TECHNICAL REFERENCE DOCUMENT DATABASE SYSTEM (TDOCS) DESIGN PLAN

Prepared for

Nuclear Regulatory Commission
Contract NRC-02-88-005

Prepared by

Christopher J. Moehle
James L. Patty
Rawley D. Johnson

Center for Nuclear Waste Regulatory Analyses
San Antonio, Texas

January 1994

CONTENTS

CONTENTS (CONT'D)

CONTENTS (CONT'D)

FIGURES

TABLES

ABBREVIATIONS

ACRS	Advanced Computer Review System
AUTOS	Agency Upgrade of Technology for Office Systems
CDS	Compliance Determination Strategy
CDM	Compliance Determination Method
CNWRA	Center for Nuclear Waste Regulatory Analyses
CREV	Corrosion Review Database
DBA	Database Administrator
DHLWM	Division of High-Level Waste Management
DOE	Department of Energy
EPS	Encapsulated Postscript
FIPS	Federal Information Processing Standards
GEOX	Geochemistry and Hydrology Index
GUI	Graphical User Interface
HLW	High-Level [Radioactive] Waste
IMS	Information Management Systems
IRIS	Improved Records Information System
KBPS	Kilobits Per Second
LAN	Local Area Network
LSS	Licensing Support System
NFS	Network File System
NIST	National Institute of Science and Technology
NRC	Nuclear Regulatory Commission
NUDOCS	Nuclear Document System
OCR	Optical Character Recognition
OITS	Open Item Tracking System
PASS/PADB	Program Architecture Support System/Program Architecture Database
QA	Quality Assurance
RDBMS	Relational Database Management System
RPC	Remote Procedure Call

ABBREVIATIONS (CONT'D)

RPD	Regulatory Program Database
SRA	Systematic Regulatory Analysis
SQL	Structured Query Language
TDI	Technical Document Index
TCP/IP	Transmission Control Protocol/Internet Protocol
TDOCS	Technical Reference Document Database System
TIFF	Targa Information File Format
USGS	United States Geological Survey
WAN	Wide Area Network

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-02-88-005. The activities reported here were performed on behalf of the NRC Division of High-Level Waste Management (DHLWM). The report is an independent product of the CNWRA and does not necessarily reflect the views or regulatory position of the NRC.

The following trademarks are used in this report:

- Galaxy is a trademark of Visix Software

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

- Open Look is a trademark of Sun Microsystems

  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 trademark of IBM

  IBM Corp.
  Armonk, NY

- Oracle is a trademark of Oracle

  Oracle Corporation
  500 Oracle Parkway
  Redwood Shores, CA 94065

ACKNOWLEDGMENTS (CONT'D)

- 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

EXECUTIVE SUMMARY

This report is part of the implementation of the Advanced Computer Review System (ACRS) (CNWRA, 1992). The ACRS will support two major applications, technical computing and document management, for the Division of High-Level Waste Management (DHLWM). This report provides the design and implementation plan for the Technical Reference Document Database System (TDOCS) for the document management application.

The plan is to take full advantage of the hardware and software used in the development of the Regulatory Program Database (RPD) V1.0 Phase 1 system (DeWispelare et al., 1993a, 1993b). The general requirements for TDOCS and RPD overlap considerably in that they provide for search and retrieval of documents through a graphical user interface (GUI) compatible with multiple platforms in a client/server environment. Common functionality includes:

- Automated document database loading
- Full-text search and retrieval
- WordPerfect document viewing with cut and paste capability
- Document database content and status reporting

Implementation of these functions can employ the same off-the-shelf but customizable software tools, including:

- 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 GUI application development environment from Visix Software, Inc., which supports access to system functions

TDOCS differs from the RPD in that it will contain a much broader set and larger number of High-Level Waste (HLW) program documents, some of which will require scanning, optical character recognition (OCR) and cleanup. Documents from existing Center for Nuclear Waste Regulatory Analyses (CNWRA) and DHLWM indexing systems will be incorporated into TDOCS for improved user access. TDOCS will be accessible by RPD users to support search and retrieval of text and images in preparing RPD records. TDOCS will also provide access to other databases, such as the Nuclear Document System (NUDOCS), for a broad set of users.

