

PROPOSED TECHNICAL REFERENCE DOCUMENT DATABASE SYSTEM (TDOCS) DESIGN

Prepared for

**Nuclear Regulatory Commission
Contract NRC-02-93-005**

Prepared by

**Center for Nuclear Waste Regulatory Analyses
San Antonio, Texas**

April 1995



PROPOSED TECHNICAL REFERENCE DOCUMENT
DATABASE SYSTEM (TDOCS) DESIGN

Prepared for

Nuclear Regulatory Commission
Contract NRC-02-93-005

Prepared by

Aaron R. DeWispelare
Joseph H. Cooper
Robert L. Marshall

Center for Nuclear Waste Regulatory Analyses
San Antonio, Texas

April 1995

CONTENTS

Section	Page
FIGURES	vii
ABBREVIATIONS	ix
ACKNOWLEDGMENTS	xi
EXECUTIVE SUMMARY	xiii
1 INTRODUCTION	1-1
1.1 BACKGROUND	1-1
1.2 OVERVIEW OF SYSTEM DESIGN	1-3
2 REQUIREMENTS REVIEW AND SUMMARY	2-1
2.1 OVERALL REQUIREMENTS	2-1
2.1.1 Provide Full-Text Search and Retrieval and Image Viewing Capabilities	2-2
2.1.1.1 General Requirements for a Full-Text Search and Retrieval Facility	2-3
2.1.1.2 Required Capabilities for Text Search	2-4
2.1.1.3 Required Capabilities for Presentation of Full-Text Search and Retrieval Query Results	2-4
2.1.1.4 Required Capabilities for Viewing of Documents Selected by Full-Text Search and Retrieval	2-5
2.1.1.5 Required Capabilities for Viewing Images and Accessing Documents Selected Through Full-Text Search and Retrieval	2-5
2.1.2 Provide Staff Productivity Enhancement, Download, and Cut and Paste Capabilities for Text and Images	2-5
2.1.2.1 Document Download Capabilities	2-6
2.1.2.2 Cut and Paste Capabilities	2-7
2.1.2.3 Hypertext Generation Capabilities	2-7
2.1.2.4 Hypergraphic Link Capabilities	2-7
2.1.3 Load and Maintain an Electronic Repository of Technical Documents	2-8
2.1.3.1 On-Demand Loading	2-9
2.1.3.2 Routine Loading	2-10
2.1.3.3 Management and Control Capabilities	2-12
2.2 DESIGN OVERVIEW	2-14
3 DOCUMENT MANAGEMENT SERVER DESIGN	3-1
3.1 SERVER HARDWARE	3-1
3.2 SERVER SOFTWARE	3-1
3.2.1 Relational Database	3-1
3.2.1.1 Considerations	3-1
3.2.1.2 Evaluation	3-3
3.2.1.3 Recommended Relational Database Management System	3-4
3.2.2 Full-Text Indexing Search/Retrieval Software	3-4
3.2.2.1 Considerations	3-4
3.2.2.2 Evaluation	3-4

CONTENTS (Cont'd)

Section	Page
	3.2.2.3 Recommended Full-Text Indexing Search/Retrieval Software.....3-6
3.3	TECHNICAL REFERENCE DOCUMENT DATABASE DATA REPOSITORIES.....3-6
	3.3.1 Directory Structures in the UNIX File System.....3-7
	3.3.1.1 Archive Subdirectories.....3-7
	3.3.1.2 Images Subdirectories.....3-9
	3.3.1.3 Data Subdirectories.....3-9
3.4	FUNCTIONAL MANAGEMENT AND CAPABILITIES OF THE DOCUMENT MANAGEMENT SERVER.....3-9
	3.4.1 Server Logon Processing.....3-10
	3.4.2 Changing Passwords.....3-10
3.5	DOCUMENT PROCESSING SERVICES.....3-10
	3.5.1 Submitting Scanned Documents.....3-11
	3.5.2 Submitting Documents for Electronic Loading.....3-13
	3.5.3 Deleting Documents.....3-13
	3.5.4 Updating Documents.....3-14
3.6	DOCUMENT SEARCH AND RETRIEVAL SERVICES.....3-15
3.7	DOCUMENT DOWNLOAD.....3-15
3.8	BATCH LOADING, INDEXING, HYPERLINKING, AND SECURITY.....3-15
	3.8.1 Batch Processing.....3-16
	3.8.2 Full-Text Indexing.....3-16
	3.8.2.1 Preparation of Documents for Full-text Indexing.....3-17
	3.8.3 Generation of Hypergraphic Links.....3-17
	3.8.3.1 Preparation of Documents for Hypergraphic Generation.....3-17
	3.8.3.2 Document Indexing and Hypergraphic Link Generation.....3-17
3.9	DATABASE ADMINISTRATION AND MAINTENANCE.....3-18
	3.9.1 User-ID Maintenance Functions.....3-18
	3.9.2 Initiating the Batch Process.....3-18
4	DOCUMENT PROCESSING CLIENTS DESIGN.....4-1
4.1	DOCUMENT PROCESSING CLIENTS HARDWARE.....4-1
4.2	DOCUMENT PROCESSING CLIENTS SOFTWARE.....4-1
	4.2.1 Multiplatform Software Packages.....4-1
	4.2.2 Open-System Environment.....4-2
	4.2.3 Hardware/Software Independent Communications Between Client and Server Platforms.....4-3
	4.2.4 Graphical User Interface Development.....4-3
4.3	FUNCTIONAL MANAGEMENT AND CAPABILITIES OF THE DOCUMENT PROCESSING CLIENTS.....4-4
	4.3.1 Password Protected Access.....4-4
	4.3.1.1 Document Processing Functionality Available to Custodians.....4-5
4.4	SCANNING, OPTICAL CHARACTER RECOGNITION, AND CLEANUP.....4-6
	4.4.1 Scanning.....4-8

CONTENTS (Cont'd)

Section	Page
4.4.1.1	Potential Contention for Scanning Resources 4-8
4.4.1.2	Color Scanning Capabilities 4-8
4.4.1.3	Factors Affecting Quality of the Scanned Image 4-8
4.4.2	Optical Character Recognition 4-9
4.4.3	Document Cleanup Following Scanning and Optical Character Recognition of Hard-Copy Documents 4-10
4.4.4	Requesting Scan and OCR Functionality 4-10
4.5	DOCUMENT SUBMISSION, DELETION, AND MAINTENANCE 4-10
4.5.1	Submitting New Documents 4-10
4.5.1.1	Entry of the Header Record 4-11
4.5.1.2	Text and Image Files 4-13
4.5.1.3	Document Submission 4-14
4.5.1.4	Document Import Option 4-15
4.5.2	Deleting Documents 4-15
4.5.3	Updating Documents 4-16
4.6	DATABASE ADMINISTRATION 4-20
4.6.1	Adding a New User-ID 4-20
4.6.2	Deleting a User-ID 4-21
4.6.3	Changing User Privileges 4-21
5	DOCUMENT SEARCH AND RETRIEVAL CLIENTS DESIGN 5-1
5.1	DOCUMENT SEARCH AND RETRIEVAL CLIENTS HARDWARE AND SOFTWARE 5-1
5.1.1	Minimum Sun Workstation Requirements for Using the Technical Documents Reference Database 5-1
5.1.2	Minimum Windows Workstation Requirements for Using the Technical Document Reference Database 5-2
5.1.3	Minimum OS/2 Workstation Requirements for Using the Technical Document Reference Database 5-2
5.1.4	Minimum Macintosh Workstation Requirements for Using the Technical Document Reference Database 5-2
5.1.5	Document Client Search and Retrieval Software 5-2
5.2	FUNCTIONAL MANAGEMENT AND CAPABILITIES OF THE DOCUMENT SEARCH AND RETRIEVAL CLIENTS 5-3
5.2.1	Searching and Retrieving 5-3
5.2.2	Viewing Documents 5-5
5.2.3	Launching Word Processing Software 5-6
5.2.4	Image Viewing 5-8
5.2.5	Hypertext Linking 5-9
5.2.6	Cutting and Pasting 5-9
5.2.7	Document Downloading 5-10

CONTENTS (Cont'd)

Section		Page
6	NETWORK ARCHITECTURE	6-1
6.1	REMOTE PROCEDURE CALL	6-1
6.2	NETWORK FILE SYSTEM	6-1
6.3	AUTOS AND ACRS	6-2
6.4	SECURITY.....	6-2
7	SUMMARY.....	7-1
8	REFERENCES	8-1

FIGURES

Figure		Page
1-1	Advanced computer review system capabilities.	1-1
2-1	Overall requirements hierarchy	2-1
2-2	Requirements hierarchy for full-text search and retrieval and image viewing	2-3
2-3	Requirements hierarchy for staff productivity enhancement	2-6
2-4	Requirements hierarchy for document loading and maintenance	2-8
2-5	Requirements hierarchy for on-demand document loading	2-9
2-6	Requirements hierarchy for routine document loading and maintenance	2-11
2-7	Requirements hierarchy for management and control capabilities	2-13
2-8	Major technical reference document database system functions	2-16
3-1	Existing Division of Waste Management server capabilities.	3-2
3-2	Directory structures in the UNIX file system.	3-8
3-3	Movement and renaming of files by the server	3-12
4-1	User platforms and functionality	4-2
4-2	Division of Waste Management logon screen	4-5
4-3	TDOCS custodian main menu	4-5
4-4	TDOCS operations pull-down menu	4-6
4-5	TDOCS change password screen	4-6
4-6	The scanning process	4-7
4-7	TDOCS custodian pull-down menu	4-10
4-8	Header entry screen	4-12
4-9	Document import entry screen	4-15
4-10	Delete document screen	4-16
4-11	TDOCS confirm screen	4-17
4-12	Update document request screen	4-17
4-13	Update document screen.	4-18
4-14	TDOCS/TDAS main menu for database administrators	4-20
4-15	TDOCS database administrator system pull-down menu	4-20
4-16	TDOCS User ID maintenance screen.	4-21
5-1	Search pull-down menu	5-3
5-2	Simple query entry screen.	5-4
5-3	Query results list.	5-6
5-4	Form query screen	5-7
5-5	TOPIC document display	5-8
5-6	Launch pull-down menu.	5-9
5-7	Selecting images for viewing	5-10
5-8	Hypertext links within a displayed document	5-11
5-9	Directory specification for document download	5-12

ABBREVIATIONS

ACRS	Advanced Computer Review System
API	Application Program Interface
ASCII	American Standard Code for Information Interchange
AUTOS	Agency Upgrade of Technology for Office Systems
CDS	Compliance Determination Strategy
CSP	Correspondence Index System
CNWRA	Center for Nuclear Waste Regulatory Analyses
DBA	Database Administrator
DNS	Domain Name Service
DOE	U.S. Department of Energy
DWM	Division of Waste Management
FIPS	Federal Information Processing Standards
FTP	File Transfer Protocol
GUI	Graphical User Interface
HLW	High-Level [Radioactive] Waste
IRM	Information Resources Management
LAN	Local Area Network
LSS	Licensing Support System
NFS	Network File System
NIS	Network Information System
NMSS	NRC Office of Nuclear Material Safety and Safeguards
NRC	Nuclear Regulatory Commission
NTD	NRC technical documents
NUDOCS	Nuclear Document System
OCR	Optical Character Recognition
ONC	Open Network Computing
PASS/PADB	Program Architecture Support System/Program Architecture Database
PC	Personal Computer
QA	Quality Assurance
RDBMS	Relational Database Management System
RPC	Remote Procedure Call
RPD	Regulatory Program Database
RRT	Regulatory Requirement Topic
SGML	Standard Generalized Markup Language
SRA	Systematic Regulatory Analysis
SQL	Structured Query Language
SwRI	Southwest Research Institute
TCP/IP	Transmission Control Protocol/Internet Protocol
TDI	Technical Document Index
TDOCS	Technical Reference Document Database System
TIFF	Tagged Image File Format

ACKNOWLEDGMENTS

This report was prepared to document work performed by the Center for Nuclear Waste Regulatory Analyses (CNWRA) for the U.S. Nuclear Regulatory Commission (NRC) under Contract NRC-02-93-005. The activities reported here were performed on behalf of the NRC Division of Waste Management (DWM). The report is an independent product of the CNWRA and does not necessarily reflect the views or regulatory position of the NRC.

The following products are discussed in this report:

- Galaxy is a trademark of Visix Software

Visix Software Inc.
11440 Commerce Park Drive
Reston, VA 22091

- OpenLook is a trademark of Unix System Laboratories, Inc.

Sun Microsystems, Inc.
2550 Garcia Avenue
Mountain View, CA 94043

- Microsoft Windows is a trademark of Microsoft

Microsoft Corporation
One Microsoft Way
Redmond, WA 98052

- Motif is a trademark of Open Software Foundation

Open Software Foundation
11 Cambridge Center
Cambridge, MA 02142

- OS/2 is a trademark of IBM

IBM Corp.
Armonk, NY

- ORACLE is a trademark of Oracle

Oracle Corporation
500 Oracle Parkway
Redwood Shores, CA 94065

- Paint Shop Pro is copyrighted by JASC, Inc.

JASC Inc.
10901 Red Circle Drive
Suite 340
Minnetonka, MN 55343

- PMJPEC is copyrighted by Norm and Ken Yee

58 Chandler Street
Boston, MA 02116

- Unix is a trademark of Unix System Laboratories, Inc.

- System 7 is a trademark of Apple Computer

Apple Computer, Inc.
20525 Mariani Avenue
Cupertino, CA 95014

- Topic is a trademark of Verity

Verity Inc.
1550 Plymouth Street
Mountain View, CA 94303

- WordPerfect is trademark of WordPerfect

WordPerfect Corporation
329 N. State Street
Orem, UT 84057

- ADI is a trademark of Verity

Verity Inc.
1550 Plymouth Street
Mountain View, CA 94303

EXECUTIVE SUMMARY

This report presents the system design for the implementation of the Technical Reference Document Database System (TDOCS), including the specification of requirements, constraints, and design. The TDOCS is a leading-edge application employing relatively new technology to capture a broad spectrum of technical materials and make them available to users in the Division of Waste Management (DWM) and Center for Nuclear Waste Regulatory Analyses (CNWRA) staffs. The TDOCS addresses the immediate and anticipated future needs of the DWM for a document management system.

Design and implementation of the TDOCS system (Johnson et al., 1993) focused on exploring particular issues, selecting appropriate off-the-shelf commercial software products, and integrating those tools with custom written code to provide the desired functionality. Therefore, the TDOCS employs relatively new, but well-understood, software tools and sound methodologies for integration of those tools. The design strategy for the TDOCS is to provide currently achievable and immediately beneficial system capabilities to meet the defined requirements and constraints of the system and application environments. TDOCS design and implementation decisions have been appropriately influenced by the requirements, policies, and procedural issues.

To support the DWM staff in precicensing reviews and independent analyses, the TDOCS system must provide capabilities for full-text search and retrieval of the text of documents, as well as retrieval and presentation of the images of figures, equations, etc., associated with the documents. Both the text and image information must also be available for downloading to the staffs' personal computers and workstations. Facilities for staff productivity enhancement, including cut and paste capabilities, hyperlinks within documents, hyperlinks between related documents, and hyperlinks between textual materials and their associated images, are also required capabilities of the TDOCS to help the staff select, correlate, and integrate information for incorporation in analyses, evaluations, and review reports. Thus, the TDOCS must address three major requirements:

- Load and maintain an electronic repository of technical document
- Provide full-text search and retrieval and image viewing capabilities
- Provide staff productivity enhancement, download, and cut and paste capabilities for document text and images

Many similarities were found at both technical and application levels between the requirements for the TDOCS and another DWM computer application, the Regulatory Program Database (RPD) system. Several strategic software products were found to fully meet the defined requirements and constraints of both TDOCS and RPD, and these products were used in the initial implementation of both systems, resulting in a considerable synergy and cost savings between the TDOCS and RPD systems. The strategic off-the-shelf software products identified in the TDOCS design are as follows:

- TOPIC, a document database system from Verity Corp., which supports full-text search and retrieval

- Oracle, a relational database management system (RDBMS) from Oracle Corp., which supports configuration control and reporting
- Galaxy, a multiplatform graphical user interface (GUI) application development environment from Visix Software, Inc., which supports access to system functions

Custom code was produced to provide the TDOCS application framework and to integrate the system capabilities provided by the software packages. This was a significant software development and integration effort, and the results have been very satisfactory.

The TDOCS is implemented using a client/server architecture. The relational database and text and image data repositories reside on a central server platform that is accessed by staff from their workstation and PC client platforms. Multiple client platforms are supported including SUN workstations, PCs utilizing either Microsoft Windows or IBM OS/2, and Macintosh computers. The TDOCS is implemented through three major modules:

- Document Management Server—code that supports the maintenance and use of the TDOCS relational and full-text data repositories
- Document Processing Clients—code that supports user interaction for document loading and maintenance
- Document Search and Retrieval Clients—code that supports user interaction for search and retrieval of documents in the TDOCS full-text repository

The TDOCS initial implementation satisfies the system requirements and complements and completes the capabilities of the Advanced Computer Review System (ACRS) by making technical reference documents available for referencing, synthesizing, and incorporating in staff analyses. TDOCS also provides the interface to the Technical Data Access System (TDAS) as part of the overall Advanced Computer Review System (ACRS) (Johnson and Murphy, 1994).

1 INTRODUCTION

1.1 BACKGROUND

The technical staff of the Nuclear Regulatory Commission (NRC), Division of Waste Management (DWM) performs technical reviews and analyses in conducting licensing activities. These technical reviews require reference to, and use of, large volumes of technical information. This information is in the form of technical data and technical reference documents which must be available for referencing, synthesizing, and incorporating in the staff analyses. The plans are that staff members will have all the necessary information available to them on or through their individual computer workstations (see Figure 1-1).

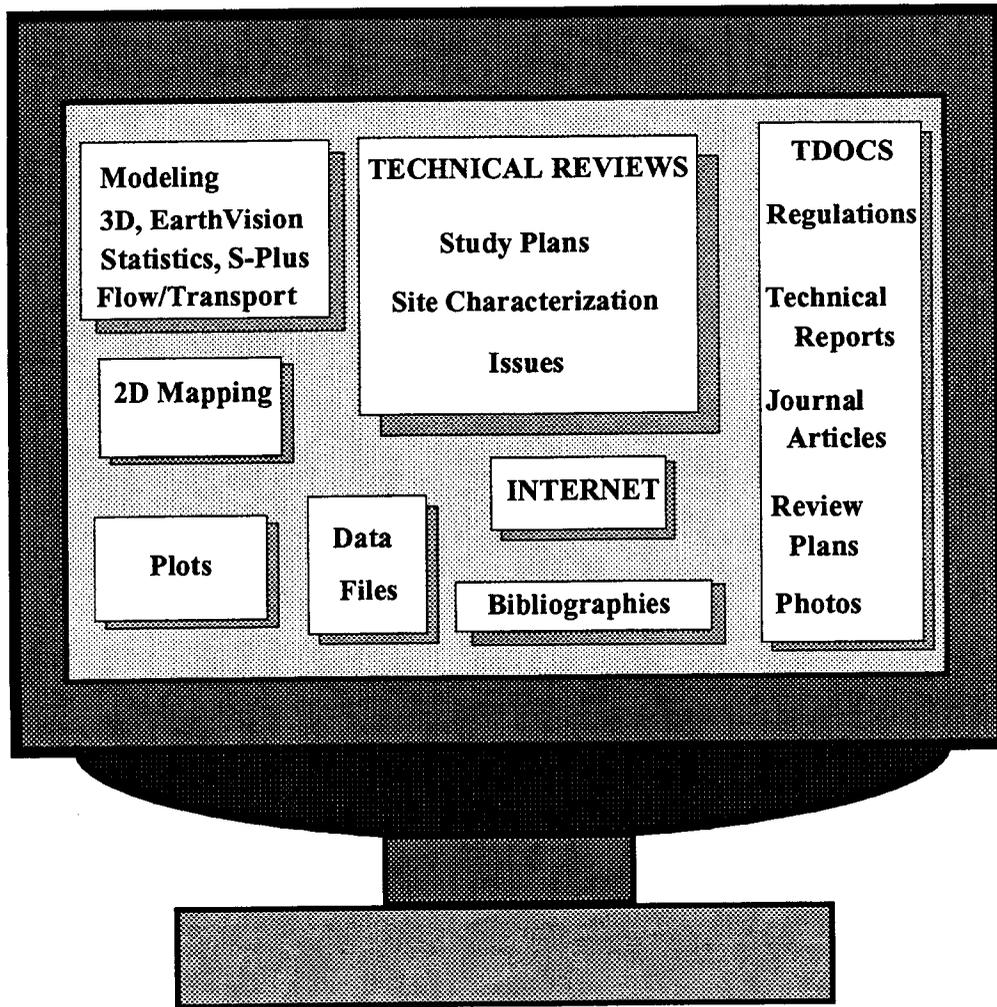


Figure 1-1. Advanced computer review system capabilities

The NRC has requested increased capability in making independent technical analyses. This requirement places additional demands on the DWM technical staff during any prelicensing or licensing review activity and makes it all the more necessary to provide enhanced computer capabilities for the management, retrieval, and visualization of technical information. The Advanced Computer Review System (ACRS) was designed and implemented to provide these capabilities for the NRC Office of Nuclear Material Safety and Safeguards (NMSS) and DWM. This system provides advanced computing and visualization capabilities. It utilizes available and emerging hardware and software technology for geographic information systems analyses, site characterization evaluations, computer modeling, and display of the results of modeling exercises in ways that permit them to be understood in their relation to other information.

A particular need for such capabilities is associated with the high-level waste (HLW) program, for which the NRC has a very stringent review schedule. Following docketing of the U.S. Department of Energy (DOE) license application for a geologic repository for HLW at the NRC, the DWM has a statutory requirement to make a construction authorization decision within 3 yr. The first 18 mo of this 3-yr period will be devoted to a technical review of the license application; and the remaining 18 mo, with a possible 1-yr extension, will be devoted to the licensing hearing. This requirement will place considerable demands on the DWM staff as it conducts its prelicense application and license application reviews. Updated computer resources will be needed to accomplish these reviews within the allotted time.

The Technical Reference Document Database System (TDOCS) was intended to complement and complete the capabilities of the ACRS. The creation and maintenance of a text and image document reference database with full-text search and retrieval capabilities were requirements of the DWM ACRS that was initiated in January 1992 and completed in August 1992. Because of resource limitations, the document storage and retrieval system (eventually named TDOCS) was separated from the ACRS design report and added as a follow-on task after the completion of the ACRS. The task on the TDOCS project was initiated in March 1993. The NRC direction to the Center for Nuclear Waste Regulatory Analyses (CNWRA) included a requirement for an NRC capability: "Given that the CNWRA is well informed on the hardware and network capabilities planned for the DWM and that there are now *off-the-shelf* software products that provide for text and image retrieval and indexing for many data bases..., it is expected that a functioning *turnkey* system can be readily planned." Following development, installation, and testing at the CNWRA, a turnkey system is to be installed also at the NRC.

The report, Technical Reference Document Database System (TDOCS) Requirements Definition (Johnson and Moehle, 1993), initiated work on developing document management capabilities. It identified overall requirements for the task of providing the DWM with the TDOCS in order to facilitate the analysis and decision making necessary to initiate the design of the system. The TDOCS includes a requirement to support user confidence in being able to find just those documents being sought and provide for increased functionality and staff productivity. The system needs to provide reliable access to, and practical use of, a state-of-the-art database of technical documents. The TDOCS represented a major activity that needed to be pursued immediately to provide the necessary technical document reference and information extraction capability for the DWM and the CNWRA technical review staffs in the HLW Program (Meehan, 1993).

As a follow-up to the requirements definition report, CNWRA representatives met with DWM management and staff on September 8-9, 1993. Discussion at these meetings reviewed matters of policy raised during the requirements definition process and established the initial implementation of the TDOCS at the CNWRA. This initial implementation at the CNWRA was also to function as the Center's document

management system for such holdings as quality assurance records, correspondence indexing, and technical document bibliographic records. Discussion also clarified that the TDOCS would not become an official repository of documents since that is the purpose of the Nuclear Document System (NUDOCS) within the NRC. In the future, when the Licensing Support System (LSS) is implemented, that system will become the official repository of documents for the HLW repository licensing process. Planning for a full-text and image NUDOCS upgrade is now being done by the Office of Information Resources Management (IRM) at the NRC.

1.2 OVERVIEW OF SYSTEM DESIGN

This report provides the TDOCS system design, including the specification of requirements, constraints, system design, and a proposed plan for implementation.

The TDOCS is an application employing relatively new technology to capture a broad spectrum of technical materials and to make them available to the DWM and CNWRA staffs. The TDOCS addresses the immediate and anticipated future needs of the DWM for a document management system, and allows end users to participate in decision making through feedback based on actual use of the system. In this way, the TDOCS assures a closer match between evolving user requirements and the long-term capabilities of the system. The ultimate requirements of the system will be modified, refined, and enhanced with actual user experience of the TDOCS system. Design of the TDOCS system was done to meet the initial requirements (Johnson et al., 1993), focusing on exploring particular issues, selecting appropriate software tools, and integrating those tools with custom written code to provide the desired functionality. Thus, the initial TDOCS implementation is a flexible system developed in response to currently documented requirements with provision to accommodate the evolving needs of the staff.

The design strategy for the TDOCS is to provide currently achievable and immediately beneficial system capabilities to meet the defined requirements and constraints of the system and application environments. Accordingly, some of the TDOCS design and implementation decisions have been influenced by remaining open requirement, policy, and procedural issues. The TDOCS employs well known and relatively mature software tools and sound methodologies for integration of those tools.

This report reviews requirements, constraints, and policies and their effects on the design, and specifies a system design for the TDOCS:

- Chapter 2 reviews the TDOCS requirements
- Chapter 3 presents the TDOCS Document Management Server design
- Chapter 4 presents the TDOCS Document Processing Clients design
- Chapter 5 presents the TDOCS Document Search and Retrieval Clients design
- Chapter 6 presents the TDOCS network architecture
- Chapter 7 presents summary
- Chapter 8 lists references

2 REQUIREMENTS REVIEW AND SUMMARY

This chapter summarizes the requirements relevant to the development and implementation of the TDOCS system, based on the TDOCS Requirements Definition report delivered in August 1993 (Johnson et al., 1993), and on follow-up discussions with the DWM. The requirements definition report considered current needs to load, access, use, and manage documents in a technical reference document database. This report drew on previous efforts by DOE, the NRC, and the CNWRA to define requirements, design, and implement other document management systems. The requirements contained in the requirements definition report are expected to be fully applicable to subsequent enhancements of the TDOCS.

2.1 OVERALL REQUIREMENTS

As indicated in the introduction, the NMSS/DWM requirement analysis demonstrated the need to provide the NRC staff with a technical document reference database system capable of storing and retrieving both text and images, for specified information and/or concepts. This system is needed to support DWM precicensing technical reviews and independent analyses. The TDOCS is intended to support the capabilities of the staff to perform technical reviews and independent analyses through activities that rely heavily on technical documents. Figure 2-1 illustrates the overall requirements for the TDOCS including:

- Providing full-text search and retrieval and image viewing capabilities
- Providing staff productivity enhancement, download, and cut and paste capabilities for document text and images
- Loading and maintaining an electronic repository of technical documents

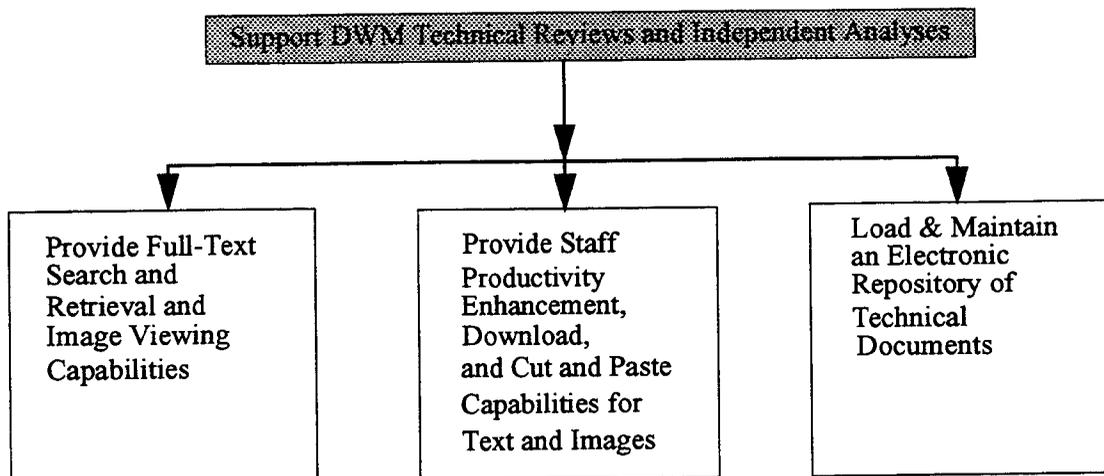


Figure 2-1. Overall requirements hierarchy

Once relevant documents have been processed and loaded, the staff must be able to find those materials quickly and reliably. The TDOCS full-text search and retrieval capabilities must permit (i) full-text search and retrieval of document text, (ii) effective presentation of the text of documents, and (iii) retrieval and presentation of the images of figures, equations, etc., associated with the text of documents.

The text of TDOCS documents and their associated images must be available for downloading to staff computers and workstations. Facilities for staff productivity enhancement, including cut and paste capabilities and hyperlinks, are also required to help the staff select, correlate, and integrate information for incorporation in analyses, evaluations, and review reports.

