Advanced Decision Support for Archival Processing of Presidential Electronic Records: Final Scientific and Technical Report

September 22, 2006 – September 21, 2009

William Underwood
Marlit Hayslett
Sheila Isbell
Sandra Laib
Scott Sherrill
Matthew Underwood

Technical Report ITTL/CSITD 09-05
October 2009

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

The Army Research Laboratory (ARL) and the National Archives and Records Administration (NARA) sponsor this research under Army Research Office Cooperative Agreement W911NF-06-2-0050. The findings in this paper should not be construed as an official ARL or NARA position unless so indicated by other authorized documentation.

Abstract

The overall objective of this project is to develop and apply advanced information technology to decision problems that archivists at the Presidential Libraries encounter when processing electronic records. Among issues and problems to be addressed are areas responsive to national security, including automated content analysis, automatic summarization, advanced information retrieval, advanced support of decision making for access restrictions and declassification, information security, and Global Information Grid technology, which are also important research areas for the U.S. Army.

The performance of the previously developed Information Extraction tool has been improved by the inclusion of additional wordlists and JAPE rules. Additional semantic categories such as facilities, legislative bills and statutes, governments, and relative temporal expressions are now annotated. An experiment with actual presidential e-records indicates a performance in recall, precision and F-measure of greater than .90.

A method for automatic document type recognition and metadata extraction has been implemented and successfully tested. The method is based on the method for automatically annotating semantic categories such as person's names, dates, and postal addresses. It extends this method by: (1) identifying about 100 types of intellectual elements of documents, (2) parsing these elements using context-free grammars defining the documentary form of document types, (3) interpreting the pragmatics of the form of the document to identify some or all of the following metadata: the chronological date, author(s), addressee(s), and topic. This metadata can be used for indexing and searching collections of records by person, organization and location names, topics, dates, author's and addressee's names and document types. It can also be used for automatically describing items, file units and record series.

Speech acts are acts of speech or writing in which one does something just by saying something, for example, "I appoint you…", "I hereby proclaim…" One hundred twenty Presidential records were analyzed with regard to the expression of speech acts with performative verbs and speech acts about the author's past or future speech acts or other's speech acts. More than 60 kinds of speech acts were discovered in the corpus. The analysis confirms that performative verbs are used to express the actions conveyed by records. A method has been formulated for identifying the speech acts occurring in e-records. It will be implemented, tested using records from the analyzed corpus and then experimentally evaluated

A method for automatically identifying the topics of e-records would facilitate automatic description of the records, and subsequent access to record collections. A corpus of fifty presidential records of various documentary forms was analyzed to determine the topic(s) of the records and possible techniques for automatically identifying the topics. The linguistics literature addressing discourse topic was reviewed. Technologies for domain-independent document summarization were also reviewed. An approach is proposed for identifying topics in presidential e-records that is a combination of domain-dependent and domain-independent methods.

A tool called the Access Restriction Checker is being developed to support archivists in archival review. Progress in implementing the Access Restriction Checker includes the interface of the prototype to the results of document type recognition and extraction of metadata about a record. Still needed is the provision to the Access Restriction Checker of the results of speech act and topic recognition.

The Presidential Electronic Records PilOt System (PERPOS) has been tested by archivists at the Bush Presidential Library in processing of Presidential records in response to FOIA requests. The results of the pilot test include: (1) the conclusion that the tool substantially supports FOIA processing, (2) the identification of additional features that would better meet the needs of archivists in FOIA processing of e-records, and (3) the adaptation of PERPOS to include some of these features.

Due to the rapid changes in computer technology, archivists must be concerned not only with the obsolescence of e-record file formats, but with the obsolescence of the operating systems, database management systems and integrated development environments of their Archival System. The Presidential Electronic Records Pilot System (PERPOS) as a case in point. Two exercises were conducted in using conversion tools to migrate Visual Basic 6 modules of PERPOS to Visual Basic 8 and to Java. The two exercises resulted in components that were functionally the same as the components written in VB6. The migration tools were judged useful, though substantial manual recoding was necessary. It was also concluded that to improve the maintainability of PERPOS, the more complex projects of the PERPOS architecture should be refactored into smaller, simpler classes.

File format identification is a core requirement for digital archives. The UNIX file command is among the most promising technologies for file type identification, but its reliability needs to be demonstrated. A database system for managing file format information and creating the magic file used by the file command is described. A graphical user interface has been developed for the file command. File signature tests have been created for more than 800 file formats. The performance of the file command and file signature tests is being evaluated on examples of the file formats that it purportedly identifies.

Table of Contents

Table of Figures

1. Introduction

1.1 Background

The increasing volume of digital records being acquired by the National Archives and Records Administration (NARA) poses significant challenges to archivist's traditional procedures for processing records. Archivists traditionally describe records at record group (or collection), series, file unit and item levels. This provides the Archives intellectual control over its holdings and supports access to the records. These descriptions include in summaries and/or metadata such information as document types, authors, addressees, topics and actions of the records. To construct complete descriptions, the archivists have to read the records at the item level.

Archivists at Presidential Libraries face the challenge of reviewing a large volume of presidential e-records for possible restrictions on disclosure. As noted in a recent NARA report to Congress:

> This year [2009], NARA received over 150 million Presidential emails from the George W. Bush Administration, as well as numerous other classified and unclassified electronic systems containing Presidential records. Presidential Library holdings in electronic form are now much larger than the paper holdings. Indeed, the email system for the George W. Bush Administration alone is many times larger than the entire textual holdings of any other Presidential Library. These electronic holdings bring new challenges to processing and making available Presidential records. The sheer volume exponentially increases what archivists have to search and isolate as relevant to a request, a lengthy process in and of itself before the review begins. Once review begins, the more informal communication style embodied in Presidential record emails often blends personal and record information in the same email necessitating more redactions. [NARA 2009, p. 25]

Assume that an email message is on average one page in length (including attachments). A Committee House Government Operations report of costs of declassification indicated that on average 59.4 pages could be reviewed per hour [HCGO 1973]. Let us assume one minute to review and make access restriction decisions for a single page. An archivist working 8 hours a day (480 min), 220 days a year, just on reviewing email could review approximately 100,000 emails in a year. To review the Bush email the first time would require 1,500 work-years. The Clinton Presidential Library holds 73,834,000 pages of paper records [NARA 2009, p. 27]. Assume a comparable number for Bush 43. Using the assumptions above, this translates into about 750 work-years for review of the paper textual records. Decision support is needed for this intellectually demanding task.

Another of the challenges that NARA faces is that the operating system and programming language technologies that are the bases of archival systems for processing of e-records are changing so rapidly. To maintain the viability of an archival processing system, it must be possible to cost effectively migrate archival processing software to new programming languages, operating systems and database systems.

File format identification is a core requirement for digital archives. Such identification is needed to insure that the files received from a creator have the expected file formats so that the archive is able to preserve the files and make them available to the public. Knowledge of the file formats is necessary to insure that viewers/players are available for the files, for conversion of legacy file formats into standard, current or persistent object file formats, and for extraction of files from archive files.

There is an opportunity to apply research is computational linguistics and advanced information system technologies to these challenges. The linguistic technologies include computational models of information extraction, documentary form, topic identification, speech act recognition, and discourse analysis. The information system technologies include rule-based reasoning, refactoring, programming language conversion, and automatic file format identification.

1.2 Purpose

The primary purpose of this report is to describe progress and results in applying computational linguistics technology to the support of archival description, review and search and retrieval of electronic records. Secondary purposes of the report are to summarize the results of pilot testing of the Presidential Electronic Records PilOt System (PERPOS) in support of FOIA processing of Presidential e-records, to summarize the results of two exercises in migrating PERPOS to a new programming language and development environment, and to describe progress in demonstrating a reliable technology for automatic file format identification.

1.3 Scope

In section 2, progress is described in improving the performance of an information extraction tool applied to presidential e-records. In section 3, a method for automatic recognition of documentary form, metadata extraction, and description of e-records is described. In section 4, a method for recognizing the speech acts carried out by e-records is described. In section 5, an analysis of the topics of presidential e-records is described. In section 6, progress in developing a tool for supporting archival review for restrictions on disclosure is described. In section 7, the results of pilot testing of PERPOS in support of FOIA processing is described. In section 8, an assessment of PERPOS for processing of national security classified records is described. Experiences in migration of some PERPOS code from Visual Basic 6 to Visual Basic 8 and to Java are also summarized. In Section 9, extensions of the UNIX file command and magic file to support reliable identification of file types is described. Section 10 summarizes the results of the research project.

2. Annotating Semantic Information in Electronic Records

Recognition of the actions and topics of e-records is dependent on the capability to recognize the proper names of persons and organizations who perform the actions or who may be the topic of a

record. In prior research, the capability to recognize and annotate person names, organization names, location names, dates, monetary amounts and percents was demonstrated [Underwood 2004, Isbell et al 2007]. The Semantic Annotator is based on Information Extraction technology developed at the University of Sheffield [Cunningham et al 2007]. The performance of the method of semantic annotation was substantially enhanced by creation of additional and enlarged wordlists and additional and refined pattern annotation rules [Underwood and Isbell 2008].

Tools for extracting files from containers, for converting the files to text formats, and for annotating their contents are installed on the PERPOS system in the Virtual Laboratory at Archives II. They were used in an experiment to evaluate the performance of the semantic annotation tool applied to e-records from the Administration of President George H. W. Bush.

A corpus of 50 records was selected from a record series accessioned into PERPOS. They are records from the Office of Legislative Affairs. The Office of Legislative Affairs provides advice and support regarding the President's legislative agenda and legislation is general, and liaison between the White House staff and members of Congress.

The experiment evaluated the performance of the semantic annotator with regard to the named entities addressed in the previous two experiments, namely, annotation of person, location, and organization names, dates, money and percents. The results are shown in Figure 1.

Annotation Type	Correct	Partially Correct	Missing	Spurious	Precision	Recall	F-Measure
Person	515	11	42	57	0.8928	0.9164	0.9044
Location	270	15	54	24	0.8981	0.8186	0.8565
Organization	509	31	31	50	0.889	0.9186	0.9035
Date	456	1	1	1	0.9967	0.9967	0.9967
Money	28	1	0	8	0.7703	0.9828	0.8636
Percent	6	0	0	0	1.0	1.0	1.0

Overall average precision: 0.9178 Overall average recall: 0.9282 Overall average F-measure: 0.9108

Figure 1. The performance of the Semantic Annotation on Corpus 3

The eight examples that were spuriously annotated as money were all instances of “mark up” of legislation. The term “mark” was annotated as Money (German Mark). This error was easily repaired by a rule that differentiates “mark up” of legislation from the monetary term “Mark”.

Most of the “Missing” annotations for locations were state abbreviations (OK, MA, ME, IN, VA) that appeared in parentheses after a legislator’s name, indicating the state they represented. These missing location abbreviations were in a list of ambiguous state abbreviations. However, they can easily be disambiguated since they appear after a person’s name and are in parentheses.