TDOCS will be implemented in three phases:

- FY94, Phase 1 — initiation of TDOCS prototype system development at the CNWRA with electronic loading of documents and investigation of scanning, OCR, and cleanup operations

- FY94, Phase 2 — TDOCS production system development at the CNWRA with scanning, OCR, and cleanup operations and user interfaces at both the CNWRA and the DHLWM

- FY95, Phase 3 — TDOCS enhanced and expanded, with increased document loading and interfaces to NUDOCS and other databases

The TDOCS design plan is provided in this report. It calls for three major components:

- A Document Management Server with database and services for loading and accessing documents and administering and maintaining the database at the CNWRA

- Document Processing Clients for defining, submitting, and deleting documents, including scanning, OCR, and cleanup at the CNWRA

- Document Search-and-Retrieval Clients for document search and retrieval, cut and paste, downloading and editing, on-demand request, reporting, and access to other databases at both the CNWRA and the DHLWM

Exploratory work has laid the foundation for TDOCS, but technical refinements, extensions, and issues remain. These include:

- How to support routine and on-demand document loading, including hyperlink access to images

- Ways to configure Topic search and retrieval based on user privileges, document types, etc.

- Defining technical requirements for scanning, OCR, and cleanup hardware and software selection and integration with Topic

- How to identify access to other databases, particularly NUDOCS

In addition, a number of policy matters remain to be resolved. These include:

- Selection of document sets such as the Techncial Document Index (TDI), correspondence, NUDOCS, etc.

- Specification of required and optional header fields

- Formulation of administration and maintenance procedures

- Decisions on remote database access at the DHLWM or dual operations at the CNWRA and the DHLWM

- Delineation of training needs

Some technical issues require further investigation by the CNWRA Information Management Systems (IMS) team before they can be resolved and implemented. Policy matters require resolution by the Advisory Groups recently formed at the CNWRA and DHLWM, in conjunction with the CNWRA IMS team.

As stated in the CNWRA FY94-95 Operations Plan for the DHLWM (CNWRA, 1993), estimates regarding scope, cost, and schedule will be updated during the design phase. Efforts to make TDOCS interfaces seamless, provide for easy access to other document databases, and support imaging functions will be based on identifiable technical alternatives and cost and schedule constraints. These efforts will be fully discussed with the Advisory Groups as work progresses. The strategy is to provide a system capability that meets stated requirements and is currently achievable and beneficial. TDOCS will be designed as a small-scale system, constrained to a relatively low volume of data with immediate but limited benefits. It will interface to the RPD and complement both the NUDOCS revision and the Licensing Support System (LSS) when those systems are implemented, providing support for DHLWM users.

Center for Nuclear Waste Regulatory Analyses

6220 CULEBRA ROAD • P.O. DRAWER 28510 • SAN ANTONIO, TEXAS, U.S.A. 78228-0510
(210) 522-5160 • FAX (210) 522-5155

TECHNICAL REFERENCE DOCUMENT DATABASE SYSTEM (TDOCS) DESIGN PLAN

Prepared for

Nuclear Regulatory Commission
Contract NRC-02-88-005

Prepared by

Center for Nuclear Waste Regulatory Analyses
San Antonio, Texas

January 1994

logo

Washington Office • Crystal Gateway One, Suite 1102 • 1235 Jefferson Davis Hwy. • Arlington, Virginia, 22202-3293

1 INTRODUCTION

1.1 BACKGROUND

Following the U.S. Department of Energy's (DOE) submittal of a license application for a geologic repository for high-level waste (HLW) to the U.S. Nuclear Regulatory Commission (NRC), the Division of High-Level Waste Management (DHLWM) has a statutory requirement to make a construction authorization decision within 3 years. This requirement will place considerable demands on the DHLWM staff as it conducts its prelicense application and license application reviews.

The NRC Overall Review Strategy for the NRC HLW Repository Program identifies assumptions and strategies that suggest the need for enhanced computer capabilities and tools to support prelicense application reviews (Youngblood, 1993). These include:

- There will be early availability of preliminary information developed and documented by the DOE during the prelicense application phase
- NRC will be able to use the results of prelicense application reviews and supporting investigations
- It is possible to streamline the acceptance review process, resulting in compliance reviews that focus less on detailed supporting information and methodologies and more on how the detailed information was used to demonstrate compliance
- It will be possible to use compliance reviews to verify the acceptability of the DOE compliance demonstrations
- In support of reviews concerns will be documented and tracked as open items.
- The NRC will develop computer models and codes as an ongoing activity in support of iterative performance assessment

These assumptions and strategies drive the two basic needs that the DHLWM Advanced Computer Review System (ACRS) is intended to address, namely enhanced technical computing and document management capabilities.

