PERPOS: An Electronic Records Repository and Archival Processing System

W. E. Underwood and S. L. Laib
Georgia Tech Research Institute
250 14th St. N.W. Atlanta, Georgia 30332-0832
william.underwood@gtri.gatech.edu
sandra.laib@gtri.gatech.edu

Abstract

An Electronic Records Repository and an Archival Processing System have been developed to support archivists in processing Presidential e-records. Accession, arrangement, preservation, review, description, and creation of finding aids are supported. The data management system includes the schema for the repository as well as metadata for arrangement, review, preservation and description. The prototype system provides an environment for the experimental application of advanced information technologies to archival processes.

Introduction

The first objective of the Presidential Electronic Records Pilot System (PERPOS) project was to support archivists in gaining intellectual and physical control over the personal computer records created and used during the administration of President George H. W. Bush. A second objective is to apply advanced information technologies to support archivist's decisions in processing Presidential e-records. A software system (also called PERPOS) was developed using a method known as evolutionary prototyping. An initial prototype was constructed to learn more about the problems of separating operating system and office application software files from user-created files and viewing personal computer (PC) files in legacy file formats. Once the prototype had been used in processing actual PC files from the White House Offices and the requisite knowledge gained, the prototype was adapted to satisfy the now better-understood needs.

Archivists who used PERPOS learned that there were some files that could not be viewed. The files included password protected or encrypted files, damaged files, and files in obsolete formats for which there were no viewers. The prototype was extended to include the capabilities to recover passwords from protected or encrypted files, to use the recovered passwords to decrypt files, to repair damaged files, and to convert obsolete file formats to current or standard formats for which there were viewers. Then the prototype was used again to process Bush PC e-records, more was learned, and the prototype was re-adapted based on archivist's recommendations [3].

Then the prototype was used again, more was learned, and the prototype readapted. This process of prototype use, learning and re-adaptation repeats until the prototype system satisfies all the needs and has thus evolved into a system. The resulting, but still evolving system, is an Electronic Records Repository and an Archival Processing System [4].

In the next section the activities of accession and systematic processing are described. Then the repository data model is discussed. The PERPOS prototype system consists of two subsystems, the Archival Repository Tool (ART) and the Archival Processing Tool (APT). These two subsystems are briefly described.

Accession and Systematic Processing

Figure 1 illustrates the dataflow of Accessioning and Systematic Processing activities that are supported by the PERPOS prototype system. The numbered, labeled circles represent systematic processing activities supported by the system. The labeled parallel lines represent the kinds of information that are created, stored and used by the activities. The labels on directed edges represent the kinds of information that are inputs or outputs of an activity or stored as a result of activities and subsequently retrieved for use by other activities. The rectangles represent entities external to the PERPOS system. Stepping through the diagram in the numerical sequence of the activities, one sees the dataflow.

