Presidential Electronic Records Pilot Operating System (PERPOS)

William Underwood

Presented at
Digital Strategies 2000
16 November 2000

Georgia Tech Research Institute

Overview

- Background: Presidential Electronic Records Project
- PERPOS Project objective and approach
- Archival processing requirements and decision support tools
- Pilot testing at the Bush Presidential Library

Presidential Electronic Records Project

Objective: Identify opportunities for applying information technology to gain intellectual and physical control over Presidential electronic records created on PCs.

Tasks:

- **Analyzed digital documents from Bush PC hard drives.**
- **Analyzed archival (Presidential Library) processes to determine opportunities for supporting archival decisions.**
- **Identified and evaluated information technologies to support archival processing.**
- **Demonstrated and recommended relevant information technologies.**

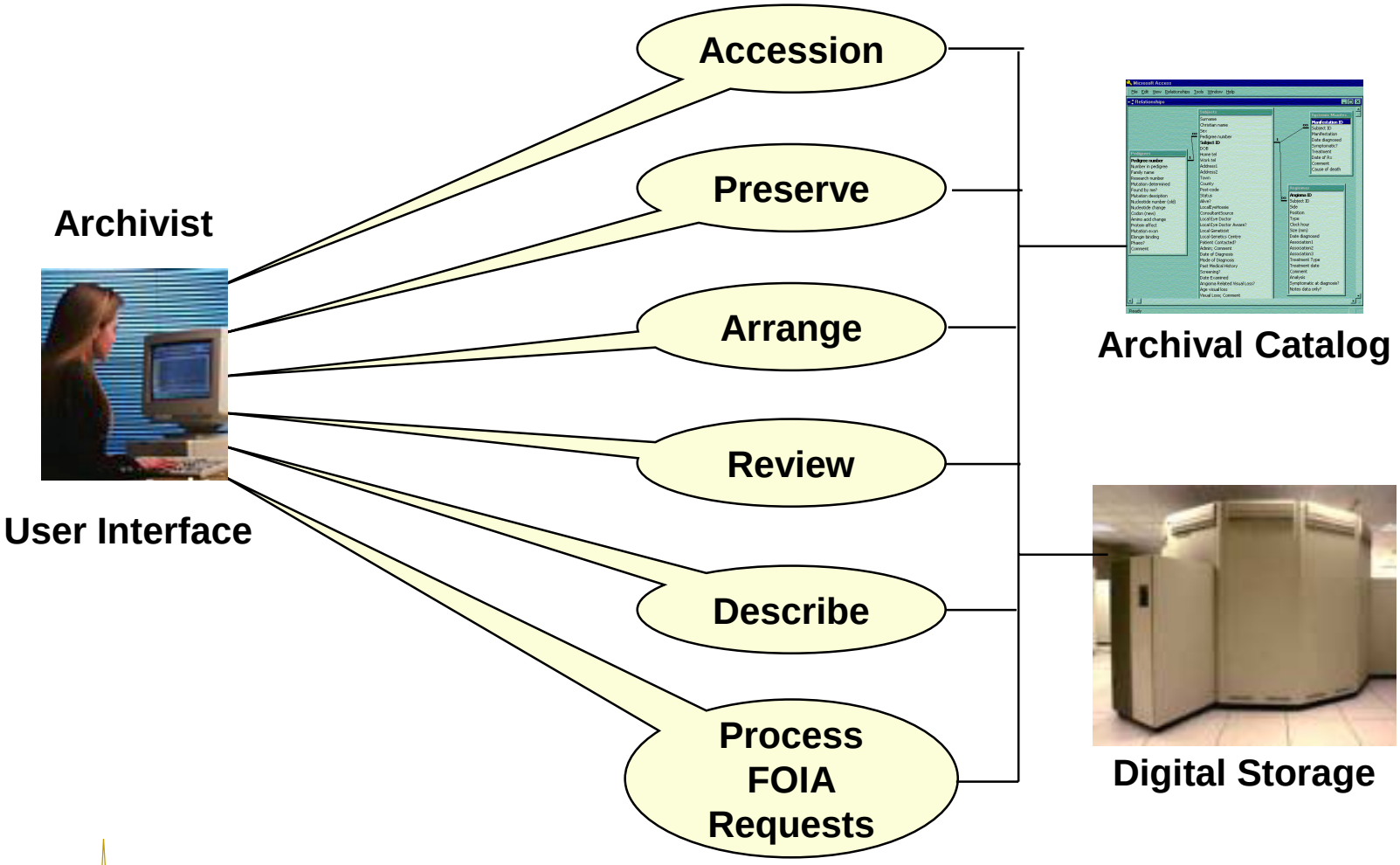
PERPOS Project Objective and Approach

Objective: Software tools for processing, managing, and accessing digital archives.

Approach:

- **Develop software tools to support archival decision making.**
- **Conduct experiments to evaluate these tools.**
- **Demonstrate their use at Archives II.**
- **Evaluate the tools in a pilot project at the Bush Presidential Library.**

Archival Processes



Preservation

