

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

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

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

> 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")
)
defrule prmr1b
        (communication_act
        ((author ?person&: ((=?person ?presidentID) | (=?person ?firstLadyId)))
                (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")

)
(defrule 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")
)
```

2.3.6 Example of 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 2005a].

Following is an example of some of the decision rules for recognizing a PRA restriction P5, 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 ?presidental_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")
```

```
                (content domestic_economic_policy_issue)
        )
=>
        (assert (review_class (type P5) (rule p5r3a) (is_waived P5r3a)))
        (printout t " review_class type p5R1")
)
```

After the formulation of additional rules and tests on sample electronic documents, an experiment will be conducted using the Access Restriction Checker on actual Bush PC records at the Bush Presidential Library.

2.4 Refinement of the PERPOS Tools to Support FOIA Processing

When a request for records under the Freedom of Information Act is received from a citizen, a search is made of accessioned records (often unprocessed) to determine which records are responsive to the request. The requestor is notified of the volume of records (in pages) that is responsive and an estimate is made of the time needed to process them. An archivist will then review just those records that might be responsive, not considering an entire record series or container, but often just the contents of some folders within several containers. While the archivist might also perfect the arrangement and perform preservation actions on those records reviewed, they often do not fully describe, or preserve and arrange the contents of an entire container or record series. The requestor is then notified of the completion of the review and the availability of the requested records. This process is called FOIA processing.

The Archival Repository Tool (ART) supports FOIA Processing by supporting indexing of e-records in holdings, adding FOIA cases, searching for e-records relevant to the FOIA request, saving the results of the search, and copying containers containing relevant records to an archivist's work area. The Archival Processing Tool (APT) supports FOIA Processing by supporting review of the relevant records in containers. Then ART is used to move the containers back to the repository and to make a FOIA Reference Collection and Finding Aid [Underwood et al 2005].

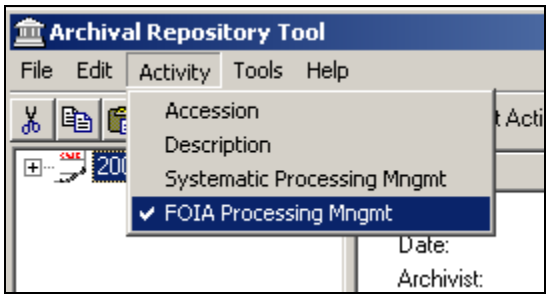
2.4.1 Index Holdings

Before one can search for electronic records relevant to a FOIA case, one must create an index of records in Holdings. One does this in ART by selecting *Index* from the *Tools* pull-down menu. A message "indexing containers" appears on the status bar. Only containers in Holdings that have been filtered will be indexed. Note that archive files that have not been extracted, password-protected files that have not been decrypted, and image and audio files will not be indexed.

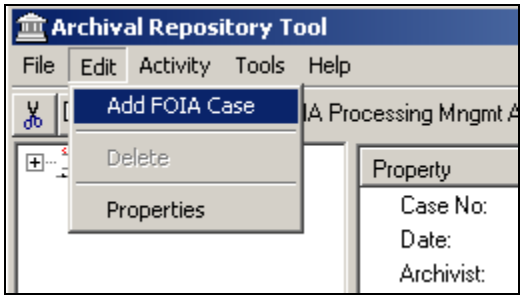
2.4.2 Add FOIA Case

FOIA requests are logged into the Presidential Library Database using a Bush Presidential Library Reference Request Form. A paper copy of this goes into a yellow folder labeled with the requestor's last name and the assigned FOIA case number.

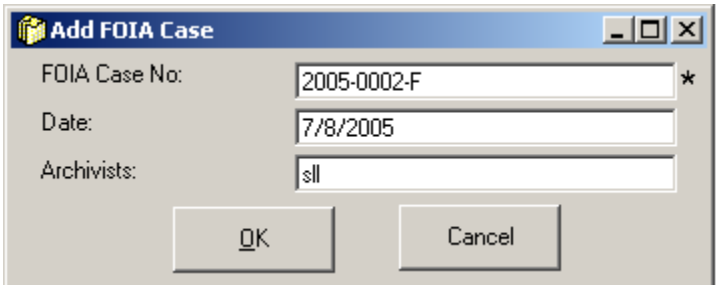
In ART one selects *FOIA Case Management* from the *Activity* pull-down menu.