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

As a follow-up to the requirements definition report, CNWRA representatives met with DHLWM management and staff on September 8-9, 1993. Discussion at these meetings reviewed matters of policy raised during requirements definition and recommended an approach to decision-making for TDOCS design and implementation. Discussion also clarified that TDOCS would not become an official repository of documents since that is the purpose of the Nuclear Document System (NUDOCS) for the agency and will be the purpose of the Licensing Support System (LSS) for the HLW program. NUDOCS is currently being upgraded to manage full text and images (Johnson and Moehle, 1993), and a modified approach has been proposed for DOE to develop, implement, and operate the LSS as a part of INFOStreams (Chilk, 1993). These meetings also led to the formation of two Advisory Groups, one at the DHLWM and the other at the CNWRA, to advise the CNWRA IMS team on design decisions and to provide feedback on implementation. One purpose of this design plan report is to provide the Advisory Groups with a set of current issues and involve them in decision-making as early as possible.

1.2 OVERVIEW OF DESIGN PLAN

The main purpose of this report is to provide a design plan, including the specification of an initial design and a timetable of implementation phases and tasks. Remaining milestones for FY94 are system implementation and preparation of a users' guide.

An iterative approach will be taken to the design and implementation of TDOCS. Design and implementation will be carried out in phases, each phase focusing on exploring particular issues and feeding back and building on previous phases. This approach differs from the standard computer system engineering approach. In that approach, tools, methodologies, and requirements are known in advance, and thus a design specification can precede an implementation specification, both of which can precede development of the system. TDOCS, however, is not a system with completely determined requirements, well-known tools, or specified methodologies. Rather, it will be flexible and integrative. The immediate need of the DHLWM for a document management system is not the only motivating factor for adopting an iterative approach. This approach allows end-users to participate in decision-making during design and to provide feedback during implementation, assuring a closer match between end product and requirements.

This report contains sections that review requirements, outline the plan for design and implementation, describe exploratory work, specify a prototype system design, and discuss production system issues. TDOCS requirements, the goals for design and implementation, are reviewed in Section 2. The iterative design plan and implementation phases are outlined in Section 3. Exploratory work in document management conducted in conjunction with the design and implementation of the RPD is described in Section 4. The prototype system design to be implemented during Phase 1 is specified in Section 5. Open technical issues and policy matters concerning the design and implementation of the production system are discussed in Section 6. Finally, action items for the IMS team and for the Advisory Groups are summarized in Section 7.

2 REQUIREMENTS

This section summarizes requirements for TDOCS based on the *Technical Reference Document Database System (TDOCS) Requirements Definition* (Johnson et al., 1993) report delivered in August 1993 and on follow-up discussions with the DHLWM. The requirements definition report considered current and evolving needs to load, access, use, and manage documents in a technical reference document database and drew on efforts to define requirements, design, and implement other document management systems on the part of the DOE, the NRC, and the CNWRA.

This section identifies requirements in terms of functions, constraints, and policies that are applicable to the TDOCS system. The major requirements to be discussed are depicted in Figure 2-1. These consist of:

- Document processing: structuring documents for use in the database
- Database loading: getting documents into the database
- Search and retrieval: accessing the documents
- Document manipulation: facilitating document use
- Administration and maintenance: assuring system functionality

The design strategy is to provide a system capability that meets these requirements, and that is currently achievable and immediately beneficial. At this point, some TDOCS requirements are still open-ended. Many requirements must be defined in consideration of alternative approaches and actions. Some are technical issues that the IMS team will investigate. Others will be resolved as policies, procedures, and responsibilities which are determined cooperatively among the CNWRA IMS team and Advisory Groups at the CNWRA and the DHLWM.

2.1 MAJOR SYSTEM FUNCTIONS

2.1.1 Document Processing

Routine, on-demand, and electronic loading of technical references will require document processing. The processing functions associated with database loading are scanning, OCR, cleanup, bibliographic header entry, and full-text indexing.