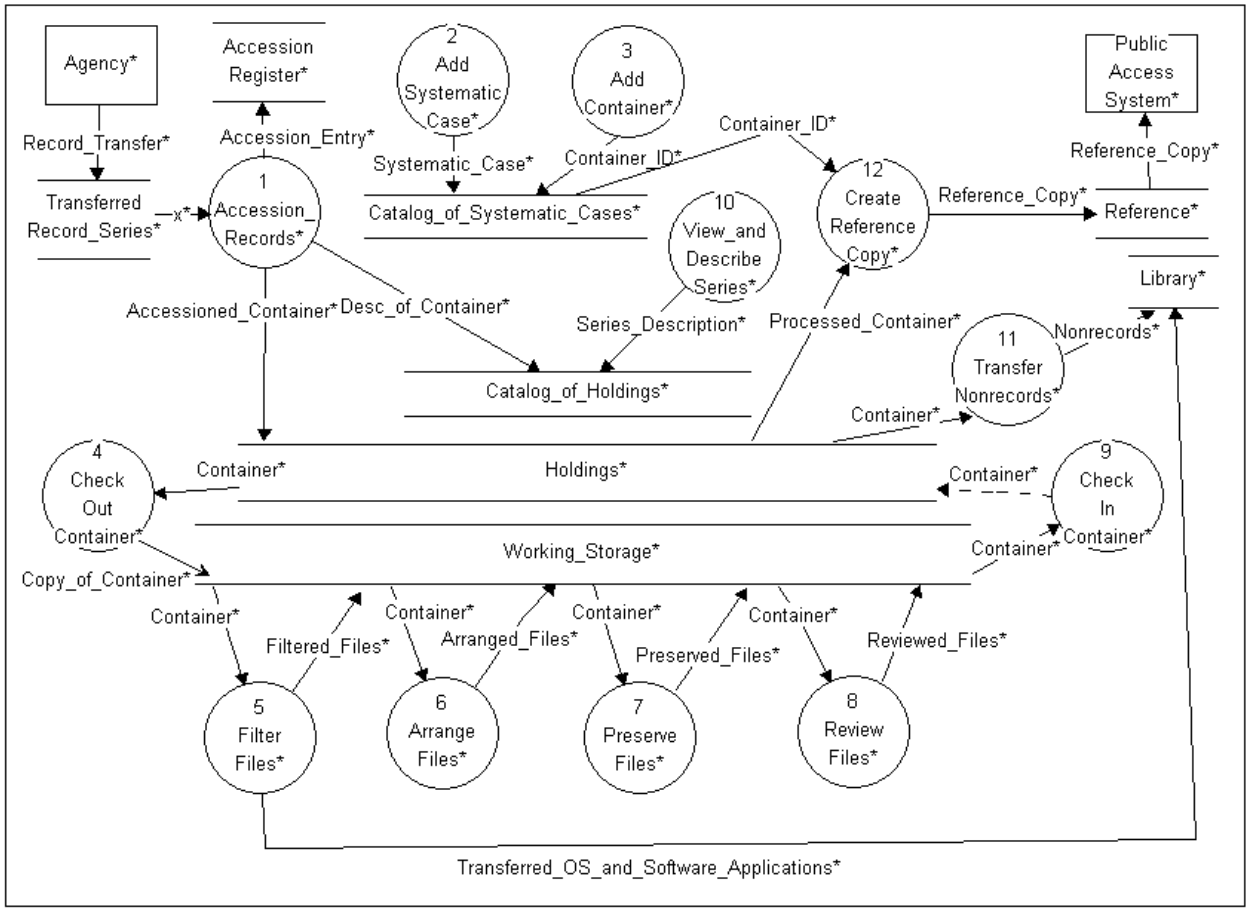

Figure 1. Dataflow Diagram for Accession and Systematic Processing Activities

Archivists at the Presidential Library must first accession the electronic records (Activity 1). Transferred files are loaded from media such as tapes or disks. An entry is made in an accession register, and files that are associated with the accession are placed into containers and stored in the Archival Repository (Holdings). Information about the accessioned containers, such as the name of the collection, office name, and record series title are also entered into the Catalog of Holdings. During accession, the archivist can open a container and browse its contents to determine this information.

The first activity in the Systematic Processing of accessioned electronic record series is for the Supervisory Archivist to schedule archival work. This involves loading the Accession register and/or Catalog of Holdings to determine which containers remain to be processed, and assigning work to archivists. Archivists start work by creating a Systematic Case (Activity 2). Once a Systematic Case has been created, the archivist accesses the Catalog and finds the containers associated with the record series they are to process. They add each container to the Systematic Case they just created (Activity 3). To process one of the containers in a Systematic case, the archivist Checks Out a copy of the container (Activity 4). This leaves the original container in the holdings area as a backup and places a copy of the container in the archivist's Work Area.