To add a FOIA case, select *Add FOIA Case* from the *Edit* pull-down menu.



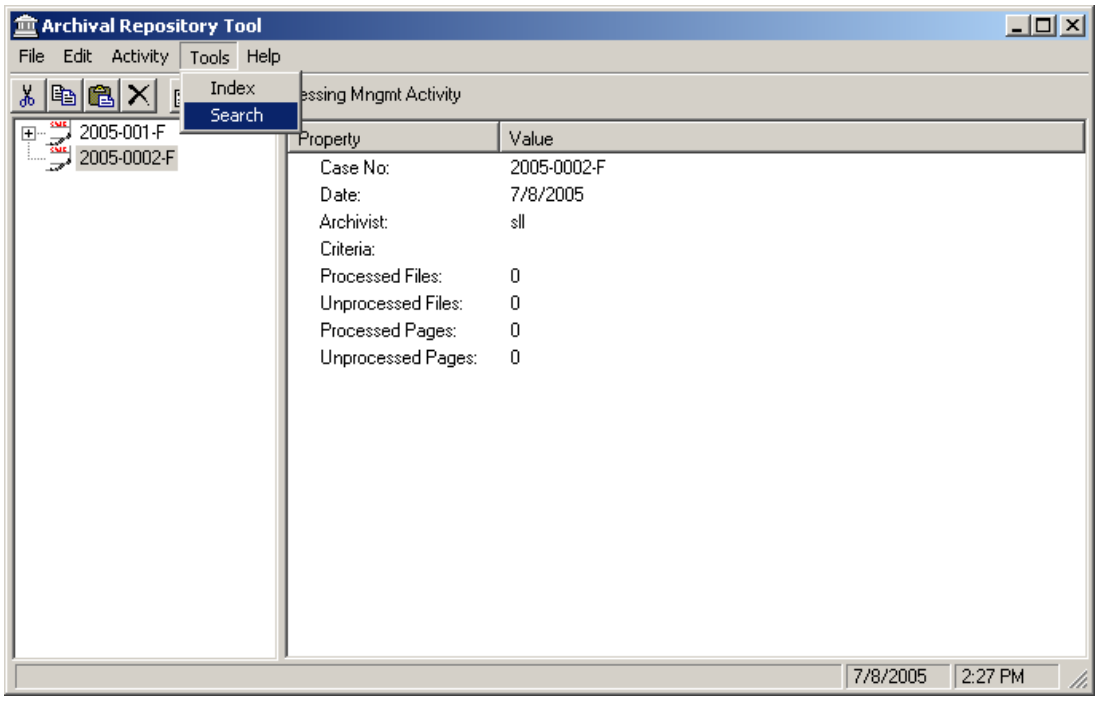
The *Add FOIA Case* dialog box appears.



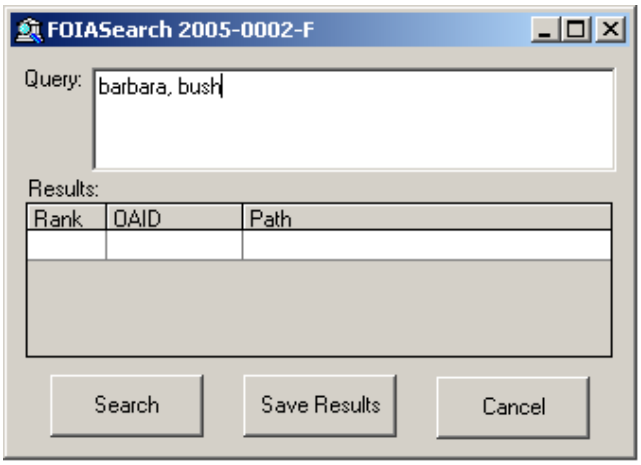
One enters the FOIA Case No (also called the Log Number). The date and Archivist's ID are automatically filled in. One selects OK. The previous FOIA case numbers and the new one are displayed in the left pane of the FOIA Case Management window.

2.4.3 Search for Relevant Records

To search for relevant records, one highlights the FOIA case number, and selects *Search* from the *Tools* pull down menu.



A dialog box similar to the following will be displayed



The archivist must translate the FOIA request into a query understood by Oracle Text. Oracle Text uses the basic Boolean operators AND (&), OR (|) and NOT (~). Parentheses can be used for grouping expressions.