2.1.1.1 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. This is a semi-automated process that is subject to errors

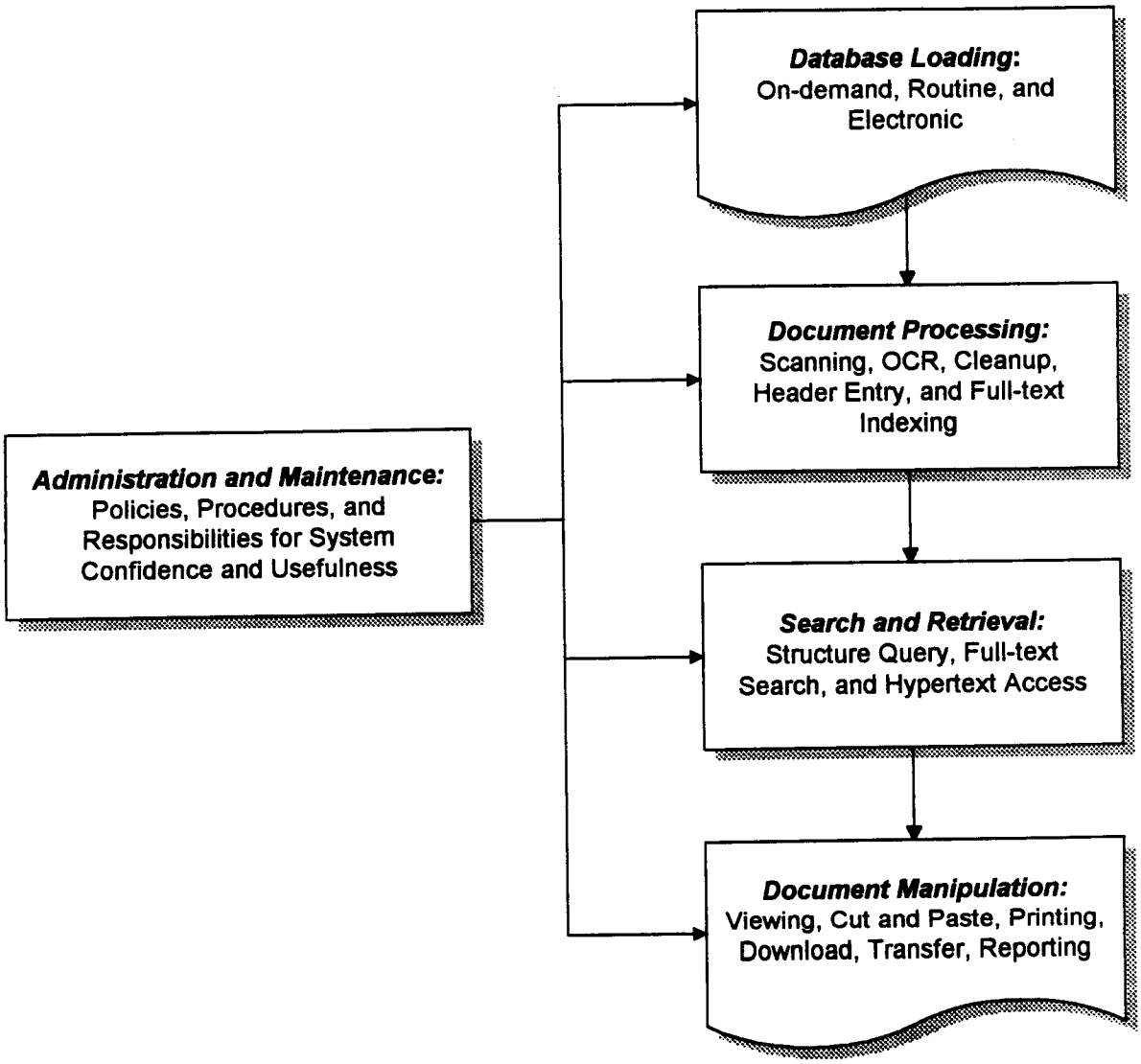

Figure 2-1. Major TDOCS requirements

that may arise from the condition of the documents, the nature of the technology, 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.

2.1.1.2 Bibliographic Header Entry

Bibliographic headers are required as part of the document processing operation to permit searching for documents by author, title, journal name, volume, issue, date, and other standard fields. Header information should be entered by the operator at the time of loading. Nontextual documents such as geologic maps, photographs, or frames captured from videotape require more descriptive headers, as this is the only method for identifying and describing them. A specialized entry screen specific to the document type should be used for the operator to "fill in the blanks" to complete the header.

2.1.1.3 Full-Text Indexing

Full-text indexing is an integral part of document processing, as users of TDOCS will rely heavily on full-text searches to find desired materials. The indexing of the documents should be accomplished automatically as part of the loading process. Header information should also be captured so that documents may be retrieved through structured header queries. Headers for material that consists solely of images should be processed, formatted, and loaded into the full-text repository for search and retrieval through the full-text system.