Sometimes accessioned PC record series will include the entire file system of a personal computer—operating system and application files as well as user-created files. PERPOS supports filtering file systems by blocking operating system and applications files and passing through user-created files (Activity 5). This activity could be considered a preservation function as it involves separating records and non-records. After each step of work, the archivist saves their work back to the work area.

Archivists should attempt to maintain the original order of files in a file system, but some files may not have been stored in the proper directory (folder). For instance, some word processing files that should have been stored in a Correspondence directory may have been stored in the root directory or in the directory including the word processing application. Archivists may need to perfect the arrangement by moving misplaced files into the proper directory. This activity is called arrangement (Activity 6).

Some PC files may be in obsolete or proprietary file formats that can no longer be viewed. Other files may be corrupted due to media deterioration or file transmission errors. Other files may be encrypted, so that there is a need to recover a password and decrypt the file so that it can be viewed. These activities are referred to as preservation (Activity 7).

Next, PC records must be reviewed for Freedom of Information Act (FOIA) exemptions on their disclosure to the public. They must also be reviewed for Presidential Record Act (PRA) restrictions on their disclosure (Activity 8). During this activity, Presidential records are opened, closed, or redacted. During review, an archivist may discover non-records such as software application documentation or sample files for a software application that were not removed during filtering. These can be marked for transfer to the Library. They may also discover personal records that were misfiled with the Presidential records. These can be marked as Personal Record Misfiles (PRMs) that can later be removed from the containers.

When the contents of a container have been filtered, arranged, preserved, and reviewed, the container of processed records is checked back into Holdings (Activity 9). This results in the processed container replacing the original container in the Holdings area.

When archivists have completed the preceding activities, they must describe the record series (Activity 10). This involves loading the containers in the record series, viewing and describing their contents, and determining the extent (number of files, number of bytes, or number of pages) of the processed container. Non-records that are in a container and that were marked for transfer to the Library can be removed at this time (Activity 11). Personal records marked as PRMs should also be removed from containers.

Since the master copy that is stored in the repository may contain records whose access is restricted in whole or in part, it is necessary to create a Reference Copy that includes just those records that are open to the public or redacted versions of records. To do this, an archivist creates a Reference Copy of the series containers that can be made available to the Public Access System (Activity 12).

Archivists can access the containers of processed record series to re-review closed or redacted e-records when access restrictions have expired. They can also access containers in Holdings for preservation actions, such as converting to new file formats when current file formats become obsolete.

Repository Data Model

The Electronic Records Repository consists of the Accession Register, the Catalog of Holdings and the containers in Holdings. Figure 2 shows the classes of digital objects in the Repository and their attributes.