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 (=) can be 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 yhr *Clarence Thomas nomination to the Supreme Court*, the expression

(justice, judge, Supreme Court*5, 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; e.g., ACCUM (dog, pet, Millie) will give the highest score to documents that contain all three terms.

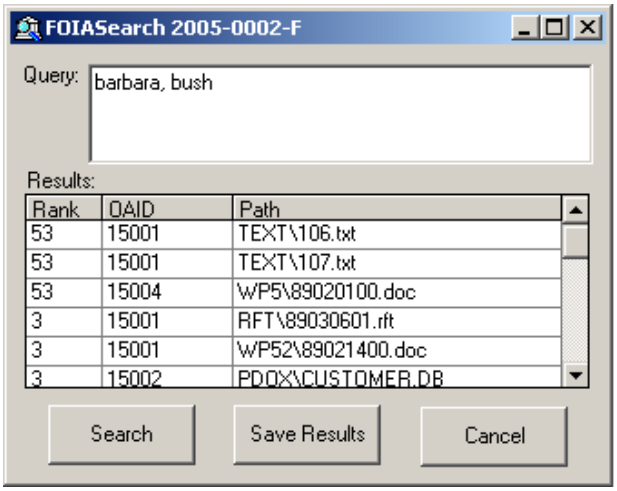
One can search for terms that are in close proximity with the NEAR operator. For example, to find all documents where Soviet is within 6 words of Revolution, the following query would be issued.

NEAR((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).

The following window shows the results of a search for "Barbara, Bush"



A list is returned of container (OAID) numbers and paths to files in those containers in descending (Rank) order of the score of e-records most relevant to the query.

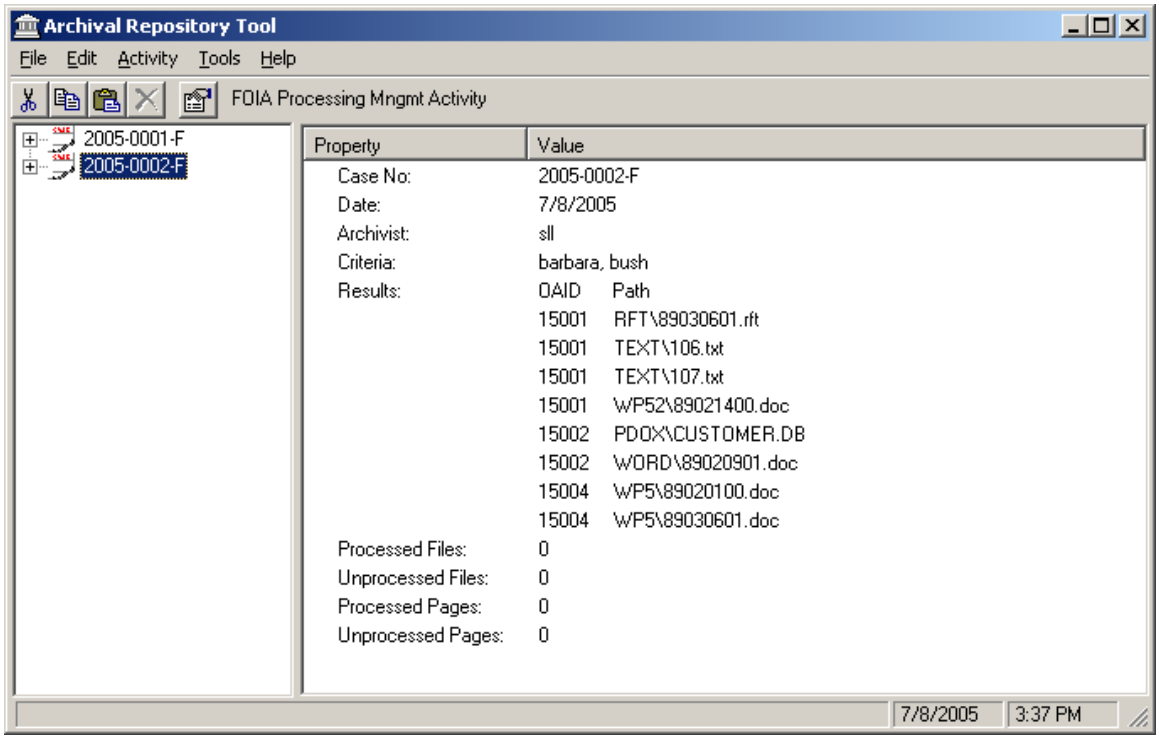
To calculate the relevance score for a document that matches a query, Oracle Text uses an inverse frequency algorithm based on Salton's formula.[1] Inverse frequency scoring assumes that for a document to score high, the query term must occur frequently in the document, but infrequently in the entire document set.