Definition: ***Digital Preservation*** is the storage, maintenance and access to digital objects/materials over the long term. This may involve one or more digital preservation strategies including technology preservation, technology emulation or digital information migration

Requirement: An archivist must distinguish user-created files (potential records) from system and software application files (non-records).

File Filtering Tool for Identifying User-created Files

- Archivist indicates collection of files to be filtered.
- File Filtering Tool
  - **identifies non-records, i.e. system and application files.**
  - **counts the number of non-document and document files.**
  - **identifies native formats of document files.**
- Archivist can view the files using Quick View Plus.
- Archivist decides which files to preserve, and copies them to a container.

Document File Filter

Document File Filter

Filter Help

Source Path:

D:\22 Browse ...

Destination Path:

C:\Work Area\22 Browse ...

Excluded Extensions:

Extension	Comment
bat	DOS batch file
com	compact executable files
dll	dynamically linked library
drv	Device Driver for various prgs
exe	executable files
fnt	font file for various programs
fon	Quattro font file
hlp	help files
prs	printer driver
sym	Harvard Graphics Symbol file

Add Delete Save Setting

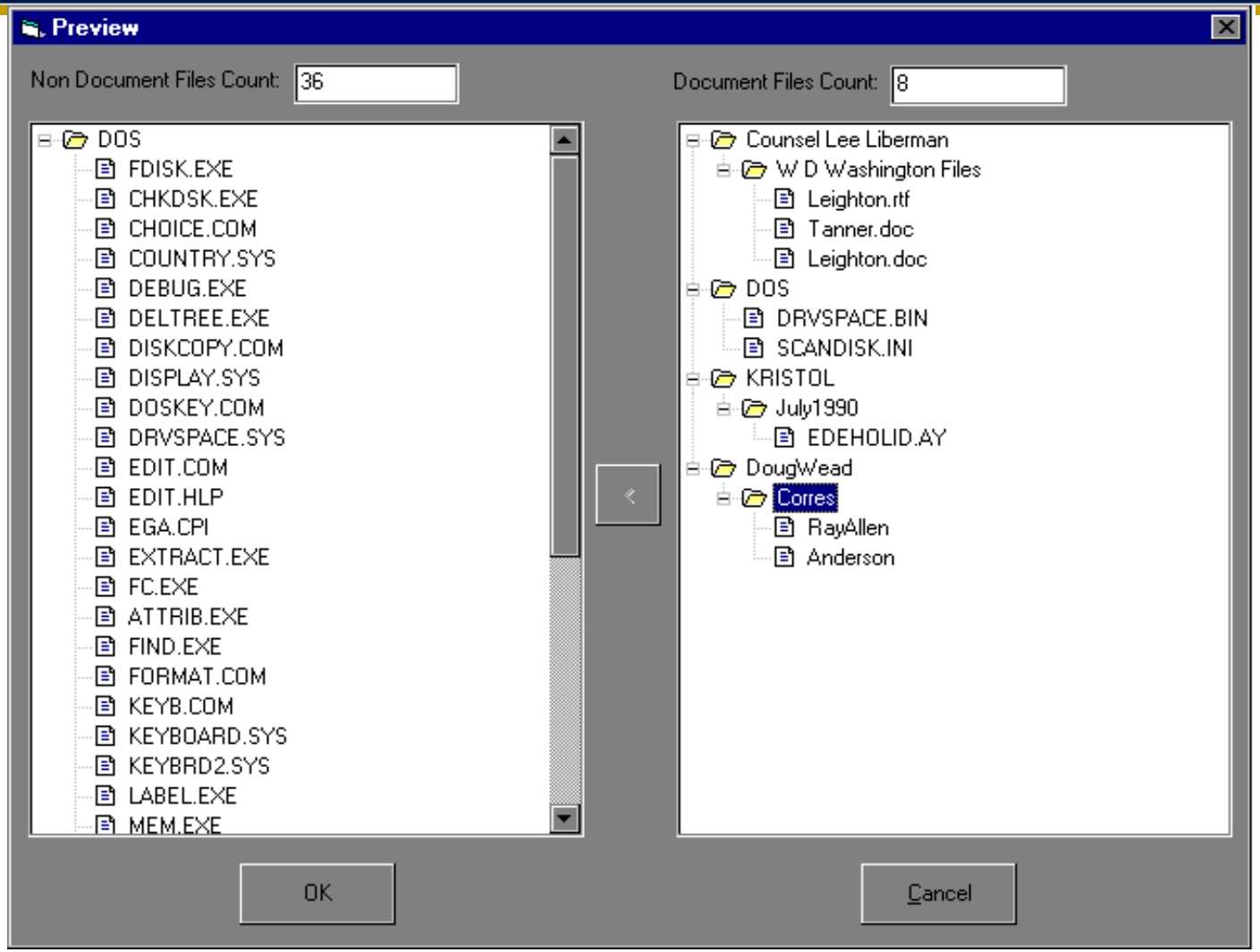
Excluded File Names:

File Name	Comment
1.pal	HG palette
10.pal	HG palette
11.pal	HG palette
123.cmp	Lotus
123.cnf	Lotus
123.dyn	Lotus
123.mu	Quattro - 123 menu tree
123.rf	Quattro resource file
123.set	Lotus
123vax.kbm	REFLECT 123 VAX keyboard

Add Delete Save Setting

Preview Transfer Exit

Preview of Non-Document & Document Files



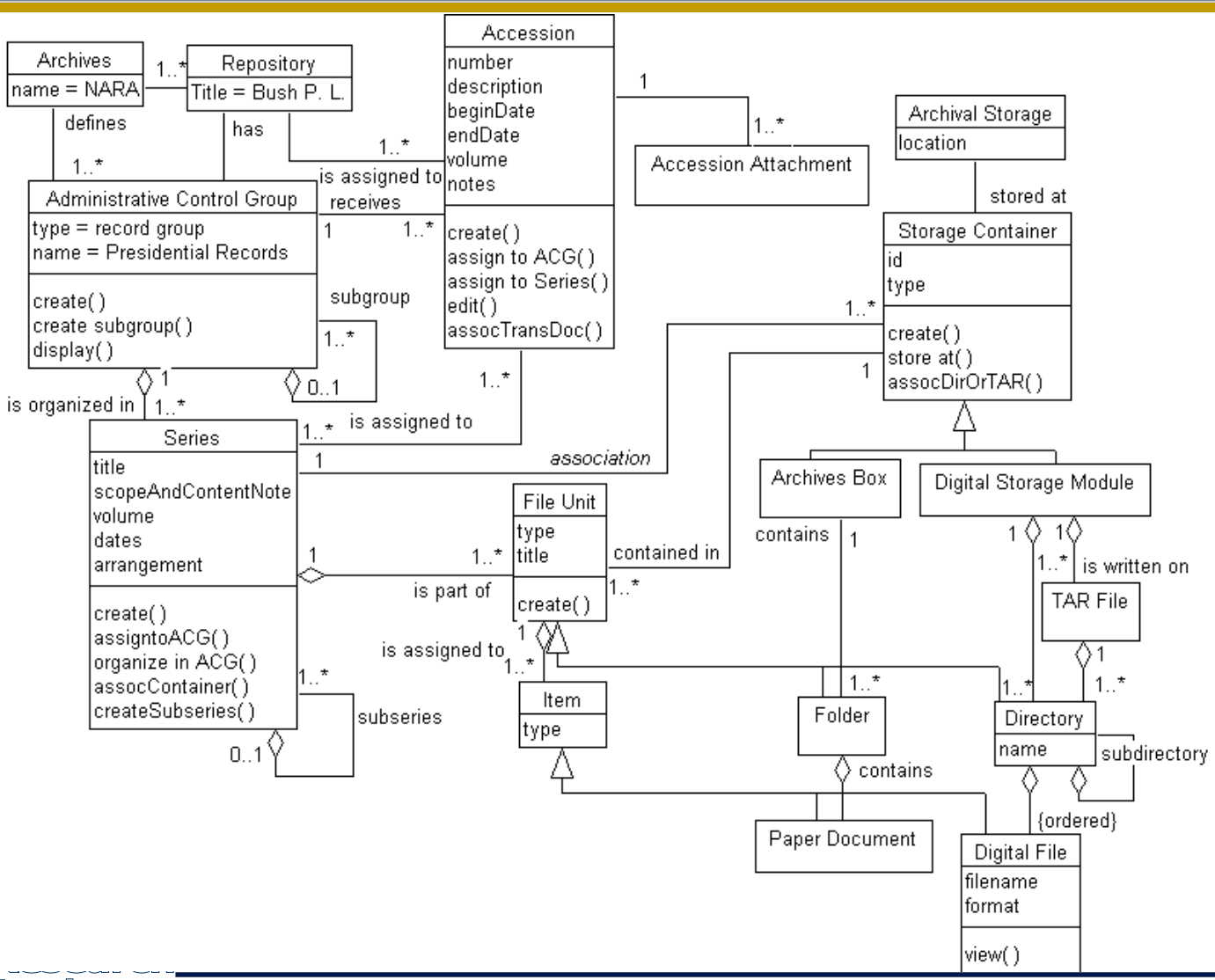
Arrangement

Definition: ***Arrangement*** is "the intellectual and physical processes and results of organizing documents in accordance with accepted archival principles, particularly provenance, at as many as necessary of the following levels: repository, collection, record group or Fonds, subgroups, series, subseries, file unit, and item."

[Society of American Archivists Glossary]

Requirement: An archivist *perfects the arrangement* of files that have no apparent logical order, e.g., miscellaneous files recorded in a root directory or in a software application directory.

Presidential Library Data Model



Description

Definition: ***Description*** is "the process of analyzing, organizing, and recording information that serves to identify, manage, locate, and explain the holdings of archives and the contexts and records systems from which those holdings were selected.

[Society of American Archivists Glossary]

Requirements:

Creating finding aids (indices and/or catalog data) for collections of digital files from a filing system.

Creating scope & content notes and directory titles.

Archival Description Tool

- Tool creates a conceptual index (map) of concepts in the document.