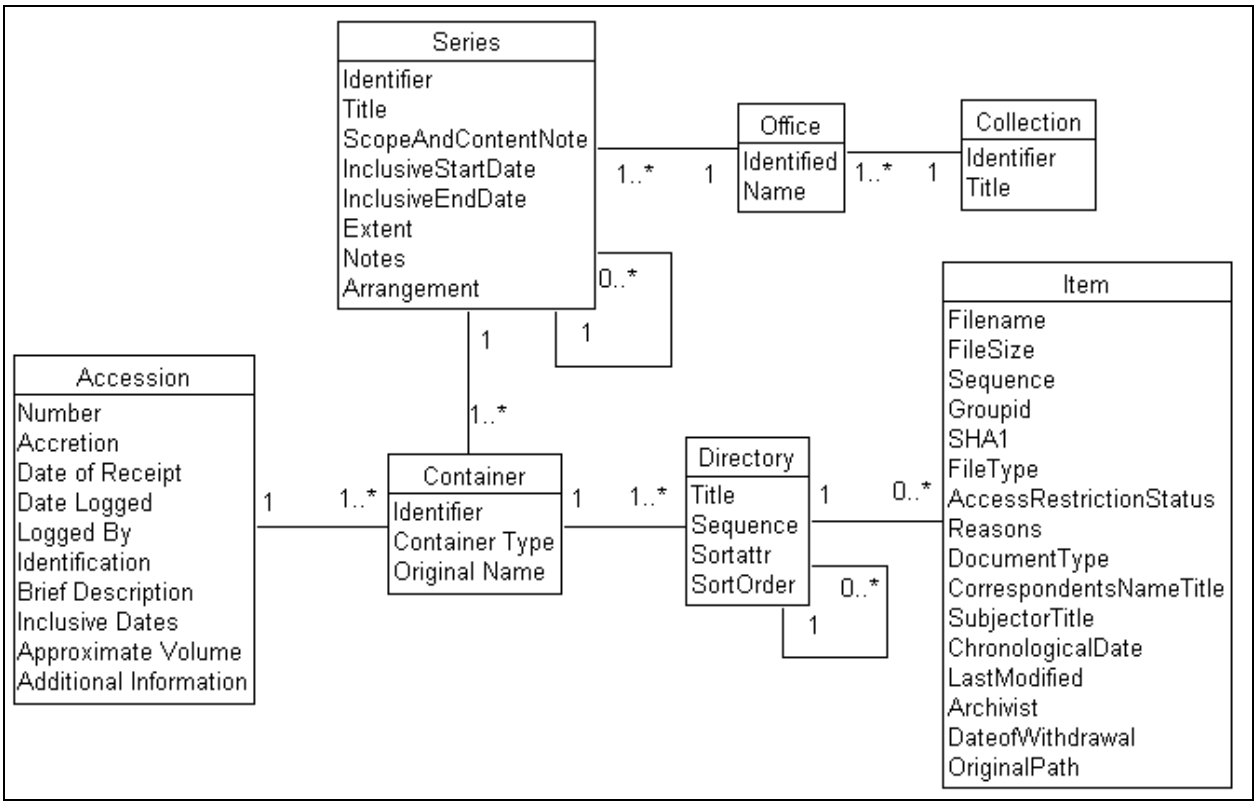

Figure 2. Repository Data Model

User Interface

PERPOS consists of two subsystems, the Archival Repository Tool (ART) and the Archival Processing Tool (APT). The Archival Repository Tool supports Accession, Systematic Case Management, FOIA Case Management and Description. Fig. 3 shows containers associated with an accession.