Document loading in the TDOCS must be driven, in part, by user requests. This requirement is based on the assumption that users have specific needs for information, and best know the documents and reviews that will satisfy those information needs. Capabilities must be provided to permit selected documents, as well as large groups of documents, to be added to TDOCS on a routine basis. Capabilities must also be provided, depending on the urgency of the need and the availability of the materials, to process and make documents available to the staff on an on-demand basis. Document loading must support a combination of: (i) loading of technical documents and references from electronic files, and (ii) scanning, optical character recognition (OCR), and cleanup of hard-copy documents.

The TDOCS system requirements discussed in the following sections are general and are intended to provide support for the DWM and CNWRA staffs in performing independent reviews and analyses.

2.1.1 Provide Full-Text Search and Retrieval and Image Viewing Capabilities

In order to support DWM technical reviews and independent analyses, the TDOCS must permit document access through full-text search and structured header queries. The requirement hierarchy for full-text search and retrieval is illustrated in Figure 2-2:

- Full-text search and retrieval permits the user to find desired materials quickly and efficiently.
- Document text retrieval and viewing permits the user to access, view, and utilize textual information.
- Image retrieval and viewing permits the user to retrieve and view full-page images and/or selected images of equations and figures.

Full-text search provides the capability to search the text of a document for words, phrases, and combinations of words. Structured header queries provide the capability to locate documents by specific attributes when some information is known about the document or document set. Full-text searching is the primary means of document retrieval, but it is only available to the extent that the documents contain textual information. Some documents that are partly or wholly non-textual in nature must be retrieved using structured document headers. A properly formatted header allows for efficient system access to the document through searching of specified data fields, and it also supports retrieval of records for which there is limited or no textual information. The content of headers and the methods employed for document retrieval may vary depending on the type of document:

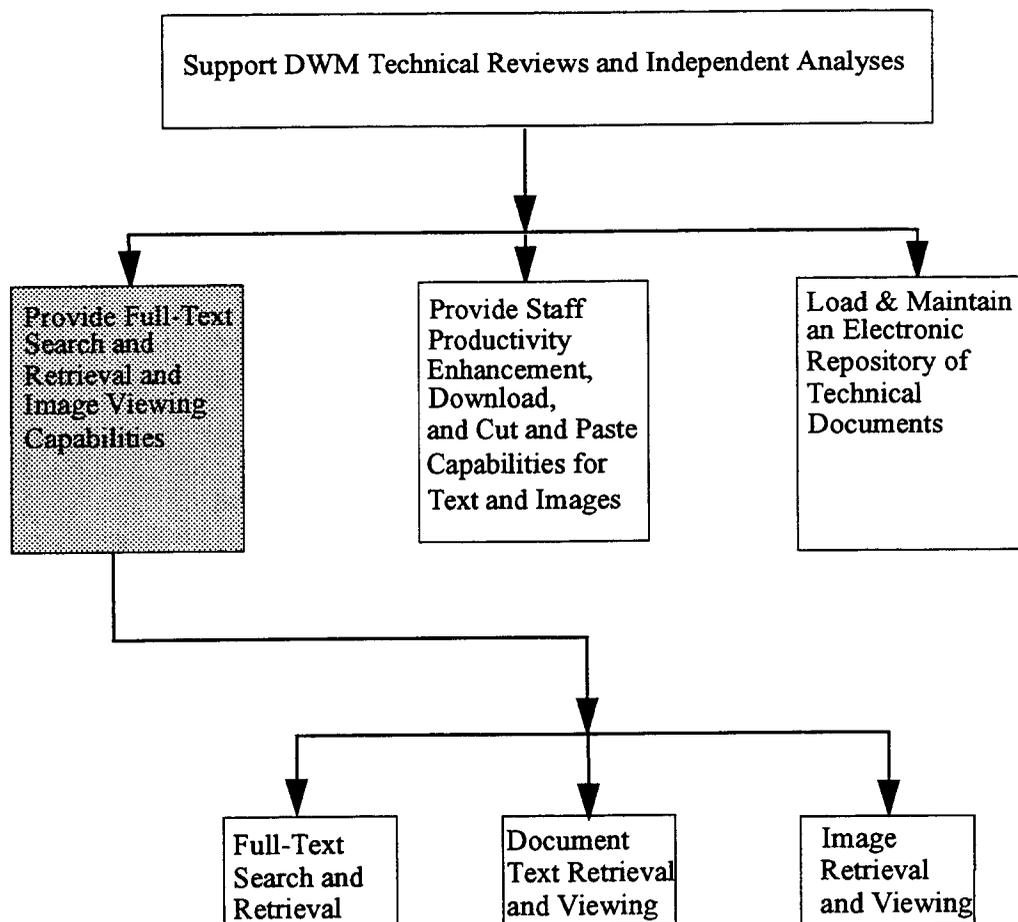


Figure 2-2. Requirements hierarchy for full-text search and retrieval and image viewing

- Non-textual/non-imageable materials, such as magnetic tapes, computer programs, etc., contain no image or textual information. The header record is the only way of retrieving information on such documents.
- Image-only materials, such as maps, engineering drawings, etc., contain an image but no textual information. The header record is the only way of retrieving such documents.
- Text and image materials such as reports include documents that are primarily textual in nature that may have associated images of figures, equations, etc. The header record provides an alternate method of accessing these documents that supplements the full-text search and retrieval capabilities.

The full-text search and retrieval system must be available on multiple hardware/software platforms to accommodate the diverse computing environments used by the DWM and CNWRA staffs. It must provide utilities for fielding headers and indexing text. It must also provide graphical user interface (GUI) support for structured header queries and full-text document searches to facilitate the user interaction and permit straightforward and efficient formulation of complex queries. The full-text search

and retrieval system must provide efficient mechanisms for document viewing to permit users to browse textual material. Cut and paste facilities must be provided to permit portions of text to be identified and copied into new work products. Hyperlinking facilities are needed to permit rapid movement between occurrences of search terms in the document and facilities for launching image viewers are needed to permit images to be accessed directly from the document text display.

2.1.1.1 General Requirements for a Full-Text Search and Retrieval Facility

The following general features are highly desirable in a full-text search and retrieval system:

- **Password-Protected Access Available**—security procedures are available to users of the document search and retrieval clients. System custodian and database administrator functions are required to be password protected and limited to authorized users.
- **Graphical User Interface**—the full-text search and retrieval facility must utilize a GUI that is user friendly and permits users to quickly tap the full power of the system. Because staff may access the system from a variety of hardware/software platforms, the look and feel of the full-text search and retrieval facilities must be relatively constant to avoid confusion or extra training when using different platforms. Equivalent functions must be presented in a consistent manner on all hardware/software platforms, varying only as needed to accommodate the user's specific environment.
- **Customization of User Privileges and Preferences**—individual staff members will have different requirements in terms of subsets of the data repository to be accessed and definitions for concept-based searches. Therefore, the full-text search and retrieval facility must permit user environments to be customized with privileges, preferences, and subsets of the documents in the full-text repository.

2.1.1.2 Required Capabilities for Text Search

The following features are highly desirable for full-text search:

- **Search Capabilities**—the full-text search and retrieval facility must provide extensive search capabilities, including simple searches, formatted searches, and concept-based searches.
- **Full-Text and Header Search Capabilities**—the full-text search and retrieval facility must support a wide range of full-text search capabilities, including wildcards and boolean operators, near spell searches, fuzzy searches, phrase searches, and proximity searches.
- **Query Save, Recall, and Edit**—the full-text search and retrieval facility must support the capability to save, recall, and edit queries.

2.1.1.3 Required Capabilities for Presentation of Full-Text Search and Retrieval Query Results

The following features are highly desirable to support presenting full-text search and retrieval query results:

- Search Result Browsing—the results of a full-text or structured header search must be displayed as a list of documents returned by the query so that the user can select specific documents for viewing.
- Search Result Ranking—to facilitate finding the desired documents in a lengthy list, the search results must be presented in a selection list according to closeness of match between search criteria and the selected documents.

2.1.1.4 Required Capabilities for Viewing of Documents Selected by Full-Text Search and Retrieval

The following features are highly desirable to support viewing documents selected by full-text search and retrieval queries:

- Concurrent, Multiple Document Viewing and Scrolling—because staff must frequently compare information in multiple documents, the system must permit them to select multiple documents from a search results list and view them concurrently in separate windows.
- In-Document Search Term Highlighting—the search terms matched in the document text must be highlighted to facilitate finding the desired information within the text of a document.
- Document Browsing—facilities must be provided to permit the user to move rapidly to the preceding or next successive highlighted search term.
- In-Document Text Search—text search facilities must be provided to enable the user to find information within the displayed document based on additional words and/or phrases.

2.1.1.5 Required Capabilities for Viewing Images and Accessing Documents Selected Through Full-Text Search and Retrieval

The following features are highly desirable to support viewing images and accessing machine readable copies of documents selected through full-text search and retrieval:

- Launch of Image Viewers—if the document has associated images, it is important for the user to be able to view those images quickly and efficiently. Therefore, facilities are required to permit the user to view the images by selecting icons embedded in the text.
- Launch of Word Processors—document text accessed through the full-text search and retrieval facilities may be needed in machine-readable form for incorporation in other work products. Therefore, the user must be able to select an icon from the document text display to launch the document into a word processor for viewing and/or editing in its original format.

2.1.2 Provide Staff Productivity Enhancement, Download, and Cut and Paste Capabilities for Text and Images

The TDOCS is intended to support the DWM and CNWRA staffs in performing independent reviews and analyses. These reviews and analyses build on prior work found in technical documents, and frequently require materials from such documents to be incorporated into new work products. Therefore,

the TDOCS must provide staff productivity enhancement facilities. The requirement hierarchy for staff productivity enhancement is illustrated in Figure 2-3.

2.1.2.1 Document Download Capabilities

To support DWM independent reviews and analyses, the TDOCS must provide a mechanism for downloading text and images. Both textual and image information must be accessible from staffs' computers and workstations. To support DWM independent reviews and analyses that incorporate and build on materials from other documents. Document download permits the user to obtain an electronic copy of an entire document. Unlike the cut and paste capabilities discussed in Section 2.1.2.2, document download does not require interactive highlighting and moving portions of the document. Therefore, downloading is the most efficient vehicle for obtaining electronic access to complete documents or large portions of documents.

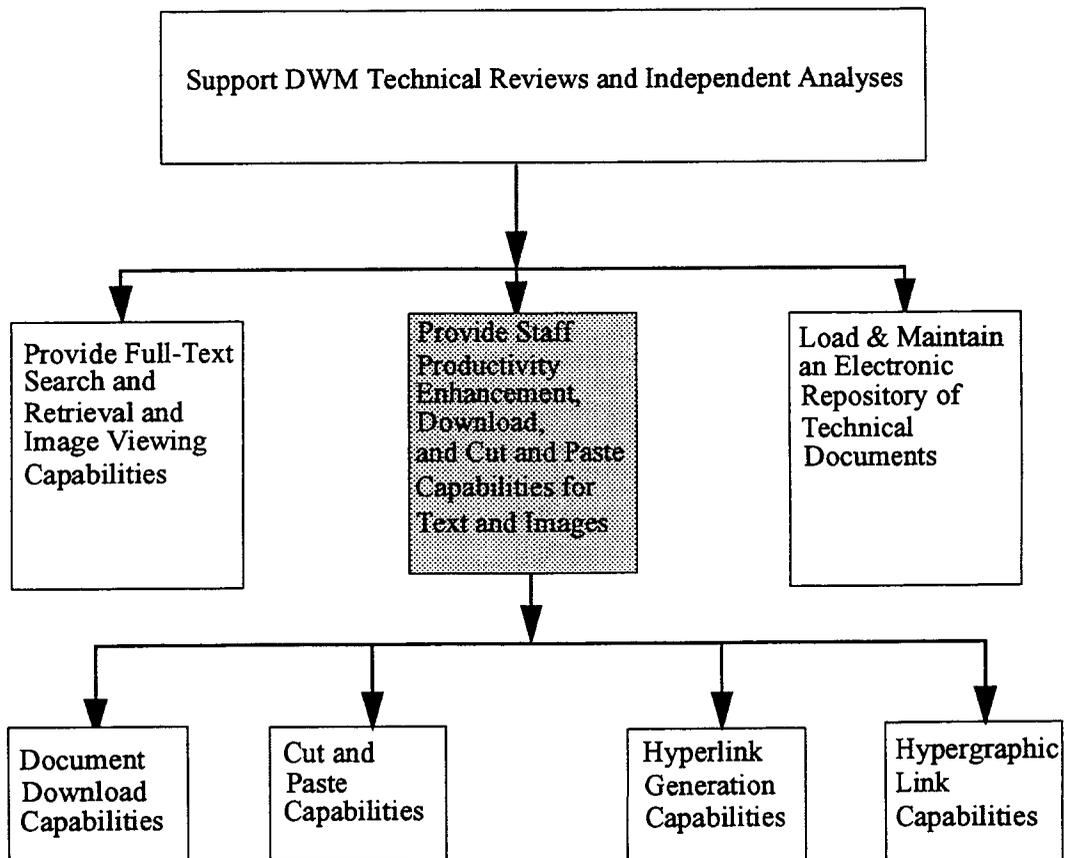


Figure 2-3. Requirements hierarchy for staff productivity enhancement

2.1.2.2 Cut and Paste Capabilities

Cut and paste capabilities provide important support for DWM staff in performing independent reviews and analyses. The TDOCS system must provide capabilities to permit text to be cut from the document viewer and pasted into another document, so that it can be saved to a local file, and/or printed. Cut and paste permits the user to highlight selected portions of text in one document and copy it to another document where it may be edited and/or combined with other materials. Cut and paste capabilities are generalized facilities that work across application and program boundaries, provided that the applications and programs adhere to standard protocols.

2.1.2.3 Hypertext Generation Capabilities

Hypertext is a facility that permits documents to be related to each other based on their content. Links are established that permit the user to move directly from document to document when viewing, and then return to the starting point in the original document. Hypertext may be used to associate related material within a document and facilitate user access on a concept basis. The TDOCS must provide capabilities to create and maintain hypertext links.

The implementation of hypertext usually takes the form of highlighted words or icons, known as launch pads, that are embedded in the document. When the user selects a launch point, the associated document is automatically displayed and positioned to the appropriate location in the text. This powerful technique permits documents to be related in highly meaningful ways so that the user may traverse a large quantity of associated materials quickly and efficiently.

Intra-document Hypertext Link Capabilities

Intra-document hypertext links provide the ability to connect related information within a single document in a way that facilitates rapid user access and traversal of the materials. Intra-document hypertext links connect text references, highlighting associated words in the text. By selecting intra-document hypertext links, the user is able to move rapidly between successive highlights within the document. The TDOCS full-text search and retrieval system must support the creation and maintenance of public intra-document hypertext links.

Inter-document Hypertext Link Capabilities

Inter-document hypertext links provide the ability to link or connect related documents for later retrieval. Inter-document hypertext links connect text references in one document to related references in other documents. Public inter-document hypertext links are shared across the system for all users. By selecting inter-document hypertext links, the user is able to move rapidly between associated documents. The TDOCS full-text search and retrieval system must support the creation and maintenance of public inter-document hypertext links.

2.1.2.4 Hypergraphic Link Capabilities

Hypergraphic links provide the ability to connect the text of documents and images in a way that facilitates retrieval and viewing of the images. Hypergraphic links connect graphic images to specified locations, called launch pads, in the text of the document. These hypergraphic links are indicated by the presence of an icon in the text. When the icon is selected, the appropriate image is displayed. The TDOCS full-text search and retrieval facilities must support the creation and maintenance of hypergraphic links.

2.1.3 Load and Maintain an Electronic Repository of Technical Documents

Document loading is the process through which materials are identified, acquired for loading in either electronic or hard-copy form, captured in an appropriate format, and loaded into the database. The process involves both automated loading of documents that are obtained in electronic form and the scanning and OCR of paper documents. The requirements hierarchy for loading technical documentary materials is depicted in Figure 2-4.

The TDOCS must support document loading on either a routine or on-demand basis. Routine loading permits staff to submit documents to a central document loading station where bibliographic headers are prepared and the documents are captured in electronic form. Electronic copies of documents may also be submitted for routine loading. In order to support management and control of the document repository, the TDOCS requires bibliographic headers for each document submitted for routine loading in either electronic or hard-copy form.

In some instances, staff may require rapid access to electronic copies of a document or portions of a document. The TDOCS must support on-demand loading to address this need and permit staff to capture several pages or even an entire document through a document scanning workstation.

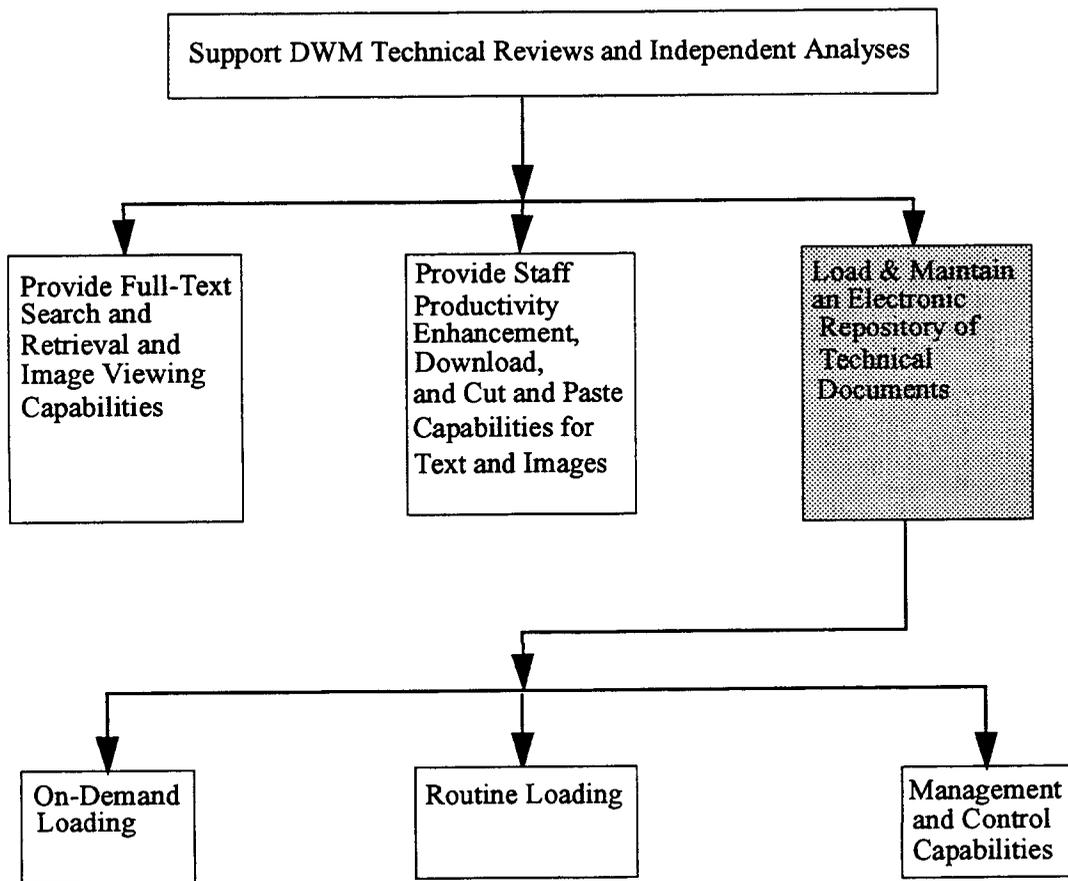


Figure 2-4. Requirements hierarchy for document loading and maintenance

2.1.3.1 On-Demand Loading

The TDOCS system must support on-demand document scanning and OCR to address immediate staff needs for limited amounts of textual and/or graphical materials. The requirements hierarchy for on-demand loading of technical documents is depicted in Figure 2-5.

On-demand scanning and OCR capabilities are not a substitute for routine loading of documents. These capabilities are simply a method for staff to obtain access to a full-text and electronic copy of a few pages of a particular document on an expedited basis. A typical request for on-demand scanning and OCR arises when a staff member needs several pages from a report or journal article to incorporate into another document. The staff member takes the relevant hard-copy pages to a scanning station where they are scanned and converted to American Standard Code for Information Interchange (ASCII) text through an OCR process. This function is accomplished either by the individual staff member or by another designated member of the staff. The result of on-demand loading is an ASCII file of the textual information, possibly accompanied by images of the figures, equations, etc., that appeared on the scanned pages. In this way, staff members are able to obtain electronic copies of small numbers of pages of relevant text and graphic materials without having to wait for submission and indexing through the routine loading process. The volume of materials processed in this manner is limited because of the availability of the scanning station due to anticipated routine loading demands.

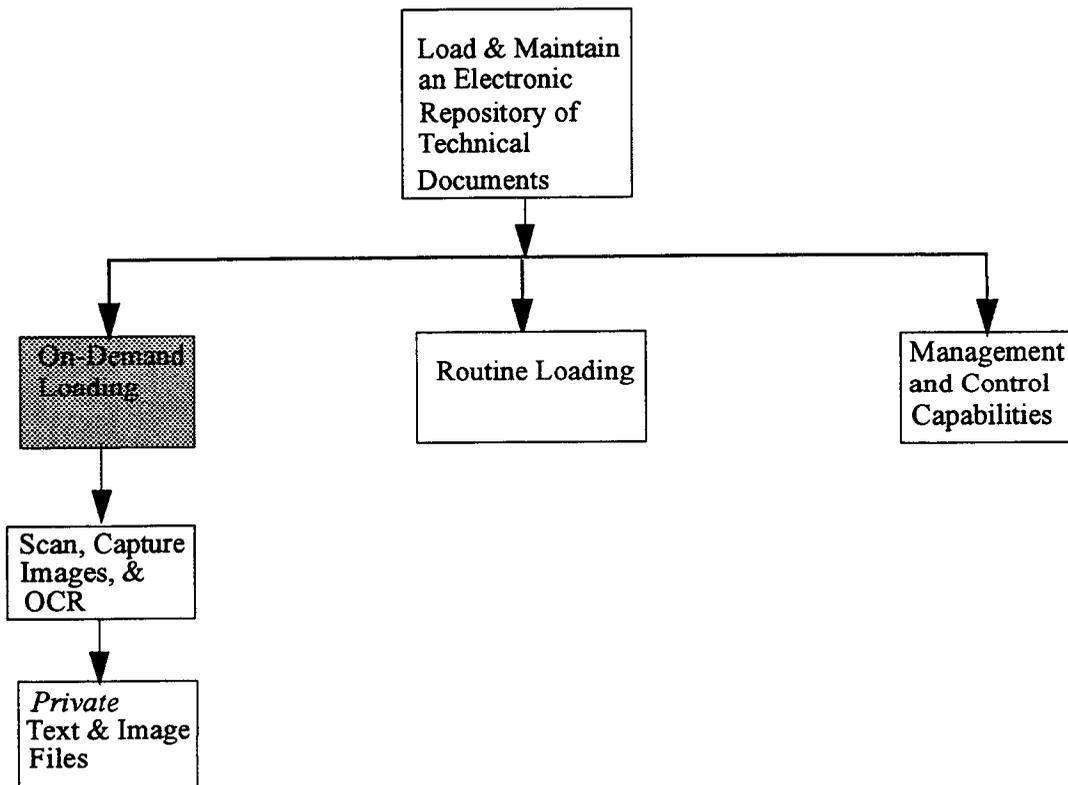


Figure 2-5. Requirements hierarchy for on-demand document loading

Materials submitted for on-demand loading must be made available to the requestor as ASCII text with associated image files. However, the materials scanned on an on-demand basis are not loaded into the permanent repository because they generally represent partial or incomplete documents. Therefore, on-demand loading must also permit concurrent requests for routine loading of the full document from which the on-demand pages were selected. If the complete document is scanned in an on-demand mode, the resulting files can be passed to a staff member who is responsible for routine loading for final processing and loading into the database through the routine loading flow.

2.1.3.2 Routine Loading

Processing through routine loading is the only method for entering materials into the *public* TDOCS repository. Requests for routine loading identify individual documents and/or groups of documents that will be obtained and loaded on a one-time, periodic, or cyclical basis. Routine loading activities must be designed to handle a substantial document volume and a large percentage of the document requirements of the staff should be satisfied through routine loading. The requirements hierarchy for routine loading of technical documents is depicted in Figure 2-6.

The TDOCS system must implement two types of routine document loading to meet the needs of the DWM and CNWRA staffs: (i) routine loading of electronic copies of documents, and (ii) routine loading of hard-copy documents.

Many documents are available in electronic form. These documents are submitted for electronic loading with a minimum of manual intervention and administrative overhead. Other documents are available in hard copy form and must be scanned and processed through OCR to obtain usable electronic copies of the text.

Loading of Electronic Copies of Documents

Wherever possible, electronic copies of materials should be obtained and loaded to avoid the labor and system overhead associated with document scanning, OCR, and cleanup operations. These materials may be received on a variety of electronic media including magnetic tape, diskettes, optical disks, or via communications facilities from other databases or on-line bibliographical services. Many of the materials to be loaded into the TDOCS system are available in electronic form. This is particularly true for materials generated internally by the DWM and the CNWRA. Such materials are normally available as full-text with embedded or accompanying image files of figures, equations, images, etc.

The availability of electronic copies of technical documents does not mean that document processing is not required. Electronic copies of documents and materials exist in a variety of formats, including word processing formats, such as WordPerfect, desktop publishing, such as FrameMaker, plain text in ASCII format, document interchange formats such as the Standard Generalized Markup Language (SGML), and others. To support a consistent and efficient environment for document search, retrieval, and presentation, the input documents must be converted to ASCII format prior to loading. Format conversion requires that the format of the input document be clearly identified. Identification of the format of electronic copies of documents submitted to the TDOCS and conversion of the text to ASCII is the responsibility of: (i) the requestor obtaining the electronic copy, or (ii) the TDOCS custodian. The output of the electronic loading process is a full-text instance of the document in the repository and image files for the submitted document in the file system.

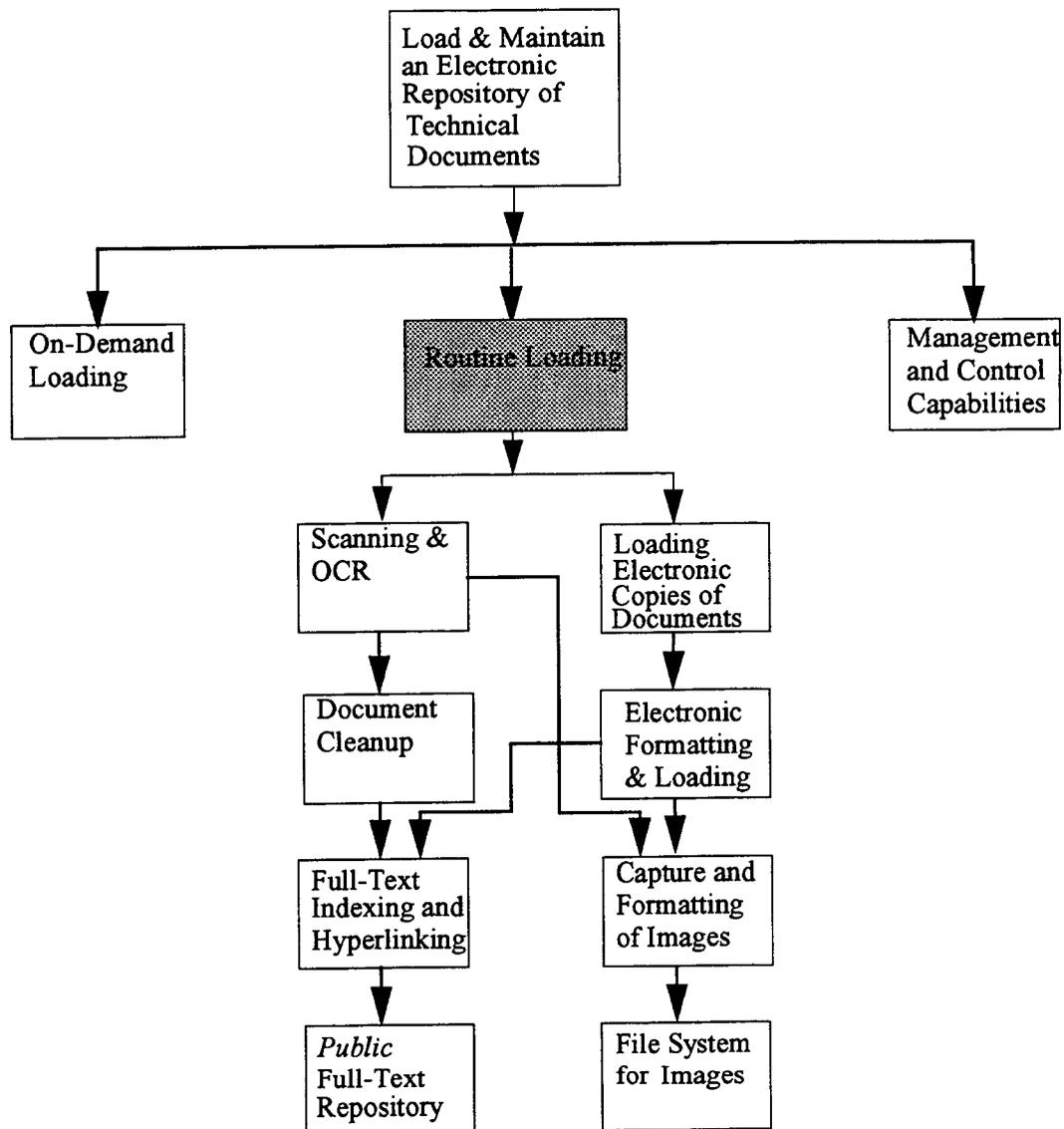


Figure 2-6. Requirements hierarchy for routine document loading and maintenance

Scanning, Optical Character Recognition, and Cleanup

The process of scanning materials and converting them to full text through OCR is required to load paper-based documents into the system. Materials that are not available in electronic format must be obtained as hard-copy documents and processed by scanning and OCR to obtain an electronic copy. This is a semiautomated process that is subject to errors that may arise from the condition of the documents, the quality of the printing, etc. Scanning and OCR errors may or may not be detected automatically. Therefore, a manual document cleanup procedure is required as part of the document capture process to ensure document accuracy. Following the completion of the scanning, OCR, and cleanup processes, an electronic copy of the document is available for loading into the system.