When the path is highlighted and selected the Archivist can view the contents of the file to determine whether this search is finding relevant documents. The Archivist can revise the query and search again.

In this search, only the first three e-records, scores of 53, were actually relevant to finding records regarding "Barbara Bush." The remaining records contained only the term "Bush", usually "George Bush."

We intend to add a field, "Score must be greater than *value*" below the query in the search dialog box. The interpretation is that the score must be greater than that value for the e-record to be retrieved.

When the archivist selects Save Results, the query and results are saved as Query 1 for this FOIA Case. When the Archivist selects OK, they are returned to the FOIA Case Management Window.



[1] $3f(1+\log(N/n)$ where f is the frequency of the work in the document, N is the number of documents, and n is the number of documents containing the query term.

The Query and containers containing files potentially relevant to this query are shown attached to the FOIA Case Number. With the FOIA case number highlighted, the information entered for that case number is displayed. With Query 1 highlighted in the left pane, the query and query results are displayed in the right pane. If the Archivist highlights the Container no, he sees the info that he would see if highlighting a container in the Description activity (including the collection, office and series names. This also includes whether the contents of the container are unprocessed, FOIA processed, or systematically processed.

To determine the number of processed and unprocessed files in the paths of the containers responsive to this query, highlight the query, and select *Calculate Number of Files* from the *Tools* pull-down menu. The number of processed and unprocessed files will be calculated and added to the query results.

If a FOIA Request is particularly complex and requires more than one query, the archivist can enter a second query and perform a second search. When the archivist is satisfied with the results, he can save the results as a second query. This process can be repeated.

2.4.4 Print Reference Search Form

The kinds of information appearing on a Library Reference Search Form can be printed for inclusion in the Yellow folder labeled with the Requestor's Name and FOIA Case Number. With the FOIA case number selected on the FOIA Case Management Window, Select Print Ref Search Form from the File pull down menu.

2.4.5 Reviewing Records for FOIA Cases

The Archival Processing Tool of PERPOS already supports systematic review of Bush PC Files. This includes the creation of withdrawal sheets for closed or redacted files. The FOIA case number of the FOIA case for which the document was actually reviewed is included on the Withdrawal Sheet.

To begin work on the review of records relevant to a request, an archivist uses the query search results for a FOIA Case with the archival tools to locate and load into their workspace a copy of a container in which there are records relevant to a request. They then open the container using the Archival Processing Tool and select FOIA Review from the Activity pull-down menu. They locate the relevant folders, and review the records, viewing them, and opening, closing or redacting them. When they have finished the review of records in a relevant folder, they store the partially reviewed container back to the archival repository (including both reviewed and unreviewed records). Then they proceed to the next relevant container. They can stop review at any time and resume review later by selecting the relevant FOIA case number to access the query results list.

When an archivist has reviewed the records in all containers and folders that are relevant to a request, they make a FOIA Reference Copy of the container for the Public Access System. Previously processed records (opened or redacted) are not included in the records of the current FOIA case, but the FOIA case number for which they were reviewed is indicated in the finding aid of the current case. The requestor is thus directed to other FOIA cases to see some records relevant to their FOIA request.

2.4.6 FOIA Description and the Manifest file of the FOIA Case

The following metadata will need to be included in the Manifest file of the FOIA Case container.

FOIA ID number
Scope and Content Note
Folder Title
OAID number
Collection, e.g., Bush Presidential Records or Quayle Vice Presidential Records
Office, e.g., Press office
Series, e.g., Files of Marlin Fitzwater
Subseries, e.g., Subject file, Chron file, Alpha File

The Manifest also indicates the folder titles and associated information for files not included in the container, but in other containers, that are part of another FOIA case or that have been systematically processed. It also indicates records that are included that were not relevant to the request, but were incidentally processed.

The Manifest File will be used to create a finding aid for the FOIA case that can be published on the Bush Presidential Library Web Site.