The table in Figure 2 shows in terms of F-measure:

1. The performance on Corpus 1 of the “vanilla” Semantic Annotator provided with the GATE distribution [Underwood 2004]
2. The performance on Corpus 2 of the Semantic Annotator after improvements were made based on an analysis of Experiment 1 [Isbell et al 2007]

3. The performance on Corpus 3 of the Semantic Annotator after improvements to the wordlists and JAPE rules [Underwood and Isbell 2008].

	Experiment 1: Default ANNIE Corpus 1	Experiment 2: GaTech IE Vers 1 Corpus 2	Experiment 3: GaTech IE Vers 2 Corpus 3
Person	0.6768	0.7269	0.9044
Location	0.7926	0.7922	0.8565
Organization	0.6342	0.7182	0.9035
Date	0.8934	0.9471	0.9967
Money	1.0000	0.9286	0.8636
Percent	0.8182	0.9273	1.0000
Overall Average F-measure	0.7490	0.8316	0.9108

Figure 2. Improvements in F-measure in three experiments

Figure 3 shows in graphical form the improvements in performance.

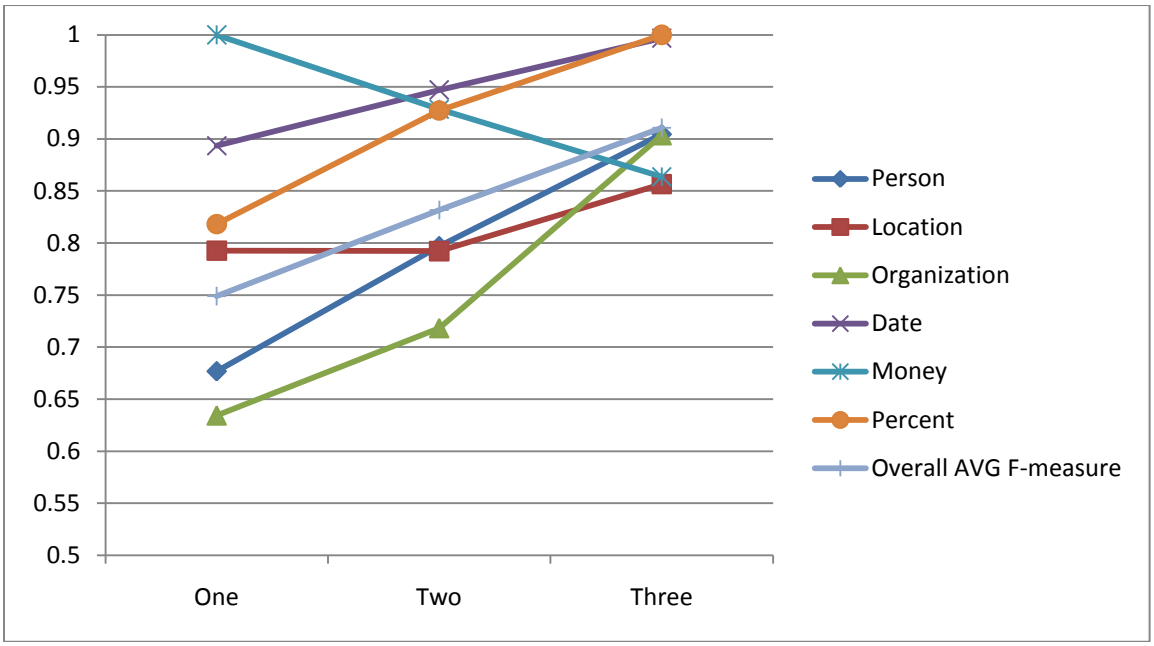

Figure 3. Graph showing improvements in the performance of the Semantic Annotator.

In the three experiments, the performance has increased in all cases with the exception of the annotation of money terms in Experiment 3. As previously explained, in experiment 3, the decrease in performance in annotation of money terms is due to the incorrect annotation of "mark" in the "mark up" of legislation.

The current performance is very good. Without this level of performance, methods for speech act, topic and document type recognition, which are dependent on the semantic annotation method, cannot achieve a high-level of performance. However, modifications were made to correct the annotations in Experiment 3 that were missed, partially correct and spurious.

3. Grammar-based Recognition of Documentary Form and Metadata Extraction

The Freedom of Information Act (FOIA) provides that citizens may request Presidential records 5-years after the end of an administration. In responding to FOIA requests, Archivists need to be able to search collections of records with high precision and recall. But at the time of responding to FOIA requests, archivists have not read all of the records, so cannot index the records and search on such attributes as person, organization and location names, topics, dates, author's and addressee's names and document types.

Archival descriptions include the names of the types of records that occur in a record series, for example, correspondence, memoranda or agenda. Record descriptions also include author's and addressee's names as well as the topics of records. Archivists cannot describe a collection until the collection has been manually read and reviewed. With increasing volumes of electronic records, it may be decades or even centuries before new acquisitions are described.

3.1 The Method

A method for recognition of documentary form and extraction of metadata has been implemented [Underwood and Laib 2008]. The method is an extension of the method for annotating semantic categories as described in the previous section.

Tokenizer
Wordlist Lookup + Wordlists
Sentence splitter
Hepple POS Tagger + Lexicon
Named Entity Transducer + Rules for Named Entities
Intellectual Element Transducer + Intellectual Element Rules
SUPPLE Parser + *Document Type Grammars*
Extract Record Metadata

The intellectual element transducer uses JAPE (Java Annotation Pattern Engine) rules to recognize and annotate the intellectual elements of a variety of document types. The intellectual elements are currently recognized in several phases. First, the key terms or phrases that compose intellectual elements are annotated. Figure 4 shows a JAPE rule for recognizing the intellectual element called *to*, the addressee caption of a memorandum.

```
//MEMORANDUM FOR -> TO
//TO: -> TO
Rule: TO
(
 (
  (Token.string == "MEMORANDUM")
  (Token.string == "FOR")
 )  |
 (
  (Token.string.string == "TO")
```

```
  (Token.string == ":")
 )
)
: to
-->
      :to.PossElement = (eletype = "to")
```

Figure 4. JAPE Rule for the intellectual element "to"

In the remaining phases, additional intellectual elements are annotated based on named entities recognized during information extraction and on annotations created during the earlier phases. The Intellectual Element Transducer currently annotates about 100 kinds of intellectual elements that may occur in fourteen documentary forms of Bush Presidential records.

The SUPPLE parser is a bottom-up chart parser that uses a context-free grammar for English to parse a sequence of word tokens (with their parts of speech) and named entities (usually proper nouns) in a sentence [Gaizauskas et al 2005]. To recognize the documentary form of an e-record, SUPPLE is provided with context-free grammars for documentary forms and a sequence of intellectual elements from the document.

Figure 5 shows an example of a White House memorandum.[1] Memoranda such as this were printed on White House stationery. The electronic copies of memoranda typically did not contain a letterhead, but in some cases a letterhead was typed at the top of the memorandum.

April 27, 1992

MEMORANDUM FOR SAM SKINNER

FROM: EDE HOLIDAY

SUBJECT: California Earthquake

Attached is a situation report from FEMA on the northern California earthquake. No deaths have been reported and 45 people are known to have suffered injuries. In addition, there has been extensive property damage. While FEMA is awaiting a request from the State before initiating any recovery activities, a joint State/Federal preliminary damage assessment is likely to begin today.

Director Stickney has requested that we forward the situation report to you.

Attachments

Figure 5. An example of a White House Memorandum

[1] Bush Presidential Library, Bush Presidential Records, WHORM Subject File, Disasters-Natural, ID#324869.

The simplified context-free grammar shown in Figure 6 defines the documentary form of the memo shown in Fig. 5 as well as the forms of other memoranda in the Bush e-record collection.

MEMO → HEAD BODY
MEMO → HEAD BODY OPTIONAL
HEAD → DATE ADDRLINE SNDRLINE SUBJLINE
HEAD → DATE ADDRLINE SNDRLINE THRULINE SUBJLINE
ADDRLINE → TO ENTITIES
SNDRLINE → FROM ENTITIES
THRULINE → THRU ENTITY
SUBJLINE → SUBJ TOPIC
ENTITIES → ENTITY ENTITIES
ENTITIES → ENTITY
ENTITY→ PERSON
ENTITY → PERSON JOBTITLE
ENTITY → ORGANIZATION
BODY → PARAS
BODY → SECTS
BODY → PARAS SECTS
PARAS → PARA PARAS
PARAS → PARA
SECTS → SECT SECTS
SECTS → SECT
SECT → SECTHDG PARAS
OPTIONAL → ATTACH
OPTIONAL → COPIES
OPTIONAL → ATTACH COPIES
ATTACH→ ATTACHMENT
ATTACH → ATTACHMENT TITLES
TITLES → TITLE TITLES
TITLE → TITLE
COPIES → CC ENTITIES
COPIES → CC ADDRESSES
ADDRESSES → ENTITY ADDRESS ADDRESSES
ADDRESSES → ENTITY ADDRESS
COPIES → JOBTITLES
JOBTITLES → JOBTITLE JOBTITLES
JOBTITLES → JOBTITLES

Figure 6. A context-free grammar for the documentary form of Memoranda

This and other context-free grammars for document types are translated into the notation used by the SUPPLE parser and semantic notations are added to the rules to enable the interpretation of the text appearing in the intellectual elements. Figure 7 shows an example of rules of the augmented grammar for memoranda.

```
%% MEMO → HEAD BODY
rule(memo(s_form:F,sem:D^[[document,D],[document_form,D,memo],
[author,D,SNDRList],[addressee,D,AddrList],[topic,D,TOPIC],
[date,D,DATE]]),
 [head(s_form:F,sem:[DATE,AddrList,AuthorList,TOPIC]),
 body(s_form:F)]).
```

```
%% HEAD → DATE ADDRLINE SNDRLINE SUBJLINE
rule(head(s_form:F,sem:[Date,ADDRList,SNDRList,TOPIC]),
 [chrondate(s_form:F,date:DATE),
 addrline(s_form:F,sem:AddrList),
 sndrline(s_form:F,sem:SNDRList),
 subjline(s_form:F,topic:TOPIC)]).
```

Figure 7. A fragment of the SUPPLE grammar for the documentary form of Memoranda

Figure 8 shows the parse tree of the structure of the document shown in Fig. 5 as recognized by the SUPPLE parser.

```
{best_parse=
(memo (head (chrondate (sem_cat "April 27, 1992"))
            (addrline (for (sem_cat "MEMORANDUM FOR"))
                  (entities (entity (person (sem_cat "SAM SKINNER")))))
            (sndrline (from (sem_cat "FROM:"))
                  (entities (entity (person (sem_cat "EDE HOLIDAY")))))
            (subjline (subj (sem_cat "SUBJECT:"))
                  (topic (sem_cat "California Earthquake"))))
      (body (paras (para
            (sem_cat "Attached is a situation report from FEMA on the
            northern California earthquake. No deaths have been reported and
            45 people are known to have suffered injuries. In addition, there
            has been extensive property damage. While FEMA is awaiting a
            request from the State
            before initiating any recovery activities, a joint State/Federal
            preliminary damage assessment is likely to begin today."))
            paras (para
            (sem_cat "Director Stickney has requested that we forward the
            situation report to you."))))
      (optional (attachment (sem_cat "Attachments"))))}
```

Figure 8. The documentary form (parse tree) of the sample Memorandum

Figure 9 shows the pragmatics of the memorandum in a quasi-logical form (qlf). This also represents the metadata of the memoranda that are needed for describing the e-record.

```
{qlf=[document(e1), document_form(e1, memo), author(e1, 'EDE HOLIDAY'),
      addressee(e1, 'SAM SKINNER'), topic(e1, 'California Earthquake'),
      date(e1, 'April 27, 1992')]}
```

Figure 9. The pragmatics of the document shown in Fig. 3

This logical notation is read as “e1 is a document,” “The documentary form of e1 is memo,” the author of e1 is EDE HOLIDAY,” etc.