Full-Text Indexing

Users of the TDOCS will rely heavily on full-text searches to find desired materials. Therefore, full-text indexing is required as an integral part of document loading and processing. The indexing of documents must be accomplished automatically as part of the loading process. Header information must be entered during the document preparation and processing steps and loaded so that documents may be retrieved through structured header searches. Headers for materials that consist solely of images must be processed, formatted, and loaded into the full-text repository for search and retrieval through the full-text system.

Full-Text Repository

A full-text repository is required to store the textual portions of documents, along with the necessary indexes, and control information to support full-text search and retrieval.

File System Repository for Image Files

When images are associated with a document, they must be captured as formatted files. Image files are distinct from the text files. They must be stored along with the text files in a separate repository with appropriate links to associate the images with the corresponding text.

2.1.3.3 Management and Control Capabilities

The TDOCS must include capabilities for management and control of the system. The requirements hierarchy for management and control capabilities is depicted in Figure 2-7.

Bibliographic Headers

The TDOCS utilizes both bibliographic header and full-text searches. Full-text search and retrieval is provided to permit users to search for occurrences of words and/or phrases in the text of documents. Bibliographic header searches are used to permit users to find documents more precisely and quickly when specific information is known, such as the title, author, date, etc. They are also used to support control of the database and reporting functions.

Bibliographic header preparation and entry must be performed for all documents loaded through the routine loading process. The bibliographic headers permit searching for documents by author, title, date, and other standard bibliographic fields. Documents that are not textual in nature, such as geologic maps, photographs, or frames captured from videotape, also require descriptive headers because these headers provide the only method for identifying and describing such materials. A specialized entry screen must be provided to help the operator complete the header by "filling in the blanks." When electronic copies of technical documents and materials are obtained for loading, they may or may not be accompanied by electronic copies of bibliographic headers. When available, these electronic copies of the bibliographic header information must be in the proper format before they can be loaded automatically into the TDOCS. To provide for automatic conversion of selected electronic headers to the standard TDOCS header format, specific conversion routines must be developed. Electronic copies of non-standard header formats must be entered manually as part of the loading process for electronic copies of documents.

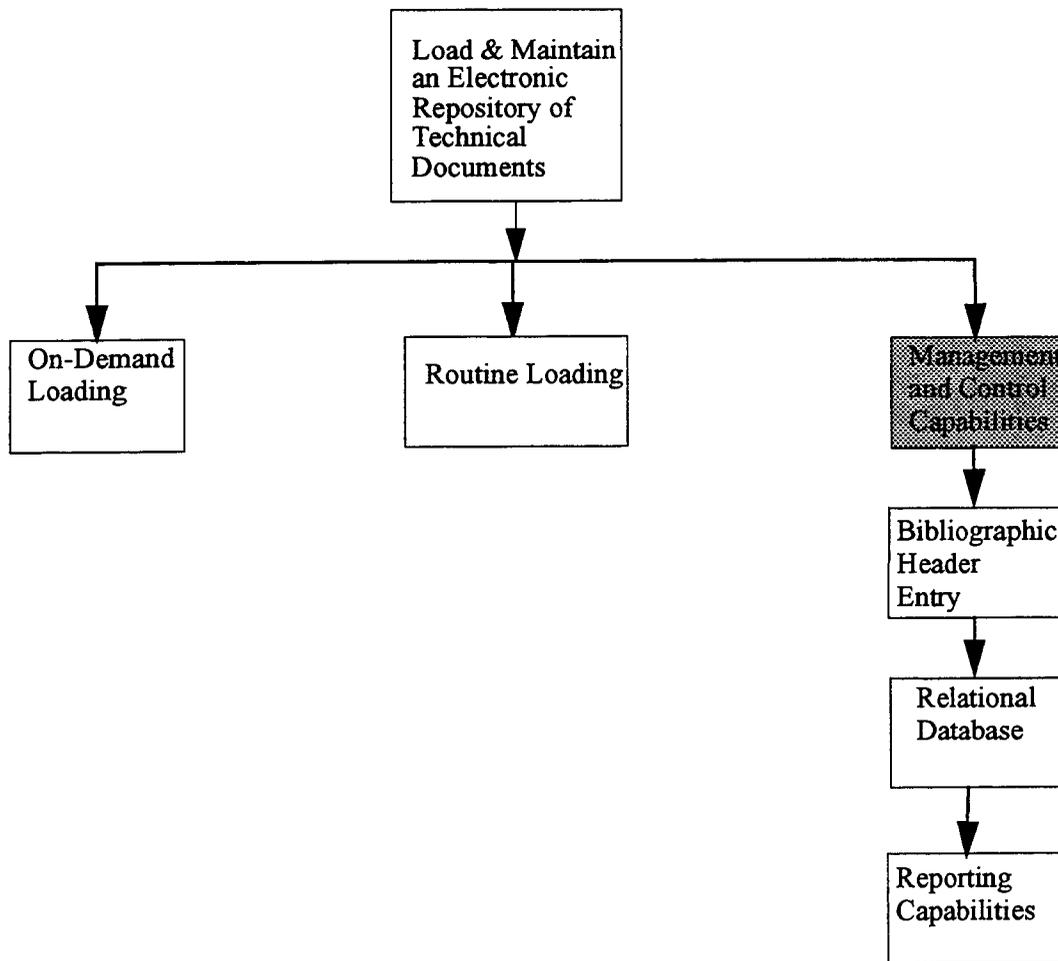


Figure 2-7. Requirements hierarchy for management and control capabilities

The preparation of bibliographic headers must serve two functions:

- Bibliographic headers are captured, validated, and entered into the relational database to support automated production of bibliographic listings and other reports.
- The captured bibliographic headers are formatted appropriately and used for loading header fields in the full-text repository to support structured searches against the header fields.

Relational Database

In order to manage and control the TDOCS repository it must be possible to list and review the status of documents. This requires the ability to search, retrieve, and report bibliographic information about documents. Bibliographic headers are captured and stored in a relational database to address these requirements.

The relational database management system (RDBMS) is one of the components of the TDOCS database. The RDBMS provides data integrity, an interface through the C programming language, and

support document loading, configuration control, and database reporting. The RDBMS may also be used in a secondary role to support downloads from other databases (e.g., from NUDOCS).

The RDBMS also serves a role in the administration and maintenance of user accounts. All users are provided accounts stored in the RDBMS. These accounts consist of user name, password, and privileges. A user's name must be unique within the TDOCS but may be the same as that used on other systems. Users with passwords are able to change their passwords at any time.

Reporting Functions

The TDOCS must support reporting functions for management and control of the document repository.

Classes of reports will include the following:

- Acquisitions and new materials reports
- Document statistical reports
- Bibliographic listings

2.2 DESIGN OVERVIEW

The constraint to support multiple hardware/software platforms has very significant implications for the design and implementation of the TDOCS. Much of the required functionality of the system is relatively independent of the users' hardware/software platforms and should be centralized on a shared facility. Other required functionality, primarily related to the user interface and local processing capabilities, is quite platform-specific. This environment suggests the need for a client/server architecture in which the user interface is distributed to individual personal computers (PCs) or workstations, while the code supporting common functionality is implemented on a database server platform.

The system architecture required to support multiple user hardware/software platforms involves a distributed processing client/server environment where much of the application code and support for the user interface is resident on the users' PCs or workstations. Database storage and management functions are resident on the server platform and operate independently and asynchronously. The resulting balanced processing load across the system is more responsive to processing demands and is able to accommodate growth by adding user workstations incrementally and upgrading the capabilities of the database server when required. Separation of function is central to the TDOCS design.

- A central server facility, resident on a high-speed computer, services requests for information from all users and handles the database and file management functions.
- Client facilities, resident on the users' PCs and workstations, handle the user interface and certain local processing functions.

This approach permits users to interact efficiently with dedicated facilities on their own PCs or workstations to enter commands, search for records, and view results while sharing a central repository of data and documents. The client/server architecture maximizes utilization of the capabilities of local hardware and the network architecture.

The TDOCS application, illustrated in Figure 2-8, uses a client/server architecture consisting of three major modules:

- Document Management Server—code that supports the maintenance and use of the TDOCS relational and full-text data repositories
- Document Processing Clients—code that supports user interaction for document loading and maintenance
- Document Search and Retrieval Clients—code that supports user interaction for search and retrieval of documents in the TDOCS full-text repository

The Document Management Server design is discussed in Chapter 3. That discussion covers the following topics:

- Server Hardware
- Server Software
- The TDOCS Data Repositories
- Functional Management and Capabilities of the Document Management Server
- Document Processing Services
- Document Search and Retrieval Services
- Document Download
- Batch Loading, Indexing, Hyperlinking, and Security
- Database Administration and Maintenance

The Document Processing Clients design is discussed in Chapter 4. That discussion covers The following topics:

- Document Processing Clients Hardware
- Document Processing Clients Software
- Functional Management and Capabilities of Document Processing Client
- Scanning, Optical Character Recognition, and Cleanup
- Document Submission, Deletion, and Maintenance
- Database Administration

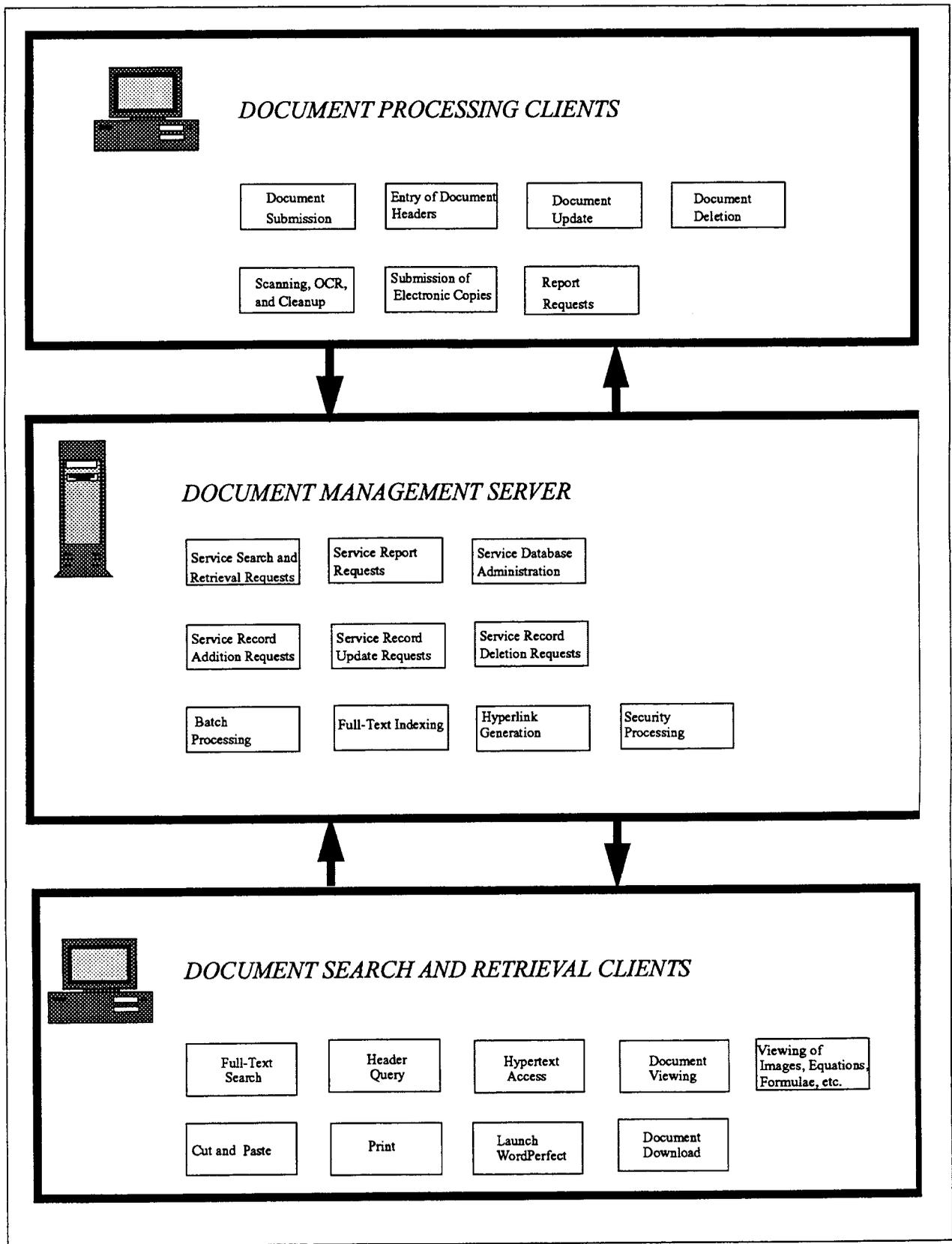


Figure 2-8. Major technical reference document database system functions

The Document Search and Retrieval Clients Design is discussed in Chapter 5. That discussion covers the following topics:

- Document Search and Retrieval Clients Hardware
- Document Search and Retrieval Client Software
- Functional Management and Capabilities of Document Processing Client

The Network Architecture is discussed in Chapter 6. That discussion covers the following topics:

- Remote Procedure Call
- Network File System
- AUTOS and ACRS
- Security

In the TDOCS client/server design, the Document Management Server resides on the NMSS TDOCS Sun server. Document Processing and Document Search and Retrieval Clients reside on Sun workstations using OpenLook or Motif, or on PCs using Microsoft Windows. Clients use Transmission Control Protocol/Internet Protocol (TCP/IP) and allow a client application to make calls to functions that are part of a server application. Network File System (NFS) allows a client system to access remote file systems. These communications protocols are available on the NRC Agency Upgrade of Technology for Office Systems (AUTOS) network and the NMSS ACRS network.

The logical design of the TDOCS application does not place restrictions on implementation. The Document Management Server may well be implemented as a composite of services and utilities. The Document Processing and Document Search and Retrieval Clients may well be combined with user privileges determining which portions of the functionality are accessible. Design dictates what must be implemented, not how implementation will be achieved.

3 DOCUMENT MANAGEMENT SERVER DESIGN

The Document Management Server consists of the TDOCS Database and TDOCS Services modules. The database serves as the repository for document headers, text, images, and indexes. The TDOCS Services modules provide access to the database, whether client-requested document submittal, server-based document loading, or administration and maintenance. Direct access to the database for administrative purposes is limited to authorized system administrators and database administrators (DBAs). All other access to the database is through the facilities of the Document Management Server.

3.1 SERVER HARDWARE

The impact of the TDOCS on existing and planned systems and configurations, namely AUTOS and ACRS, must be minimized. The present DWM server capabilities are illustrated in Figure 3-1. These server capabilities must accommodate both the initial and production implementations of the TDOCS. When installed at the DWM, the TDOCS will reside on the SPARC-1000 server and will be accessible through the existing network and communications lines. As document volumes grow, tradeoff decisions may require additional or enhanced hardware for data storage, high-resolution display of images, and additional communication lines.

3.2 SERVER SOFTWARE

The TDOCS server requires software to support relational database, full-text search and retrieval, and application functionality.

3.2.1 Relational Database

A fully Structured Query Language (SQL)-compliant relational database facility is necessary to support TDOCS document loading, configuration control, and reporting functions.

3.2.1.1 Considerations

The TDOCS design conforms to the open systems and SQL standards to support a wide diversity of present and anticipated hardware and software configurations, and ensure compatibility with current and future SQL database syntax. Standardized and extensible gateways and connectivity approaches are implemented to permit smooth and consistent integration of diverse operating environments and networks.

The RDBMS is a major component of the server database. The RDBMS provides data integrity, an interface through the C programming language, and capability for password-protected access. The primary role for the RDBMS is to store document headers for configuration control and database reporting. Database reports are drawn from the RDBMS on the database server, formatted, and transferred via a remote procedure call (RPC) to the client workstation. The RDBMS also serves a role in the administration and maintenance of user accounts. All users are provided accounts stored in the RDBMS. The RDBMS could also be used to support downloads from other databases (e.g., from NUDOCS).

To comply with the requirements of the task and to expedite the implementation of TDOCS, commercially available software was evaluated and used where possible. The specific application requirements of TDOCS precluded finding commercial software capable of supporting the entire system, but major functional areas were addressed through appropriate, commercially available, software

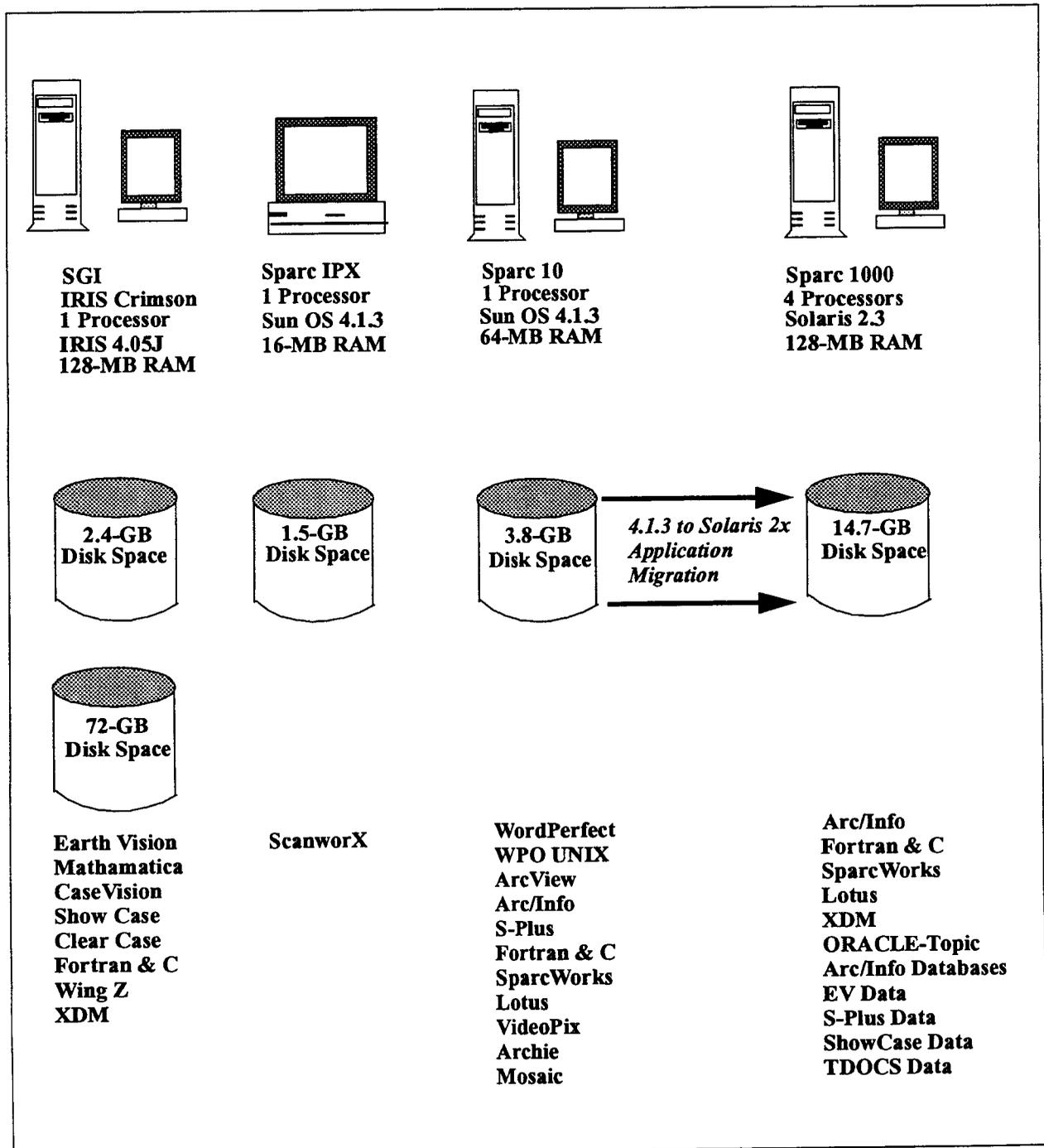


Figure 3-1. Existing Division of Waste Management server capabilities

packages. The use of commercially available database software is feasible and supported by prior experience with the development of the Program Architecture Support System/Program Architecture Database (PASS/PADB), which utilized SQL-DS, and the Regulatory Program Database (RPD) and NUDOCS, which utilized ORACLE. Client application code to support interactive and *ad hoc* queries is accessed from PC and workstation platforms. Due to the specific nature of the application, its dependency on the structure of the relational database tables, and its interaction with the search and retrieval software, application code was developed using the C programming language. This C application code accesses the

facilities of the commercially available off-the-shelf database and full-text search and retrieval tools through their respective application program interfaces (APIs).

3.2.1.2 Evaluation

An analysis of relational database products was performed by the CNWRA as part of the initial design of the TDOCS and RPD applications (DeWispelare et al., 1993). This analysis resulted in selection of the ORACLE relational database software for the TDOCS and RPD applications. The TDOCS database server application code was written in the C language to utilize the facilities of the ORACLE database software. This code is specific to the ORACLE API and is not adaptable to other relational database products without modification. Therefore, the ORACLE relational database is required to utilize the existing code of the TDOCS application. Other relational database products, even if functionally similar to ORACLE, cannot be utilized with the TDOCS application without code modification to the TDOCS database server modules. The following paragraphs summarize the evaluation of relational database products performed by the CNWRA as part of the initial design of the TDOCS and RPD applications.

In the three years prior to the implementation of the TDOCS and RPD applications, UNIX workstation-based SQL-compliant relational databases were developed by Oracle Corp. (ORACLE), Informix Software (Informix), Sybase Corp. (SYBASE), and Relational Technology Inc. (Ingress).

An evaluation of the features offered by ORACLE, Ingress, Informix, and SYBASE (Rodgers, 1990; Khoshafian, et al., 1992) produced the following observations:

- Only the ORACLE and SYBASE RDBMSs implement cost-based optimizers.
- ORACLE is the only UNIX database management system that has been certified as Federal Information Processing Standards (FIPS) compliant.
- The ORACLE and Informix database systems implement datatypes compatible with those of IBM's SQL/DS product.
- Of the four database management systems evaluated, only ORACLE supports all three levels of locking (row, page, and table).

The evaluations indicated that ORACLE offered the best range of functionality to support the TDOCS relational database. ORACLE Versions 6.04 for Sun and 6.036 for IBM AIX were obtained directly from Oracle Corp. for an in-house test and evaluation at the CNWRA. Performance measurements (large-table load times, index creation, two-table joins [merging], etc.) were comparable between the two machines.

From the evaluations, the performance of the ORACLE database on the UNIX workstations was considerably better than that of the IBM SQL/DS database on the Southwest Research Institute (SwRI) mainframe. Long-running maintenance functions (table loads, index creations) were accomplished on the workstations in a fraction of the time required for the mainframe. For example, unloading a 480,000-row table (27 MB of data) from SQL/DS took 37 min. Loading the table into ORACLE on the Sun workstation was accomplished in under 5 min. While loading and unloading a table are not identical processes, they are comparable in terms of I/O operations and timings. Indexing the same table was accomplished in ORACLE in 7 min versus 39 min for SQL/DS. The performance observations reflect the increased

processing power of the Sun and IBM workstations over that of the IBM mainframe, as well as the effects of the heavy load on the IBM mainframe.

3.2.1.3 Recommended Relational Database Management System

The CNWRA evaluation recommended that the ORACLE database software be used to implement the relational database portion of the TDOCS application. ORACLE provides the feature set needed, maximizes past SQL experience by CNWRA staff, and has the necessary compatibility with existing NRC and CNWRA systems. It can be interconnected with the selected search and retrieval software (TOPIC) and the application environments of both the NRC UNIX workstations and PCs by C programming.

3.2.2 Full-Text Indexing Search/Retrieval Software

A highly capable full-text search and retrieval facility is necessary to support TDOCS document search and retrieval, viewing, browsing of document text, and display of associated images.

3.2.2.1 Considerations

The full-text database server tool consists of: (i) a specialized database of indexed text and document header fields, and (ii) full-text search and retrieval software. The RDBMS provides for storage of header information in relational database tables. The file system provides for storage of documents and images. The full-text search and retrieval facility provides for password-protected indexing, and search and retrieval of headers and document text.

The full-text search and retrieval software indexes and stores the text of documents and headers in the full-text database. It also provides the GUI software for structured header queries and full-text document searches, as well as the mechanisms for document viewing, *cut and paste*, hyperlinking, and the launching of WordPerfect and image viewers. Requests for these search and viewing capabilities are made through the full-text search and retrieval software.

Off-the-shelf commercially available software products to perform the TDOCS full-text-search and retrieval functions were evaluated during the design phase. Extensive discussions were held with personnel from the Department of Defense (Air Force Intelligence Command), where a similar application had been successfully installed and used in a comparable computer system environment. Due to the specific nature of the application, its dependency on the functions of the search and retrieval software, and its interaction with the relational database software, application code was developed using the C programming language. This C application code accesses the facilities of the commercially available off-the-shelf RDBMS and full-text search and retrieval tools through their respective APIs.

3.2.2.2 Evaluation

An analysis of full-text search and retrieval products was performed by the CNWRA as part of the initial design of the TDOCS and RPD applications (DeWispelare et al., 1993). This analysis of products resulted in the selection of Verity's TOPIC full-text search and retrieval software for the TDOCS and RPD applications. The TDOCS database server and workstation client application code was written in the C language to utilize the facilities of the TOPIC full-text search and retrieval software. This code is specific to the TOPIC features and API, and it is not readily adaptable to other full-text search and retrieval

products without modification. Therefore, the TOPIC full-text search and retrieval product is required to utilize the existing code of the TDOCS application. Other full-text search and retrieval products, even if functionally similar to TOPIC, cannot be utilized with the TDOCS application without code modification to the TDOCS database server and search and retrieval client modules. The following paragraphs summarize the evaluation of full-text search and retrieval products performed by the CNWRA as part of the initial design of the TDOCS and RPD applications.

The full-text search and retrieval system provides the repository function for full-text documents and the access methods to search for and retrieve documents from that repository. The TOPIC system from VERITY Corporation and Zyindex-based system from CERTREC were evaluated in-house at the CNWRA. TOPIC was chosen for evaluation based on support for the required client platforms (Sun, 1991) and the experiences of a TOPIC user at a local U.S. Air Force installation. The CERTREC¹ implementation of Zyindex was evaluated as part of a previous requirement to support a CNWRA user who needed NUREG-0800 on-line.

Numerous documents were loaded into TOPIC, including Part 60 of Title 10 of the Code of Federal Regulations, NUREG-0800, Systematic Regulatory Analysis (SRA) products such as Regulatory Requirement Topics (RRTs) and Compliance Determination Strategies (CDSs), CNWRA correspondence, and CNWRA Technical Document Index (TDI) headers. Search times across approximately 1,200 documents were consistently small (<1 sec for first *hit* and <15 sec total search time). Search and retrieval were accomplished from Sun, Macintosh, and OS/2 clients.

The Zyindex-based product was used for retrieval of NUREG-0800. Performance was acceptable. However, the Zyindex character interface was judged inferior to the TOPIC windowed interface by end users. Further consideration of Zyindex was ruled out due to inability to support Sun workstation clients.

The TOPIC full-text search and retrieval system has a number of features and capabilities that are either totally absent or severely limited in other alternative packages. These features and capabilities are utilized by the TDOCS implementation and/or planned future enhancements.

- A library of pre-defined searches and interest profiles that serve as a knowledge base for organization of expertise
- Concept based search capabilities that result in superior selection of search results and provides more precision while maintaining completeness
- GUI-based client support for required DWM/CNWRA staff hardware/software platforms
- Security at multiple levels including collection, document/file, and document content
- Ability to use SQL databases as document repositories
- Extensive annotation capabilities including cross-reference, search term, multi-media, and application launch links in documents

¹ Personal communication made under a non-disclosure Memorandum of Understanding between Southwest Research Institute and the United States Air Force.

- Links to compound documents containing text, images, video, audio, etc.
- Document storage capacities greater than 100 gigabytes

3.2.2.3 Recommended Full-Text Indexing Search/Retrieval Software

The TOPIC software product was selected and employed by the CNWRA design for use as the indexing and full-text search and retrieval components of TDOCS. There are several advantages to using the TOPIC software:

- It can be implemented using the NMSS/DWM UNIX client/server advanced computer environment and associated PCs
- It allows the storage of documents in WordPerfect format using WordPerfect *filters*
- It provides for a robust concept-based search mechanism
- It can be interconnected with the selected relational database software (ORACLE) and with the display environments of both the NRC UNIX workstations and PCs by C programming.