2.4.7 Estimating the Number of Pages to be Reviewed

An archivist responding to a FOIA request should enter the estimated number of pages to be processed. To do so one has to be able to estimate how many pages are in a file. There a number of factors contributing to the analysis of how many pages there are in a file. Different document types will generate very different numbers of pages per file. 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. Our approach is to estimate for the document types that occur on the Bush hard drives, the

Average number of bytes per page for file type $t = (\Sigma_{i=1,n} \text{filesize}(i)/\text{number of pages}(i))/n$

To determine the number of pages of e-records relevant to a FOIA request query, the archivist selects *Calculate number of pages* from the *Tools* pull down menu. The following is computed.

$$\text{Est. no of pages in files of type } t = \sum_{i=1,\ \text{number of files of file type } t} \text{filesize}(i)/\text{average no of bytes per page if file type } t$$

Then the

$$\text{Estimated number of pages to be reviewed} = \sum_{t \in \text{ file types in files to be reviewed}} \text{Est. no of pages in files of file type } t$$

If some of these files have already been processed or are being processed in another FOIA case, then the number of pages corresponding to those files are also estimated and displayed. The displayed results can be recorded on the Reference Search Form.

Then the FOIA case is assigned to the appropriate FOIA Queue. The request will be placed in a long/complex or short/simple queues. Determining which queue is appropriate for a request involves judging the complexity of the request as measured by the number of pages and the concentration of documents and folders within the collection. The FOIA Access Restriction Checker being prototyped in the research task described in section 2.4 could conceivably attempt a review of all the documents and indicate the types of PRA and FOIA restrictions that might apply and thus contribute to the estimation of complexity of reviewing the records relevant to the FOIA request query.

2.4.8 Pilot Evaluation

Two archivists at the Bush Presidential Library are using the PERPOS tools. The file systems from about 150 offices have been accessioned and they will soon all be filtered. 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. This copy can be read by an archivist when it is necessary to re-review the documents. An additional benefit is that redacted documents in html or PDFD format can be indexed and searched, whereas this is more difficult for redacted Tiff images.

Laura Spencer and Stephanie Oriabure, Archivists at the Bush Presidential Library, reported at the NAGARA Annual Conference on the results of their pilot testing of the PERPOS tools.

2.5 Evaluation of Advanced Technologies for Information Assurance

Electronic record archives, and especially those that are connected to the Internet, are at risk of attack by hackers and other risks such as worms and denial-of-service attacks. During the year, GTRI 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 [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 currently used on 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 2.

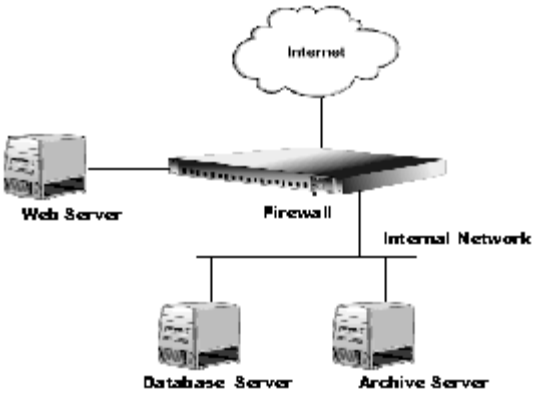


Figure 2. PERPOS Network Configuration

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.

3. Summary of Progress

Refinements have been made to the information extraction technology to address the layout segmentation problem, the domain knowledge problem and the Title Plus Caps, Name Format, and the Title+Title=Person problems. An experiment will be conducted to determine whether the refined information extractor has improved performance. A grammatical induction method is being developed that will use the annotated e-records to learn the documentary form of the document types occurring in the Bush e-record collection. The induced grammars will be used by a document type identifier to determine the document type of an e-record. 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.

Experiments will be conducted this coming year to evaluate the performance of three advance document retrieval technologies. The experiment will be conducting using the Bush administration e-record collection at the Bush Presidential Library. One of the technologies is Oracle Word search, which is a Boolean query with relevance ranking technology. It is already installed at the Bush Presidential Library and configured to search the Bush e-record collection. The second technology is Oracle XML DB with XQuery which will be used with copies of the Bush e-records that will be annotated using the information technology previously described. The third document retrieval technology is Sun's NOVA natural language-based passage retrieval system. Extensions will be made to provide a Boolean query capability.

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. This coming year, an experiment will be conducted using the Bush Administration e-records to evaluate the performance of the Access Restriction Checker.