From the metadata shown in Fig. 9, the following item description can be automatically generated.

> A memorandum dated April 27, 1992 from Ede Holiday to Sam Skinner regarding California Earthquake.

Suppose that there was a directory in *Edith E. Holiday's Files* titled *Petrolia, Calif. (Cape Mendocino) Earthquake.* The following file unit (directory, folder) description can be automatically generated from the descriptions of items in directory.

> This file unit contains materials relating to the 1992 Petrolia, California Earthquake. It includes memoranda, situation reports and correspondence.

Grammars have been developed for the following 14 document types.

Formal Letter	Presidential Determination
White House Casual Letter	Executive Order
White House Memorandum	Presidential Proclamation
Action-Decision Memorandum	National Security Directive
White House Referral	National Security Review
Recommended Telephone Call	Memorandum of Conversation
White House Press Release	Memorandum of Telephone Conversation

A corpus of 112 documents with examples of each of these 14 document types was constructed from paper Presidential records that are public records or have been reviewed and disclosed to the public. They include records from presidential administrations from Reagan to Obama. The records were scanned, OCRed and converted to file formats typical to the period in which they were created, e.g., DisplayWrite, Word Perfect 5, Word 98. This corpus simulates the digital records created by Presidential administrations. The method for documentary form recognition was applied to this corpus. Approximately 80% of the documentary forms were correctly recognized. There were a few cases of incorrectly extracted metadata. The failures in recognition and metadata extraction were primarily due to failures of the Semantic Annotator to correctly annotate some person names and job titles. With modifications to the word lists and JAPE rules of the Semantic Annotator, the Documentary Form Recognizer was able to correctly recognize all of the documentary forms and extract the correct metadata for all 112 documents in the corpus.

The document type recognizer is run inside the GATE (General Architecture for Text Engineering) graphical user interface [Cunningham et al 2007]. To conduct experiments with the Bush Presidential e-records, the recognizer was interfaced to PERPOS. PERPOS provides the facility for converting files in legacy file formats to plain text or html. GATE provides an Application Programmers Interface (API) that allows GATE to be run inside a Java program. A GATE persistent pipeline application was created that can be loaded inside a JAVA program. A JavaBean (DocParser) was created to perform the parsing of documents. The DocParser JavaBean was then packaged and registered using the Java packager command. This created the DocParser.dll file that was registered as the DocParser Bean Control. Next, the PERPOS Archival Processing Tool (APT) was modified to use the DocParser Bean Control for recognizing documentary forms and extracting metadata..

A Describe Records function was added to PERPOS. It applies to a container of records in a record series. The container is a Java Archive (JAR) file with a manifest. The document type and metadata extracted from a record are associated with the section in the manifest corresponding to the record. An item description is also created and inserted in the manifest.

The method is being experimentally evaluated by applying it to series of presidential e-records that have been accessioned into PERPOS. The table in Figure 10 shows the results of one of the early experiments on a small series of e-records [Underwood 2009d]. For this series, the method recognized the document type and extracted the metadata of two-thirds of the records. The records are predominantly White House Memoranda and White House Casual (Informal) Letter. Those memoranda not recognized include a memos without a subject and memos through two people, rather than a single person as specified by the grammar.

Document Type	**Number of Documents**	**Recognized**	**Not Recognized**
Memorandum	49	43	6
Draft Memorandum	1		1
Casual Letter	65	62	3
Casual Letter Template	12		12
Letter with no internal address	3		3
Recommended Telephone Call	1	1	
Photo Opportunity	5		5
Agenda	1		1
Talking Points	1		1
List of Names and Job Titles	1		1
Address for Envelope	1		1
Presidential Photograph Record	1		1
Video Script	2		2
List of Quotes	1		1
Schedule Proposal	2		2
Note	5		5
Address .List	2		2
White Paper	1		1
Presidential Remarks	3		3
Status of Congressmen on Legislation	1		1
Tabular Report	1		1
Total	159	106	53

Figure 10. Results of the method applied to record series 113

The failure to recognize some of the documents as White House Casual Letters was due to the Semantic Annotator failing to recognize a postal address, a person's name or job title. This problem can be addressed by improving the performance of the annotator.

Suppose that the documentary form recognizer fails to recognize the documentary form of a record. How might one automatically determine that the record was one of the documentary

forms that should have been recognized? One might use intellectual elements and partial parses of the entire set of documentary forms as features and use pattern classification techniques [Santini 2004, Kim and Ross 2007] to learn and then recognize document types. This could enable proper classification, but not metadata extraction.

3.2 Induction of Grammars for Documentary Forms

The question arises, could grammatical induction be used with samples of a particular documentary form to induce a grammar automatically rather than manually? This would eliminate the manual effort needed to construct grammars from large samples, and could provide a method for automatically refining the grammar when examples of a documentary form were encountered that did not fit the current grammatical model. It would also facilitate the extension of documentary form recognition to a larger number of documentary forms.

Underwood and Harris [2006] demonstrated that it is possible to induce a grammar for the documentary form of White House memoranda and correspondence from sequences of intellectual elements from samples of these document types. The method used is state-based search. The initial state is a grammar with a rule for each sequence of intellectual elements in the sample documents. A set of operators is constructed that create simpler grammar rules by substitution, recursion and generalization. A search strategy uses these operators to generate alternative grammars. A function is created to evaluate the quality of the generated grammars. The search terminates when the evaluation function indicates no improvement in the quality of the induced grammars.

One of the obstacles to progress in this research was that samples of document types were created from OCRed paper documents and the intellectual element recognizer had not been created. Now, the intellectual element recognizer and document type recognizer have been interfaced to PERPOS. There are hundreds of thousands of e-records in the PERPOS repository that can be used in grammatical induction experiments. The intellectual element recognizer is being modified to output the intellectual elements of a record for use in grammatical induction rather than in recognition.

Underwood and Laib [2009] are continuing the investigation of induction of grammars for documentary forms. The research has not progressed to the point that there are experimental results to report, so the approach and some research issues will be summarized.

The grammatical induction method is implemented using a library of Common LISP functions created by Stolcke [1994]. The library support induction of probabilistic context-free grammars. An evaluation function and alternative operators on grammars are being added to the library.

The grammars that are being automatically induced do not describe the documentary form in the same way as the grammars created manually. It may be that the person creating the grammar may have been able to make generalizations from that could not be made from a small sample by the automatic method of grammatical induction. It may be that the induction method and person are using different evaluation criteria.

When records are encountered that should have been recognized as one of the defined documentary forms, but are not, is there a method for inducing a grammar that also describes the new instances? We are exploring applying the induction method to the manually constructed grammar plus rules for each of the sequence of intellectual elements of records of the same documentary form that are not recognized using the grammar. The resulting grammar typically keeps the overall structure of the manually constructed grammar and adds some new rules with non-mnemonic non-terminal names.

Even if it is possible to induce new rules for existing grammars, it is necessary to manually merge the new rules into the grammar used by the bottom-up phased parser. Is it possible to use the merged grammars for all document types as the baseline grammar when inducing grammar rules for the documentary forms of records that should be, but are not recognized? If so, this would facilitate incorporation of the new rules into the grammar used in parsing and extracting metadata.

4. Recognizing the Acts Carried Out by Records

Among the challenges facing archivists at Presidential Libraries and the National Archives are the tasks of reviewing and describing terabyte- and petabyte-sized collections of electronic records. To describe and review Presidential records, archivists must be able to recognize the acts performed by the records. The kinds of acts carried out by Presidential records include proclaiming, directing, ordering, declaring, reporting, certifying, prohibiting, delegating, designating, authorizing, appointing, nominating, resigning, and pardoning, to name a few.

Records in Presidential Libraries are available to the public once they have been reviewed for any restrictions on disclosure as specified by the Presidential Records Act (PRA) and the Freedom of Information Act (FOIA). PRA restriction a(5) “Confidential Advice” is an example of the kinds of restrictions on disclosure than an archivist must identify. This restriction on disclosure applies to "confidential communications requesting or submitting advice, between the President and his advisers, or between such advisers." Records that provide such advice may explicitly express the advice as a recommendation, suggestion, proposal, or advice. Or they may express the advice implicitly, for example, as “You should …”

After review of the records in a record series, archivists summarize the contents of the series (and sometimes items and file units) in scope and content notes. These notes often include descriptions of the actions conveyed by records.

4.1 Speech Acts

Performative verbs are verbs whose action is accomplished merely by saying them or writing them, for example, "I promise, "I recommend," "I advise," I nominate," or "I appoint." Vanderveken [1990] defines the semantics of 271 performative verbs. Wierzbicka [1987] defines the semantics of about 230 speech act verbs, not all performative.

A speech act can be represented by indicating a *speaker* (author), who is the utterer (writer) of a message and a hearer (*addressee*) who is any of the immediate intended recipients of the speaker's (writer's) communication, the *propositional content* that consists of a subject and a predicate which expresses something about the subject, and its *illocutionary force* (purpose). According to Searle's taxonomy of elementary illocutionary acts, there are only five *illocutionary points* that speakers can attempt to achieve in expressing a propositional content with an illocutionary force. These are the *assertive, commissive, directive, declarative* and *expressive* illocutionary points [Searle 1969, 1979].

4.2 Analysis of Speech Acts Expressed in Presidential E-Records

Underwood [2008] discussed the relevance of speech act theory to the comprehension of the actions conveyed by records, and to the archival description and review of these records. Speech acts include common acts such as asserting, promising, requesting, ordering, and congratulating. They also include acts that can only be performed by persons with the power and authority to do so such as proclaiming, declaring, directing, pardoning, appointing, nominating, and counseling.

A corpus of 120 Presidential records was analyzed to determine the occurrence of explicit and implicit speech acts and assertions about the writer's or others' speech acts. Instances of 76 speech acts previously defined by Vanderveken [1990 chap. 6] were discovered in the corpus. Instances of 32 speech acts that were not defined by Vanderveken were discovered and defined. The report gives examples from the corpus of each of these speech acts.

In the sample corpus, it was found that 63 performative verbs are used to express speech acts such as recommend, nominate, and proclaim. Additional speech acts carried out by the records are not expressed with the use of performative verbs, but by declarative, interrogative and imperative sentences and by section headings, titles, and captions used in documentary forms such as referrals and recommended telephone calls. The mode (e.g., performative verb, implicit speech act, section heading) used to indicate the speech act is known as the illocutionary force indication device (IFID).