3.3 TECHNICAL REFERENCE DOCUMENT DATABASE DATA REPOSITORIES

To implement TDOCS using the client/server architecture discussed in Section 2.2, the relational, full-text, and image data repositories must reside on a single central server. The decision to implement the TDOCS using this client/server architecture with the server facilities resident on a UNIX-based platform implied that the server portion of the relational database and full-text search and retrieval facilities must be able to operate in a UNIX environment. The effect of this was to impose a major constraint on the selection of RDBMS and full-text search and retrieval software.

The server database is the primary repository for document headers, text, images, and indexes. This database resides on a server computer that provides sufficient capacity to support the database storage functions and the projected number of users. Information in the TDOCS is stored in several repositories to support different system requirements:

- A relational database management system for control information
- A file system for text, word processing, and image files
- A full-text index to support search and retrieval capabilities

Although multiple software tools are used to implement this module, it is accessed and maintained as a single database. Headers are stored in the RDBMS and indexed in the full-text search and retrieval system. Document text and images are physically stored in the file system, and the textual information is indexed in the full-text repository.

In addition to the RDBMS and the full-text repositories, *archive* text and image repositories reside in the file system of the central server platform. The UNIX file system provides for

password-protected access as well as group and user level read/write privileges on directories and files. Access is controlled according to privileges assigned to individual users or groups of users. Password-protected access and read/write privileges are used to protect documents, headers, and their indexes from unauthorized modification or deletion. The UNIX system also provides many of the utilities necessary for system administration and maintenance.

The primary function of the file system is the storage of documents. Both the RDBMS and full-text search and retrieval facilities make extensive use of the UNIX file system for storage. Text and images are stored in partitions that correspond to sets of documents. These partitions are formed as directory structures that parallel the partition structure used to store indexes in the full-text search and retrieval facility. Examples of document sets include NRC technical documents (NTD) at the DWM and TDI, quality assurance records (QA), and correspondence (CSP) at the CNWRA.

3.3.1 Directory Structures in the UNIX File System

Document storage in the public TDOCS repository, unlike private files which result from on-demand scanning and which are taken back to an individual's own computer, requires support for text, images, full-text display, and hyperlinks. Accordingly, parallel directory structures are provided in the UNIX file system, as illustrated in Figure 3-2:

- *archive* directories—contain the text of the document as originally submitted to the system
- *image* directories—contain full-page images as well as any images of figures, equations, etc.
- *data* directories—contain copies of the document text formatted to support full-text display and hyperlink functionality

Under each of the primary directories (i.e., *archive*, *images*, and *data*), there are multiple partitions (directories) corresponding to the document sets within TDOCS. Under each document set partition, there are multiple sub-partitions (subdirectories). The purpose of these sub-partitions is to accommodate a limitation that permits a maximum of 1,024 file name arguments to be passed to a command. Therefore the first 1,024 documents added to a document set are stored under the first subpartition; the next 1,024 documents are stored under the second subpartition, etc. The structures of the *archive*, *images*, and *data* directories are parallel down to the subpartition level. Below this level, the structures diverge to accommodate the requirements for each type of information.

3.3.1.1 Archive Subdirectories

The partitions and subpartitions under the *archive* directory contain the original text. This text may have been entered in ASCII or in a word processing format such as WordPerfect. Therefore, two subdirectories are created under each *archive* subpartition to accommodate ASCII and word processing text:

- *txt*—files stored in ASCII format
- *bin*—word processing files stored in word processing (binary) format

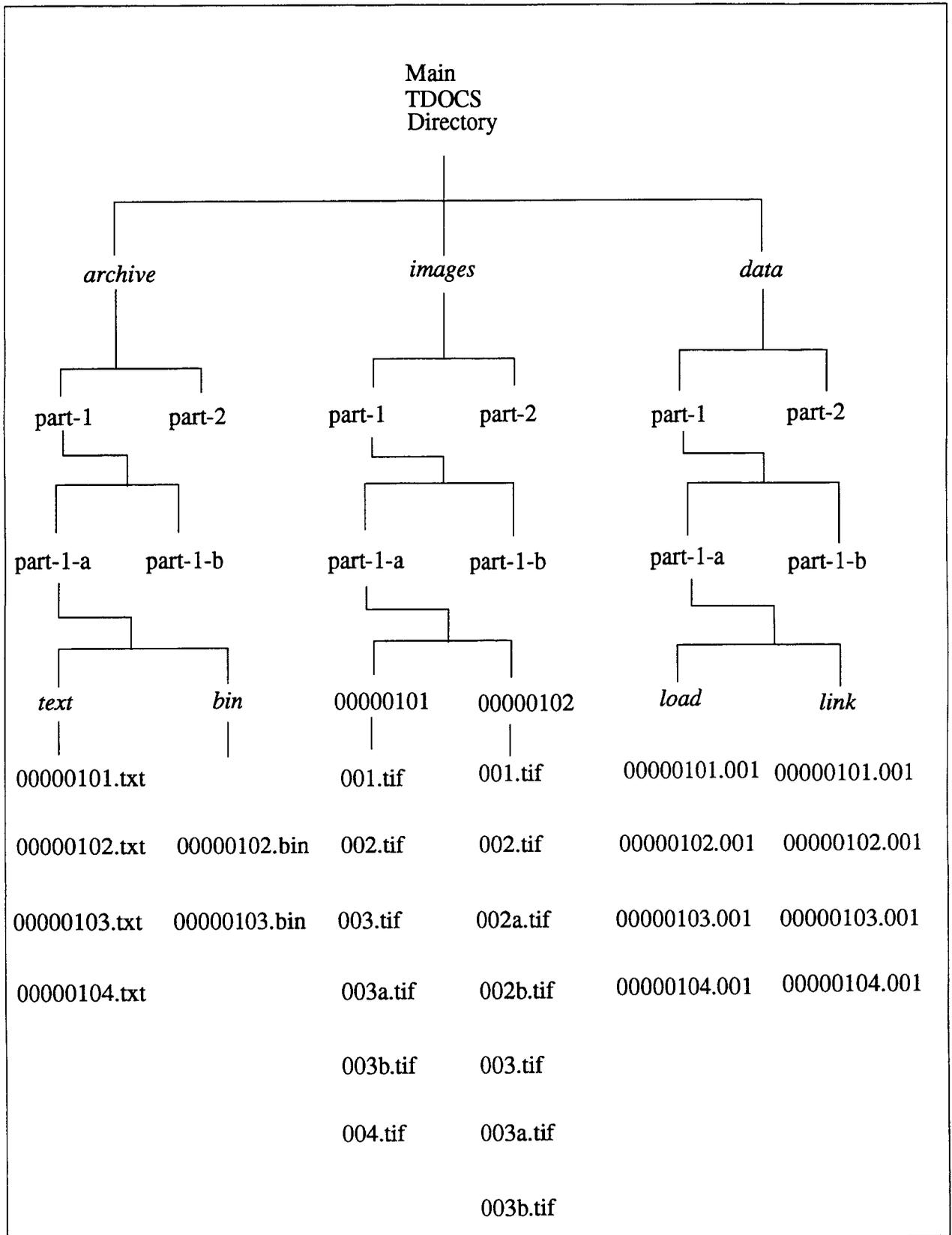


Figure 3-2. Directory structures in the UNIX file system

The Document ID, a unique internally generated number associated with each document, is used as the file name of the ASCII and word processing files under the *archive* directory. An appropriate file type is used as the file extension to indicate the format of the text. ASCII files always have a file name extension of *txt*. Word processing files always have a file extension of *bin*.

An ASCII copy of each submitted document is required to support loading and indexing of the document for full-text search and retrieval. When a document is submitted in word processing format, it must be converted to ASCII for this purpose, but the word processing file is also stored. The ASCII copy of the text is stored in the *txt* directory, and the corresponding word processing file, if any, is stored in the *bin* directory.

3.3.1.2 Images Subdirectories

The partitions and subpartitions under the *images* directory may contain the full-page images and/or images of figures, equations, etc. An *images* directory will exist for a document if and only if images are submitted with that document. Thus, a document submitted with images will have the submitted image files stored in a separate subdirectory that is named with the unique internally generated Document ID for that document. The submitted image files may represent full-page images and/or individual figures. The *images* subdirectory for a scanned document will normally contain an image file for each page. These full-page image files use their respective page numbers as file names (e.g., *001*, *002*, *003*, etc.). Image files for figures, equations, etc., are named with a composite of the page number and an alphabetic suffix indicating the image sequence within the page (e.g., *001a*, *001b*, *002a*, *003a*, etc.). The file name extensions for image files indicate the format of the file (e.g., *gif*, *tif*, *pcx*, etc.).

3.3.1.3 Data Subdirectories

The partitions and subpartitions under the *data* directory contain processed copies of the document text that are used to support full-text display and hyperlinks. Under each *data* subpartition, there is a *load* subdirectory and a *link* subdirectory. Each document has a copy of the text in both the *load* and *link* subdirectories, and these files have identical file names that consist of the system-generated document number, and a file name extension of *001*. The file extension permits future implementation of version control, if and when it is required.

The *load* subdirectory contains a single file for each document. The document text is prefixed with a formatted ASCII copy of the document header and includes the document text and embedded hyperlink labels. The files in the *load* subdirectory are used to support viewing of the document.

The *link* subdirectory contains a single file for each document. The text for each document is prefixed with a formatted ASCII copy of the document header and includes the document text and embedded hyperlink labels. The files in the *link* subdirectory are input to the full-text indexing process.

3.4 FUNCTIONAL MANAGEMENT AND CAPABILITIES OF THE DOCUMENT MANAGEMENT SERVER

The document management server manages all access to the TDOCS database and provides services to the client functionality. Major capabilities include the following:

- Logon processing

- Password maintenance

3.4.1 Server Logon Processing

For database integrity considerations, certain users will have passwords that allow them to access custodian and DBA functionality. The main purpose of logging in to the TDOCS is to distinguish custodians and DBAs from other users. Users other than custodians and the DBAs will logon simply by activating their icons. These users will have read privileges and will not have access to the custodian and DBA functionality.

When the TDOCS system is started from the client platform, a logon screen appears for custodian and DBA users that permits entry of the *User-ID* and *Password*. This information is formatted into an RPC message and sent to the server where it is validated. The server uses this information to attempt a logon to the relational database. If the logon fails, a reply including error status codes is sent to the client indicating that the User-ID and/or password was incorrect. If the logon is successful, the server retrieves the appropriate record from the user privileges table in the relational database. The user privilege record contains a user-class code that defines the permissions for the user. If no errors are detected, a reply is sent to the client including the user class code and status to indicate that the logon was successful. For a further discussion of the client logon process, see Section 4.3.1.

Custodian users are presented with a user interface that includes functionality for submitting, deleting, and updating documents. DBA users are presented with a user interface including system functions for user account maintenance.

3.4.2 Changing Passwords

Custodian and DBA users may change their passwords at any time by selecting the *Change Password* entry from the *Operations* pull-down menu (see Section 4.3.1). The client code accepts duplicate entry of the password in two fields, which are not displayed as they are entered. The duplicate password entries are validated and compared. If they are not valid or are not identical, an error message is displayed to the user by the client. When valid duplicate passwords have been entered, the client formats an RPC message and sends it to the server. The server updates the password for the user in the relational database security facilities. Upon completion of the change password transaction, the server sends a reply including status codes to the client and the client displays an advisory message to the user.

3.5 DOCUMENT PROCESSING SERVICES

Document Processing Services include the following facilities for entering and maintaining documents in the TDOCS relational database and full-text repositories:

- Submitting scanned documents
- Submitting documents for electronic loading
- Deleting documents
- Updating documents

3.5.1 Submitting Scanned Documents

When a request to submit a document is initiated by the client, text and image files for the document being submitted are stored in the *submit* directory on the client platform and an RPC message is sent to the server to initiate document submission processing. The RPC transaction includes a flag indicating the type of transaction, the document header information, and an indication of the types of files that have been stored in the *submit* directory on the client platform. For a discussion of client processing to submit new documents, including the structure of file names and file formats, see Section 4.5.1.1.

The client displays the header entry screen to the custodian and accepts input. The custodian user enters the header information and signals completion of the update by: (i) selecting the *Update* push-button, (ii) selecting the *Import* push-button, (iii) selecting the *Clear* push-button, or (iv) selecting the *Close* push-button. If either the *Clear* or *Close* push-button is selected, the transaction is terminated without storing the header or image and text files in the relational database and UNIX file system.

The *Update* push-button is used to indicate that only the header information is being submitted. If the *Update* push-button is selected, the client formats and sends a header update RPC message to the server, including the transaction code and the updated header information. The server stores the header information in the relational database and terminates the transaction without storing text or image files. The server sends a reply to the client indicating completion of the transaction and the client displays an appropriate advisory or error message.

The *Import* push-button is used to indicate that the header, text, and image information is being submitted. If the *Import* push-button is selected, the client displays another input screen that permits the custodian to specify the types and locations of files being submitted. When the *OK* push-button is selected from this screen, the client formats a header and text submission RPC message and sends it to the server, including the header and information about the types of files stored in the *submit* directory on the client platform. The server uses file transfer protocol (FTP) to copy the files from the client *submit* directory into the appropriate client-named *incoming* temporary storage directory on the server. A separate *incoming* directory is provided on the server for each client platform to temporarily store submitted files. Providing a separate directory to store the files submitted by each client prevents file name conflicts and possible overwrites of input files in the case of multiple concurrent document submissions from different clients.

The record submission transaction must include the following:

- A Flag—indicating the type of transaction and the type of files stored in the client's *submit* directory
- A Document Header—including all of the required header fields, arranged in the RPC message as pairs of field identifiers and data

The record submission transaction may also include the following:

- A text file—if the document submission includes text, one or more files must be stored in the *submit* directory of the client platform that contains all of the document text, with an appropriate file name extension to indicate the format of the text file. An ASCII text file with a *txt* extension is required for all submissions that include text. A word processing file may

also be submitted in addition to the ASCII text file. Word processing files may have extensions other than *txt*, *gif*, *pcx*, or *tif*.

- Full-page images—a separate image file for each page of text may be stored in the *submit* directory on the client platform, with an appropriate extension to indicate the format of the image (e.g., *gif*, *tif*, or *pcx*).
- Images of figures, equations, etc.—a separate image file for each graphical element may be stored in the *submit* directory on the client platform with an appropriate extension to indicate the format of the image (e.g., *gif*, *tif*, or *pcx*).

The client stores the text and image files for the document in the *submit* directory on the client platform. The server uses FTP to copy these files from the client *submit* directory into an *incoming* temporary storage directory named with the client name. Figure 3-3 illustrates the process by which the server moves and renames files submitted by the client.

Once the files have been stored temporarily in the client-named *incoming* subdirectory, the server generates a unique internal document identifier and uses that document identifier to change the names of the input files in the following ways as they are copied from the client-named *incoming* directory to the server partition's document set *incoming* directory:

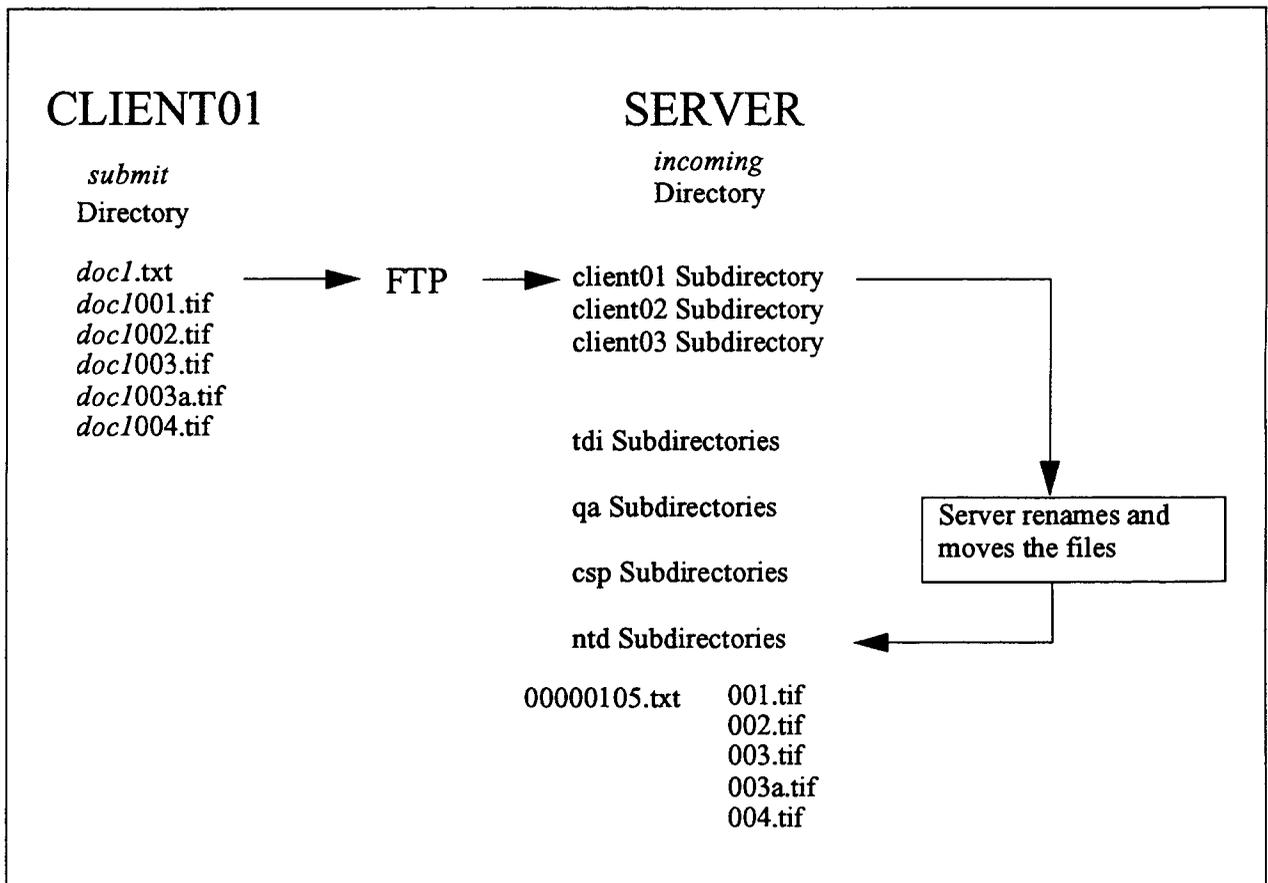


Figure 3-3. Movement and renaming of files by the server

- Text file—the name of the text file is changed to the internal document identifier as the file is copied to document set *incoming* subdirectory. The text file extension is also changed to indicate the format of the text (e.g., *data.txt* would be changed to *docname.txt*; *data.wpf* would be changed to *docname.bin*).
- Image files—the names of image files are changed using the page number (e.g., *data001.tif* would be changed to *001.tif*; *data001a.tif* would be changed to *001a.tif*). Image file extensions are not changed because they already indicate the format of the image file, as required by many image viewers.

The server validates the input, checking for the presence of the expected file types and examining the document header. If errors are found, the server sends a reply including status and error messages to the client. If no errors are detected, the header information is used to format a document header record in the relational database. This record in the relational database is marked to cause the document to be inserted into the full-text repository during the next execution of the batch process. Following processing of a successful transaction, the server sends a reply including status codes to the client. The client erases all files from its *submit* directory and notifies the user that the transaction was completed successfully.

3.5.2 Submitting Documents for Electronic Loading

Many of the documents to be loaded into the TDOCS are available in electronic form. This is particularly true for documents generated internally by the DWM and the CNWRA, which are normally available as full-text with embedded or accompanying image files of figures, equations, images, etc. Wherever possible, electronic copies of documents should be obtained to avoid the labor and system overhead associated with document scanning, OCR, and cleanup operations. Electronic copies of documents must be converted to ASCII format prior to loading. An ASCII copy of the text is required for submission, but the original word processing format may also be submitted. Once the electronic copy of a document has been obtained, reformatted, as required, and stored in the *submit* directory on the client platform, the document may be added to the TDOCS repository using the document submission facilities described in Section 4.5.1.

3.5.3 Deleting Documents

When a request to delete a record is initiated by the client, the document set and external document identifier are entered by the custodian through an entry screen. The client code formats an RPC record validation message and sends it to the server. The record validation message contains a flag indicating the transaction type, the document set, and the external document number. The server checks the validation message and retrieves the header record for the document from the relational database. If an error condition is found, the server sends a reply including status codes to the client. Several conditions result in an error being returned to the client:

- No record is found in the relational database for the specified document identifier
- The document record has already been deleted
- The document record has already been flagged for batch processing (e.g., record addition or record change)

If the retrieval of the record to be deleted is successful, an RPC reply, including the document title, is formatted and sent to the client so that confirmation of the deletion request may be obtained from the user. The client presents the confirmation information to the custodian user and waits for confirmation. If the custodian confirms the deletion validation, a delete request RPC message is formatted and sent to the server, including the internal document number and document set.

When the server receives the deletion request, the selected record is flagged for deletion and updated in the relational database. During the next execution of the batch process, the record is removed from the full-text repository, the text and image files are deleted, and the document header record is marked as deleted in the relational database. For a discussion of client processing for deleting documents, see Section 4.5.2.

3.5.4 Updating Documents

When a request to update a document is initiated by the client, a screen is presented to the custodian, and the document set and external document number are entered. The client formats and sends an RPC message to the server, indicating the transaction type, the document set, and the external document number. The server retrieves the document header from the relational database. If the record is not found or if it has been marked as deleted, the server sends a reply including status and error codes to the client, and an error message is displayed. If the record is retrieved successfully, the server formats an RPC reply, including the document header information, and sends it to the client for update. For a discussion of client processing for updating document records, see Section 4.5.3.

The client displays the header information to the custodian in an entry screen and accepts updates. The custodian updates the header information and signals completion of the update by: (i) selecting the *Update* push-button, (ii) selecting the *Import* push-button, (iii) selecting the *Clear* push-button or (iv) selecting the *Close* push-button. If either the *Clear* or *Close* push-button is selected, the transaction is terminated without updating the relational database or replacing the image and text files.

The *Update* push-button is used to indicate that only the header information is being changed. If the *Update* push-button is selected, the client formats and sends a header update RPC message to the server, including the transaction code and the updated header information. The server updates the header information in the relational database and terminates the transaction without altering the text or image files. An RPC reply, including status codes, is sent to the client, and an appropriate advisory or error message is displayed.

The *Import* push-button is used to indicate that the header, text, and image information is being changed. If the *Import* push-button is selected, the client displays another input screen that permits the custodian to specify the types and locations of files being submitted. When the *OK* push-button is selected from this screen, the client formats a header and text update RPC message and sends it to the server, including the updated header and information about the types of files stored in the *submit* directory on the client platform. The server uses FTP to copy the files from the client *submit* directory into the appropriate client-named *incoming* temporary storage directory on the server. A separate *incoming* directory is provided on the server for each client platform to temporarily store submitted files. The header and text update transaction includes the same information required for record submission transactions (see Sections 4.5.1.1 and 4.5.1.2).

The server validates the input, checking for the presence of the expected file types and examining the document header. If errors are found, a reply including error and status codes is sent to the client. If no errors are detected, the header information is used to update the document header record in the relational database. The files that have been stored temporarily in the client-named *incoming* subdirectory are moved to the appropriate document set *incoming* directory and are renamed using the internal document identifier. The document record in the relational database is marked to cause the document text and image files to be updated in the full-text repository during the next execution of the batch process. Following processing of a successful document update transaction, the server sends a reply including status codes to the client, and the client erases all files from the *submit* directory and notifies the user that the transaction was completed successfully. If the transaction was not successful, the server sends a reply including the error codes to the client and the client displays error messages to the custodian.

3.6 DOCUMENT SEARCH AND RETRIEVAL SERVICES

While the full-text indexes reside on the server platform with the full-text repository, all full-text search and retrieval processing is actually performed by the client. The full-text repositories and indexes are mounted through NFS so that they are directly accessible from the client platform. When the client full-text search and retrieval services are selected, the user interface is initiated and displayed. As the first query is executed by the client, the full-text indexes are read from the full-text repository on the server and paged into virtual memory on the client platform. Depending on the client platform being utilized and the size of the indexes being read, this can be a relatively slow process. However, once the indexes are in the virtual memory of the client platform, subsequent queries are executed very rapidly.

The queries performed by the full-text search and retrieval client produce result lists, from which the user may select documents for display. The display and browsing facilities for selected documents are also functions performed by the Document Search and Retrieval Clients. For further information on the functions of the Document Search and Retrieval Clients, see Section 5.2.

3.7 DOCUMENT DOWNLOAD

Document download capabilities are provided to permit users to access documents and copy the associated text and image files to a local hard disk or diskette. Document download facilities are accessed from the full-text search and retrieval document display facilities of the Document Search and Retrieval Clients. The desired document is located using the full-text search and retrieval facilities. When the document is displayed, the user may select *Download* from the *Launch* pull-down menu. The client displays an input screen to permit the user to specify the disk drive and directory where the text and image files should be stored. The client formats an RPC message, including the document identifier and the full path of the target directory, and sends it to the server. The text and image files for the desired document are retrieved from the *archive* and *images* directories in the TDOCS repository. The server uses FTP to transfer the text and image files to the specified directory on the client platform and sends a reply including status information to the client. For a further discussion of the client processing for document downloading, see Section 5.2.7.

3.8 BATCH LOADING, INDEXING, HYPERLINKING, AND SECURITY

The maintenance of the full-text repository is a time-consuming and potentially disruptive activity. Therefore maintenance of the repository is performed overnight when the full-text search and retrieval facilities are not being used. As the full-text repository is updated, a new set of text indexes is

created. Users must access the newly created full-text indexes in order to search, retrieve, and view the newly added and changed document records. This involves logging off the full-text system and logging on again. Since the batch process runs overnight, the newly added and updated records are automatically available to users when they logon the next morning. A log of batch activities is prepared by the server and e-mailed to the system administrator.

3.8.1 Batch Processing

During the day, custodians may enter requests for submitting, deleting, and updating records as described in Section 4.5. As those requests are processed by the server, the affected records are flagged in the relational database, and input files are copied to the appropriate document set subdirectories under the server's *incoming* directory. The batch process is normally started automatically as a scheduled process during the evening hours.

The batch process is driven by flagged records in the relational database. Therefore, the batch process examines each flagged record and performs the indicated functions in the following sequence:

- Record Updates—the replacement text and image files are moved to the appropriate subdirectories in *archive*, *images*, *load*, and *link* directory structures on the server.
- Record Submissions—the new text and image files are moved to the appropriate subdirectories in *archive*, *images*, *load*, and *link* directory structures on the server.
- Record Deletions—the text and image files are removed from the appropriate subdirectories in *archive*, *images*, *load*, and *link* directory structures on the server.

When the manipulation of files associated with added, deleted, and updated records has been completed, the batch process initiates the utility functions to rebuild the full-text indexes and hyperlinks. If batch processing is successful, the input files in the document set subdirectories under the *incoming* directory are deleted.

3.8.2 Full-Text Indexing

Users of the TDOCS rely heavily on full-text searches to find desired documents, and full-text indexing is an integral part of document loading and processing. Full-text indexing involves identifying individual words and updating indexes so that the full-text search and retrieval software can locate, retrieve, and display requested documents. Documents submitted for routine loading may be entered through scanning, OCR, and cleanup or may be submitted as electronic copies. Documents submitted for routine loading by either method are processed for full-text indexing. However, no full-text indexing occurs for documents submitted for on-demand loading because these materials are simply scanned, converted to text, and returned to the requestor as machine-readable files.

Since full-text indexing occurs at the word level, misspelled words are also indexed. This increases the importance of document cleanup following the OCR process, because OCR errors are propagated to the full-text indexes and can seriously compromise the effectiveness of the full-text search and retrieval facilities. The indexing of documents is accomplished automatically as part of the nightly batch loading process. Header information entered during the document preparation and processing steps is reformatted and loaded with the text so that documents may be retrieved through structured header

searches. Headers for documents that consist solely of images are also processed, formatted, and loaded into the full-text repository for search and retrieval through the full-text system.

3.8.2.1 Preparation of Documents for Full-text Indexing

During batch processing, the document text must be prepared for full-text indexing. The header information for the document is retrieved from the relational database, formatted with appropriate key words to permit the full-text indexing parser to identify specific fields, and prefixed to the text.

3.8.3 Generation of Hypergraphic Links

A policy decision has been made to preserve the full-page images created by the scanning process through routine loading. Full-page images are associated with the corresponding text through hypergraphic links. During full-text indexing, an icon is inserted at the top of each page that serves as a *launch pad* for the full-page image. Creation of these hypergraphic links is performed automatically during the batch process.

3.8.3.1 Preparation of Documents for Hypergraphic Generation

As part of the document preparation phase of the batch processing, patterns must be inserted in the text to identify the locations of hypergraphic links and associate them with the appropriate image files. A full-page image is normally associated with each page of text. The batch process can identify these full-page images by the patterns of their file names in the *images* subdirectory (e.g., *001.tif*, *002.tif*, etc.). The batch process prepares hypergraphic links to the files containing the full-page images by inserting hypergraphic link patterns, including the appropriate file names, after each page break. This results in a hypergraphic link being created for each full-page image at the top of the corresponding page of text.

If a binary formatted file (e.g., a WordPerfect representation of the text) exists, a hypertext link is generated and inserted at the top of the first page. This causes an icon and message to appear at the top of the first page for launching the WordPerfect software to display, edit, or print the text.