2.1.2 Database Loading

Selected documents, rather than the entire document backlog, should be added to TDOCS by means of routine loading, on-demand loading, or electronic downloading of technical references. These database loading requirements, along with their associated document sources, are depicted in Figure 2–2.

2.1.2.1 Routine Loading

Routine loading will be the primary method for entering materials into TDOCS. Requests for routine loading will identify groups of documents that will be obtained and loaded on a periodic or cyclical basis. For example, technical documents received by the DHLWM from the DOE would be submitted for routine loading. Similarly, certain technical journals may be identified for routine loading. The routine loading activities should be designed to handle a substantial document volume.

2.1.2.2 On-Demand Loading

On-demand loading will support immediate staff needs for limited amounts of textual and/or graphical materials. A typical request for on-demand loading would arise when a staff member needs several pages from a report or journal article to incorporate into another document. The required pages could be scanned by the technical staff or given to a secretary for on-demand loading. In this way, staff would be able to load a small number of pages of relevant materials, so that these materials could become immediately available on the system. It is expected that the volume of materials loaded in this manner will be rather small.

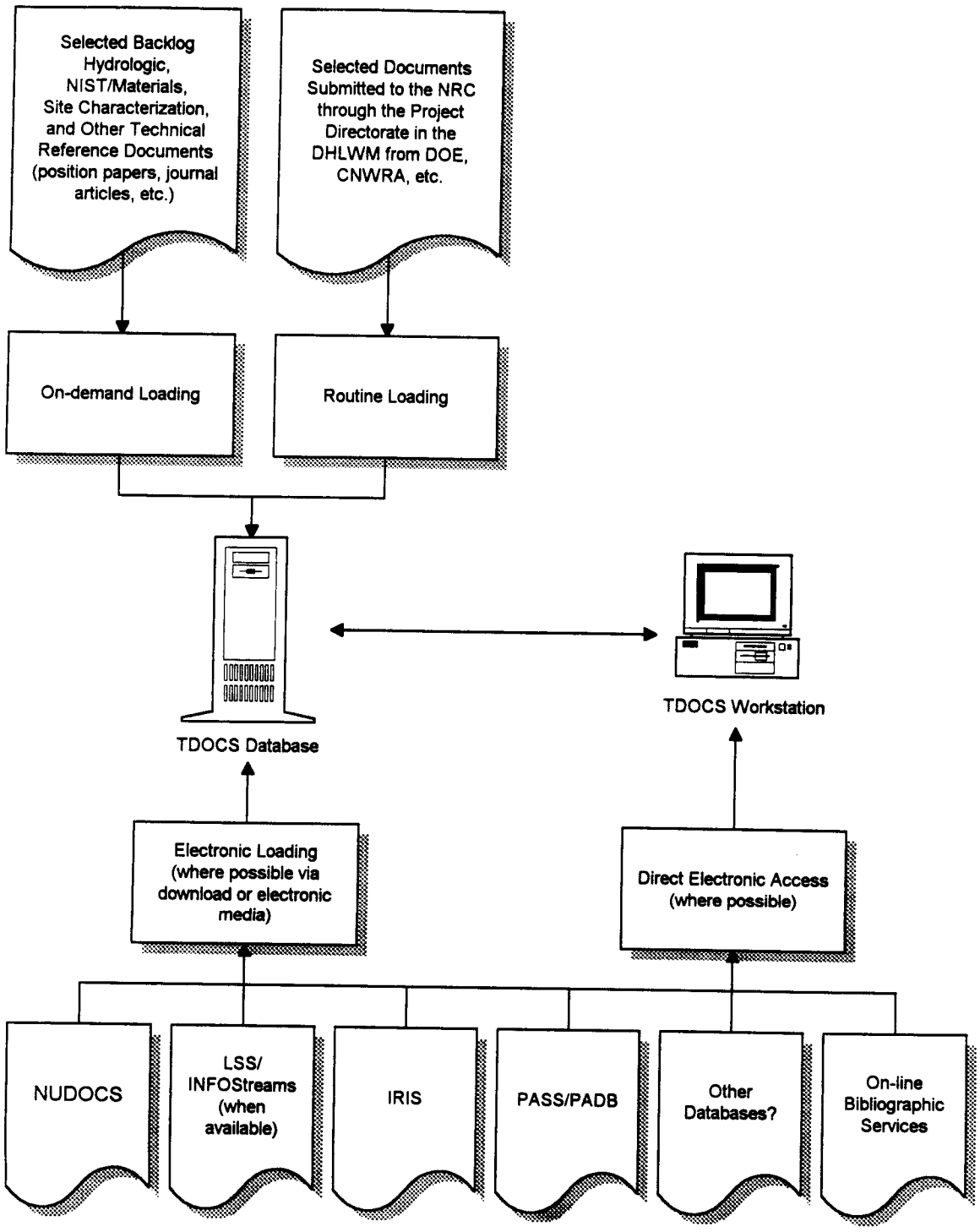

Figure 2-2. TDOCS document sources

As soon as the on-demand scanning process is completed, the scanned material should be made available to the staff. However, the materials scanned on an on-demand basis should not be loaded into the permanent repository, because they would generally represent partial or incomplete documents. Therefore, on-demand loading should always be coupled with a concurrent request for routine loading of the full document from which the on-demand pages were selected.

2.1.2.3 Electronic Loading

Most materials to be loaded into the system should be available in electronic form. This is particularly true for materials generated internally by the DHLWM and the CNWRA. Such materials are anticipated to be available as full text with embedded or accompanying bitmapped files of figures, equations, images, etc. 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 in a variety of electronic media including magnetic tape, diskettes, optical disks, or via communications facilities such as other databases or on-line bibliographical services.

2.1.3 Search and Retrieval

Once relevant documents have been processed and loaded, the staff must be able to find material quickly and reliably. A broad range of functions is required to facilitate search and retrieval of materials stored in TDOCS.

2.1.3.1 Document Access

In order to ensure confidence in retrieval, TDOCS should support three types of document access: full-text search, structured header queries, and hyperlinks between documents. 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 where some information is known about the document or document set. Hyperlinks allow users to establish and use electronic relationships between documents and/or related elements within a document.

2.1.3.2 Search Confidence