Our hypothesis is that the speech acts identified in records support record (or item) description. To test this hypothesis, item descriptions (or scope and content notes) were manually constructed for each of the records in the corpus of 120 Presidential records. The descriptions indicate the document type, the author and addressee of the record, and the act (s) and topic(s) of the record. The action in the scope and content note was created on the basis of performative sentences, implicit speech acts, structural features of the record indicating the speech act, or indirect speech acts expressed in the records. Shown below are a few examples of the descriptions.

> Signature Memorandum from Boyden Gray to the President recommending the nomination of Ronald B. Leighton to be a US District Judge.
>
> Letter from President Bush to President Mikhail Gorbachev suggesting an informal meeting.

Memorandum from President Bush to Boyden Gray requesting an analysis of the War Powers Resolution.

Letter from Susan Black to President Bush expressing appreciation for nomination and commitment to serve.

Referral Memorandum from Sally Kelley to FEMA requesting appropriate action to a letter from Beryl Anthony to the President.

The following table summarizes the results of the test.

Illocutionary Force Indicating Device (IFID)	**Number of records in which the IFID was used to create the record's description**
Explicit Performative Sentence	77
Implicit Performative Sentence	31
Speech Act Indicated by Textual Structure	11
Indirect Speech Act	1
Total	120

Almost two thirds of the actions expressed in the scope and content notes could be determined from explicit performative sentences. About one fourth of the actions could be determined from implicit performative sentences. About one tenth of the actions could be determined from speech acts indicated by structural features. Only scope and content note required the recognition of an indirect speech act. All actions expressed in the scope and content notes could be determined from the speech acts identified in the records.

4.3 Method for Recognizing the Action Conveyed by a Record

A method was proposed for recognizing the speech acts conveyed by the sentences of a record. It is outlined below.

File Conversion to Plain Text (or HTML)
Document Reader
English Tokenizer
Wordlist Lookup + enhanced wordlists
Sentence Splitter
Hepple POS Tagger + lexicon
Semantic Tagger + Named Entity Rules
Document Element Tagger + Document Element Rules (DER)
SUPPLE Parser + Document Type Grammars
Extract Record Metadata

Orthomatcher
Pronominal Coreferencer + rules for pronominal coreference
Morphological Analyzer
Supple Parser + grammar for English + interpretation rules
Speech Act Transducer

The steps prior to the dashed line annotate semantic categories in the digital record such as person names, organization names, location names, postal addresses, and dates. The three steps just prior to the dashed line annotate the intellectual elements of the documentary form of the record, recognize and interpret the documentary form of the record by parsing the intellectual elements using manually constructed grammars for the documentary form, and extract metadata from the record indicating the author(s), addressee(s), topic and date of the record. These steps have been previously implemented and experimentally evaluated [Underwood and Isbell 2008; Underwood and Laib 2008].

In the steps after the dashed line, the Orthomatcher creates references from a proper noun of one type to another proper noun of the same type. For example, *Ms. April Franklin* recognized as a person's name would be referenced to mentions of *April*, and *Ms. Franklin* appearing later in the document. The results of the Orthomatcher are needed for pronominal coreference. Pronominal coreference involves finding the proper antecedent for the following types of pronouns:

personal: I, we, you, me, him, her
possessive: my, your, our
reflexive: myself, yourself

Pronominal coreference resolution is important to speech act recognition because without finding the proper antecedent one cannot know whether it is the author of a document who is performing the speech act or whether the author is commenting on the speech act of some other person(s). The knowledge of the author(s) and addressee(s) of documents provided by document type recognition is used to determine the referents of first person (I, we, my, our) and second person pronouns (you, your).

The Morphological Analyzer takes as input a document with the parts of speech identified for each token. It identifies the lemma and an affix of each token and adds them as features of the Token annotation. The results of the morphological analyzer are needed by the SUPPLE parser.

To recognize speech acts one also needs to be able to parse the sentences in the document to identify the sentences that have first person pronouns combined with performative verbs. To accomplish this, the SUPPLE parser and an English grammar is used [Gaizauskas et al 2005].

The transduction of the syntactic structure of a sentence into an annotation of the speech act of the sentence is accomplished by the Java Annotation Pattern Engine (JAPE) and so-called JAPE rules. The JAPE rules are processed by the Java Annotation Pattern Engine in phases. The rules in the first phase are processed before the rules in the second phase, and so on. Five phases are anticipated.

1. Annotation of speech acts expressed in performative sentences
2. Annotation of implicit speech acts
3. Annotation of speech acts indicated by text structure
4. Annotation of indirect speech acts
5. Annotation of the primary speech act(s) performed by the record

The first phase matches sentence patterns for performative sentences with sentences in a record, and when there is a match, constructs a representation of the illocutionary force and proposition of the sentence. The pronouns *I, you, we, your, my* and *our* will have been referenced to author's names and addressee's names by an enhanced Pronominal Coreferencer.

The second phase of the speech act transducer identifies declarative sentences that are not performative sentences, imperative sentences and interrogative sentences. It associates them with the illocutionary forces *assert, request,* and *ask(2),* respectively, and constructs a representation that includes the proposition of the illocutionary force.

Speech acts are sometimes represented by text structure rather than in sentences. For instance, the speech act of recommending may be indicated by a section heading or a run-in paragraph head. The procedure for document type recognition produces annotations of a record representing its text structure. The third phase of speech act recognition examines that structure to determine whether it is indicating a speech act, and if so constructs a representation of that act.

The fourth phase of speech act transduction uses rules to recognize indirect speech acts. Indirect speech acts are commonly used to make requests and to reject proposals. A few occurrences of indirect speech acts were discovered during the analysis of the corpus. They primarily consisted of making requests by asking a question or making an assertion. For instance, "Would you prepare for me a short analysis of the War Powers Resolution?" is a request for action in the form of a question.

The first four phases of the speech act transducer produce annotations of the speech acts performed by individual sentences and sequences of sentences in the record. The last phase contains JAPE rules that transform these annotations into annotations of the speech act(s) performed by paragraphs, sections, and the entire record. These annotations will include elements such as the following.

[paragraph(e1), act(e1, F1), proposition(e1, P1)]
[section(e2), act(e2, F2), proposition(e2, P2)]
[document(e3), act(e3, F3), proposition(e3), P3)]

These annotations are lists of predicates that can be saved as an XML annotated copy of the original record and passed to software modules for automatic record description or checking for access restrictions.

4.4 Automatic Recognition of Performative Sentences

The implementation of the method needed for recognizing the performative sentences in e-records was initiated (phase 1 of the speech act transducer). The lexicon provided with GATE contains about 17,000 words. Some performative verbs and their nominalizations are not included in the lexicon. Additional performative verbs from Vanderveken [1990] and their nominalizations were added.

The performative sentences in the 120 record corpus were analyzed to determine the sentence patterns in which a performative verb occurred. JAPE rules have been constructed that match these patterns against the performative verbs and parse trees created by the SUPPLE parser. If there is a match, a speech act representation is created that includes the name of the illocutionary force and the proposition of the speech act. These rules were tested against 50 records selected from the corpus of 120 records.

Vanderbeken has defined 271 performative verbs. The JAPE rules that we have created apply to only 62 performative verbs—41 from Vanderbeken and 21 identified by the author (Those defined in section 3.1 of Underwood [2008]). The Public Papers of the Presidents [1991-2005] were searched for examples of the other 231 performative verbs defined by Vanderveken. Examples were identified for an additional 111 performative verbs defined by Vanderveken. Examples were also found of 28 additional performative verbs not defined by Vanderveken [Underwood 2009b]. These additional examples of performative sentences will be analyzed to refine the sentence patterns in the JAPE rules used to recognize performative sentences.

Next, rules will be constructed for recognizing implicit speech acts and speech acts indicated by document structure, for example, section headings and captions. Eventually, an experiment using e-records from the Bush Presidential Records Collection will be conducted to evaluate the performance of the speech act recognition method.

5. Topics of Discourse and Archival Description

Archival descriptions summarize the contents of record series, folders and individual records. They are useful to researchers when browsing a catalog describing an archival collection. They are also useful to archivists when they use document retrieval systems to locate electronic records relevant to a FOIA request. Text-based document search and retrieval does not obviate the need for archival descriptions. If there were no description of the items, the archivist or researcher would have to open each item in the results set of a search and read part of it to determine whether the file was a record that was actually relevant to the FOIA request.

To adequately describe a record, an archivist needs to determine the topic(s) of the record. A collection cannot be described at the record level until all the records have been read and the topics identified. To provide earlier access to a collection, a tool is needed to automatically identify the topics of records for use in creating the archival description.

Underwood [2009a] surveyed the Linguistics literature for theories and models of discourse topic. The current literature in automatic summarization was also reviewed. Finally, a method for identifying topics of Presidential e-records was outlined that would support archival description.

5.1 Discourse Topic

The topics of discourse are what the parts or the whole of a discourse are about. The macrostructure theory of discourse introduced by van Dijk [1977, 1980] explains discourse topic as follows. Syntactic rules for clauses and sentences and the meaning of the words of the sentence and sequences of sentences in a paragraph form the microstructure of a text. Propositions can be derived from this microstructure. Macrorules translate these sequences of propositions into a smaller set of more general propositions by deleting propositions that are less important for the overall meaning of the text, by generalizing propositions and by constructing new propositions that replace the meaning of sequences of propositions. The result is a set of macropropositions that summarize the text. The macropropositions make up the macrostructure of the text. Macrostructures are textual structures that form the global meaning of a text. Some indicators of macropropositions are titles of text, summaries, section headings and topical sentences. These are the discourse topics. Van Dijk and Kintsch [1983] proposed a situational model that incorporates a reader's background knowledge and pragmatic, rhetorical structures such as narrative and argumentative schema into the macrostructure theory of discourse.

One of the problems with the macrostructure theory of discourse is that the macrorules that apply to the microstructure are underspecified. The propositional macrostructures cannot be automatically computed [Kintsch 2002]. However, Kintsch has shown that macrostructures can be derived from the text using latent semantic analysis. The meaning of sentences is represented as a vector in a semantic space. Those vectors that have the highest typicality scores and relate most to the overall text are identified as the macropropositions. This refinement of the theory has been computationally implemented in the Construction-Integration (CI) model [Kintsch 1998]. The CI model consists of two phases. In the construction phase an approximate semantic model is locally constructed based on the text and the reader's background knowledge. In the integration phase, irrelevant and redundant information is filtered out of the initial model. An activation network is used to boost strong links between propositions and dampen weak links.

In Grosz and Sidner's Theory of Discourse Structure (GST) [1986], discourse structure is a tree of recursively embedded discourse segments. Each discourse segment is characterized by a primary intention, which is called the discourse segment purpose (DSP). There are two intention-based relations that hold between the DSPs of two discourse segments: *dominance* and *satisfaction precedence*. When a discourse segment purpose DSP1 provides part of the satisfaction of a discourse segment purpose DSP2 it is said that there exists a dominance relation between DSP2 and DSP1, i.e., DSP2 *dominates* DSP1. If the satisfaction of DSP2 is conditioned by the satisfaction of DSP1, it is said that DSP1 *satisfaction-precedes* DSP2.