When additional image files are submitted for figures, the batch process can identify them by the patterns of their file names in the *images* subdirectory (e.g., *001a.tif*, *002a.tif*, *002b.tif*, etc.). The user may explicitly designate the locations in the text for the hypergraphic links to these figure images by inserting [CTL]O characters in the text. If the locations of the hypergraphic links to figure images are not explicitly designated, the links will be created at the top of the appropriate pages. The batch process prepares hypergraphic links for these figure images by inserting textual labels for the icons and hypergraphic link patterns, including the appropriate file names, at the designated locations indicated by the [CTL]O characters or at the top of the appropriate pages following the page break. This results in a hypergraphic link being created for each figure image at the designated location or at top of the appropriate page of text.

3.8.3.2 Document Indexing and Hypergraphic Link Generation

The prepared text files, including the formatted header and hypergraphic links, are stored in the *link* directory. When the full-text indexing utilities run, the files in the *link* directory are processed, and output is stored in the *load* directory. The files in the *load* directory are used to support document retrieval and display.

3.9 DATABASE ADMINISTRATION AND MAINTENANCE

Certain functionality related to the maintenance of user accounts is restricted to DBAs. This functionality is accessed through the *System* pull-down menu of the Document Processing Clients (see Section 4.6).

3.9.1 User-ID Maintenance Functions

User-ID maintenance capabilities are accessed by selecting the *User-IDs* entry from the *System* pull-down menu of the Document Management Client. The client displays an entry screen that supports adding new users, deleting users, and changing passwords and privileges.

To enter a new user, the DBA enters the user name, User-ID, password, and privilege level in the appropriate fields and selects the *Add* push-button. The client formats an RPC message and sends it to the server. The server uses the information in the RPC message to create a new user's record in the user privileges table, and a new record in the user names table. The password is stored in the RDBMS security facilities. If the transaction is successful, the server formats and sends an RPC reply to the client indicating successful completion. If errors are detected, the server formats and sends an RPC reply to the client, including the error codes, and the client displays an error message to the DBA.

To update an existing user record, the DBA: (i) selects the desired User-ID from a scrollable list; (ii) updates the user name, User-ID, password, and privilege level in the appropriate fields; and (iii) selects the *Update* push-button. The client formats an RPC message and sends it to the server. The server uses the information in the RPC message to update the user's record in the user privileges table, and the user's record in the user names table. The password is updated in the RDBMS security facilities. If the transaction is successful, the server formats and sends an RPC reply to the client indicating successful completion. If errors are detected, the server formats and sends an RPC reply to the client including the error codes and the client displays an error message to the DBA.

To delete an existing user record, the DBA selects the desired User-ID from a scrollable list and selects the *Delete* push-button. The client formats an RPC message and sends it to the server. The server uses the information in the RPC message to delete the user's record in the user privileges table, and the user's record in the user names table. The password user is deleted from the RDBMS security facilities. If the transaction is successful, the server formats and sends an RPC reply to the client indicating successful completion. If errors are detected, the server formats and sends an RPC reply to the client, including the error codes, and the client displays an error message to the DBA.

3.9.2 Initiating the Batch Process

Batch processing is normally started as a timed process during the evening hours. When the batch process is completed, a log is sent to the system administrator by e-mail.

4 DOCUMENT PROCESSING CLIENTS DESIGN

The Document Processing Clients provide the user interface and the client portion of the application processing for all TDOCS document management activities. Functionality supported by the Document Management Clients includes the following:

- GUI support
- Password protected access for custodian and DBA users
- Scanning, OCR, and cleanup
- Document submission, deletion, and maintenance

4.1 DOCUMENT PROCESSING CLIENTS HARDWARE

As illustrated in Figure 4-1, there is a requirement that the TDOCS must support multiple user hardware/software platforms. This requirement is driven by the necessity to use existing equipment and facilities. Since multiple platforms are in use by individual members of the DWM and CNWRA staffs at the present time, the TDOCS must accommodate those platforms and must also make provision for handling additional platforms in the future.

Utilizing a distributed processing architecture and standard communications protocols, the initial version of the TDOCS was designed and implemented to support multiple client platforms. Supported hardware/software platforms included PCs running Microsoft Windows and Sun Workstations running OpenLook or Motif at the DWM; and PCs running Microsoft Windows and IBM's OS/2, Macintosh computers running System 7, and Sun Workstation running OpenLook or Motif at the CNWRA.

4.2 DOCUMENT PROCESSING CLIENTS SOFTWARE

Supporting four distinct client platforms in a client/server environment requires major design decisions about how to handle the multiple platform support. Two alternatives are available to the developer: (i) prepare separate implementations for each platform based on a common design, or (ii) select one platform as the primary implementation vehicle and then adapt and transport the code to other platforms. The latter approach was selected for initial TDOCS and RPD system implementations at the CNWRA.

4.2.1 Multiplatform Software Packages

Multiplatform considerations also apply to the selection of the full-text search and retrieval product. The full-text facilities are resident on the central server. However, the bulk of the interface and requestor code for the full-text search and retrieval facilities executes on the client platforms. Therefore, support for requestor code running on the four required hardware/software platforms was a major consideration in the selection of the full-text search software product for initial implementation of the TDOCS and the RPD system. Subsequent investigations have shown that, since the initial implementation of the TDOCS and RPD applications, there has been little change in the availability of full-text search and retrieval programs that support the required platforms. Therefore, the rationale for selection of TOPIC for

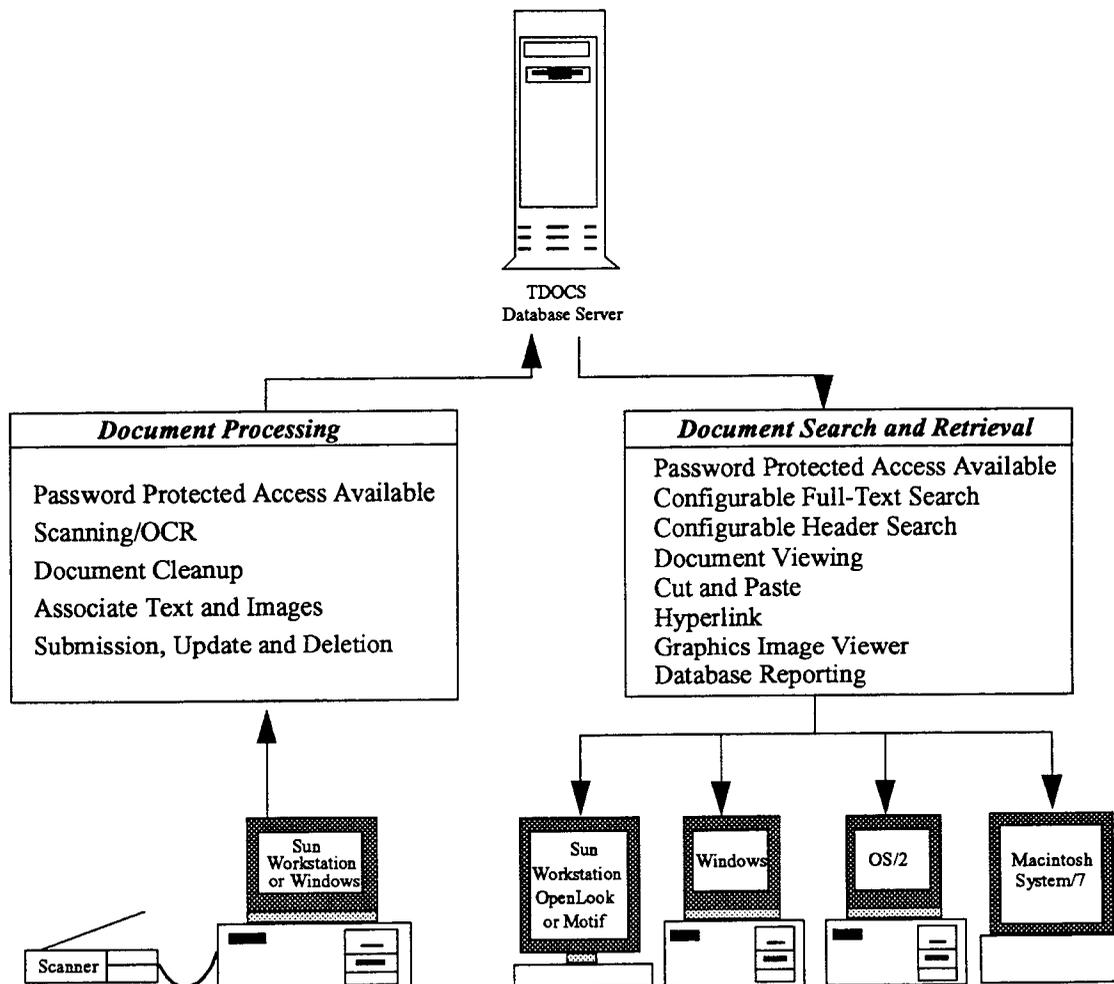


Figure 4-1. User platforms and functionality

the full-text search and retrieval function was considered to be valid for the initial implementation of the TDOCS as well.

4.2.2 Open-System Environment

The constraint to support client/server architecture involves the need to support cross-platform communication and data transfer capabilities. Client facilities are implemented and customized for the individual users' hardware/software platforms and environments. The UNIX operating system environment is particularly well adapted to support such an open-system environment because it provides the greatest compatibility, portability, and support for future enhancement of the system. Therefore, UNIX was selected for the database server platform operating system environment. However, the operating system environments for the client platforms were determined by the existing hardware and software of the individual users.

4.2.3 Hardware/Software Independent Communications Between Client and Server Platforms

The use of standard vendor-independent communication and file sharing protocols such as TCP/IP, RPC, and NFS are required because of the need for an implementation with a high degree of hardware/software independence. This need results from the requirement to support multiple platforms in a client/server architecture. Experience with the development of the TDOCS and RPD systems at the CNWRA demonstrated the feasibility and effectiveness of this approach.

Communication between the client and server facilities in a cross-platform open-system environment requires that hardware/software independent communications be accommodated. This is accomplished through the RPC message and data transfer protocol. When the client facility requires services, such as retrieval of a record from the server facility, an RPC message is formatted and transmitted to the server. When the server finishes processing the request, a reply including status information and data records is sent to the client. Document downloading and editing are accomplished through a combination of RPC and NFS. These communication and file sharing protocols were utilized in both the TDOCS and RPD initial implementations, and they are also utilized in the existing AUTOS and ACRS systems at the NRC.

4.2.4 Graphical User Interface Development

The TDOCS is implemented using a GUI tool that supports the required user interface for all DWM and CNWRA PC and workstation platforms, including Microsoft Windows, IBM OS/2, SUN OpenLook, MOTIF, and Macintosh System 7. The GUI supports menus, buttons, selection lists, multiple windows, clipboards, and dialogue boxes for user interaction and feedback. The relatively straightforward concept of a GUI requires a certain level of sophistication in its implementation to achieve compatibility between platform-specific implementations. The *look and feel* of the GUI for each type of workstation platform is implemented in a consistent and intuitive manner, varying only in details that are specific to the particular implementation environment. Effective use of this approach requires that a highly flexible and capable software development tool be available that permits transporting the user interface code to the different platforms.

The choices of GUI development tools that support all four of the required hardware/software platforms are quite limited. The initial evaluation of such products during the design of the initial TDOCS and RPD applications resulted in the selection of the Galaxy GUI development tool from Visix because of its overall unique functionality and ability to support the required platforms (DeWispelare et al., 1993). Subsequent reviews of the capabilities of GUI development tools have confirmed the initial decision to utilize the Visix Galaxy product for this functionality in the initial implementation of the TDOCS.

The Galaxy development tool employs an unusual concept in addressing multi-platform application requirements. Rather than developing the code for each platform, the application is written to use a virtual platform with generic capabilities. The transformation from the *virtual* platform to the *look and feel* of the target environment is accomplished by the Galaxy software at execution time. Therefore, selection of each different *look and feel* (e.g., Motif versus OpenLook) is accomplished by a parameter setting when the code is executed.

Development of multiplatform client code involves additional factors beyond the actual user interface that involve specific hardware/software dependencies: (i) the handling of application launches

must conform to the operating system requirements of the client platform; (ii) interprocess communication using RPC requires cognizance of the hardware architecture of the client environment; and (iii) compiler incompatibilities between diverse platforms introduce technical challenges. However, the advantage of using a cross-platform GUI tool is that the user interface portion of the application code can be designed and implemented in a generic manner without regard for platform specific considerations.

4.3 FUNCTIONAL MANAGEMENT AND CAPABILITIES OF THE DOCUMENT PROCESSING CLIENTS

Available functionality in the TDOCS system is adapted to the requirements of the individual user and is controlled by user privileges that define which portions of the client functionality are accessible. TDOCS users are assigned permissions by their user classes. User classes for custodians and DBAs are assigned through the user account and password maintenance functions available only to the DBA.

There are three broad classes of TDOCS users:

- **Users**—search for documents, access and download documents and images, print reports, and utilize document manipulation facilities
- **Custodians**—perform all of the functions available to users and are also able to perform database custodian functions to add, delete, or update records
- **Database Administrators**—perform all of the functions available to users and custodians, and are also able to perform database administration functions, including adding, changing, and deleting user accounts, preferences, and permissions

4.3.1 Password Protected Access

Password protected access is available for implementation on all platforms, if desired or required due to security requirements. It is mandatory for custodian and DBA clients. When the TDOCS system is started from a password protected client platform, the first screen that appears is the *DWM Logon Screen* (Figure 4-2). This screen permits entry of the *User-ID* and *Password*. The entered *User-ID* and *Password* are formatted into an RPC message and sent to the server. The server attempts a logon to the relational database to validate the *User-ID* and *Password*. If the logon is not successful, the server sends a reply including error status codes to the client and an advisory message is displayed. If the logon to the relational database is successful, the server uses the *User-ID* as a search argument to retrieve the appropriate record from the *User Privileges* table. This table contains a user class code which is sent to the client in an RPC reply. For a further discussion of the server logon processing, see Section 3.4.1.

When a TDOCS user with a password performs a logon, the server code retrieves a user class code and sends it to the client in an RPC reply. This user class code is interpreted by the client to set permissions and condition the display of menus so that only the functionality appropriate to that user is available. In the following sections, custodian functionality will be illustrated unless otherwise noted.

Users on clients without password protection access TDOCS in a read-only mode. The initial menu is accessed on PCs by activating the TDOCS program icon, or on Sun workstations by selecting the *TDOCS* entry from a menu on the desktop.

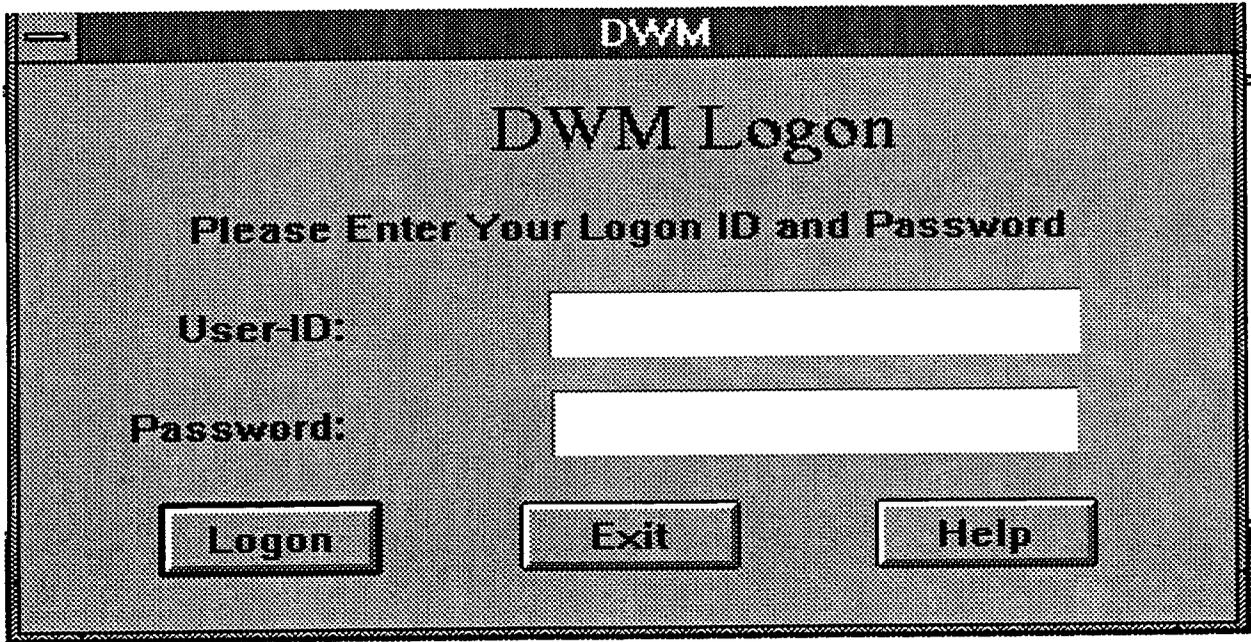


Figure 4-2. Division of Waste Management logon screen

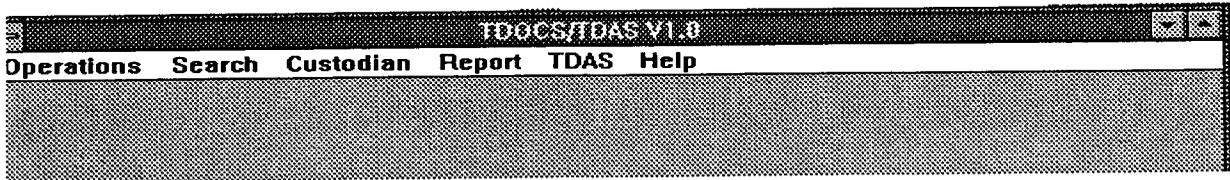


Figure 4-3. TDOCS custodian main menu

4.3.1.1 Document Processing Functionality Available to Custodians

TDOCS custodian may access all of the functionality available to the TDOCS user class. Additionally, they may perform functions that permit them to add, delete, or update documents in the TDOCS database. The document processing client capabilities of the TDOCS are resident on the custodians PC or workstation to facilitate interaction and respond to user input.

Following a successful logon sequence, the custodian *Main Menu* is displayed (Figure 4-3). This menu contains entries of all of the functionality available to custodian, including *Operations*, *Search*, *Custodian*, *Report*, *TDAS*, and *Help*. The *System* entry for DBA functionality is not displayed in the *Main Menu* for custodians.

Selecting the *Operations* entry from the *TDOCS/TDAS Custodian Main Menu* causes the *Operations* pull-down menu to appear (Figure 4-4).

Selecting the *Change Password* entry from the *Operations* pull-down menu causes the DWM *Change Password* screen to appear (Figure 4-5). When the *Change Password* option is selected, the client accepts duplicate entry of the password in two entry fields. The contents of the two fields are validated and

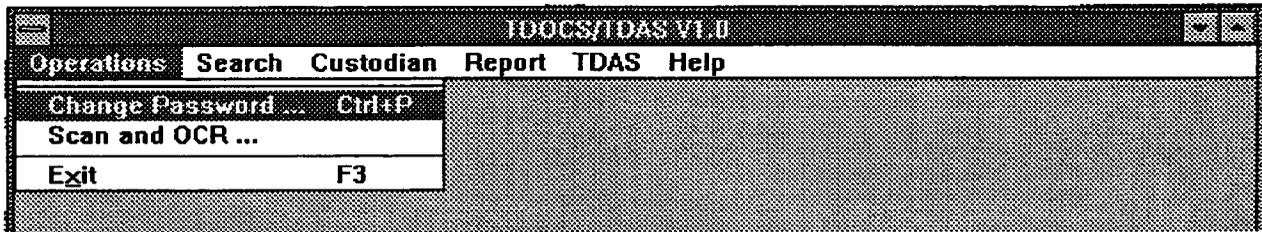


Figure 4-4. TDOCS operations pull-down menu

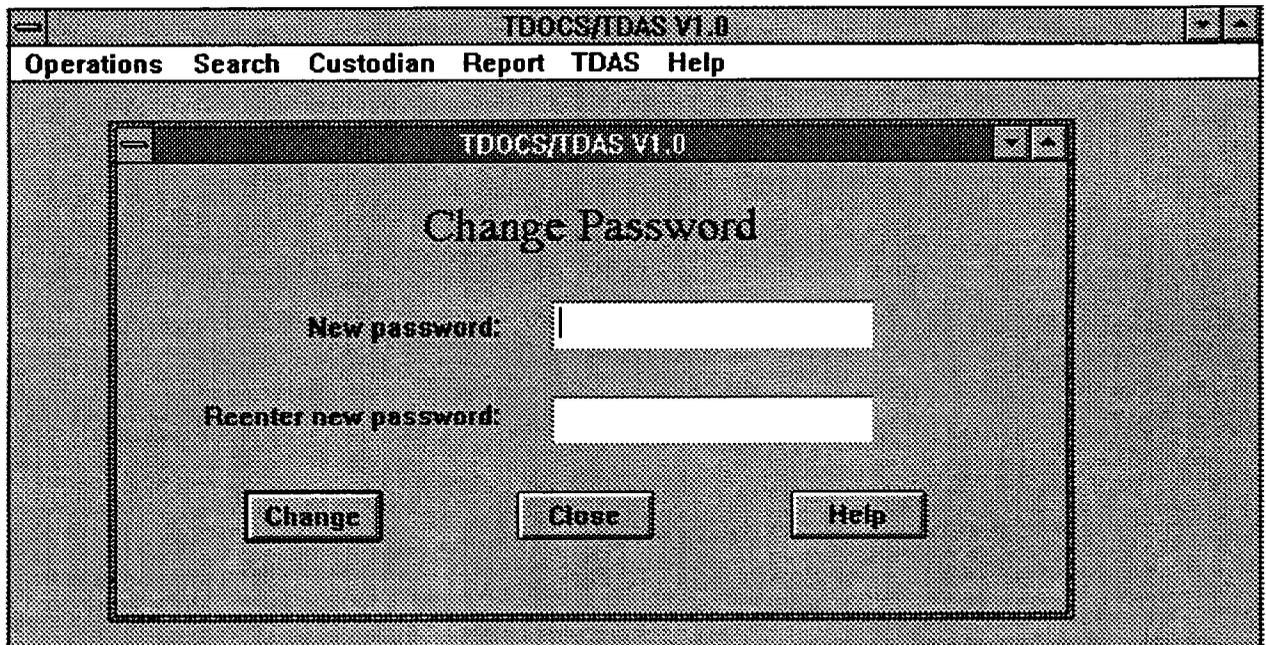


Figure 4-5. TDOCS change password screen

compared. If they are not valid or are not identical, an error message is displayed. When valid duplicate password entries have been entered, the client formats an RPC message and sends the User-ID and password to the server. The server changes the user's password in the relational database security facilities. For a further discussion of the server processing for changing a user password, see Section 3.4.2. When the server has completed the change password transaction, it sends a reply with status to the client and an advisory message is displayed to the user.

4.4 SCANNING, OPTICAL CHARACTER RECOGNITION, AND CLEANUP

Documents that are not available in electronic format are obtained as hard-copy documents and processed by scanning and OCR to obtain an electronic copy. The process of scanning materials and converting them to full text through OCR is required to prepare paper-based documents for loading into the system. This is a semiautomated process that is subject to errors that may arise from the condition of the documents, the quality of the printing, etc. Scanning and OCR errors may or may not be detected automatically. Therefore, a manual document cleanup procedure is required as part of the document capture process to ensure document accuracy. Following the completion of the scanning, OCR, and

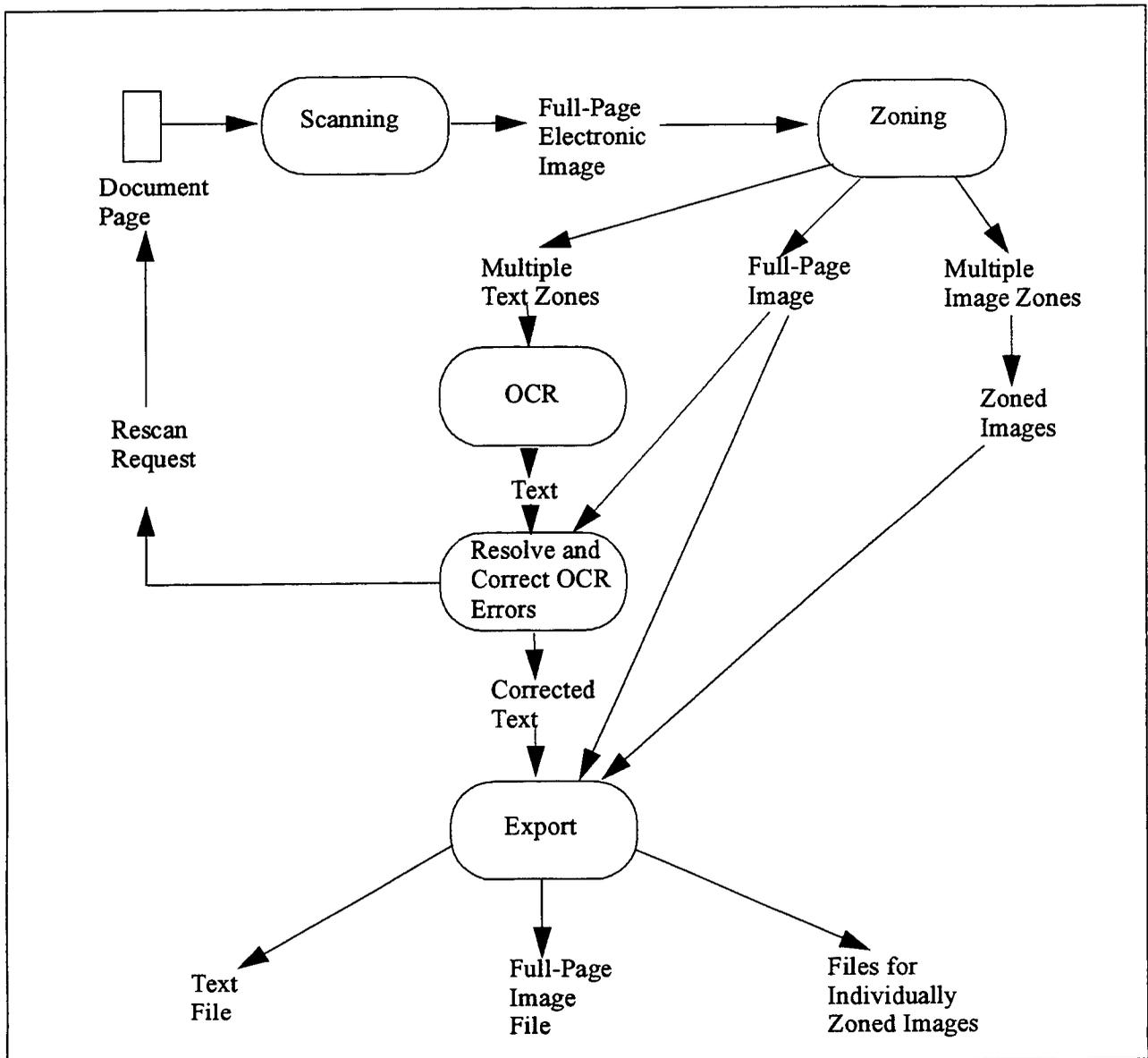


Figure 4-6. The scanning process

cleanup processes, an electronic copy of the document is available for loading into the system. Figure 4-6 illustrates the data flow and relationships between the scanning, OCR, and document cleanup functions. The document is examined visually to determine if exception conditions must be accommodated.

- A user-supplied document name is entered and options for text and full-page images are selected.
- The pages are placed in the document feeder or on the scanner bed and scanned.
- Pages are zoned as required to designate areas of text versus areas that contain images.
- Pages of text are accessed through a special editor to correct scanning/OCR errors.

- The corrected text is exported to the file system along with any associated images as a collection of related files for loading into the full-text and image repositories or for individual use.

4.4.1 Scanning

Documents submitted to the TDOCS in hard-copy form must be scanned to obtain an electronic copy. Scanning is performed on both a routine and on-demand basis. For routine loading, such scanning is performed by the TDOCS support staff at a central facility where bibliographic headers are entered and documents are prepared for scanning. To satisfy the requirements for on-demand loading, a scanning facility is provided that is accessible to the technical staff and their secretarial support. Technical staff may perform scanning for on-demand loading in much the way that they would utilize a copy machine, scanning one or more pages, graphical images, etc., and obtaining the machine readable copy of the scanned material on a diskette or a network hard disk. The requirements and intended functionality for routine versus on-demand scanning are quite different. On-demand scanning is intended to support low levels of activity for materials that are needed too urgently to wait for the routine scanning and loading processes. The quality of the scanning and the management of the machine-readable files is the responsibility of the technical staff performing the on-demand loading. Therefore, some additional training requirements are expected in conjunction with on-demand loading.

4.4.1.1 Potential Contention for Scanning Resources

The need to perform on-demand loading occurs concurrently with routine loading, and there is potential for scheduling conflicts. A request for on-demand loading could be significantly delayed if the scanning station were being used to capture a very large document. Furthermore, the physical requirements for a routine loading facility are different from those of a walk-up facility for on-demand loading. Therefore, depending on the volume of routine scanning and the level of contention for resources, multiple scanning stations may be required in the future.