- Tool extracts information from this conceptual map and summarizes it in the form of a scope and content note.
- Tool uses the conceptual map for each subdirectory to create extended directory (folder) title (variant title) as an extension of the directory name.

Review Process

Definition: ***Review*** is the process of identifying and segregating materials that are to be temporarily closed to researchers. All Presidential records are reviewed under the provisions of the Presidential Records Act (PRA) and the Freedom of Information Act (FOIA).

Requirement: Page-by-page review for FOIA exemptions and PRA restrictions is required.

PRA Restrictions

- a(1) National security
- a(2) Appointments to Federal offices
- a(3) Exempted by Statute
- a(4) Confidential commercial info
- a(5) Confidential advice
- a(6) Personal privacy

FOIA Exemptions

- b(1) National security
- b(2) Agency guidelines / agreements about investigations
- b(3) Exempted by statute
- b(4) Confidential commercial info
- b(5) Agency communications regarding litigation
- b(6) Personal privacy
- b(7) Law enforcement investigations
- b(8) Financial institution reports
- b(9) Geological info about wells

Issues with the Review Process

- The volume of digital documents accessioned by National Archives and Presidential Libraries is increasing dramatically.
- FOIA and PRA review is a bottleneck because the current procedures are manual and intellectually intensive.
- Many records are closed in an archives because there has not yet been sufficient time for page-by-page review to locate PRA restrictions or FOIA exemptions that may apply.