In Mann and Thompson's Rhetorical Structure Theory (RST) [1988], discourse structures are represented with rhetorical structure trees. Sibling nodes in the trees represent contiguous text. Contiguous non-overlapping text spans are either a nucleus or a satellite. The nucleus expresses what is more essential to the writer's purpose than the satellite. A rhetorical relation is a relation that holds between two non-overlapping text spans. Each textual span can be connected to another span by only one rhetorical relation. Rhetorical relations can be assembled into rhetorical structure trees (RS-trees) on the basis of five structural constituency schemata. Arrows in the tree are labeled with the name of the rhetorical relation that holds between the units over which the

relation spans. These relations have such names as Motivation, Enablement, Sequence, and Contrast.

Marcu [1997] developed a method of text summarization called rhetorical parsing. It uses a technique to determine the rhetorical relations among text segments (discourse units) based on discourse cues and punctuation. For instance, the discourse cue word *although* indicates either a *Concession* or *Elaboration* relation with a neighboring discourse unit. The rhetorical trees that this method produces indicate changes of topic in the text. The rhetorical relations also indicate the central discourse unit in a section of related discourse units.

5.2 Computational Models of Summarization

DARPA, ARDA and NIST have sponsored a number of conferences in which research groups apply their text summarization technologies to common corpora and summarization tasks and submit the results for evaluation. These conferences include the TIPSTER Text Summarization Conference [Mani et al 1998], the Document Understanding Conferences [DUC 2001-2007] and the Text Analysis Conference [TAC 2008].

Most current systems for summarization are domain independent and share the following three steps:

1. [Topic segmentation] Identify the discourse units.
2. [Topic identification] Extract from each unit the most important sentence.
3. [Summary formation] Combine the sentences to form a summary

For text segmentation, a technique such as latent semantic analysis can be used to determine inter-sentence similarity and boundaries of text segments by divisive clustering [Choi 2000, Choi et al 2001]. Sentence selection for topic identification often involves weighting sentences based on co-occurrence with words appearing in titles, headings and in the first sentences of paragraphs. Techniques from natural language generation are then used to process the sentences into a coherent and readable summary. The current literature of text summarization techniques is reviewed in Underwood [2009a]. Suffice it to say that most systems rely on shallow text processing techniques and statistical methods. It is widely believed that natural language summarization based on syntactic, semantic, and pragmatic methods will not scale up to large collections.

An ongoing issue is performance evaluation. The primary metric used in the Document Understanding Conferences was *coverage*. Coverage measures the overlap between a human summary (the model) and a target summary (either automatically generated or human authored). The model summary is split into clauses and a human compares the model and target summaries. For each clause in the model, the human judges the percent (0 to 1.00 in increments of .20) of the content of the model clause that is contained in the target summary. The coverage for the entire summary is the average coverage of model clauses.

Another of the metrics used in the DUCs is the ROUGE metric (Recall-Oriented Understudy for Gisting Evaluation). This metric calculates n-gram overlaps between automatically generated summaries and human summaries. A high level of overlap is interpreted to indicate a high level of shared concepts between the summaries.

5.3 Analysis of the Topics of Presidential E-records

Much of the reported research in computational models of summarization (or abstracting) has applied the models to scientific documents or newswires. Scientific documents are an expository, assertive discourse that focuses on explanations or interpretations of topics. Newswires are narratives of the particulars of an act or an event or a sequence of events.

There are Presidential records that are narratives, for example, a memorandum describing the events of a legislative hearing, or a White House Press Release describing an act of the President. There are also Presidential records that are expository, for example, a memorandum interpreting or explaining a bill before Congress. However, there are also Presidential records that convey acts, for example, Presidential directives or declarations, memoranda that recommend a course of action, speeches that express attitudes. The archival description (or summary) of such records should indicate the speech act(s) as well as the topics of the records. Hence, the computational model must also recognize the speech acts.

A corpus of 50 Presidential records of 15 document types was analyzed in terms of the discourse structure and discourse topics [Underwood 2009a]. The analysis includes an identification of the speech acts expressed by the author(s) of the record. The analysis also includes consideration of anaphoric reference and rhetorical relations. Attempts are made to create short archival descriptions of these records that include an indication of the primary topic(s) of the record.

Some of the initial findings of this analysis are that the identification of the topic and/or act of a record is dependent on documentary form. For instance, there is a structure to Presidential Proclamations, Presidential Determinations, Action-Decision Memoranda, White House Referrals, and Recommended Telephone Calls that enables easier identification of the actions and topics than in memoranda and correspondence.

The difficulties arise with informal and formal correspondence, memoranda and white papers. In the case of memoranda, there is a subject line. In most cases the subject line might suffice as the topic. But in some cases one can improve on the expression of the topic by identifying those sentences (or propositions) that include the terms or phrases from the subject line.

Informal correspondence is often a reply to a letter from a citizen expressing attitudes toward an issue or asking a question. The reply usually identifies the topic in the first line of the correspondence. Even in formal correspondence there is usually a speech act performed whose proposition indicates the primary topic of the record.

It may be possible to extend the document type recognition, metadata extraction, and speech act recognition technology described in the previous sections of this report to include topic

identification that is based on documentary form and recognition of speech acts. This domain-dependent method could apply to many of the document types, while a domain-independent method could be applied to the remaining document types.

6. Checking for Restrictions on Disclosure of Presidential E-records

Review of Presidential electronic records for possible access restrictions is an intellectually demanding task that requires page-by-page review of Presidential records. Due to the increasing volume of Presidential electronic records, the need to review these records, and the cost of the limited human resources that can be applied to the review process, the review process is an archival processing bottleneck. The objective of this research is to apply language processing and rule- based reasoning technology to the development of a tool to support archivists in review of Presidential Records for Presidential Record Act (PRA) restrictions and Freedom of Information Act (FOIA) exceptions.

During the PERPOS Project, a representative sample of 150 Presidential e-records was analyzed to determine features of those records that were important in determining whether they were records that had no restrictions, were personal record misfiles, or were subject to PRA restrictions. During that project an access restriction checker was designed and prototyped [Harris et al 2005].

Figure 11 illustrates the components of the Access Restriction Checker. When assistance in checking for access restrictions is requested, metadata such as document type, author(s), addressee(s), and dates are accessed from the manifest of the container that includes the record. This metadata has been previously determined during automatic description and saved in the manifest. These facts about an e-record are asserted into the working memory of the Java Expert System Shell (JESS) [Friedman-Hill 2004]. The JESS Inference Engine checks the antecedent criteria of rules for identifying possible access restrictions against the facts in the working memory. If the criteria include background knowledge, it is sought in a long-term memory of background knowledge. The background knowledge includes names and titles of the President's advisors, names and titles of White House staff, names, titles and dates of nominations of Presidential nominees to Federal office and names of Senators and Congressmen of the 101st and 102nd Congress. If all the criteria of a rule are satisfied, the consequent part of the rule asserts facts, which may include conclusions about possible access restrictions, into the working memory. The conclusions are then displayed to the review archivist.

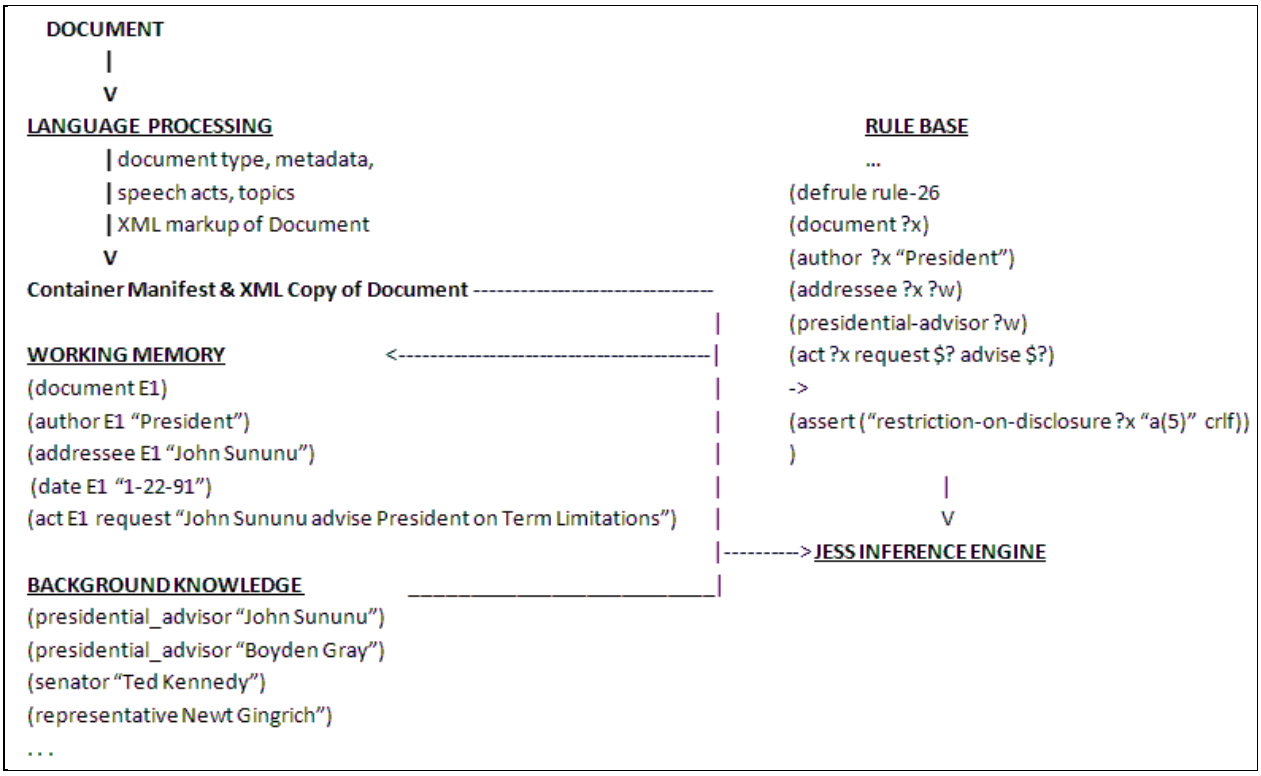

Figure 11. Components of the Access Restriction Checker

The technologically limiting factors in checking for access restrictions are the language processing capabilities to identify semantic categories such as social security numbers, birth dates, home addresses and telephone numbers, and to recognize document type and extract metadata such as author's and addressee's names. The annotation of semantic categories and recognition of a number of document types and the extraction of record metadata have now been achieved. These capabilities have been interfaced to PERPOS and a batch processing capability has been added to PERPOS so that during automated archival description of the files in a container, the recognized document types and extracted metadata can be stored in the manifest of the container. A next step is to save the annotated document in the container so that the language processing does not have to be re-performed while checking for access restrictions.

The capabilities to recognize speech acts and identify the topic(s) of records are needed to check for such restrictions on disclosure as a(2), appointment to Federal office, and a(5), confidential advice. When these capabilities are further advanced, the results of these tasks can be stored in the manifest or in an annotated copy of the record.

7. Pilot Testing of FOIA Processing Using PERPOS

The Presidential Electronic Records PilOt System (PERPOS) is a prototype archival repository and archival processing system for Presidential e-records [Underwood et al 2006]. It is a research environment for discovering and addressing issues in review, description, preservation, and

search and retrieval of electronic records. It also provides an environment for investigating advanced technologies supporting archival decisions in processing Presidential e-records.