4.4.1.2 Color Scanning Capabilities

The TDOCS will support color scanning as part of its capability (Johnson et al., 1993). Color scanning requires significantly more time than black and white scanning and the resulting images require much more storage capacity. Therefore, color scanning, is normally performed on an exception basis as part of on-demand loading, and the color images are returned to the requestor as files. Documents to be loaded into the TDOCS with color images must be scanned separately and loaded through the electronic loading facilities.

4.4.1.3 Factors Affecting Quality of the Scanned Image

A number of factors may adversely affect the quality of the scanning process. Therefore, it is important to utilize good quality original documents or first generation copies with appropriate contrast as input for scanning.

- Smudges, lines, marks, and marginal notes may be misinterpreted by the scanning hardware and software.
- Misformed characters and characters that touch each other cause OCR difficulties.

- Input documents must have adequate contrast. Inadequate contrast can result in unsatisfactory OCR.
- Documents must be properly oriented on the scanning device. If the documents are misaligned, rotated, or skewed, the scanning and OCR processes may produce unsatisfactory results.
- It is very important for documents being scanned to be flat. When pages from a book or magazine are scanned, they must be held tightly against the glass of the scanner bed to prevent distortion resulting from the curvature of the paper. Folds and creases in the input document can also introduce distortion.
- Documents printed in multiple columns may cause OCR difficulties if the columns are very close together. In such documents, the columns may not be recognized and space characters may be inserted between the left- and right-hand column text on each line. This results in interleaving of the columns and unusable text.
- Documents that contain both text and graphical images may require manual processing to define zones around figures, equations, etc., in order to distinguish these graphical elements from the text.

Since there are many factors that may adversely affect the quality of the scanned image, it is important that the TDOCS support staff be trained adequately to minimize factors and conditions that have an adverse impact on the image quality.

4.4.2 Optical Character Recognition

Once a document has been scanned, it must be processed for OCR to obtain an electronic copy of the text. This process, in which the scanned image is processed to recognize the textual characters, follows the scanning of the document and may be performed in a semiautomatic or assisted mode.

OCR is an imperfect process, and many factors can affect the quality of the scanning and, consequently, the effectiveness of OCR. In particular, poor document quality resulting from the use of second- or third-generation copies as input to the scanning process usually has a very adverse effect on OCR. When the individual characters of the text are sufficiently close to being indistinguishable or touch, the OCR process can fail to recognize the combined characters. Such errors are common in documents that are second- or third-generation copies of originals documents.

Unexpected orientations of the text and documents printed in closely spaced columns can cause the OCR process to produce unsatisfactory results. Documents with low contrast and documents that cannot be made completely flat for scanning also introduce distortion. In such documents, the individual characters may be indistinct or may touch or be distorted to a degree that makes fully automated OCR impossible.

4.4.3 Document Cleanup Following Scanning and Optical Character Recognition of Hard-Copy Documents

Scanning and OCR are imperfect processes that are affected by the condition of input documents (e.g., smudges, imprecise character definition, marginal notes, underlining, etc.). Therefore, a manual document cleanup procedure is required as part of the document capture process to ensure document accuracy. OCR is a time-consuming process that is often performed serially with scanning. A document is scanned and immediately subjected to OCR processing. This approach is normally taken so that the original hard-copy document will be readily available for reference if cleanup is required.

4.4.4 Requesting Scan and OCR Functionality

Scanning, OCR, and cleanup functionality is started by selecting the *Scan and OCR* option from the *Operations* pull-down menu (see Figure 4-4). Selecting this option initiates the scanning and OCR software. Scanning facilities are accessed through the *Operations* pull-down menu rather than the *Custodian* pull-down menu so that these facilities will be available to all users for on-demand loading. The scanning facilities for on-demand loading and routine loading are identical, but the document submission facilities used in routine loading are available only to the custodian. For further information on the use of the scanning and OCR software, see the scanning appendix to the TDOCS User Guide (DeWispelare et al., 1995).

4.5 DOCUMENT SUBMISSION, DELETION, AND MAINTENANCE

The TDOCS Document Processing Client modules provide support for submission, deletion and maintenance of document text and images. This functionality is accessed through the *Custodian* entry in the *TDOCS/TDAS Main Menu*. When the *Custodian* entry is selected, the *Custodian* pull-down menu is displayed (Figure 4-7.) This menu includes options to *Submit TDOCS Record*, *Delete TDOCS Record*, and *Update TDOCS Record*

4.5.1 Submitting New Documents

The custodian requests a maintenance function to submit a TDOCS record by selecting the *Submit TDOCS Record* entry from the *Custodian* pull-down menu in *TDOCS/TDAS Main Menu*. The client responds by displaying an entry screen for the header information.

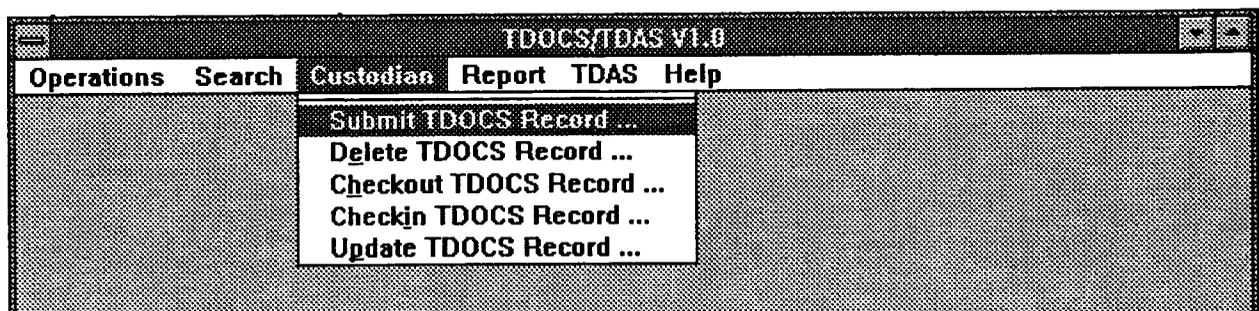


Figure 4-7. TDOCS custodian pull-down menu

4.5.1.1 Entry of the Header Record

A header record is needed for each document that is entered into the TDOCS relational database and full-text and image repositories. The header information must be associated with the text and image files produced by the scanning and OCR or electronic loading processes. This is done by entering header information through the document submission screen. The user interaction for this process is controlled by the Document Processing Client modules. When the header information has been entered, it is associated with the text and image files stored in the *submit* directory on the client platform by sending an RPC message to the server, including the header and an indication of the types of files.

The *Submit a Record* entry screen (Figure 4-8) includes both mandatory and optional fields. All entry fields are mandatory except those marked with an asterisk at the left of the entry field. Some fields are specific to a particular document set (e.g., *Addressee(s)* is appropriate only for correspondence records, and *Duration* applies to quality assurance records). Therefore, fields that are not applicable to the selected document set will be grayed.

The header record contains the following information:

- **Document Set**—a mandatory field that identifies the type of document. The content of this field is set by selecting an entry from the *Document Set* drop-down list. For NRC technical documents, the *NTD* entry should be selected.
- **Author(s)**—a mandatory field that contains the name of the author of the document. If there are multiple authors, the name of each author is entered. The names of multiple authors are separated by semicolons. Author names are entered in mixed case with the last name first, followed by a comma, followed by the initials. Each author name may be up to 50 characters in length.
- **Title**—a mandatory field, up to 500 characters in length, that contains the title or subject of the document. the title is entered in mixed case, exactly as it appears in the document.
- **Publication**—a mandatory field, up to 500 characters in length, that contains publication information, such as publisher location, publisher name, pages, journal name, etc.
- **Document Date**—a mandatory field that contains the document's creation or publication date. The format of the date is DD MON YYYY, where DD is a two-digit number representing the day of the month, MON is a three character alphabetic abbreviation of the month, and YYYY is a four-digit year.
- **Code(s)**—an optional field that contains the subject code or other external code assigned to the document. Multiple codes may be entered in this field, separated by semi-colons. Each code may be up to 30 characters in length.
- **Number(s)**—an optional field that contains the report number, project number, or other identifying number, such as, revision number or software vendor and version identifier associated with the document. Multiple numbers may be entered in this field, separated by semicolons. Each number may be up to 100 characters in length.

TDOCS

Submit a TDOCS Record

Document Set: Document Number:

Document Header:

Author(s)	Tschoepe, E.; Lyle, F.; Dancer, D. M.; Interrante, C. G.; Hair, P. K.	...
Addressee(s)		...
Title	Field Engineering Experience with Structural Materials	...
Publication	San Antonio, Texas; Southwest Research Institute	...
Document Date	01 Jun 1994	
Code(s)	462.2	
Number(s) *		
Duration		
Note *		...

*Optional Values

Figure 4-8. Header entry screen

- Note—an optional field, up to 500 characters in length, that is provided to accommodate any other relevant information about the document.

The *Clear* push-button may be selected at any point during the header entry process to discard a partial entry and clear the screen. The *Close* push-button may be selected to close the header entry screen and terminate the document submission process. When all of the header fields have been entered, the custodian selects the *Submit* or *Import* push-button at the bottom of the screen to accept the header entry. The client software validates the header fields. If any errors are detected, the header information is displayed again along with appropriate advisory messages. If no errors are detected in the header entry, the client formats an RPC message including the header information and sends it to the server to add the document record.

4.5.1.2 Text and Image Files

While it is possible to add a header-only document record (e.g., a record describing a magnetic tape), most documents have associated text and image files. These files are obtained in one of two ways:

- Hard-copy documents may be captured through scanning, OCR, and cleanup to obtain the text and image files.
- Electronic copies of the document may be obtained, including text and image files.

When the document submission transaction is processed, the server expects to find the text and image files in the *submit* directory on the client platform. The text and image files may be stored in the *submit* directory in several ways:

- The text and image files created by the scanning, OCR, and cleanup processes may be exported directly into the *submit* directory.
- The custodian may manually move the text and image files to the *submit* directory.
- The *Import* push-button at the bottom of the *Submit a Record* input screen may be selected to direct the system to retrieve text and image files from a specified directory and move them to the *submit* directory.

Once the header information has been entered and the text and image files have been stored in the *submit* directory, the document may be submitted for loading into the TDOCS database. The client formats an RPC message and sends it to the server to initiate document submission processing. The RPC message includes the document header, a flag indicating the type of transaction, and an indication of the types of files that have been stored in the *submit* directory on the client platform.

- Document header—a valid document header with all required fields, formatted as pairs of field identifiers and data in the RPC message
- Text file—a single file is required for textual documents containing all of the text of the document in ASCII format with a file name extension of *txt*
- Binary file—a file containing the text in binary format with a file name extension other than *txt*, *gif*, *tif*, or *pcx* to indicate a format such as WordPerfect
- Full-page images—including a separate image file for each page of text with an appropriate extension (e.g., *gif*, *tif*, or *pcx*)
- Images of figures, equations, etc.—including a separate image file for graphical element with an appropriate extension (e.g., *gif*, *tif*, or *pcx*)

TDOCS permits several alternative scanning and OCR software packages to be used. Each of these software packages employs its own unique conventions for naming files for submitted text and images. Therefore, the design of the initial TDOCS implementation utilizes its own internal file naming conventions, and renames output of the scanning and OCR software so that it will adhere to the TDOCS file naming conventions. The first four characters of the document name are used as the base file name for

both text and image files. Since there can be many image files associated with each document, TDOCS appends a suffix to the four-character document base name to form a unique file name for each image. If more than four characters are entered for the base document name, they are truncated. If fewer than four characters are entered, they are not padded, and the resulting file name is shortened.

The following structure is used by TDOCS for text file names:

- *dddd.txt* where:

- *dddd* is the user-specified document base name

The following structure is used for the output image file names:

- *dddnnna.ttt* where:

- *dddd* is the user-specified document base name

- *nnn* is a sequence number corresponding to the page sequence of the scanned document

- *a* is an alphabetic suffix sequentially assigned to distinguish additional images within a single page.

- *ttt* is a three-character file type for the image file (e.g., *tif*, *gif*, *pcx*, etc.)

Using this convention, the server code can properly infer the contents of each file stored in the *submit* directory of the client platform. For example, a document with the base name, *data*, might have files with the following names:

- *data.txt*—The entire text of the document in ASCII format
- *data001.tif*—The full-page image of the first page of the document in TIF format
- *data001a.tif*—The image file for the first non-textual feature on page 1 that was captured as an image in TIFF format

4.5.1.3 Document Submission

When the record submission request is received by the server, validation of user authorities and permissions are validated through the security features of the RDBMS. The server validates the input. If errors are found, a reply including status and error codes is sent to the client. If no errors are found, the server: (i) adds the bibliographic header information to the relational database, (ii) flags the record to cause the batch process to add the text and image files to the TDOCS repository, and (iii) copies the text and image files from the client *submit* directory to the server *incoming* directory. The files are renamed as they are copied to conform to the file naming conventions in the server directories. Following the processing of a successful document submission transaction, the server sends a reply, including status, to the client. The client erases all files from the *submit* directory and notifies the user that the transaction was completed successfully. For a description of server processing for submission of a new document, see Sections 3.5.1 and 3.5.2.

4.5.1.4 Document Import Option

If the user selects the *Import* push-button at the bottom of the *Submit a Record* screen, the client displays the *Import Document* screen (Figure 4-9). This screen has an upper *Scan* pane and lower *Electronic* pane. The *Scan* or *Electronic* pane is enabled by selecting the appropriate radio button.

If the *Scan* option is selected, the system initiates the scanning, OCR, and cleanup facilities when the *OK* push-button is selected. The *Zoned Images* option may be selected to indicate that both full-page and zoned images are to be stored in the *submit* directory. If the *Electronic* option is selected, the system uses the information provided to locate and copy text and image files to the *submit* directory from a designated location when the *OK* push-button is selected. The *Zoned Images* option may be selected to indicate that both full-page and zoned images are to be stored in the *submit* directory.

4.5.2 Deleting Documents

When the custodian requests a maintenance function to delete a record, the client code handles the user interaction as information uniquely identifying the record is entered. A *Delete Document* screen (Figure 4-10) is presented to the user to permit entry of the document set and document identifier. A request is formatted and transmitted to the server along with the identifying information.

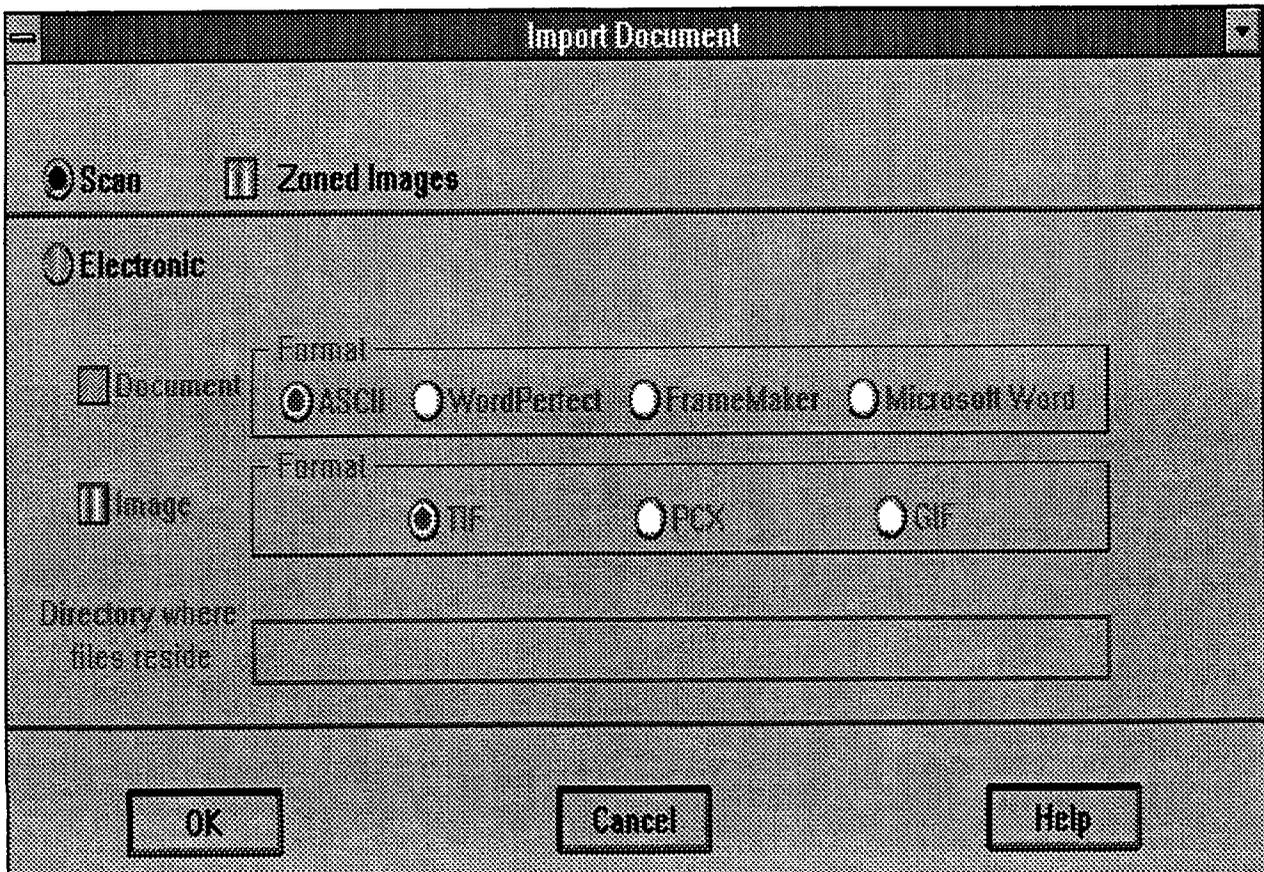


Figure 4-9. Document import entry screen

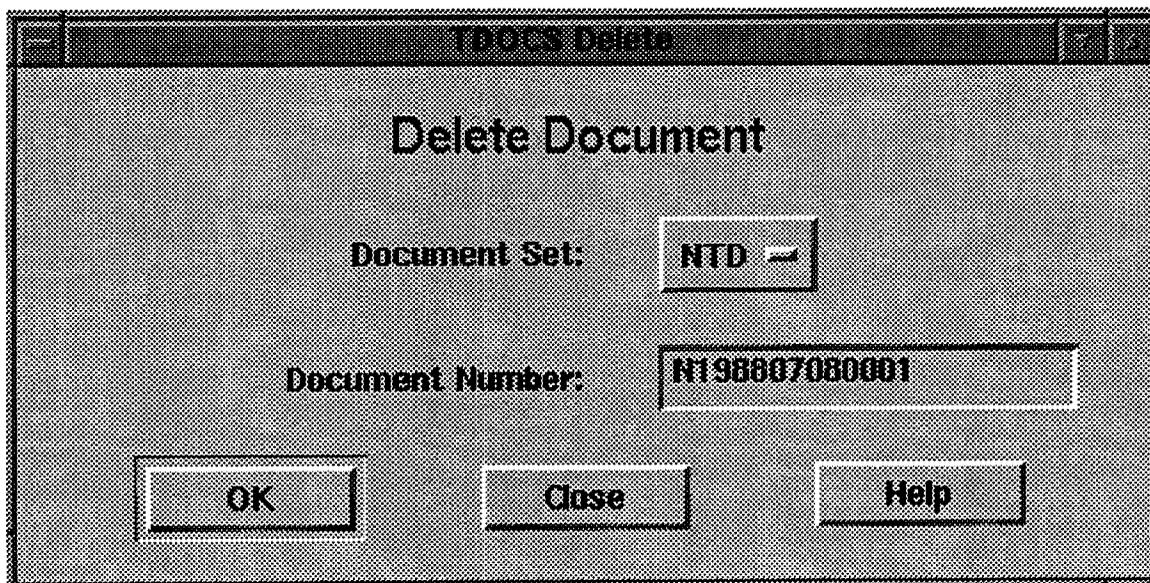


Figure 4-10. Delete document screen

The server validates the identifying information and checks for the presence of the record to be deleted. If errors are found or the record to be deleted is not present, a reply including status and error codes is sent to the client. For a further discussion of the server processing for deleting records, see Section 3.5.3).

If the record is found in the relational database, the server sends an RPC reply containing the document title. This information is used to present the *TDOCS Confirm* screen to the custodian (Figure 4-11). The custodian visually reviews the information on the *TDOCS Confirm* screen and either accepts or cancels the deletion transaction. If the *Cancel* option is selected, the client terminates the deletion request. If the custodian selects the *OK* push-button, the server flags the bibliographic header information in the relational database for deletion. Following processing of a successful document deletion transaction, the server sends a reply, including status codes, to the client, and an advisory message is displayed to notify the custodian that the deletion request has been processed. During the next execution of the batch process, records flagged for deletion in the relational database are examined. The text for these documents is removed from the full-text repository and the text and image files are removed from the UNIX file system.

4.5.3 Updating Documents

TDOCS document updates are always handled as a complete replacement of the existing document header record and, optionally, the document text and image files. The header record is updated in place in the relational database. If new text and image files are submitted, the existing files are deleted and the new files are added to the TDOCS database.

When the custodian requests a maintenance function to update a record, the client code handles the user interface interaction as the information uniquely identifying the record is entered. The *Update Document* screen (Figure 4-12) is displayed and the document set and the document identification number

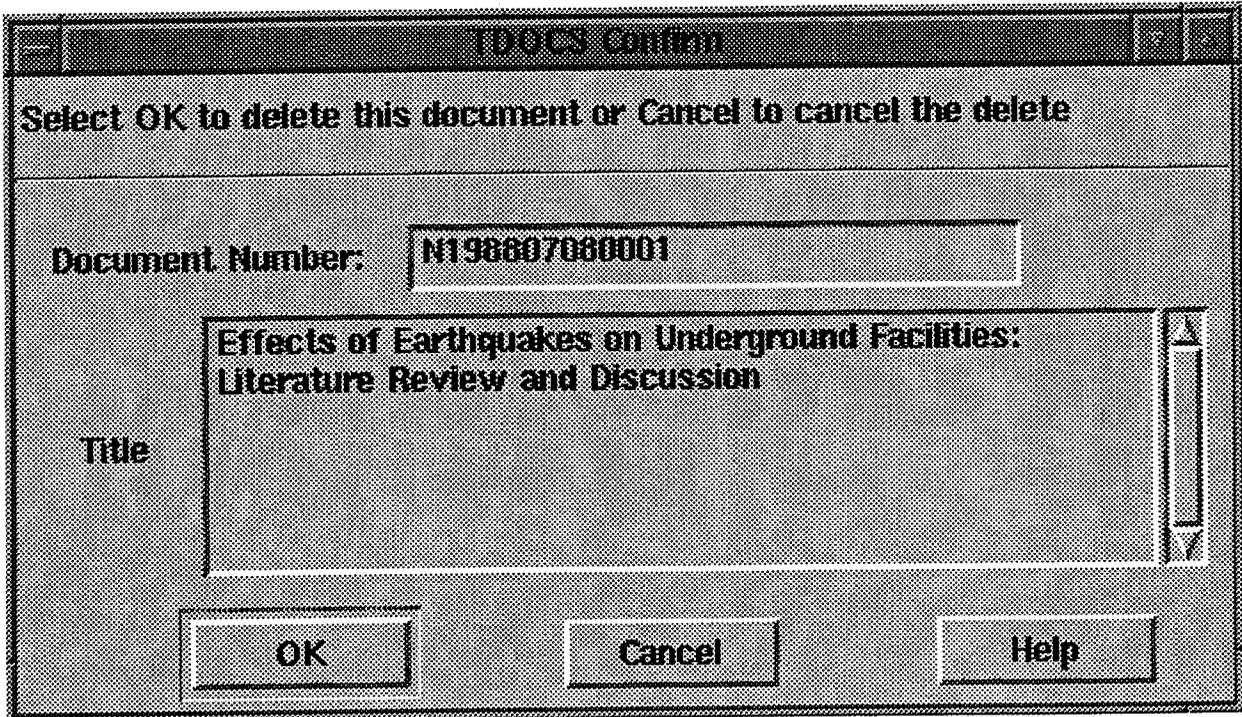


Figure 4-11. TDOCS confirm screen

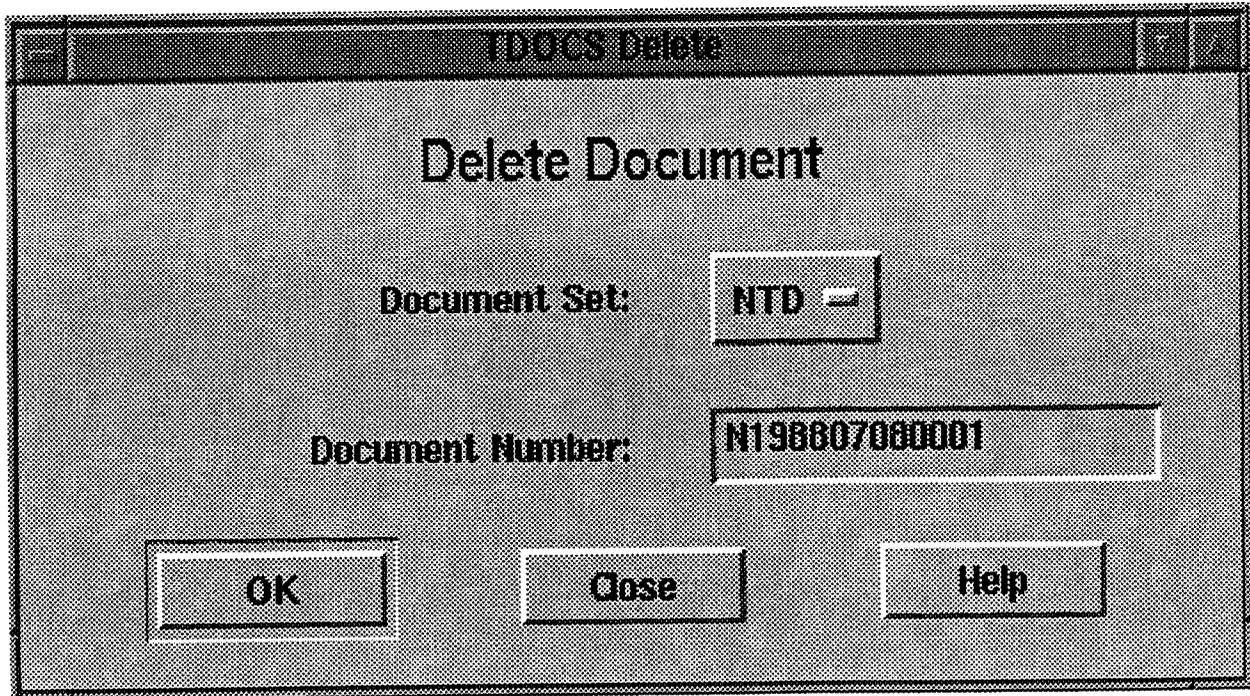


Figure 4-12. Update document request screen

is entered. The client formats an RPC message including a record update request code, the document set, and the document identifier, and sends it to the server. The server retrieves the document header from the relational database. If the record is not found or if it has been marked as deleted, a reply, including error status codes, is sent to the client and an error message is displayed. If the record is retrieved successfully, the server formats an RPC reply including the document header record and sends it to the client. The client displays the document header and accepts updates (see Figure 4-13). The custodian updates the header screen and signals completion of the update by: (i) selecting the *Update* push-button, (ii) selecting the *Import* push-button, (iii) selecting the *Clear* push-button, or (iv) selecting the *Close* push-button. If the *Clear* or *Close* push-button is selected, the client terminates the transaction.

If the *Update* push-button is selected, the client code formats a header update RPC message, including the transaction code and the updated header information, and sends it to the server. The server updates the header information in the relational database and terminates the transaction without altering the text or image files. An RPC reply, including status, is sent to the client code, and an advisory message is displayed to indicate successful completion of the transaction or any error conditions encountered.

The screenshot shows a graphical user interface window titled "TDOCS" with a main heading "Update a TDOCS Record". The interface includes the following elements:

- Document Set:** A text box containing "NTD".
- Document Number:** A text box containing "N198607000001".
- Document Header:** A section containing several input fields:
 - Author(s):** "Carpenter, D. W.; Chung, D. H." with a dropdown arrow on the right.
 - Addressee(s):** An empty text box with a dropdown arrow on the right.
 - Title:** "Effects of Earthquakes on Underground Facilities: Literature Review and I" with a dropdown arrow on the right.
 - Publication:** "unknown" with a dropdown arrow on the right.
 - Document Date:** "01 Jun 1986".
 - Code(s):** "313".
 - Number(s) *:** An empty text box.
 - Duration:** An empty text box.
 - Note *:** An empty text box with a dropdown arrow on the right.
- *Optional Values:** A label positioned below the "Number(s) *" and "Duration" fields.
- Buttons:** A row of five buttons at the bottom: "Update", "Clear", "Import", "Close", and "Help".

Figure 4-13. Update document screen

If the *Import* push-button is selected, the client code displays the *Import Document* screen (see Figure 4-9) to permit the custodian to indicate the directory containing the input files and the types of files being submitted. The client copies the files from the designated directory to the client's *submit* directory and formats a header and text update RPC message that includes the updated header information and information indicating the types of files being submitted. For a discussion of server processing for updating document records, including input file names and formats, see Section 3.5.4.