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.

GTRI 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

[Baron 2004] J. R. Baron. Towards A Federal Benchmarking Standard For Evaluating Information Retrieval Products Used In E-Discovery: A Modest Proposal. Presentation to Sedona Conference Working Group One Annual Meeting, Electronic Document Retention & Production, Phoenix, AZ, October 15-16, 2004.

[Duranti 1998] L. Duranti, *Diplomatics: New Uses for an Old Science* (Lanham, Md.: Society of American Archivists and Association of Canadian Archivists in association with Scarecrow Press, 1998).

[Harris 2005] Segmenting Textual Documents in Support of Information Extraction and Document Type Learning. Working Paper 05-8. ITTL/CSITD, Georgia Tech Research Institute, 2005.

[Harris and Underwood 2004] B. Harris and M. Underwood. Factual Knowledge Needed for Information Extraction and FOIA Review, PERPOS Technical Report 04-7, December.

[Harris et al 2005] B. Harris, E. Whitaker, R. Simpson. Access Restriction Checker, Working Paper 05-7, ITTL/CSITD, Georgia Tech Research Institute, 2005.

[Hong 2003] T. W. Hong, Grammatical Inference for Information Extraction and Visualisation on the Web. PhD Dissertation, Department of Computing, Imperial College of Science, Technology, and Medicine, London, UK, June 2003.

[Iwanska 1992] L. Iwanska. A General Semantic Model of Negation in Natural Language: Representation and Inference. In *Proceedings of the Third International Conference on Principles of Knowledge Representation and Reasoning (KR92)*, San Franciso, Calif.: Morgan Kaufmann, pp. 357-368.

[Kau and Nguyen 2005] J. Kau and S. Nguyen, PERPOS Information Assurance. Technical Report 05-1, ITTL/CSITD Georgia Tech Research Institute, Atlanta, Georgia, May, 2005.

[Underwood and Underwood 2002] W. E. Underwood and M. G. Underwood. Evaluation of Document Retrieval Technologies to Support Access to Presidential Electronic Records. PERPOS TR ITTL/CSITD 02-03, December 2002.

[Underwood 2004] M. G. Underwood. Recognizing Named Entities in Presidential Electronic Records**,** PERPOS Technical Report ITTL/CISTD 04-4, June, 2004 (Revised Nov 2004).

[Underwood and Hayslett-Keck 2004] William Underwood, Marlit Hayslett-Keck. A Corpus of Presidential, Federal and Personal Records for use in Information Extraction,

Description and FOIA/PRA Review Experiments, PERPOS Technical Report 04-5. CSITD/ITTL/GTRI, June 2004.

[Underwood et al 2005] W. E. Underwood, M. Hayslett-Keck and S. Laib. The PERPOS Tools: User's Guide (Version 3.0) PERPOS Technical Report ITTL/CSITD 05-2, Revised March, 2005

[Underwood 2005a] W. E. Underwood. The Knowledge and Reasoning Required to Determine Confidential Advice. PERPOS Working Paper 05-3, ITTL/CSITD, Georgia Tech Research Institute, 2005.

[Underwood 2005b] The Knowledge and Reasoning Required to Recognize Appointments to Federal Office. PERPOS Working Paper 05-5, ITTL/CSITD, Georgia Tech Research Institute, 2005.

[Underwood 2005c] Distinguishing Personal Record Misfiles from Presidential Records. PERPOS Working Paper 05-6, ITTL/CSITD, Georgia Tech Research Institute, 2005.

[Underwood and Harris 2005]. Learning and Classifying Documentary Forms. PERPOS Working paper 05-8, 2005.

Conference and Workshop Presentations

W. Underwood. Semantic Technologies Applied to FOIA Review. Partnerships in Innovation: Serving a Networked Nation. National Archives and Records Administration and the University of Maryland, College Park. November 2004

W. Underwood. NLP Technology Applied to E-Discovery, Sedona Conference, WG1 Mid-Year Meeting, Cambridge, MD April 21-22, 2005.

W. Underwood, "Document Type Recognition and Content Summarization," Persistent Archive Testbed Working Meeting, February 17-18, 2005, San Diego

L. Spencer, S. Oriabure and W. Underwood. Launching E-Records with a PERPOS: The Presidential Electronic Records PilOt System. NAGARA Annual Meeting 2005, Richmond, Virginia, July 20-23, 2005.