PERPOS version 3.1 was installed at the Bush Presidential Library in College Station, Texas. Two archivists tested the functionality of PERPOS as it supports FOIA Processing. A report was written that records their observations, suggestions and conclusions [Carter et al 2007]. The results of the pilot test are:

- The conclusion that the prototype substantially provides the functions needed to support FOIA processing of e-records;
- The identification of a number of program bugs, which, were fixed;
- The identification of additional features that would better meet the needs of archivists in FOIA processing of e-records.

Adding all of the features that were needed was beyond the scope of this task. However, a few of the features added to PERPOS as a result of the pilot test are summarized below. They are documented in the Reference Manual for PERPOS version 3.2 [Underwood et al 2007].

Preprocess Containers. In PERPOS Version 3.1, when e-records were accessioned, the Archival Processing Tool would identify the file type and calculate the Secure Hash Code of each file as it was added to the tar file container. The result was that an archivist might have to wait a minute or more between accessioning each container of e-records. To speed up accessioning, the identification of the file types and calculating the SHA-1 of the contents of each record is performed after accession. A new option called *Preprocess Containers* has been added to the Tools menu in the Archival Repository Tool.

Now, identification of the file types and calculation of the SHA-1 is done in batch mode. Preprocessing loops through all newly accessioned containers identifying the file types and calculating the SHA-1 for each file in the container. Preprocessing is necessary for filtering. Filtering compares a file's file type to a list of blocked file types and compares the file's SHA-1 is to a list of SHA-1s for operating system and application software files. The file type is also used to identify the proper tools to use to view, extract, convert, or redact a record. With this update to the Archival Repository Tool, containers that have not been preprocessed cannot be checked out for processing

Automatically Filter Containers. One of the obstacles that archivists at the Bush Presidential Library faced in pilot testing was the need to have filtered copies of the Bush hard drives before performing searches for records relevant to FOIA requests. The archivists have now filtered about 50 of the 550+ copies of Bush PC hard drives. They now have the confidence that the files that are blocked by the filter are just operating system and software applications and those that pass through the filter are all of the user-created files (plus some office software applications that are not yet included in the filter). The archivists are now confident that if they discover some operating system or software application files during review, they can mark them for transfer to the corresponding container of operating system and software applications. The files marked for transfer can then be removed from the reviewed containers, leaving just the user-created files. An

option has now been added to the Archival Repository Tool to automatically filter all containers that have been accessioned and that have not yet been filtered.

Enhanced FOIA Search. The archivists indicated that when searching the collection of records in the PERPOS repository they needed the capability to limit the search to e-records in a particular collection, for example, Presidential or Vice Presidential, or a particular office or record series. Figure 12 shows the new user interface to FOIA Search that provides this capability.

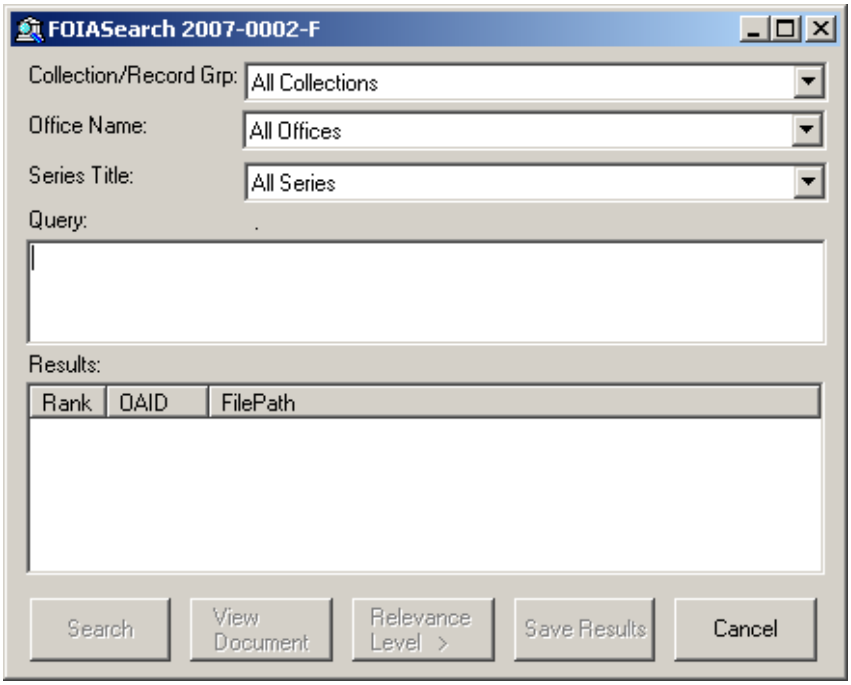

Figure 12. New user interface to FOIA Search

8. Processing National Security Classified Records and Migration of Archival Systems

8.1 Assessment of PERPOS for Processing National Security Classified Records

As indicated in the previous section, (PERPOS is a prototype archival repository and archival processing system for Presidential e-records. It is a research environment for discovering and addressing issues in review, description, preservation, and search and retrieval of electronic records. PERPOS was developed to support archivists in processing sensitive, unclassified Presidential records.

The question arose, what are the security requirements that would need to be met to certify and accredit an archival system for processing classified presidential e-records? The Information Assurance staff at the National Center for Critical Information Processing and Storage (NCCIPS)

assessed the PERPOS software as to its possible certification and accreditation for processing records at the TOP SECRET level of classification and to indicate the course of action for this task. PERPOS Version 3.1 was installed on a Dell Optiplex GX620 and this was delivered to NCCIPS. Source code and a Reference Manual for PERPOS were also provided to NCCIPS and the operation of the system was demonstrated to NCCIPS staff.

The NCCIPS IA staff's initial assessment indicates that PERPOS can be sufficiently secured to warrant certification and accreditation [NCCIPS 2007]. However, their report detailed concerns and recommendations in three areas:

- Design Level Concerns and Recommendation
- System Level Concerns and Recommendations
- Certification & Accreditation Concerns and Recommendations

As concerns Design Level concerns and recommendations, PERPOS was developed using Visual Basic 6 (VB6) and the Visual Studio development environment and a methodology known as evolutionary prototyping. Visual Basic is a language and environment that supports rapid application development. The NCCIPS report points out that Microsoft's Visual Studio has evolved into a development environment that no longer supports development of source code in the original VB6 syntax. They recommended that the code be migrated to a current development environment. Version 3.1 of PERPOS operated on a Windows 2000 platform. The NCCIPS report observed that the Windows 2000 platform was nearing its "end of life" for system-level and security-level support, and recommended that it be migrated to a more current version of the Windows operating system or to Linux.

PERPOS, version 3.1, is a stand-alone system that supports the accession, storage, preservation, review and description of e-records at the SENSITIVE level. The NCCIPS report points out that if PERPOS must support the policies and procedures that govern the transitions from highly classified to lower classification levels and publicly releasable records, then multilevel-security should be incorporated in the design so it has a reasonable chance of certification.

As mentioned earlier, PERPOS was developed using a development methodology known as evolutionary prototyping (evolutionary systems development or operational prototyping). The NCCIPS IA staff concluded that "Evolutionary prototyping is a valid development strategy, especially given the uniqueness of the domain of knowledge the system is designed to automate."

As concerns System Level concerns and recommendations, NCCIPS initial system-level assessments were performed with security scanning tools. NCCIPS IA staff identified several high-risk issues in the areas of user accounts, password requirements, and patches/service packs. More than 70 medium risk issues were identified in the areas of client signing, Federal Information Processing Standards compliance, LanManager configuration, password settings, anonymous user rights, user account policies and guest accounts. There were also many low risk issues identified.

The NCCIPS IA staff identified a number of ports that were open that should be blocked. The NCCIPS IA staff noted that all the lettered drives in the system are shared at the root level. They

note that file sharing should be restricted to subdirectories that do not have system files or system roles.

As regards Certification & Accreditation concerns and recommendations, the NCCIPS IA staff recommends Type Accreditation. Using this approach, one system is taken through the entire certification and accreditation process. Once the Type Accreditation is obtained, similar instances of the same system may be stood up in other locations utilizing the same Authority to operate. They recommend that the test and development system upon which the coding and testing is performed be more securely configured. This will reduce the level of rework that may be required once security measures are applied to support accreditation.

A report was prepared that responded to the concerns and recommendations of the NCCIPS report [Underwood 2007]. That report, in particular, addresses multilevel security issues.

The system level security issues were addressed in part by implementing security policies for user accounts, password requirements, patches/service packs, and virus detection updates. A procedure was defined for blocking open ports during installation and hardware configuration. Shared files were relocated to subdirectory Documents and Settings\All Users\Archival Tools that did not include any system files. These security features and procedures are documented in a PERPOS Administrator's Guide [Laib et al 2007]. NCCIPS' recommendation that the PERPOS code be migrated to a current development environment is discussed in the next section.

8.2 Migration of PERPOS Hardware and Software Platform

Laib and Underwood [2008] describe the user interfaces, class structure and data models of the PERPOS system. An exercise in migration from Visual Basic 6 (VB6) to Visual Basic 2005 operating in the Net Framework was performed. There are three steps to the migration from VB6 to VB8. First, use a tool like the Microsoft Code Advisor for Visual Basic 6 that gives advice on migration to Visual Basic 2005. Second, the Microsoft Upgrade Wizard included in VB2005 is run to translate as much code to VB2005 as possible. Third, the code is refactored (or cleaned up) manually or using a refactoring tool.

An exercise in migration from VB6 to JAVA using VB Converter was also performed. There are three steps in the conversion. First, run the VB Analyzer to determine unsupported controls and special controls that are used. Second, run the converter and correct any syntax errors. Third, run the converted component in a Java-enabled browser and correct any errors.

The two exercises resulted in components that were functionally the same as the components written in VB6. The migration tools were judged useful, though substantial manual recoding was necessary.

It is concluded that to improve the maintainability of PERPOS, the more complex projects of the PERPOS architecture should be refactored into smaller, simpler classes. Migration tools are available to support the migration of the Visual Basic 6 code to Visual Basic 2005 in the .NET framework or to Java. Due to interoperability of VB6, VB2005 and Java through the Common

Object Model (COM), it is possible to migrate the VB6 code incrementally, rather than all at once. Since all our research prototypes are currently being implemented in Java, we are inclined to migrate the VB6 code for PERPOS to Java.

Visual Basic 6 only supports interface inheritance. Moving to an object-oriented programming language such as VB 2005 or Java will enable us to achieve true class inheritance, that is, both interface and implementation inheritance.

The PERPOS data models are implemented in Oracle, Microsoft Access and a text data file. The persistent databases, such as the Archival Repository Tool database, should be migrated to Oracle for consistency and scalability.

9. File Format Identification

File format identification is a core requirement for digital archives. Such identification is needed to insure that the files received from a creator have the expected file formats so that the archive is able to preserve the files. Knowledge of the file formats is necessary to insure that viewers/players are available for the files, for conversion of legacy file formats into standard, current or persistent object file formats, for extraction of files from archive files, and for repair of damaged files.

There are several promising technologies for automatic file format identification, but their utility in archival applications needed to be demonstrated. Among these promising technologies is the UNIX file command [OpenBSD 2009a]. Underwood [2009c] describes extensions to the file command and magic file that enhance their utility for file format identification in archival systems.