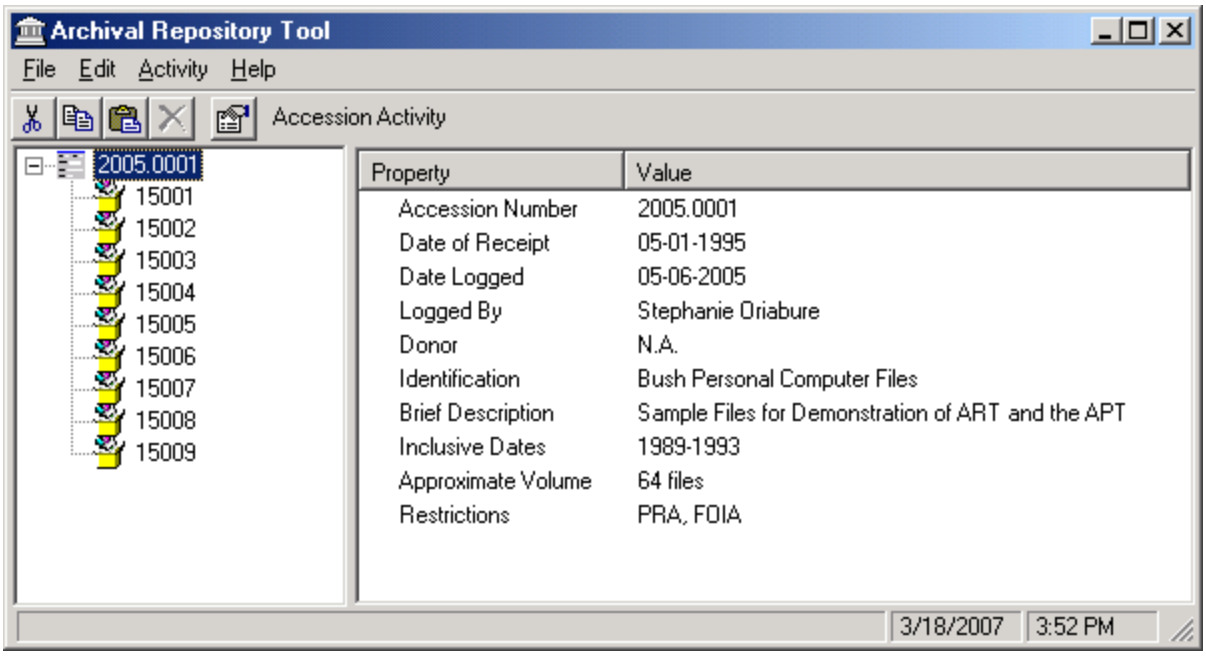

Figure 3. Containers Associated with an Accession.

Fig. 4 shows the arrangement of records in the repository after the accession of some e-records. It shows that the collection *Bush Presidential Records: Staff and Office Files* consists of *Offices* that consist of *Record Series* that consist of *containers*.