User confidence is an important factor in determining the success of document databases and search-and-retrieval systems. The system must ensure that all documents relevant to the task at hand, and only those documents, are found. In order to achieve a high level of user confidence, TDOCS should provide a variety of query and search techniques, such as wildcards, Boolean operators, near-spell searches, phrase searches, proximity searches, cross-partition searches, and result ranking.

2.1.3.3 Concept-Based Search

Concept-based searches provide a means of defining the terms associated with a particular concept in a way that causes the search facility to automatically find documents containing any of the associated terms. A concept-based search can be constructed from various combinations of the techniques mentioned in section 2.1.3.2 on Search Confidence; examples are provided in Section 5.4.

2.1.3.4 Query Save, Recall, and Edit

Search and query are seen as iterative processes. A means of supporting user confidence is through the saving, recalling, and editing of search and query criteria. Staff should be able to progressively refine criteria until they are satisfied that the resulting set of documents represents what they are looking for. Staff should be able to save search and query criteria for later recall and reuse.

2.1.3.5 Search Result Browsing

Search result browsing is another means of supporting user confidence. The results of a structured header query or full-text search should be displayed as a list of documents found so staff can select specific documents for viewing in a separate window. Text search facilities should be provided to enable the user to find information within the displayed document.

2.1.3.6 Concurrent, Multiple Document Viewing and Scrolling

Concurrent, multiple document viewing would be useful for comparing multiple documents or extracting materials from multiple documents to support analyses. Staff should be able to select multiple documents from a search result list and view them concurrently in separate windows.

2.1.3.7 In-Document Match Highlight and Browsing

In-document match highlighting and search capabilities can also support user confidence. When a document is selected for viewing from a search list, the search terms matched in the document text should be highlighted. Facilities should also be provided to permit the user to move rapidly to successive highlighted terms.

2.1.3.8 Hyperlink Creation

Providing for hyperlinks allows the staff to create their own links between documents. This capability could be even more important as staff become familiar with the documents contained in TDOCS and begin to see associations among them.

2.1.4 Document Manipulation

Once documents have been found, there are a number of document manipulation functions that need to be performed. These include the ability to cut and paste from one document to another, document printing, report generation, document download and transfer via e-mail, and image manipulation and enhancement.

2.1.5 Administration and Maintenance

For the purpose of system administration and maintenance, TDOCS should implement a standard set of tools. System tools should support password access, user privileges, user accounts, document tracking, configuration control, monitoring, reporting, and backup and recovery. In addition, a set of policies, procedures, and responsibilities should be specified as guidelines for system administration and maintenance.