A *file type* (or *file format class*) is class of files with the same file format. A *file format signature* is invariant data in a file format that can be used to identify the file type (or format) of a file. In the UNIX operating system (including flavors such as BSD, Linux and Solaris), file signatures are referred to as *magic numbers*. The tests for magic numbers are stored in a text file known as the magic file [OpenBSD 2009b]. The magic file for version 4.21 of the file command contains tests for approximately 2000 file types.

9.1 File Format Library

A File Format Library (database) has been created to manage information about file formats. This information includes file format name, MIME type, PRONOM Universal Identifier and file signature tests. There is a one-to-one correspondence between file formats and file signature tests. Precedence relations between file signature tests are explicitly expressed in the database. Published specifications for file formats are also collected in the library and are used to determine file signatures for the formats. When specifications have not been published for a file format, samples for files in those formats have been collected and analyzed to determine possible file signatures. File signature tests have been created for more than 800 file formats. Sample files

for more than 500 of the file formats in the library have been created or collected for testing of the file signatures. These examples are included in the library

The Library includes links to file format software resources that are needed in archival processing of digital records. These include: file viewers/players, archive extractors, file format converters, password recovery software and repairers for damaged files.

The File Format Library supports the creation of a magic file from the file signature tests in the Library. The metadata for file formats including file signature tests can be easily changed in the database rather than in the magic file. This is a substantial improvement in the flexibility of the UNIX file command and magic file.

9.2 File Format Identifier

The GTRI File Type Identifier is a graphical user interface to the file command and the magic file created from the File Format Library. The file command and magic tests have been applied to examples of 500+ file formats from the File Format Library. These tests have led to refinement of the file signature tests and discovery of the precedence relationships among file signature tests.

Future efforts will be directed toward creating file signatures for file formats that actually occur in the digital collections of NARA, for example, in presidential e-records that are acquired at the end of each administration. Also samples of the file formats will be collected for testing of the file command-based file format identifier, as this is a primary way of demonstrating the reliability (accuracy) of a file format identification technology.

The National Archives (TNA) of the UK provides a public registry of file format information (PRONOM). This information includes file signature patterns expressed as regular expressions. TNA also provides a tool (DROID) that uses these file signature patterns for file format identification [Brown 2006]. This approach to file type identification is also promising and seems to be primarily limited by the small number of file signature patterns in the PRONOM registry. GTRI is collaborating with TNA to enhance the content of the registry and the performance of the DROID file format identifier.

9.3 A File Format Identifier for TPAP

The NARA Transcontinental Persistent Archives Prototype (TPAP) is a collaborative research testbed in which the challenges inherent in preserving, protecting, and providing access to the electronic records are being addressed.[2] TPAP is based on the integrated Rule Oriented Data System (iRODS), a second generation data grid system providing a unified view and seamless access to distributed digital objects across a wide area network. GTRI is developing for TPAP an archival service based on then GTRI File Type Identifier.

[2] www.renci.org/focus-areas/humanities-arts-and-social-science/nara-tpap

Icommands are command-line functions for accessing iRODS data and metadata. Icommands are used for batch jobs and scripting. The UNIX file command source code is being modified to make iRODS client library calls in place of UNIX file I/O CALLS. The modified file command is called *ifile*. Initially, the file command uses the magic file provided with the file command. The magic file will be replaced with the GTRI signature file and scripts will be developed to save the results of file format identification in a jar manifest or the iRODS metadata catalog iCAT.

11. Summary of Results

The performance of the Semantic Annotator tool for annotating person's names, organization names, location names and dates has been improved by the inclusion of additional wordlists and JAPE rules. Additional semantic categories such as social security numbers, telephone numbers, postal addresses, facilities, legislative bills and statutes, governments, and relative temporal expressions are now annotated. An experiment with actual presidential e-records indicates a performance in recall, precision and F-measure of greater than .90.

A method for automatically recognizing document types and extracting metadata from e-records has been developed. This metadata can be used for indexing and searching collections of records by person, organization and location names, topics, dates, author's and addressee's names and document types, and for automatically describing items, file units and record series. In an experiment, the Document Type Recognizer successfully recognized the documentary form and extracted the metadata of two-thirds of the documents in a series of presidential e-records.

Speech acts are acts of speech or writing in which one does something just by saying something, for example, "I appoint you…", "I hereby proclaim…". One hundred twenty Presidential records were analyzed with regard to the expression of speech acts with performative verbs and speech acts about the author's past or future speech acts or other's speech acts. Sixty three different performative verbs were discovered in the corpus. The analysis confirms that performative verbs are used to express the actions carried out by records. A method has been formulated for identifying the speech acts occurring in e-records. It will be implemented and tested using records from the analyzed corpus and then experimentally evaluated

A corpus of fifty presidential records of various documentary forms was analyzed to determine the topic(s) of the records and possible techniques for automatically identifying the topics. The linguistics literature addressing discourse topic was reviewed. Current technologies for domain-independent document summarization were also reviewed. The result is a suggested approach to a combination of domain-dependent and domain-independent methods for identifying topics in presidential e-records.

Progress in implementing an Access Restriction Checker includes the interface of the prototype to the results of document type recognition and extraction of metadata about a record. Still needed is the provision to the Access Restriction Checker of the results of speech act and topic recognition.

The PERPOS prototype archival repository and archival processing system has been tested by archivists at the Bush Presidential Library in processing of Presidential records in response to FOIA requests. The results of the pilot test include (1) the conclusion that the tool substantially supports FOIA processing, (2) the identification of additional features that would better meet the needs of archivists in FOIA processing of e-records, and (3) the adaptation of PERPOS to include some of these features.

The question was posed as to the requirements for certification and accreditation of an archival system for processing security classified presidential-records. While PERPOS is an experimental prototype and was not designed to process security classified records, it supports many of the functions needed for reviewing such records. The Information Assurance staff of the National Center for Critical Information Processing and Storage assessed whether PERPOS could be sufficiently secured to warrant certification and accreditation. They concluded that it could be sufficiently secured, identified areas of concern, and made numerous recommendations.

Due to the rapid changes in computer technology, archivists must be concerned not only with the obsolescence of e-record file formats, but with the obsolescence of the operating systems, database management systems and integrated development environments of their Archival System. The Presidential Electronic Records Pilot System (PERPOS) as a case in point. Two exercises were conducted in using conversion tools to migrate Visual Basic 6 modules of PERPOS to Visual Basic 8 and to Java. The two exercises resulted in components that were functionally the same as the components written in VB6. The migration tools were judged useful, though substantial manual recoding was necessary. It was also concluded that to improve the maintainability of PERPOS, the more complex projects of the PERPOS architecture should be refactored into smaller, simpler classes.

File format identification is a core requirement for digital archives. The UNIX file command is among the most promising technologies for file type identification, but its reliability (accuracy) needs to be demonstrated. A database system for managing file format information and creating the magic file used by the file command is described. The metadata for file formats including file signature tests can be easily changed in the database rather than in the magic file. This is a substantial improvement in the flexibility of the UNIX file command and magic file. A graphical user interface has been developed for the file command. File signature tests have been created for more than 800 file formats and the reliably of the file command and file signature file is being evaluated on examples of the file formats it purportedly identifies.

12. Dissemination of Results

Journal Articles

W. Underwood. Grammar-Based Recognition of Documentary Forms and Extraction of Metadata. Accepted for publication in *The International Journal of Digital Curation*.

Conference Proceedings (Peer Reviewed)

W. Underwood, Automatic Metadata Extraction for Archival Description and Access" *SAA Research Forum Proceedings,* Society of American Archivists, San Francisco, August 26, 2008.

W. Underwood. Speech Acts and Electronic Records. *Proceedings of DigCCurr2009*, Chapel Hill, NC, April 2-3, 2009.

W. Underwood. Grammar-Based Recognition of Documentary Forms and Extraction of Metadata. 5th International Digital Curation Conference, London, December 2-4 2009.

Conference (Forum and Workshop) Presentations

W. Underwood and S. Laib, PERPOS: An Electronic Records Repository and Archival Processing System, an International Symposium on Digital Curation (DigCCurr2007). Chapel Hill NC, April 18-20, 2007.

W. Underwood, S. Isbell and M. Underwood. Grammatical Induction and Recognition of the Documentary Form or Record Types, an International Symposium on Digital Curation (DigCCurr2007). Chapel Hill NC, April 18-20, 2007.

W. E. Underwood. Metadata Extraction, Archival Description and FOIA Search in PERPOS. OCLC Western Digital Forum in San Diego, August 9-10, 2007.

B. Clement and W. Underwood. Evolution of a Prototype Archival System for Preserving and Reviewing Electronic Records. Archives 2008: R/Evolution & Identities. Society of American Archivists. San Francisco, August 26-30, 2008

W. Underwood. Natural Language Processing Applied to Archival Description. WVU/NETL/ERA Workshop on Digital Preservation of Complex Engineering, Morgantown, West Virginia, April 20, 2009.

Technical Reports

S. Laib, M. Underwood & W. Underwood. PERPOS Version 3.1: Installation Guide. Technical Report ITTL/CSITD 06-07, December 2006

S. Isbell, M. Underwood and W. Underwood. Semantic Annotation of Presidential E-Records. Technical Report, ITTL/CSITD 07-01, ITTL, GTRI, August, 2007.

W. Underwood, S. Laib and M. Hayslett-Keck. Reference Manual for PERPOS: An Electronic Records Repository and Archival Processing System, Version 3.2. PERPOS TR ITTL/CSITD 07-02, Information Technology and Telecommunications Laboratory, Georgia Tech Research Institute, September 2007.

D. Carter, B. Clement, S. Laib and W. Underwood. Results of Pilot Testing of FOIA Processing Using PERPOS. Technical Report ITTL/CSITD 07-04, ITTL, GTRI, June 2007.

W. Underwood. A Path to Certification and Accreditation of a Trusted PERPOS for Processing Classified Presidential E-Records. Technical Report ITTL/CSITD 07-05, ITTL, GTRI, May 2007. (Not for General Distribution)

W. Underwood, S. Isbell, S. Laib & M. Underwood. Advanced Decision Support for Archival Processing of Presidential Electronic Records: Annual Technical Status Report (October 2007-June 2007) Technical Report ITTL/CSITD 07-06, August 2007

S. Laib, M. Underwood and W. Underwood. PERPOS Version 3.1: Administrator's Guide. PERPOS Technical Report ITTL/CSITD 07-07, GTRI, December 2007.

S. Laib & W. Underwood. Issues in Migrating PERPOS to a New Development Environment. Technical ITTL/CSITD 07-010, October 2007 (Revised November 2008)

W. Underwood and S. Isbell. Semantic Annotation of Presidential E-records, Technical Report ITTL/CSITD 08-01, Georgia Tech Research Institute, May 2008.

W. Underwood and S. Laib. Recognition of Documentary Forms. TR 08-02, Information Technology and Telecommunications Laboratory, Georgia Tech Research Institute, Atlanta, Georgia, May 2008.

W. Underwood. Recognizing Speech Acts in Presidential E-records, Technical Report ITTL/CSITD 08-03, Georgia Tech Research Institute, October 2008.