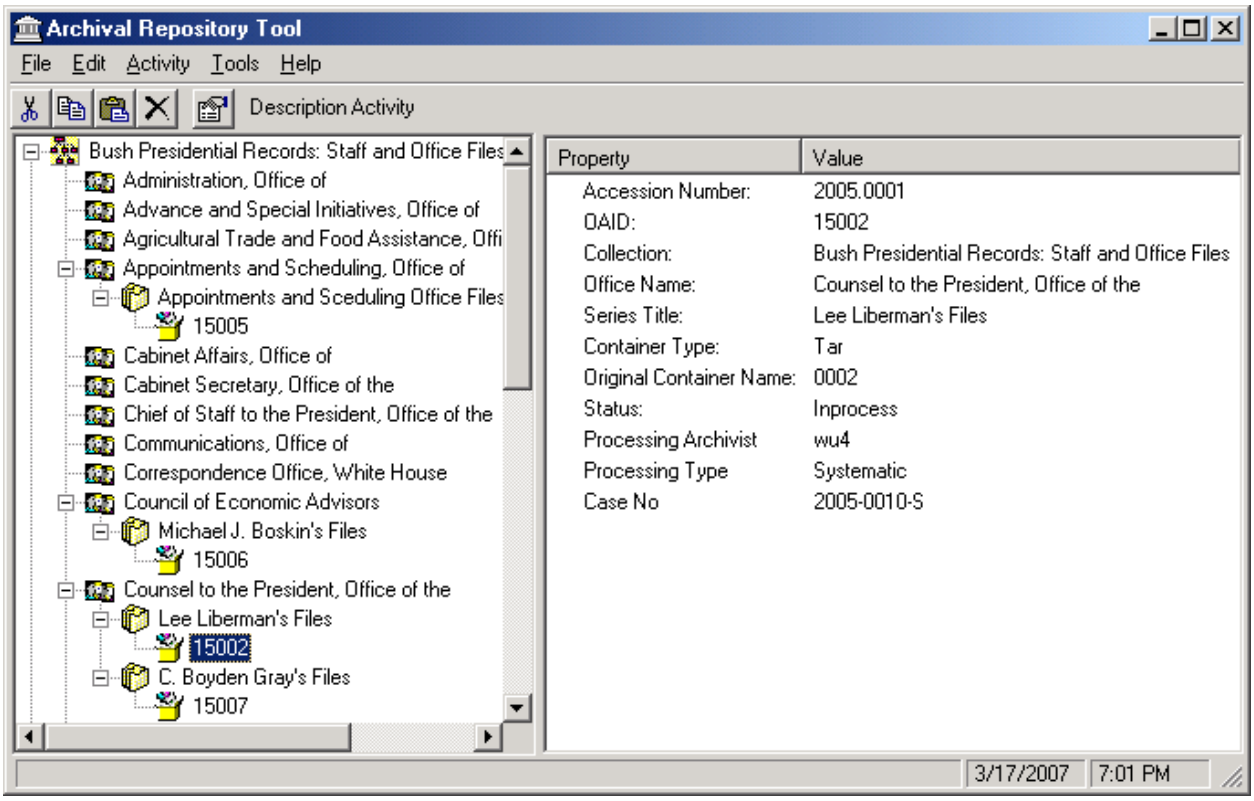

Figure 4. The Arrangement of Collections as Offices and Containers of E-records.

The Archival Processing Tool (ART)

The Archival Processing Tool supports the activities of perfecting the arrangement of a file system, converting legacy file formats to current or standard formats, recovering passwords for encrypted files and decrypting these files, repairing damaged files, and reviewing e-records to determine whether there are restrictions on disclosure.

Fig. 5 shows an example of the review activity. The left windowpane shows that a container consists of directories (or folders) and files. When the name of a file is highlighted the attributes of the file are displayed in the right windowpane. This figure shows that the e-record with filename 106.rft has been closed because or Presidential Record Act restriction a(5), Confidential Advice.