FOIA/PRA Review Tool

- Archivist indicates document(s) to check.
- Software checks document(s) for FOIA exemptions, or PRA restrictions, or both.
- Software highlights the suspect portion of the text, indicate the exemption and/or restriction that applies, and explain the line of reasoning that led to the conclusion.
- Archivist reviews the highlighted text and decides whether it can be redacted, or whether the entire document should be withdrawn.

PRA/FOIA Rules

IF: Pattern "[0-9][0-9][0-9]-[0-9][0-9]-[0-9][0-9][0-9][0-9]" occurs in document,

THEN: Conclude there is evidence of a person's Social Security Number.
Create next possible restriction in document (P6,F6) at text location (p,l).

IF: Document is a memo, and
Addressee is President, and
Author has position Counsel to President, and
Illocutionary Act (Recommendation, Advice, Suggestion),

THEN: Conclude there is evidence of confidential advice to President.
Create next possible restriction in document (P5) at text location (p,l).

IF: Possible restrictions in document.

THEN: Highlight text locations of possible restrictions.

An Example

MEMORANDUM FOR THE PRESIDENT

FROM: C. BOYDEN GRAY
COUNSEL TO THE PRESIDENT

SUBJECT: Nomination of Ronald B. Leighton to be United States District Judge for the Western District of Washington.

The Attorney General, Samuel K. Skinner, Nicholas E. Calio, C. Boyden Gray, and Constance Horner recommend that you nominate Ronald B. Leighton to be United States District Judge for the Western District of Washington. Since 1976,

law firm of
on & Daheim as an
red a B.A. degree,
and a J.D. degree
Hastings College
tried, has two
. The American Bar
alifications and
on.

mmendation

PRA Checker

Text:

The Attorney General, Samuel K. Skinner, Nicholas E. Calio, C. Boyden Gray, and Constance Horner recommend that you nominate Ronald B. Leighton to be United States District Judge

Ignore

Ignore All

Next Sentence

Explanation:

1. Appointment to Federal Office? PRA a(2)
   (a) "United States District Judge" is a Federal Office.
2. Confidential Advice? PRA a(5)
   (a) MEMORANDUM FOR THE PRESIDENT
   (b) "C. Boyden Gray" is an advisor to President.

Options...

Cancel

Some FOIA Research Issues

- Knowledge acquisition is a bottleneck.

  Can domain facts and concepts be learned from the documents themselves?

  Can cases of restrictions be learned from a corpus of documents in which restrictions have been previously identified?

- Handling waivers of PRA restrictions.

- Discourse analysis problem: Text whose interpretation is dependent on the content of other documents on same subject or matter.

Processing FOIA Requests

The National Archives and Presidential Libraries are responsible for processing FOIA requests for the records they maintain.

Requirement: An archivist must perform a search of the holdings of the Archives to determine what if any electronic records are responsive to the FOIA request.

Problem: The large volume of unprocessed electronic records make it difficult to determine if there are records responsive to a FOIA request.

Text-based Document Retrieval

- FOIA searches require content-based retrieval.
- Experimentally evaluate search tool and a document retrieval tool based on Natural Language Processing (NLP) techniques and conceptual indexing.
- Evaluate precision and recall (retrieval metrics).
- Sample document collection will be a subset of user-created files from the Bush hard drives.
- Queries will be past FOIA requests previously processed by the Bush Presidential Library.

Approaches to Document Retrieval

Boolean search (pattern matching)
Statistical methods of search (vector space, probabilistic)
Content-based search (syntax / semantics)
    Thesaurus expansion
    Conceptual Indexing
    Lexical subsumption
Structured document search
    Database search
    XML tagged documents
Hybrid forms of search

Summary

Objective: Software tools for processing, managing, and accessing digital archives.

Approach:

- **Develop software tools to support archival decision making.**
- **Conduct experiments to evaluate these tools.**
- **Demonstrate their use at Archives II.**
- **Evaluate the tools in a pilot project at the Bush Presidential Library.**