2.2 SYSTEM CONSTRAINTS

Various system constraints are imposed by the Agency Upgrade of Technology for Office Systems (AUTOS) and ACRS systems in use by the DHLWM, by the RPD and by NRC policies. For the most part, these constraints do not minimize but rather maximize the potential benefits of TDOCS. These constraints are discussed in the following sections.

2.2.1 Graphical User Interface

TDOCS must be implemented using graphical user interfaces (GUI) for all workstation platforms. These include Microsoft Windows, IBM OS/2, Sun OpenLook, and Apple System 7. These interfaces must support menus, buttons, selection lists, multiple windows, clipboards, and dialogue boxes for user interaction and feedback.

2.2.2 Acceptable Response Times

TDOCS must be implemented to provide reasonable response times for document query and search. In order to support this requirement, a full-text database should be partitioned for selective search, nonsyntactic queries and searches should be detected, search and retrieval progress feedback should be displayed, and query and search cancellation should be provided. In addition, every effort will be made to be responsive to the user. During development, the IMS team will work closely with the Advisory Groups who will have access to the system. Following delivery, one of the tasks of the system administrator should be to not only continually monitor and tune the system but also respond to user complaints.

2.2.3 Multiple Platforms

TDOCS must be designed and implemented to support multiple platforms. These include Microsoft Windows and Sun Workstations running OpenLook or Motif at the DHLWM and personal computers running IBM's OS/2, Macintoshes running System 7 and Sun Workstation running OpenLook or Motif at the CNWRA.

2.2.4 Client/Server Architecture

TDOCS must be designed and implemented using a client/server architecture. Documents, indexes, and headers will reside in a server's file system and database. They will be accessed for distributed processing by staff at client workstations.

2.2.5 Maximize Use of Commercially Available, Off-the-Shelf Software

TDOCS implementation must maximize use of commercially available, off-the-shelf software packages where prudent in design. Packages should be selected based on their support for major functional areas, extensibility in software customization, adherence to industry-wide standards, provision for a suitable application program interface, and support for multiple platforms in a client/server architecture.

2.2.6 Impact on Existing/Planned Systems and Configurations

The impact of TDOCS on existing and planned systems and configurations, namely AUTOS and ACRS, must be minimized. A number of tradeoff decisions may require additional or enhanced hardware for data storage, high resolution display of images, and additional communication lines.

2.2.7 Expandability to Meet Evolving Needs

TDOCS must be designed and implemented to ensure expandability to meet evolving needs. Expansion is expected for access to other databases and on-line services, and eventually for coexistence with NUDOCS and LSS.

2.2.8 Meeting Policies and Standards

TDOCS must adhere to NRC policies and standards for software development as laid out in the NRC Software Quality Assurance Program and Guidelines (NUREG/BR-0167). Implementation must follow specified life-cycle activities.

2.3 SYSTEM POLICIES

For efficient design and implementation, various policy matters need to be established and/or clarified. Many of these matters relate directly to maximizing user confidence. These include specification of:

- A selected set of documents
- Storage and use of proprietary and copyrighted materials
- Loading procedures and responsibilities
- Retention of images and textual materials
- Use of bibliographic headers
- System administration, maintenance, and training
- DHLWM/CNWRA TDOCS database synchronization

These policy matters will be addressed through CNWRA recommendations and discussion with TDOCS Advisory Groups during system design. Implementation of such policies will be subject to NRC management concurrence.

3 DESIGN PLAN

3.1 TASKS AND SCHEDULE

The tasks and schedule for completing the TDOCS design and implementation during FY94 and FY95 are shown in Table 3-1. At least three phases will be needed for full system implementation.

3.1.1 Phase 1, FY94 Prototype System

The TDOCS prototype system implementation phase (Phase 1) in FY94 will be based on RPD implementation in FY93 and the more recent exploratory work done to lay the foundation for TDOCS. Following RPD implementation, a Verity Corp. consultant was contracted to review the current customization of Topic and make recommendations concerning the integration of document scanning, OCR, and cleanup with document loading; the hyperlink launching of document and image viewers; and end-user training. Another consultant was contracted to provide additional training in Galaxy and review of the current user interfaces. These functions and tools, with refinements and extensions, will transfer to the design and implementation of TDOCS. The concluding work for Phase 1 will be delivery of the TDOCS design plan and the first meeting of the IMS staff with the Advisory Group at the DHLWM. The meeting agenda will include (i) comments from the DHLWM on the design plan report, (ii) formulation of specific tasks for the Advisory Group with emphasis on selection of document sets and specification of headers, and (iii) demonstration of the prototype.