W. Underwood, S. Isbell, S. Laib & M. Underwood. Advanced Decision Support for Archival Processing of Presidential Electronic Records: Annual Technical Status Report (July 2007-June 2008) Technical Report ITTL/CSITD 08-06, August 2008

W. Underwood. Examples of Performative Sentences in Presidential Records. Working Paper ITTL/CSITD 09-01, September 2009

W. Underwood. Extensions of the UNIX File Command and Magic File for File Type Identification. Technical Report ITTL/CSITD 09-02, September 2009

W. Underwood. Recognizing Topics in Presidential E-records. Technical Report ITTL/CSITD 09-04, Georgia Tech Research Institute, In Process.

Presentations and Demonstrations

W. Underwood. Demonstration of PERPOS and Briefing of Research Results to ERA Contactor Personnel, some Staff members of Presidential Libraries (NL), and ERA Technical Staff, February 2007.

W. Underwood. Advanced Decision Support for FOIA Processing of Presidential E-records. ERA Independent Advisory Committee, Archives I, Washington D.C. April 6, 2007.

W. E. Underwood. NLP Technologies Applied to Archival Tasks, NARA Administration Building, Allegany Ballistics Laboratory, Rocket Center, West Virginia, December 14, 2007

W. E. Underwood. NLP Technologies Applied to Archival Tasks, NARA, Archives II, College Park, April 2008

W. Underwood. Evolution of a Prototype Archival System for Preserving and Reviewing Electronic Records Emory University, Atlanta, Georgia, Faculty from the Maryland Institute for Technology in the Humanities, the Harry Ransom Center of the University of Texas, and Emory University Library. Participants in a NEH planning grant to explore born digital archives and scholarly research, December, 2008.

W. Underwood. File Format Identification and Archival Processing. Presentation to Representatives from NARA, The National Archives of the UK, Harvard University Library, and the Canadian Archives as a part of the Digital Formats Registry Initiative. Archive I, Washington, D.C. February 6, 2009

W. Underwood. Advanced Decision Support for Archival Processing of Presidential E-Records: Results and Demonstration To be presented in November, 2009 at Archive II, College Park, Maryland

References

[Brown 2006] A. Brown, Automatic Format Identification Using PRONOM and DROID, DTTP-01, The National Archives, March 2006.

[Carter et al 2007] D. Carter, B. Clement, S. Laib and W. Underwood. Results of Pilot Testing of FOIA Processing Using PERPOS. Technical Report ITTL/CSITD 07-04, ITTL, GTRI, June 2007.

[Choi 2000] F. Y. Y. Choi. Advances in domain-independent text segmentation. Proceedings of the North American Chapter of the Association for Computational Linguistics. Seattle, May 2001, pp. 26-33.

[Choi et al 2001] F. Y. Y. Choi, P. Wiemer-Hastings & J. Moore. Latent Semantic Analysis for Text Segmentation. *Proceedings of Empirical methods in natural language Processing*, 2001. Pp.109-117.

[Cunningham et al 2007] H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan, C. Ursu, M. Dimitrov, M. Dowman, N. Aswani, I. Roberts, Y. Li, and A. Shafirin. Developing Language Processing Components with GATE Version 4, The University of Sheffield, December 1, 2007.

[DUC 2001-2007] Proceedings of the Document Understanding Conference. http://www-nlpir.nist.gov/projects/duc/pubs.html

[Friedman-Hill 2004] E. J. Friedman-Hill. Jess, The Rule Engine for the Java Platform, Version 6.1p7. SAND98-8206 (revised), Sandia National Laboratory, Livermore, CA, 7 May 2004.

[Gaizauskas et al 2005] R. Gaizauskas, M. Hepple, H. Saggion, M. A. Greenwood, and K. Humphreys. SUPPLE: A Practical Parser for Natural Language Engineering Applications, *International Workshop on Parsing Technologies*, Vancouver, Oct. 2005.

[Grosz and Sidner 1986] B. Grosz and C. Sidner, 1986, "Attention, Intentions, and the Structure of Discourse", *Computational Linguistics*, Vol. 12, No. 3, pp. 175-204.

[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.

[HCGO 1973] House Committee on Government Operations, Executive Classification of Information — Security Classification Problems Involving Exemption (b)(1) of the Freedom of Information Act, 93d Cong., 1st sess., House Rept. 93-221, 1973, p. 50 http://www.fas.org/sgp/library/quist/chap_6.html

[Isbell et al 2007] S. Isbell, M. Underwood and W. Underwood. Semantic Annotation of Presidential E-Records. Technical Report, ITTL/CSITD 07-01, ITTL, GTRI, August, 2007.

[Kim and Ross 2007] Y. Kim, and S. Ross. "The Naming of Cats": Automated Genre Classification, *The International Journal of Digital Curation*. Vol. 2, No. 1, June, 2007, 49-61.

[Kintsch 1998] W. Kintsch. *Comprehension. A paradigm for cognition*. Cambridge: Cambridge University Press, 1988.

[Kintsch 2002] W. Kintsch. On the notions of theme and topic in psychological process models of text comprehension. In M. Louwerse & W. van Peer (Eds.) *Thematics: Interdisciplinary studies* (Amsterdam: Benjamins, 2002, pp. 157-170.

[Laib et al 2007] S. Laib, M. Underwood and W. Underwood. PERPOS Version 3.1: Administrator's Guide. PERPOS Technical Report ITTL/CSITD 07-07, GTRI, December 2007.

[Laib & Underwood 2008] S. Laib & W. Underwood. Issues in Migrating PERPOS to a New Development Environment. Technical ITTL/CSITD 07-010, October 2007 (Revised November 2008)

[Mani et al 1998] L. Mani, T. Firmin, D. House, M. Chrzanowski, G. Klein, L. Hirschman, B. Sundheim, & L. Obrsi. The TIPSTER SUMMAC Text Summarization Evaluation. Tech. rep. MTR 98W0000138, MITRE, McLean, Virginia, 1998.

[Mann and Thompson 1988] W. Mann and S. Thompson. Rhetorical Structure Theory: Toward a functional theory of text organization.

[Marcu 1997] D. Marcu (1997). The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts.. PhD Thesis, Department of Computer Science, University of Toronto, December 1997. Also published as Technical Report CSRG-371, Computer Systems Research Group, University of Toronto.

[NARA 2009] Report on Alternative Models for Presidential Libraries Issued in Response to the Requirements of PL 110-404, September 2009.

[NCCIPS 2007] National Center for Critical Information Processing and Storage. NARA PERPOS Initial Certification and Accreditation (C&A) Assessment, IA-PERPOS-001 v1.0 April 25, 2007. (NOT FOR General Distribution)

[OpenBSD 2009a] OpenBSD Reference Manual, file(1) http://www.openbsd.org/cgi-bin/man.cgi?query=file&sektion=1

[OpenBSD 2009b] OpenBSD Programmer's Manual magic(5) http://www.openbsd.org/cgi-bin/man.cgi?query=magic&sektion=5&apropos=0&manpath=OpenBSD+Current&arch=

[Public Papers of the Presidents 1991-2005] U.S. Government Printing Office via GPO Access http://www.gpoaccess.gov/pubpapers/search.html

[Santini 2004] M. Santini. *State-of-the-art on automatic genre identification*. Technical Report ITRI-04-03) University of Brighton, UK, Information Technology Research Institute (ITRI), 2004

[Searle 1969] J. R. Searle. *Speech Acts*. Cambridge University Press. 1969.

[Searle 1979] J. R. Searle. *Expression and Meaning*. Cambridge University Press. 1979

[Stolcke 1994] A. Stolcke. Boogie: A Manual for Bayesian Object-oriented Grammar Induction and Estimation. International Computer Science Institute, Berkeley, June 1994

[TAC 2008] *Proceedings of the First Text Analysis Conference* (TAC 2008)**.** National Institute of Standards and Technology, Gaithersburg, Maryland**,** November 17-19, 2008

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 2007] W. Underwood. A Path to Certification and Accreditation of a Trusted PERPOS for Processing Classified Presidential E-Records. Technical Report ITTL/CSITD 07-05, ITTL, GTRI, May 2007 (NOT FOR General Distribution)

[Underwood 2008] W. Underwood. Recognizing Speech Acts in Presidential E-records, Technical Report ITTL/CSITD 08-03, Georgia Tech Research Institute, October 2008.

[Underwood 2009a] W. Underwood. Topics of Discourse and Archival Description of Presidential E-records. Technical Report ITTL/CSITD 09-04, Georgia Tech Research Institute, In Process.

[Underwood 2009b] W. Underwood. Examples of Performative Sentences in Presidential Records. Working Paper ITTL/CSITD 09-01, September 2009

[Underwood 2009c] W. Underwood. Extensions of the UNIX File Command and Magic File for File Type Identification. Technical Report ITTL/CSITD 09-02, September 2009

[Underwood 2009d] W. Underwood. Grammar-Based Recognition of Documentary Forms and Extraction of Metadata. 5th International Digital Curation Conference, London, December 2-4 2009.

[Underwood and Harris 2006] W. E. Underwood and B. Harris. Inferring and Recognizing the Documentary Form of Record Types. PERPOS TR ITTL/CSITD 05-8, August 2006.

[Underwood et al 2006] W. Underwood, S. Laib and M. Hayslett-Keck. Reference Manual for PERPOS: An Electronic Records Repository and Archival Processing System, Version 3.1. PERPOS TR ITTL/CSITD 06-02, Information Technology and Telecommunications Laboratory, Georgia Tech Research Institute, September 2006.

[Underwood et al 2007] W. Underwood, S. Laib and M. Hayslett-Keck. Reference Manual for PERPOS: An Electronic Records Repository and Archival Processing System, Version 3.2. PERPOS TR ITTL/CSITD 07-02, Information Technology and Telecommunications Laboratory, Georgia Tech Research Institute, September 2007.

[Underwood and Isbell 2008] W. Underwood and S. Isbell Semantic Annotation of Presidential E-records, Technical Report ITTL/CSITD 08-01, Georgia Tech Research Institute, May 2008.

[Underwood and Laib 2008] W. Underwood and S. Laib. Recognition of Documentary Forms. TR 08-02, Information Technology and Telecommunications Laboratory, Georgia Tech Research Institute, Atlanta, Georgia, May 2008.

[Underwood and Laib 2009] W. Underwood and S. Laib. Induction of Grammars for Documentary Forms, Working Paper 09-03, GTRI, In Process.

[Vanderveken 1990] D. Vanderveken. *Meaning and Speech Acts. Vol 1, Principles of Language Use*. Cambridge University Press, 1990.

[van Dijk 1977] T. van Dijk. Sentence Topic and Discourse Topic. *Papers in Slavic Phil*ology 1, 1977, pp. 49-61.

[van Dijk 1980] T. A. van Dijk. *Macrostructures*. Hillsdale, NJ: Erlbaum, 1980.

[van Dijk and Kintsch 1983] T. A. van Dijk and W. Kintsch. *Strategies of discourse comprehension*. New York: Academic Press, 1983.

[Wierzbicka 1987] A. Wierzbicka. *English Speech Act Verbs: A semantic dictionary* Academic Press 1987.