The server utilizes FTP to copy the files from the *submit* directory on the client platform to the appropriate *incoming* directory on the server platform. A separate *incoming* directory is provided on the server for each client platform to store submitted files. The record update transaction must include the following:

- A Flag—indicating the type of transaction and the type of files stored in the client's *submit* directory
- The document header—a valid document header with all required fields, formatted as pairs of field identifiers and data in the RPC message
- Text File—a single file is required for textual documents containing all of the text of the document in ASCII format with a file name extension of *txt*
- Binary File—a file containing the text in binary format with a file name extension other than *txt*, *gif*, *tif*, or *pcx* to indicate a format such as WordPerfect
- Full-Page Images—including a separate image file for each page of text with an appropriate extension (e.g., *gif*, *tif*, or *pcx*)
- Images of Figures, Equations, etc.—including a separate image file for graphical element with an appropriate extension (e.g., *gif*, *tif*, or *pcx*)

The server uses the document's unique internal identifier to change the names of the input files in the following ways when they are copied into the server's *incoming* directory:

- Text file—the name of the text file is changed to the internal document name. The original text file extension is retained (e.g., *data.txt* would be changed to *docname.txt*; *data.wpf* would be changed to *docname.wpf*).
- Image files—the names of the image files are changed (e.g., *data001.tif* would be changed to *001.tif*; *data001a.tif* would be changed to *001a.tif*).

The server validates the input, checking for the presence of the expected file types and examining the format of the document header. If errors are found, a reply including status and error codes is sent to the client. If no errors are detected, the header information is used to update the document header record in the relational database. This record in the relational database is marked to cause the document text and image files to be updated in the full-text repository during the next execution of the batch process. Following processing of a successful transaction, the server sends a reply including status codes to the client code. The client erases all files from the *submit* directory and notifies the user that the transaction was completed successfully.

4.6 DATABASE ADMINISTRATION

Database administration facilities are only available to authorized DBAs. To access these facilities, the user must perform a logon with an appropriate User-ID and password. After the logon process has been completed, the system displays the *TDOCS/TDAS Main Menu* for the database administrator (Figure 4-14). This screen contains a menu bar at the top of the screen with functions for *Operations, Search, Custodian, Report, TDAS, System, and Help*.

DBA facilities are accessed by selecting the *System* entry in the *TDOCS/TDAS Main Menu*. This causes the *System* pull-down menu to appear (Figure 4-15). The *System* pull-down menu includes facilities to maintain user accounts.

The *User IDs* entry in the *System* pull-down menu entry permits the DBA to perform maintenance on the list of authorized users of the system. When the *User-IDs* entry in the *System* pull-down menu is selected, the *User ID Maintenance* entry screen is displayed (Figure 4-16). This facility permits users to be added, deleted, or changed.

4.6.1 Adding a New User-ID

When the *User ID Maintenance* screen is displayed, the DBA may enter a new user account. The new *User-ID* is entered along with the initial password and the appropriate user privilege level is selected from the *Privilege* drop-down list. When the *Add* push-button at the bottom of the *User ID Maintenance* screen is selected, the client validates the information and displays a message if error conditions are found (e.g., duplicate *User-ID*). If no errors are found, the client formats an RPC message, including the information for the new user, and sends it to the server. The server adds the new user record to the relational database and sends a reply, including status codes to the client.

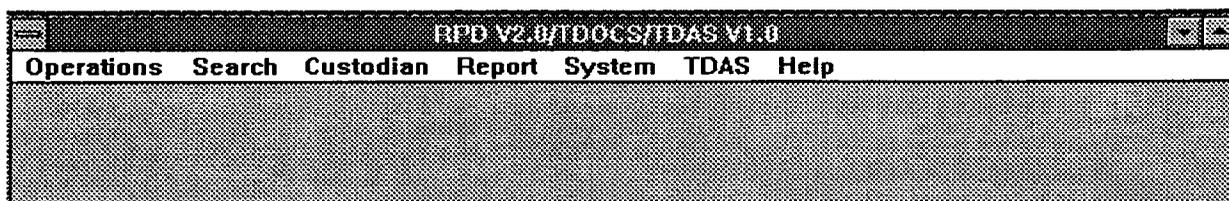


Figure 4-14. TDOCS/TDAS main menu for database administrators

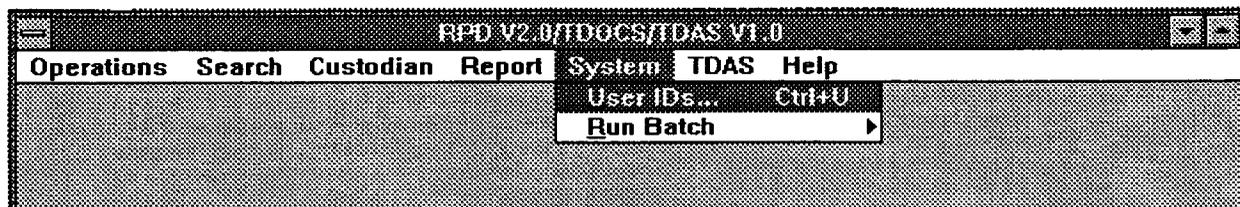


Figure 4-15. TDOCS database administrator system pull-down menu

4.6.2 Deleting a User-ID

When the User ID Maintenance screen is displayed (Figure 4-16), the DBA may delete user accounts. The User-ID to be deleted is selected from the scrollable list view of User-IDs and privileges. When the correct User-ID has been highlighted, the DBA selects the *Delete* push-button at the bottom of the *User ID Maintenance* screen to delete the user account record. The client formats an RPC message including the User-ID to be deleted and sends it to the server. The server deletes the user record from the relational database and sends a reply, including status codes, to the client.

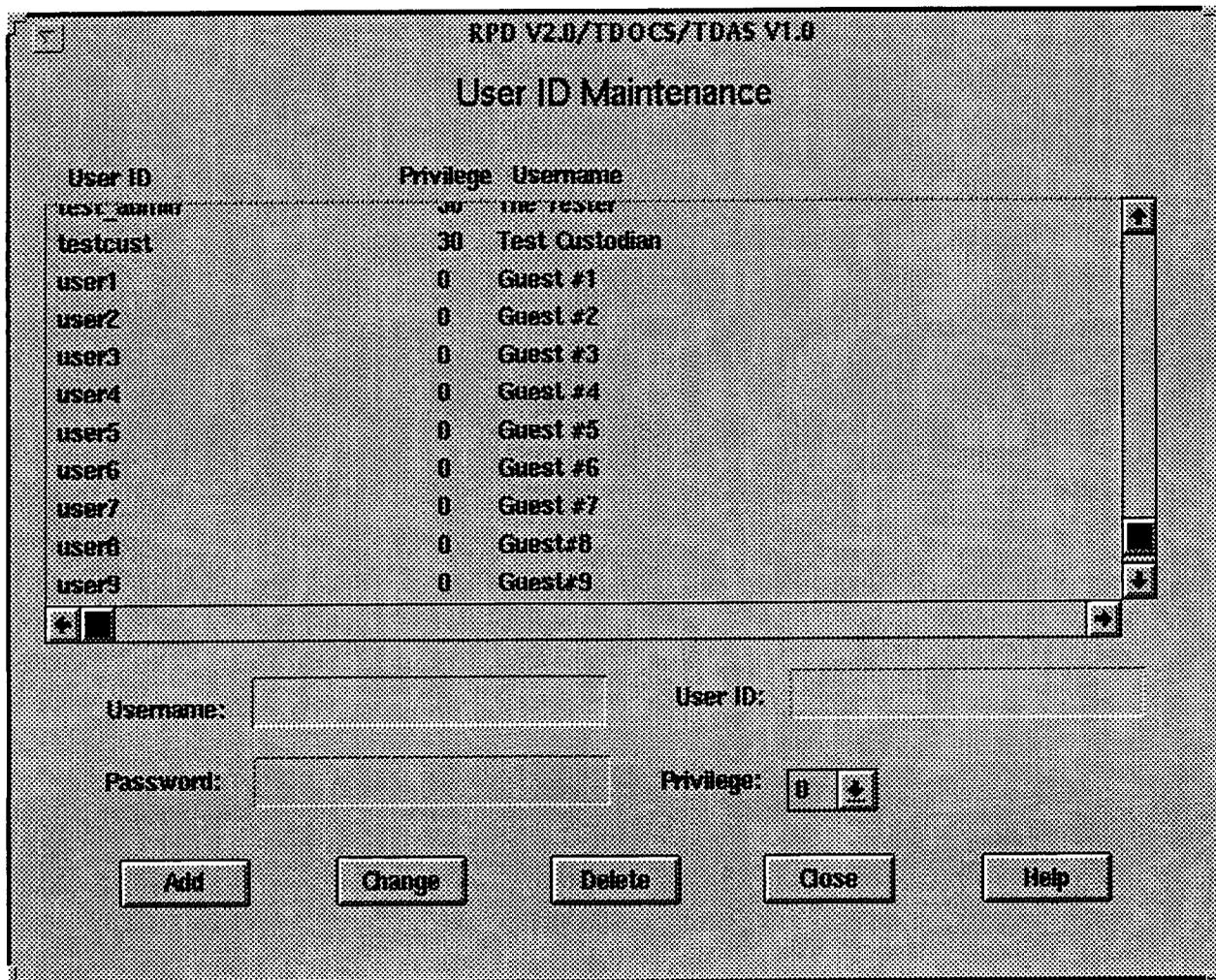


Figure 4-16. TDOCS User ID maintenance screen

4.6.3 Changing User Privileges

The DBA may also change user accounts. The User-ID to be updated is selected from the scrollable list view of User-IDs and privileges. When the correct User-ID has been highlighted, the DBA updates the displayed information including the User-ID, password, and the appropriate user privilege level selected from the *Privilege* drop-down list. When all of the information has been changed and

visually verified, the DBA selects the *Change* push-button at the bottom of the *User ID Maintenance* screen to update the user account record. The client formats an RPC message and sends it to the server. The server updates the user record in the relational database and sends a reply including status codes to the client.

5 DOCUMENT SEARCH AND RETRIEVAL CLIENTS DESIGN

All TDOCS users have access to the full-text search and retrieval facilities of the system through the Document Search and Retrieval clients. These facilities, resident on the user's individual workstation or PC, access the server full-text repository and indexes through NFS. Once initiated, the Document Search and Retrieval permits users to perform several document search and retrieval functions:

- Full-text searches
- Structured header searches
- Document viewing and browsing
- Image display
- Launch of word processors
- Document download

5.1 DOCUMENT SEARCH AND RETRIEVAL CLIENTS HARDWARE AND SOFTWARE

TDOCS runs in any of four different hardware/software environments:

- OpenLook or MOTIF using Sun hardware
- Microsoft Windows using IBM PC or compatible hardware
- IBM OS/2 using IBM PC or compatible hardware
- SYSTEM 7 using Macintosh Quadra hardware

The functionality of TDOCS and the general appearance of the user interface is the same, regardless of the hardware/software environment. However, the specific details of the screens may vary slightly to conform to the standard *look and feel* of the specific hardware/software environment. All hardware/software environments employ a mouse to permit the user to quickly select options and traverse menus. The following sections describe the minimum requirements necessary to achieve acceptable system functionality and performance for each hardware and software environment.

5.1.1 Minimum Sun Workstation Requirements for Using the Technical Documents Reference Database

- Sun 4 Architecture (IPX or faster)
- 32 MB RAM
- 10 MB free hard disk space

- Sun OS 4.1.3 or Solaris 2.3

5.1.2 Minimum Windows Workstation Requirements for Using the Technical Document Reference Database

- Intel 80486 CPU-based computer
- 8 MB RAM
- 10 MB free hard disk space
- Microsoft Windows Version 3.1 or later
- NetManage Chameleon NFS

5.1.3 Minimum OS/2 Workstation Requirements for Using the Technical Document Reference Database

- Intel 80486 CPU-based computer
- 8 MB RAM
- 10 MB free hard disk space
- IBM OS/2 Version 2.1 or later
- IBM TCP/IP Version 2.0 or later, configured for NFS

5.1.4 Minimum Macintosh Workstation Requirements for Using the Technical Document Reference Database

- Macintosh Quadra
- 8 MB RAM
- 10 MB free hard disk space
- System 7 or later

5.1.5 Document Client Search and Retrieval Software

Client software for the Document Search and Retrieval clients is determined by the selection of the commercially available off-the-shelf software tools for full-text search and retrieval. As discussed in Section 3.2.2, the TOPIC full-text search and retrieval was selected and utilized in the TDOCS initial implementation.

5.2 FUNCTIONAL MANAGEMENT AND CAPABILITIES OF THE DOCUMENT SEARCH AND RETRIEVAL CLIENTS

The TDOCS utilizes both bibliographic header and full-text searches. Full-text search and retrieval is provided to permit users to search for occurrences of words and/or phrases in the text of documents. Bibliographic header searches are used to permit users to find documents more precisely and quickly when specific information is known, such as the title, author, date, etc. They are also used to support control of the database and reporting functions.

All TDOCS users have access to the facilities of the Document Processing Clients. When a user selects the *Search* entry in the *TDOCS/TDAS Main Menu*, the Search pull-down menu is displayed (Figure 5-1).

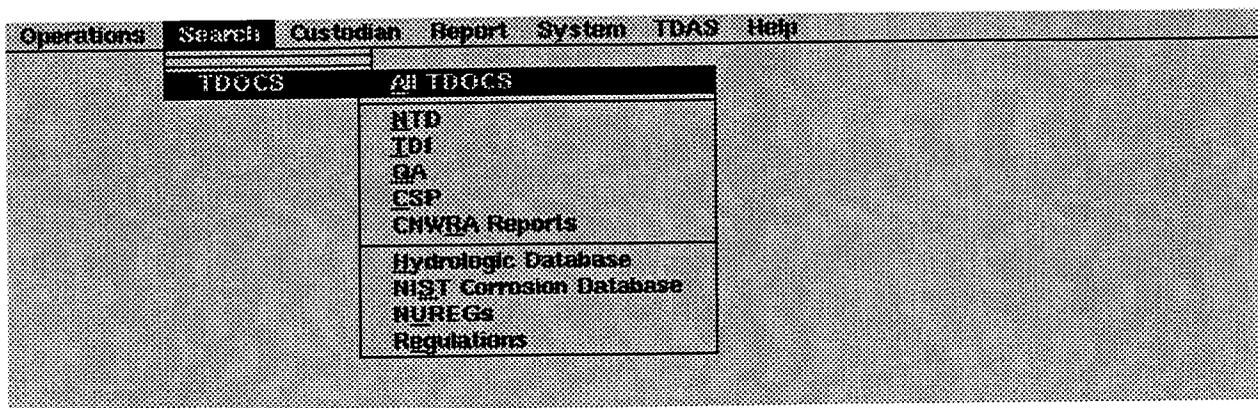


Figure 5-1. Search pull-down menu

This pull-down menu permits the user to select the document sets available to be searched. When an option is selected from this pull-down menu, the client configures an appropriate preferences file and starts the TOPIC full-text search and retrieval software using those preferences. The effect of this is to limit the scope of searches to the specified document set(s).

5.2.1 Searching and Retrieving

When the TOPIC full-text search and retrieval software is initiated, a simple query entry screen is displayed (Figure 5-2). This screen contains two panes:

- An upper pane that permits entry of the query parameters
- A lower pane that permits display of the query results

The upper pane permits the user to enter words and/or phrases to be used in searching for records. TOPIC queries are formulated in a very intuitive manner. Words, phrases, or expressions may be entered to find the desired materials in the full-text repository.

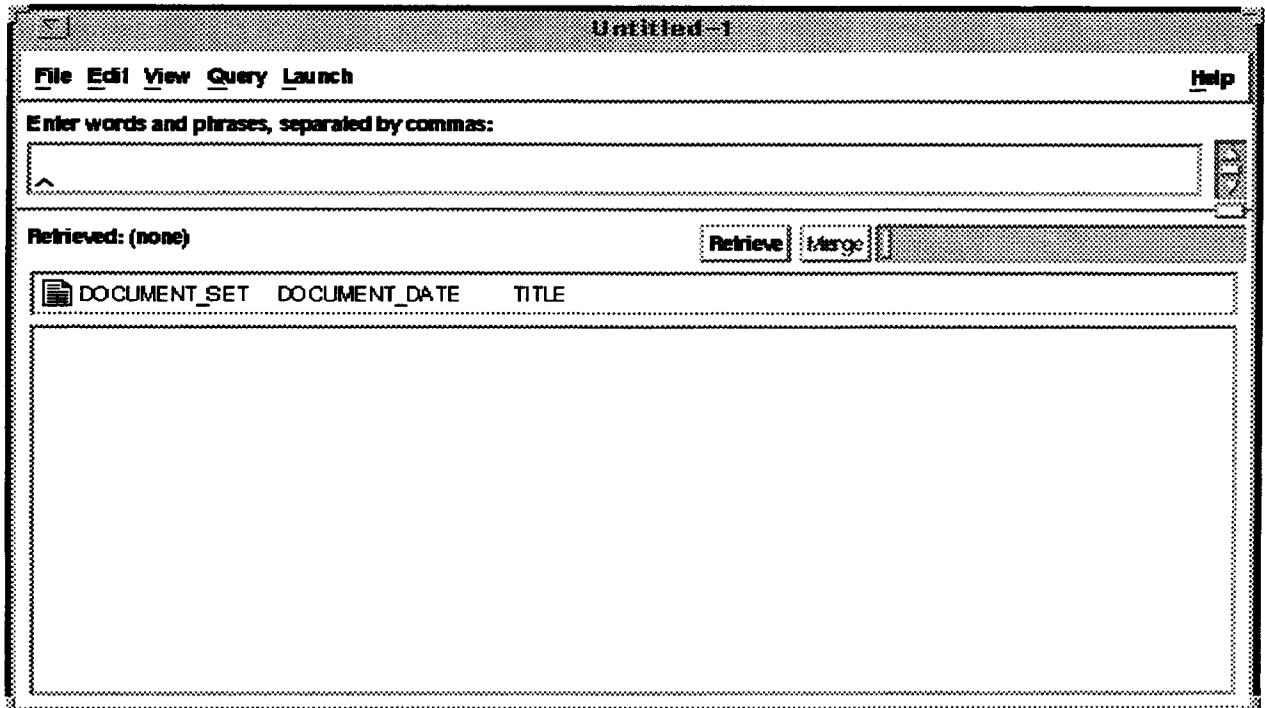


Figure 5-2. Simple query entry screen

- **Words**—a single word or multiple words separated by commas may be entered to formulate a TOPIC query. Each word will be used individually to search for documents containing that word. When multiple words are entered, separated by commas, TOPIC will retrieve documents containing any of the specified words. Words are considered to be stems in that endings are ignored. For example, if the word **market** is entered, TOPIC will select documents that contain “market,” “markets,” “marketing,” or “marketed.” If a specific word is desired, it should be enclosed in double quotation marks to prevent it from matching as a stem. Words are considered to be case insensitive. Thus, words may be entered for search purposes in lowercase, uppercase, or mixed case.
- **Phrases**—multiple words may be entered without separating them with commas. The system will treat them as phrases and retrieve only documents that contain corresponding phrases. If more than one phrase is desired, the individual phrases should be separated with commas. For example, if the search argument **RADIOACTIVE WASTE, HIGH-LEVEL WASTE** is entered, the system will search for documents containing either or both phrases. Within phrases, TOPIC treats words as stems. Thus, if the search phrase, **HIGH-LEVEL WASTE**, is entered, TOPIC will select documents that contain the phrases: (i) “high-level waste...,” (ii) “high-level wastes...,” (iii) “higher level waste...,” or (iv) “higher-level wastes...”. If a specific phrase is desired, the phrase should be enclosed in double quotation marks.
- **Expressions**—combinations of phrases and words may be indicated by using special words such as “and,” “or,” and “not” to combine the words and phrases into expressions. When one of these reserved words is entered, TOPIC will change its appearance to indicate that the reserved word is being interpreted as a logical operator rather than as a word. For example, if

the expression, **WASTE AND CONTAINMENT** is entered, the system will interpret the word “and” as an operator, and will display the query terms as **WASTE <AND> CONTAINMENT**. This query would only retrieve documents that contained both the words “waste” and “containment.” If a reserved word such as “and,” “or,” or “not” must be used as a word in a phrase, it should be enclosed in quotation marks to prevent the system from interpreting it as an operator. For example, **WASTE “AND” EMPLACEMENT** would be interpreted as a phrase rather than an expression because “and” is enclosed in quotation marks.

The lower pane is used to display the results of the query. As the query is executed, several pieces of status information are displayed: (i) a colored bar at the right side of the window, under the entry field, will fill from left to right to indicate how far the query has progressed; (ii) a message at the left side of the window, under the entry field, which has the form “**Retrieved: 25 of 150**”, will update continuously to indicate how many documents have been selected and how many have been examined by the query; and (iii) the results list in the lower pane of the window will fill from top to bottom with a description of the selected records (Figure 5-3).

When the first query is executed, the Document Processing Client accesses the appropriate full-text indexes using NFS and pages them into virtual memory on the client platform. Therefore, the first query may require more time to execute than would normally be expected due to the transfer of the indexes through NFS. However, subsequent queries are actually accessing the indexes in the client’s virtual memory and they execute quite efficiently.

TDOCS also supports the ability to search for words and/or phrases in specific *header* fields, such as title, author, date, etc. This type of query is called a form query, because the input screen looks like a highly structured form. When using a form query, the upper pane of the *TOPIC Query Entry* screen (Figure 5-4) contains a list of available fields, each with its own entry field. For each header field and its associated entry field, there is a selection labeled *Require*. If this option is selected for a field, it will be required for the search to be satisfied. Otherwise, the field will be considered optional.

5.2.2 Viewing Documents

As a query runs, two pieces of information are updated in the status area immediately below the *Query Entry* pane. When the bar is completely filled with color, the search has been completed, and a brief description of the selected documents appears in the results list located immediately below the status area (see Figure 5-3, which illustrates a typical results list following a simple query).

The results lists contains the following information about each selected document:

- **DOCUMENT_SET**—a code indicating the type of document (e.g., NTD, TDI, CSP, QAR, etc.)
- **DOCUMENT_DATE**—the date of the document
- **TITLE**—the title of the document appears in this field. If the title is too long to display within the results list pane, the horizontal scroll bar indicator at the bottom of the results list may be used to shift the display left or right as needed.

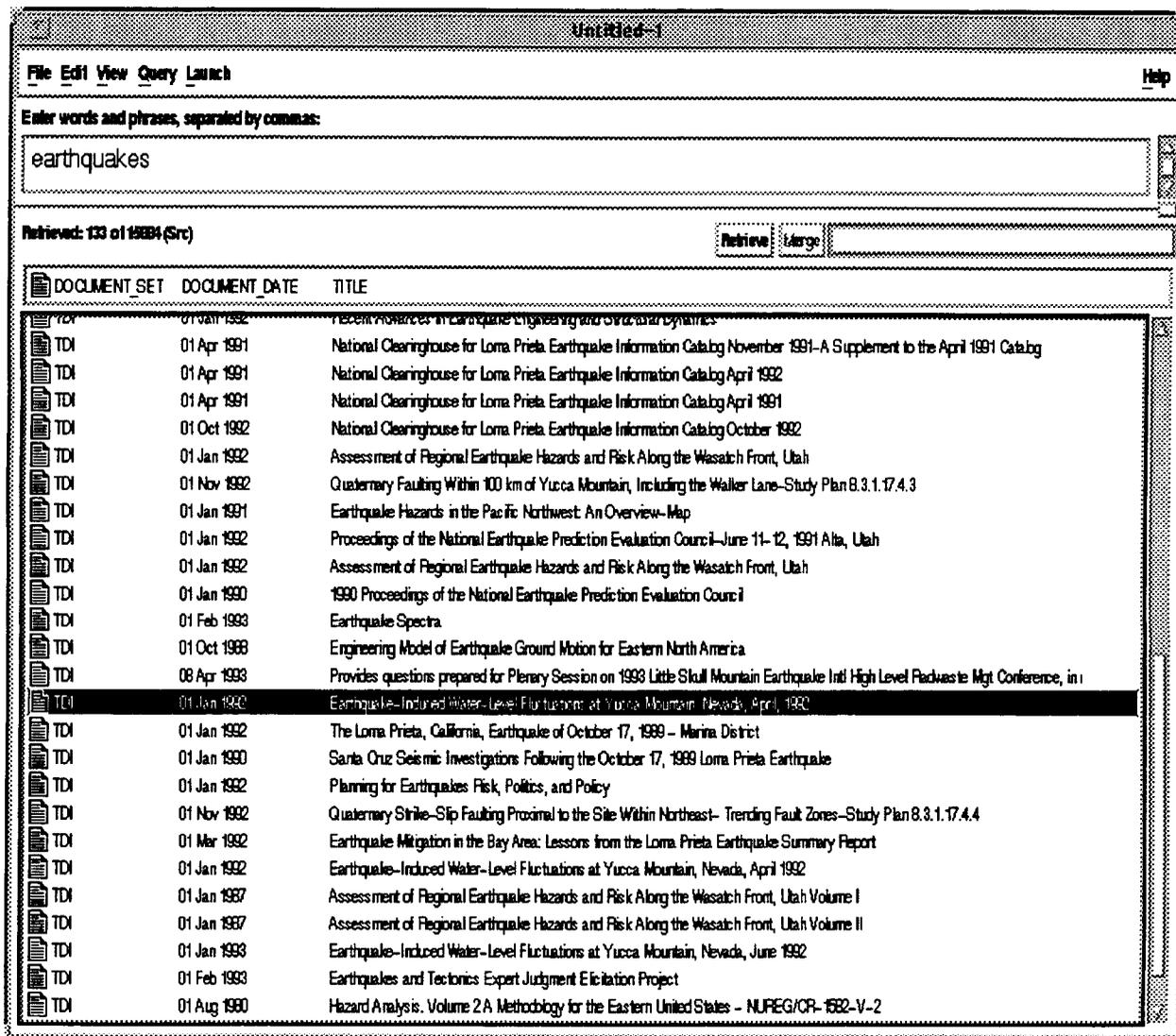


Figure 5-3. Query results list

Documents may be selected from the results list for viewing. The selected document is retrieved from the full-text repository on the server through NFS and displayed in a separate window. More than one document at a time may be displayed, and each document is displayed in its own window. The search term(s) used to retrieve the document will be highlighted, and "navigation" facilities are provided to permit the user to rapidly *jump* or *move* through the document display to the next or previous occurrence of the search term (Figure 5-5). Facilities are provided to permit the user to search for additional words and/or phrases that were not used as original search terms.

5.2.3 Launching Word Processing Software

Word processing software may be used to access the text of documents in the TDOCS repository. The full-text search and retrieval document display facilities of the Document Search and Retrieval Clients are used to locate and display the desired record. A document that has been selected for

Untitled-1

File Edit View Query Launch Help

DOCUMENT SET
 Require

DOCUMENT NUMBER
 Require

AUTHOR
 Require

TITLE
 Require earthquakes

DOCUMENT DATE
 Require

Retrieved: 22 of 15004 (Src) Refresh Range

DOCUMENT SET	DOCUMENT DATE	TITLE
TDI	01 Jan 1990	Earthquake Prediction-Nine Major Earthquakes in China-(1966-1976)
TDI	01 Jan 1988	Earthquakes
TDI	01 Jan 1991	Catalogue of Earthquakes and Volcanic Eruptions
TDI	01 Jan 1990	The mechanics of earthquakes and faulting
TDI	01 Jan 1992	Planning for Earthquakes Risk, Politics, and Policy
TDI	01 Feb 1993	Earthquakes and Tectonics Expert Judgment Elicitation Project
TDI	01 Jan 1988	Catalog of Earthquakes in Southern Alaska for 1985
TDI	18 Sep 1991	Extends invitation for participation in upcoming Workshop on Earthquakes & Tectonics for EPRI HLW performance assessment project c
TDI	10 Sep 1970	TECTONIC STRESS AND THE SPECTRA OF SEISMIC SHEAR WAVES FROM EARTHQUAKES
TDI	25 Dec 1989	FORWARDS STATE OF NV, 'EARTHQUAKES IN NV & HOW TO SURVIVE THEM.'
TDI	01 Jan 1982	A Comparison of Ground Motion from Earthquakes and Underground Nuclear Weapons Tests at NTS
TDI	01 Feb 1981	PROCEEDINGS WORKSHOP ON SEISMIC PERFORMANCE OF UNDERGROUND FACILITIES'A COMPARISON OF GROUN
TDI	01 Apr 1988	Reported effects of Selected Earthquakes in the Western North American Intermontane Region, 1652-1983, on underground workings an
TDI	01 Apr 1988	Reported Effects of selected Earthquakes in the Western North American Intermontane Region, 1652-1983, on underground workings an
TDI	01 Jan 1988	Location Refinement of Earthquakes in the Southwestern Great Basin, 1931-1974, and Seismotectonic Characteristics of Some of the I

Figure 5-4. Form query screen

display may be viewed and/or edited using WordPerfect, but may not be saved directly into the TDOCS because updates to the records in the TDOCS are performed through the document submission process under the control of custodian or DBA. However, records from the TDOCS may be freely accessed, modified, incorporated into other work products, and saved on a diskette or local hard disk.

When a document has been selected for display through the TDOCS search facilities, an icon and descriptive line (i.e., **Double-Click the icon to launch WordPerfect 5.1**) appears at the top of the document. When this icon is selected, the system will start a WordPerfect session in another window using the selected document. Alternatively, the user may select the *WordPerfect* option from the *Launch* pull-down menu (Figure 5-6). The user may perform any WordPerfect functions with this document. If the results of the WordPerfect session are saved, this action will not modify the record in the TDOCS. The text in the server's *archive* directory is flagged as *Read-Only*. Therefore, the user may read, edit, and print the document, but the updated file may not be stored back in the TDOCS *archive* directory. If the document is to be saved, the user must specify a local directory and file name.