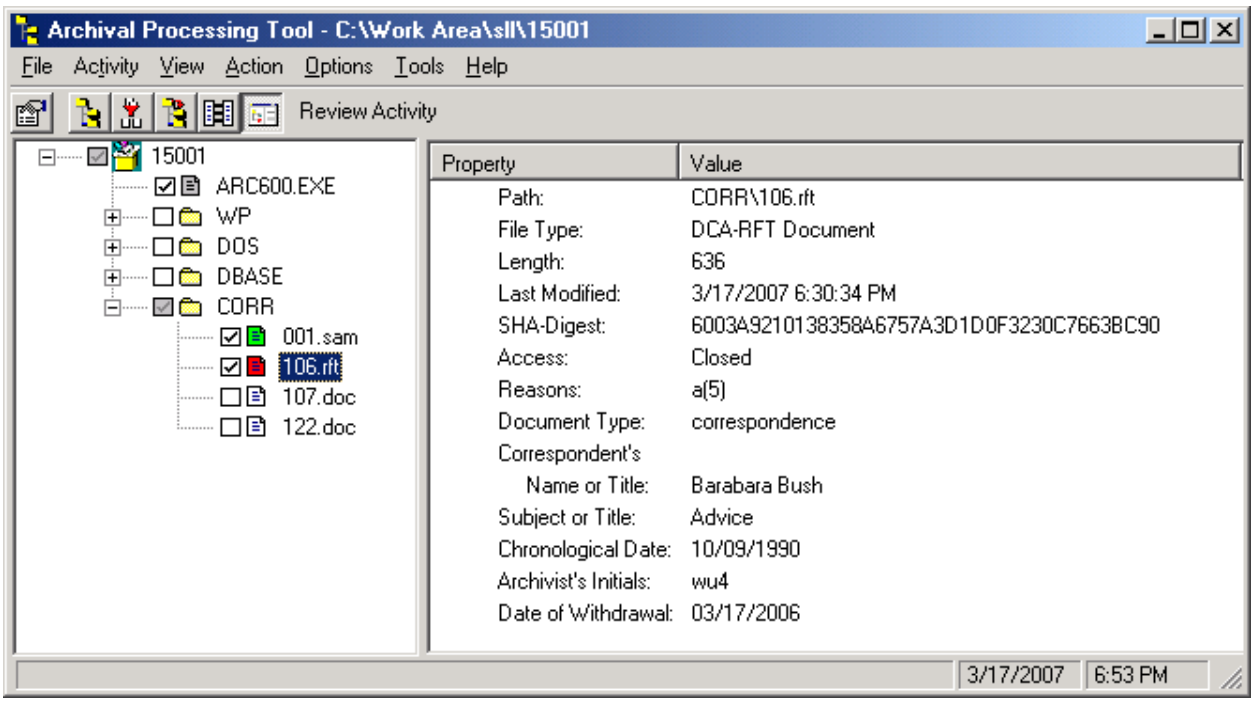

Figure 5. Reviewing Presidential E-records

FOIA Processing

Five years after the end of the Bush Presidential Administration, Archivists at the Bush Presidential Library began to respond to Freedom of Information Act (FOIA) requests, and that continues today. This requires the capability to index and search the repository of Presidential e-records for records relevant to the request. It also requires the capability to support review of the relevant records including redaction of copies of the records. In addition, finding aids must be created for FOIA collections, and the collections must be made available to the requestor.

Archivists learned that the number of files or file size as measured in bytes is not an accurate measure of how much work needs to be performed during review of records related to a FOIA request. These are also not good measures of how much work has been performed each day in responding to a FOIA request. A method was developed for estimating the number of pages represented by files of different formats, and accumulating these estimates as measures of the

number of pages in the results set of a FOIA search. This is a much better measure of review work to be performed. The prototype was extended to include these capabilities [1].

Results

The method of evolutionary prototyping has been used to develop a prototype system to support archivists in gaining intellectual and physical control of personal computer records from the Presidential administration of George H. W. Bush. Archivists at the Bush Presidential Library used the initial prototype and learned additional requirements for processing these e-records. The prototype has been iteratively adapted to meet newly learned requirements. The archivists at the Bush Presidential Library now have an Electronic Record Repository and Archival Processing System that supports Systematic and FOIA Processing [2]. They are using the prototype for experiments in FOIA Processing, and more has been learned that will be used to further adapt the prototype system.

The prototype system also provides an environment for experimental application of advanced information technologies to archival processes. These advanced technologies include content extraction, grammatical induction and recognition of documentary form [5], automatic summarization, knowledge-based reasoning, and knowledge acquisition. These technologies are being applied to:

- Automatic extraction from e-records of document type, correspondents' names, subject or title, and chronological date for inclusion in withdrawal information.
- Automatic summarization of the scope and content of record series.
- Assistance in reviewing Presidential records for PRA restrictions on disclosure.
- Improvement of the precision and recall of FOIA Search.

Acknowledgements

The ERA Program of the National Archives and Records Administration and the Army Research Laboratory sponsored this research under Cooperative Agreements DAAD19-03-2-0018 and W911NF-06-2-0050. The developers of this prototype acknowledge the members of the Bush Presidential Library Staff for their contributions to the development of the PERPOS system.

References

1. S. Laib and W. Underwood. FOIA Processing in the Presidential Electronic Records Pilot System. PERPOS TR 06-05, lTTL, Georgia Tech Research Institute, July 2006.
2. L. Spencer, S. Oriabure and W. Underwood. Launching E-Records with a PERPOS: The Presidential Electronic Records PilOt System. NAGARA Annual Meeting 2005, Richmond, Virginia, July 20-23, 2005.
3. W. Underwood. The Presidential Electronic Records PilOt System: Results of Laboratory Experiments and Use by Archivists. PERPOS TR 03-01, ITTL, Georgia Tech Research Institute, Nov. 2003.
4. W. Underwood, S. Laib and M. Hayslett. Reference Manual for PERPOS: An Electronic Records Repository and Archival Processing System, Version 3.1. PERPOS TR ITTL/CSITD 06-2, ITTL, Georgia Tech Research Institute, September 2006.
5. W. Underwood, S. Isbell and M. Underwood and S. Laib. Grammatical Induction and Recognition of the Documentary Form of Record Types. International Symposium in Digital Curation (DigCCurr2007), Chapel Hill, NC, April 18-20, 2007.