The internal document number of the selected file and its format code are stored in the full-text fielded information for the document. The client uses this information to build the file name and path for the document text file in the server's *archive* directory. The client starts the WordPerfect software using the file name and path, and the user may edit, display, save to a local directory, and/or print the document.

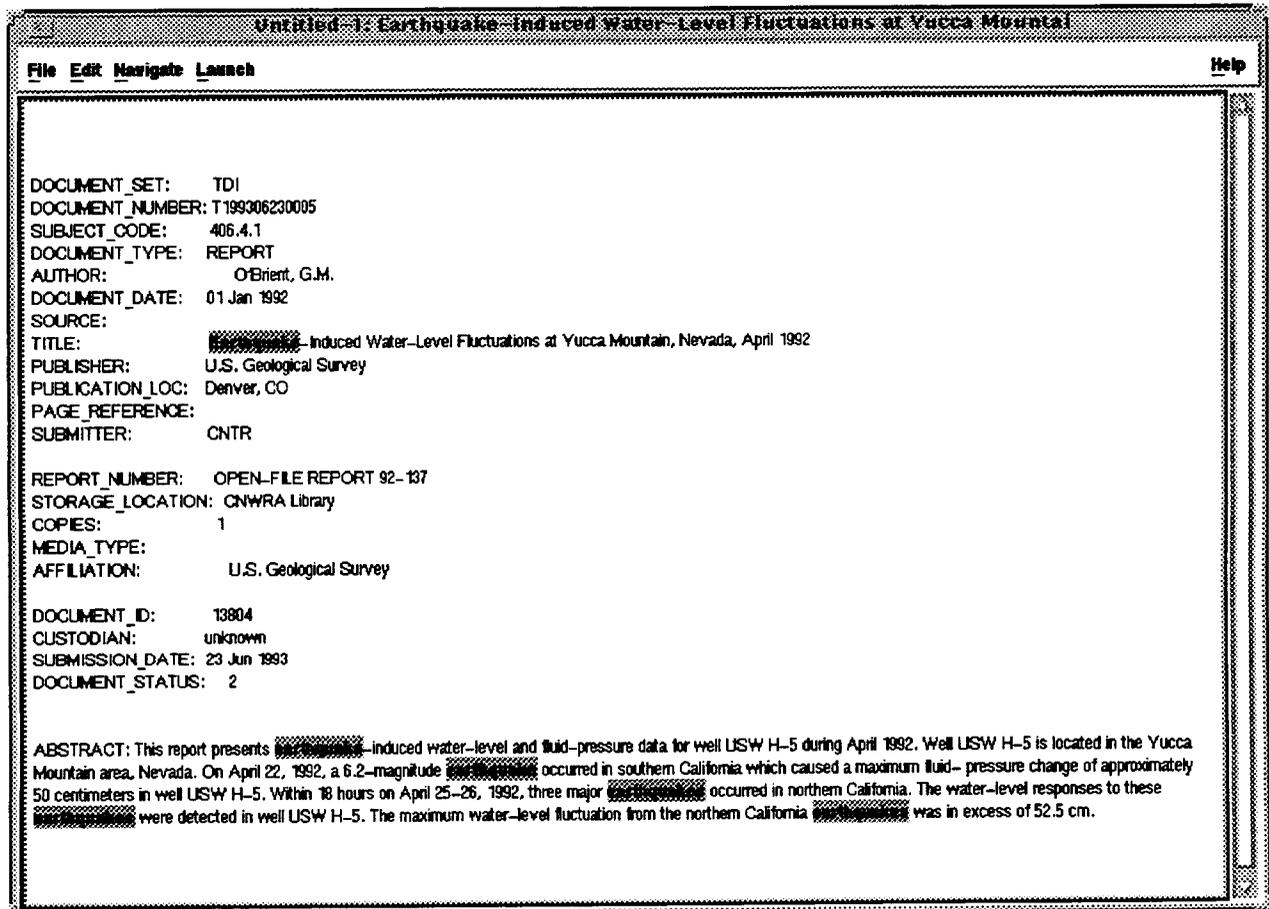


Figure 5-5. TOPIC document display

5.2.4 Image Viewing

Image viewing is accomplished by selecting the hyperlink icon for the desired image from the document display. The exact nature of the image viewer depends on the specific hardware/software platform. The following image viewers were utilized in the TDOCS initial implementation:

- Sun workstations—Sun XV image viewer
- Microsoft Windows platforms—Paint Shop Pro image viewer
- IBM OS/2 platforms—PMJPEG image viewer

When the icon is selected, the image viewer is started and the name of the image file is passed to it (Figure 5-7). The image file is displayed in a separate window for viewing and may be zoomed, and/or scrolled vertically and horizontally. When the user finishes viewing the image, the image viewer is closed, and control is returned to the display of the document text.

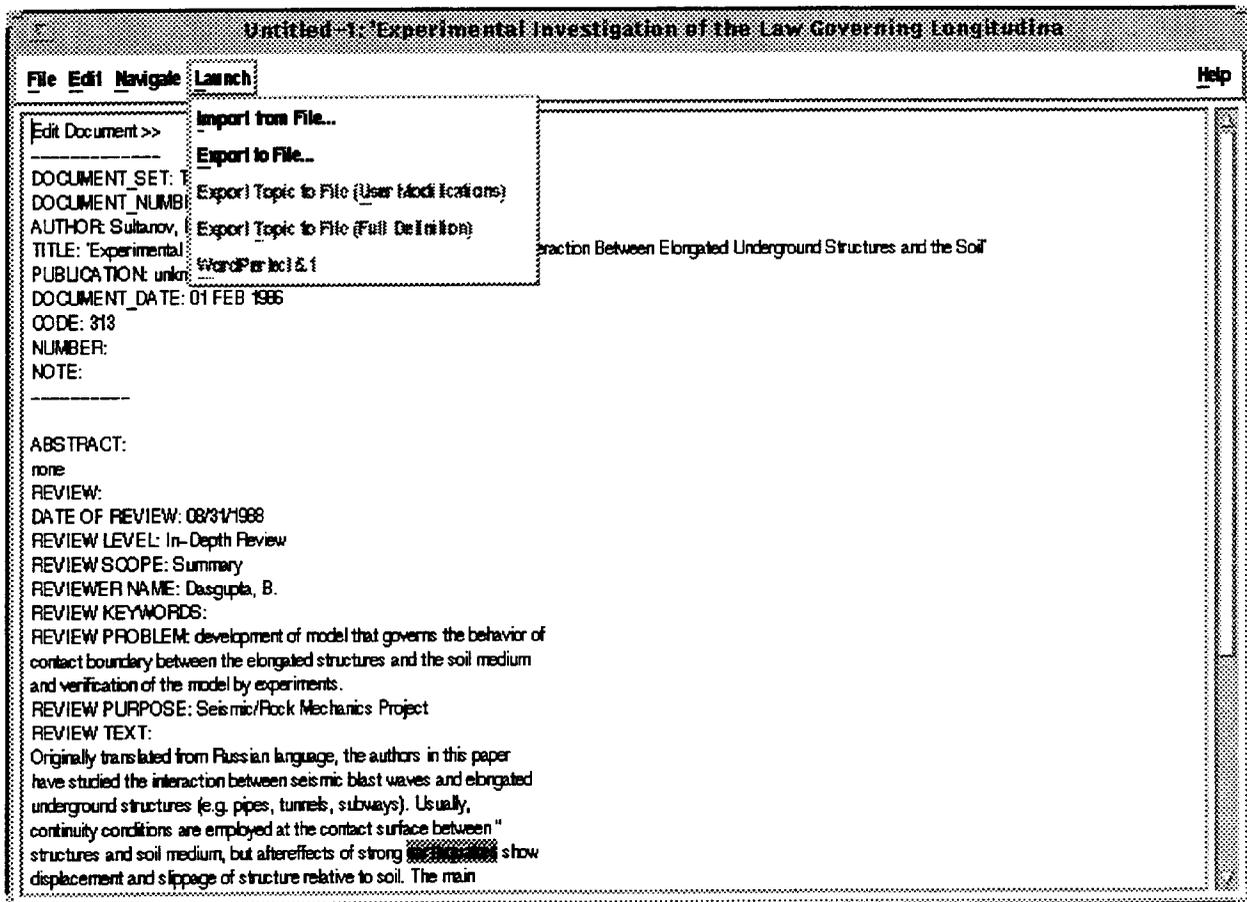


Figure 5-6. Launch pull-down menu

5.2.5 Hypertext Linking

When a document is selected for display, the document text appears in a window. The search and retrieval facilities support the creation of hyperlinks to permit related materials to be associated by embedding hyperlink patterns in the document text prior to loading. When a hyperlink exists in the text, it appears as a highlighted word or phrase. If the user selects the hyperlink, the document display is repositioned to display the associated text. If the associated text is in a different document, the target document is accessed, displayed in a separate window, and positioned to the appropriate location in the text. Figure 5-8 illustrates a document display with hypertext links.

5.2.6 Cutting and Pasting

TDOCS supports cut and paste to permit users to extract limited amounts of text from a document and incorporate that text in new work products. When a portion of the displayed text has been highlighted, the user may select *Cut* from the *Edit* pull-down menu. This causes the selected text to be stored in the system clipboard. Then the user may move to the target window, (e.g., a word processing document), select the desired location, and *Paste* the text from the clipboard to the new document.

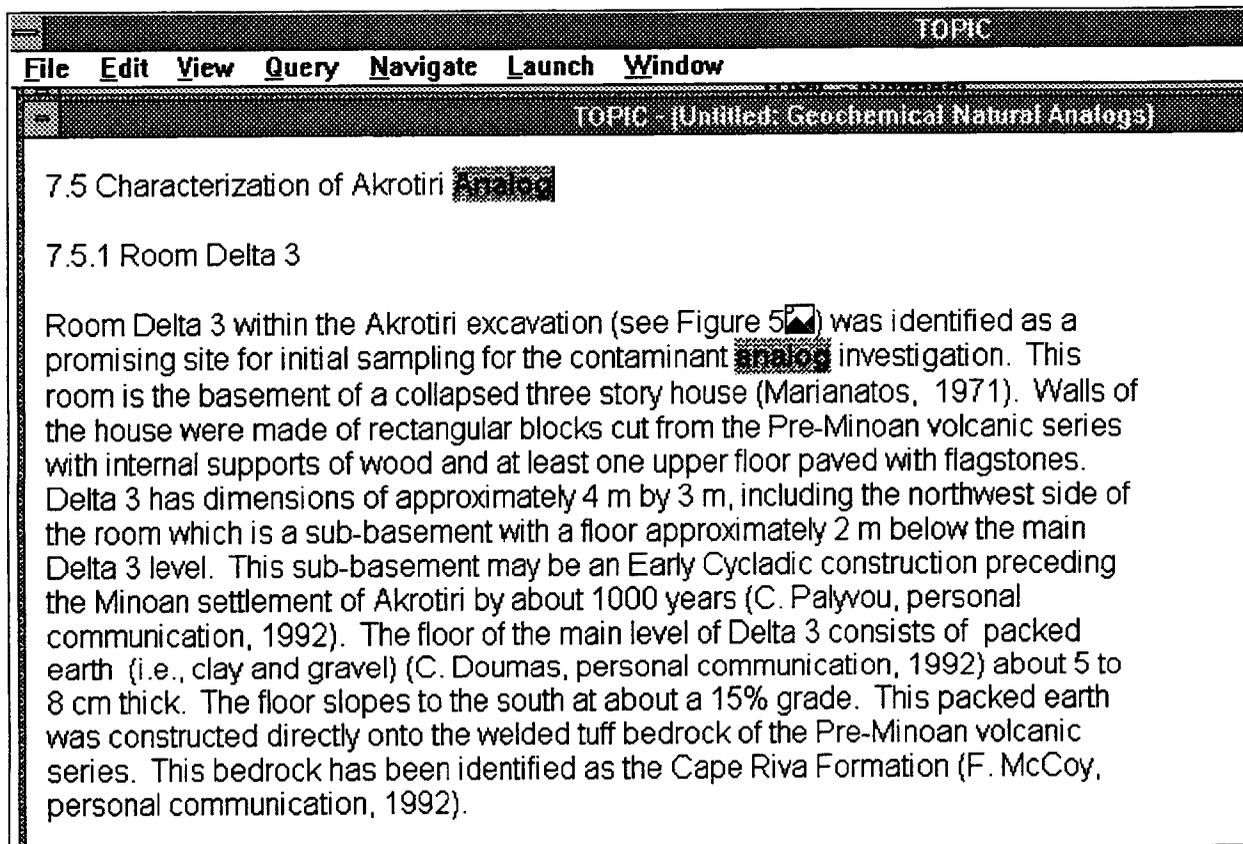


Figure 5-7. Selecting images for viewing

5.2.7 Document Downloading

Document download capabilities permit users to access documents in the TDOCS repository and copy text and image files to a local hard disk or diskette. The full-text search and retrieval document display facilities of the Document Search and Retrieval Clients are used to locate and display the desired document. When the document is displayed, the user selects the *Export to File* option from the *Launch* pull-down menu (see Figure 5-6).

The client displays an input screen and waits for the user to enter the disk drive and directory where the text and image files should be stored (Figure 5-9). If the *Cancel* push-button is selected, the download transaction is terminated. The user enters the full path of the directory where the files are to be stored, and selects the *OK* push-button. The client formats an RPC message, including the document identifier and the full path of the target directory, and sends it to the server. The text and image files for the desired document are retrieved from the *archive* and *images* directories in the TDOCS repository. The server uses FTP to transfer the text and image files to the specified directory on the client platform and then sends a reply including status information to the client. For a further discussion of the server process for document downloading, see Section 3.7.

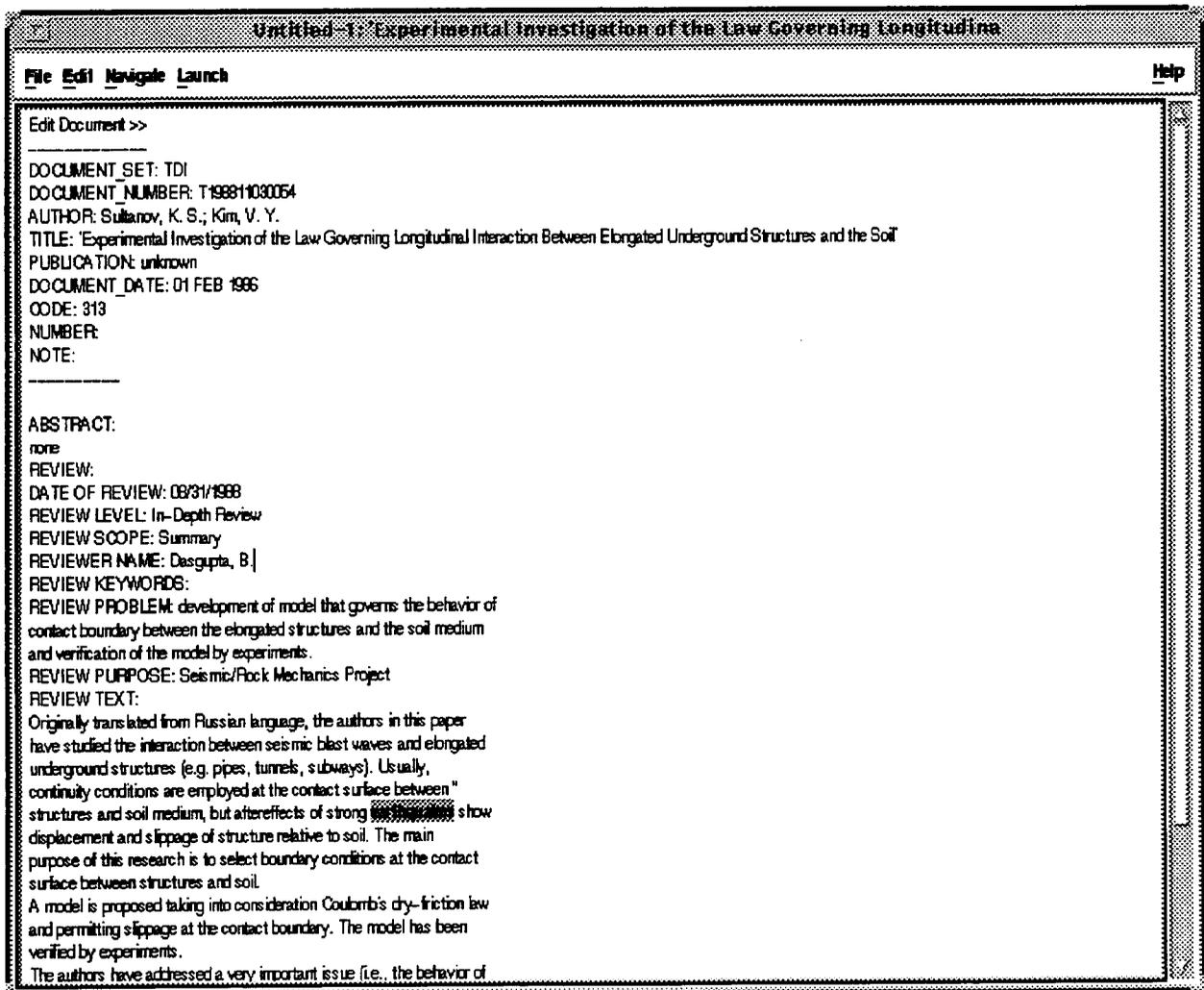


Figure 5-8. Hypertext links within a displayed document

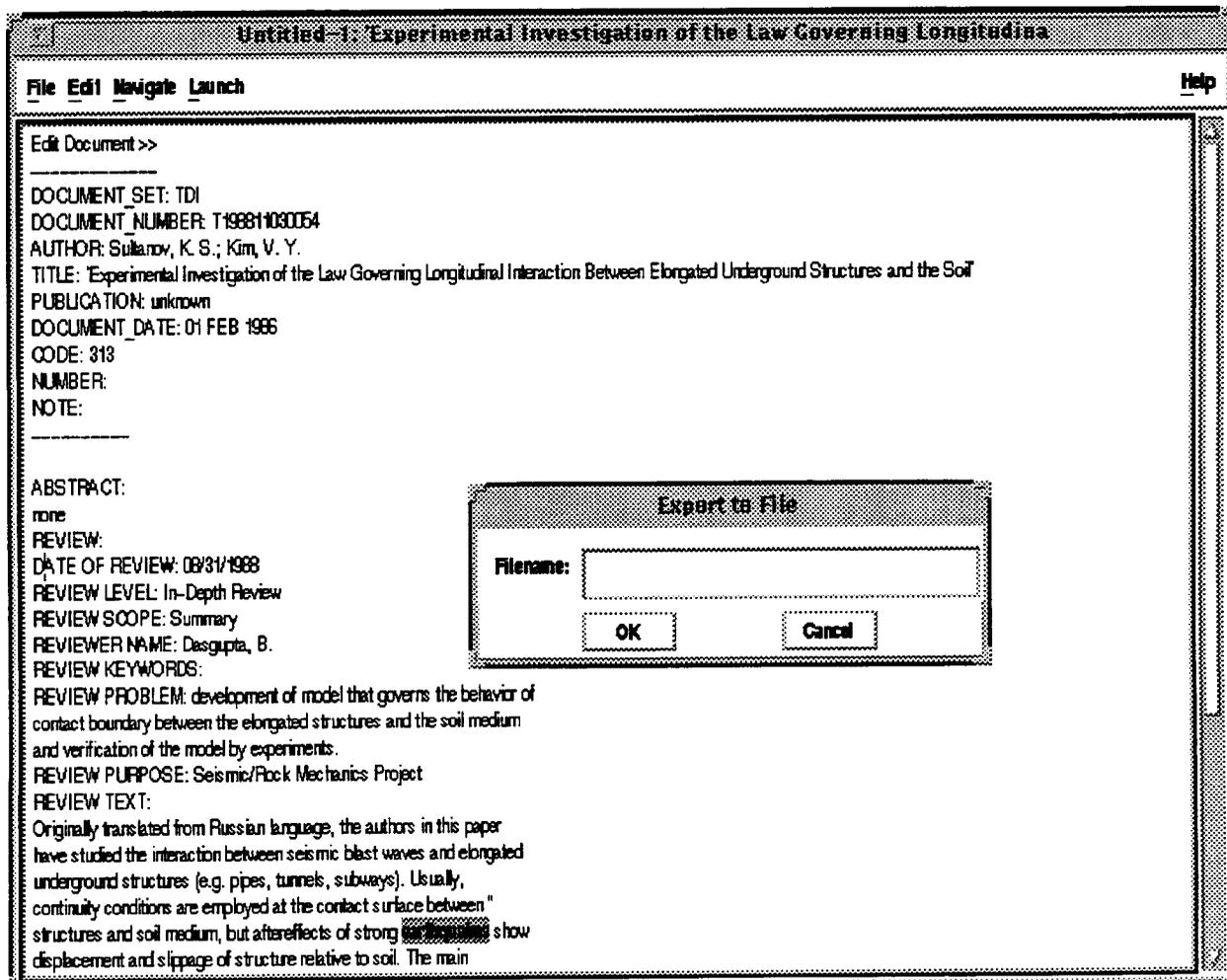


Figure 5-9. Directory specification for document download

6 NETWORK ARCHITECTURE

The network architecture of the TDOCS application is the mechanism by which the Document Manager Server, Document Processing Clients, and Document Search and Retrieval Clients exchange information. The network architecture depends on and must be compatible with the ACRS, AUTOS, and CNWRA local area network (LAN) systems. It also depends upon the NRC and CNWRA firewall systems for security protection.

The network architecture consists of an RPC mechanism and an NFS capability. The TDOCS modules use the RPC for capability to request another computer to perform a service and return the results. The TDOCS modules rely on the NFS capability to retrieve a file from another computer. The RPC and NFS systems, in turn, utilize and depend on the proper installation, configuration, and operation of the TCP/IP protocol suite on the TDOCS computers. The TCP/IP protocol suite utilizes and depends on the infrastructure that the ACRS, AUTOS, and CNWRA LAN systems provide for physical interconnection of all of the computers that execute modules of the TDOCS system (e.g., ethernet, token ring, routers, concentrators, etc.).

6.1 REMOTE PROCEDURE CALL

RPC is a mechanism used in the TDOCS application by which a program makes a function call to a block of code (procedure) executing on another computer. The remote procedure usually has access to resources (databases, compute cycles, vector processor, etc.) that are not available to the calling computer. The calling program on the first computer passes parameters into the function call used by the remote function and receives return values from the remote procedure.

An example of the RPC mechanism is found in the Document Processing Client module that requests the Database Server to delete a TDOCS document. The user enters the document set and document number through the *Delete Document* screen of the Document Processing Client to uniquely identify the document to be deleted. The Document Processing Client makes a function call, passing the document set and the document number. The RPC mechanism transports the name of the function to be performed and the arguments for that function to the Database Server machine. The requested block of code is executed on the Database Server machine. During execution of the code, the function checks to make sure that the document number is valid for the document set, retrieves the title of the document, and returns the title of the document to the Document Processing Client via the RPC mechanism. The Document Processing Client code presents the title of the document in a confirmation screen so that the user can confirm the deletion or cancel the deletion operation.

The specific RPC mechanism that is used in the TDOCS system is the Sun Open Network Computing (ONC) RPC. ONC RPC is part of the standard UNIX distribution and is included in most PC and Macintosh TCP/IP implementations. The operating systems and TCP/IP protocol in use at the NRC are Sun ONC compliant.

6.2 NETWORK FILE SYSTEM

NFS provides a mechanism whereby files on remote computers appear as local files on a client computer just as if they were located on a local hard drive. A PC usually sees the file system from a remote computer as a new drive letter. On a UNIX computer, the file system from the remote computer appears as

a branch in the local file system. In either case, the NFS system takes care of data conversion (i.e. byte ordering) when files are copied between local and remote file systems. NFS depends on the RPC mechanism for querying the remote computer and transferring data back and forth between the two computers. For example, a *dir* command (*ls* for UNIX) is mapped into a remote function that performs the equivalent of the directory command on the remote computer and returns the results to the local machine.

The Database Server exports the TDOCS database file system to clients using the NFS capability. The document clients interact with the exported files using the TOPIC full-text search and retrieval software to find and access documents in the full-text repository.

6.3 AUTOS AND ACRS

The TDOCS application depends on the facilities of the ACRS and AUTOS systems and the CNWRA LAN to provide the software and communications infrastructure that underly the application. Both ACRS and AUTOS provide the physical media used by the application for communication over TokenRing or Ethernet LANs interconnected by routers and/or concentrators. At the network level, both systems provide the TCP/IP protocol suite required to support TDOCS. Proper configuration of TCP/IP parameters and capabilities, such as Domain Name Service (DNS) and Network Information System (NIS) are necessary to allow the TCP/IP applications to function properly.

6.4 SECURITY

The TDOCS application depends upon the security facilities provided by the AUTOS, ACRS, and CNWRA LANs. These facilities include a firewall system at the NRC and a second firewall system at the CNWRA that protect the connection between the NRC and CNWRA networks and the Internet.

7 SUMMARY

This report documents system design for the TDOCS application and reviews requirements, constraints, and policies and their effects on the design. The design strategy for the TDOCS application was to provide currently achievable and immediately beneficial system capabilities, using well known and relatively mature software tools and sound methodologies for integration of those tools, and to meet the defined requirements and constraints of the system and application environments. The TDOCS application has been developed and implemented. The initial implementation addresses the major requirements identified for the system:

- Load and maintain an electronic repository of technical documents
- Provide full-text search and retrieval and image viewing capabilities
- Provide staff productivity enhancement, download, and cut and paste capabilities for document text and images

To comply with the requirements of TDOCS and to expedite the implementation of the initial application, commercially available software was evaluated and used for the relational database and full-text search and retrieval capabilities. Due to the specific nature of the application, an interface code was developed using the C programming language to access the facilities of the commercially available off-the-shelf database and full-text search and retrieval tools. Software packages selected for the implementation included the following:

- Relational database—ORACLE
- Full-text search and retrieval—TOPIC
- Graphical user interface—Galaxy

The TDOCS application is implemented using a client/server architecture. The TDOCS Database Server resides on a central server platform that is accessed by staff from their Workstations and PC client platforms using the RPC protocol. Multiple client platforms are supported including SUN workstations, PCs utilizing either Microsoft Windows or IBM OS/2, and Macintosh computers. The TDOCS application is implemented through three major modules:

- Document Management Server—code that supports the maintenance and use of the TDOCS relational and full-text data repositories
- Document Processing Clients—code that supports user interaction for document loading and maintenance
- Document Search and Retrieval Clients—code that supports user interaction for search and retrieval of documents in the TDOCS full-text repository

The client/server architecture of the TDOCS application enables it to respond to increasing numbers of users and expanding volumes of data, since much of the processing load is distributed to the client workstations. As additional users are added to the system, the processing capacity to handle those users is

increased by utilizing the resources of their PC's or workstations. Because the application is localized on the client platform, network traffic is minimized and substantial increases in utilization of the system may be accommodated within the current network. In the future when necessary, the network capabilities can be enhanced to further expand system capacity. The server platform is sized to currently anticipated volumes of documents, but it may be readily upgraded in both processing capabilities and data storage capacity to accommodate growth. This design provides for open-ended growth within the context of the specified requirements. The critical factors involved in sizing and throughput are proportionally scalable to address system demands, increased numbers of users, and additional document volumes.

The TDOCS initial implementation satisfies the system requirements and complements and completes the capabilities of the ACRS by making technical reference documents available for referencing, synthesizing, and incorporating in staff analyses.

8 REFERENCES

- DeWispelare, A.R., J.H. Cooper, and R.L. Marshall. 1995. *Technical Reference Document Database User's Guide Scanning Appendix*. San Antonio, TX: Center for Nuclear Waste Regulatory Analyses (to be published).
- DeWispelare, A.R., R.D. Johnson, R.L. Marshall, and J.H. Cooper. 1993. *Development Plan for PASS/PADB System Design Version 3.0*. San Antonio, TX: Center for Nuclear Waste Regulatory Analyses.
- Johnson, R., and R. Murphy. 1994. *Internet Access Plan for External Databases*. San Antonio, TX: Center for Nuclear Waste Regulatory Analyses.
- Johnson, R.D., and C. Moehle. 1993. *Meeting Report on Defining TDOCS Requirements*. San Antonio, TX: Center for Nuclear Waste Regulatory Analyses.
- Johnson, R.D., J.H. Cooper, C. Moehle, and E. Harloe. 1993. *Technical Reference Document Database System (TDOCS) Requirements Definition*. San Antonio, TX: Center for Nuclear Waste Regulatory Analyses.
- Khoshafian, S., et al. 1992. *A Guide to Developing Client/Server SQL Applications*. San Mateo, CA: Morgan Kaufman.
- Meehan, B. 1993. Addition of Two New Subtasks Under Center Operations Element, Task 6 of the HLW FY93/94 Operations Plan Under Contract No. NRC-02-88-005. Correspondence to W. C. Patrick, President, CNWRA. Washington, DC: Nuclear Regulatory Commission.
- Rogers, U. 1990. *UNIX Database Management Systems*. Englewood Cliffs, NJ: Yourdon Press, Prentice Hall.
- Sun. 1991. *Sun Catalyst: A Catalog of International Third-Party SPARCware Solutions*. Mountain View, CA: Sun Microsystems Computer Corporation.