3.1.2 Phase 2, FY94 Production System

The TDOCS production system implementation phase (Phase 2) in FY94 will incorporate scanning with automated indexing, and installation of document search-and-retrieval clients at the DHLWM. A second meeting will be held with the DHLWM Advisory Group to adopt policy, procedures, and operating plans for the initial production system. Phase 2 will be concluded with implementation of the production system and delivery of the *TDOCS Users' Guide*.

3.1.3 Phase 3, FY95 Enhancements and Expansions

TDOCS enhancements and expansions with full document loading will continue in FY95. Based on feedback from experience with TDOCS by the DHLWM Advisory Group, revisions will be made to the TDOCS production system. A more complete document loading plan will be developed. The need for parallel database operations at both the DHLWM and the CNWRA will be reviewed, and, if necessary, plans will be made to install the TDOCS database at DHLWM. Specifications for a TDOCS interface to NUDOCS and other databases will be developed. Dialogue with the DHLWM Advisory Group and feedback from them will be utilized to support revisions to enhance and expand TDOCS.

Table 3-1.TDOCS tasks and schedule

Completion Date	Task Description
Phase 1, FY94 Prototype System	
1/28/94*	Deliver TDOCS Design Plan (IM 5702-155-401).
2/15/94	Load electronic documents and headers (Technical Document Index (TDI), Correspondence and Quality Assurance (QA)) and develop document search and retrieval clients at the CNWRA.
2/15/94	Test remote 56 kbps access at the DHLWM.
2/15/94	First Advisory Group Design Meeting, agenda—selection of document sets and types and specification of headers.
Phase 2, FY94 Production System	
3/15/94	Investigate scanning, OCR, and cleanup operations and analyze Automated Document Input software from Verity Corp.
4/15/94	Revise prototype for production to include scanning and automated document indexing.
4/15/94	Second Advisory Group Meeting, agenda—to adopt policy, procedure, and operations plan for initial TDOCS production system.
5/15/94	Revise production system to meet requirements from second meeting and include scanning.
6/30/94*	Develop *TDOCS Users' Guide*
6/30/94*	Install TDOCS clients at DHLWM and provide training on TDOCS at DHLWM
Phase 3, FY95 TDOCS Enhanced and Expanded with Document Loading	
10/15/94	Review remote access and decide on separate or parallel TDOCS operations.
11/15/94	Preparation of document loading plan and procedures.
11/15/94	Revise TDOCS production system based on user feedback from Advisory Group.
4/01/95	Specify interface to NUDOCS, and develop and implement.
2/05/95	If parallel TDOCS operation is needed, install database at DHLWM.
9/15/95	Specify interface to other databases, and develop and implement.

* Deliverables in Operation Plans for the Division of High-Level Waste Management for FY94-95, Revision 4, Change 2.

4 EXPLORATORY WORK

Exploratory work on TDOCS began in conjunction with the design and implementation of the RPD, Version 1.0, of which Phase 1 has been recently completed. While RPD and TDOCS differ in many respects, they overlap significantly in general requirements to provide for search and retrieval through a GUI that is compatible with multiple platforms in a client/server environment. This section compares TDOCS and RPD for common functions and tools, describes the work that has been done for RPD, and discusses revisions for the TDOCS prototype system.

4.1 COMMON FUNCTIONS AND TOOLS

As illustrated in Figure 4-1, TDOCS requirements for scanning, OCR, and cleanup, and access to other databases differ from RPD requirements. Similarly, TDOCS has no requirements for parsing and report generation as does RPD. Nevertheless, the two systems share similar core functions, such as:

- GUI to document search, definition, checkin, deletion, and reporting
- Network file system (NFS) and remote procedure call (RPC) client/server interface
- Configuration control and reporting
- Document loading, indexing, hyperlinking, security, search, retrieval, and launch

Implementation of these functions for TDOCS can employ the same off-the-shelf but customizable software tools, including:

- 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 GUI application development environment from Visix Software, Inc., which allows multiplatforms

These functions and tools, with refinements and extensions, will transfer to the design and implementation of TDOCS. They are discussed in detail in Section 5.

4.2 REGULATORY PROGRAM DATABASE

Development and implementation of the RPD was predicated on the ability to acquire and integrate appropriate, commercially available, off-the-shelf software. This off-the-shelf software included two broad groups of products: (i) strategic software products and (ii) system support software. The strategic software included products to support enhanced system usability through GUIs, full-text search

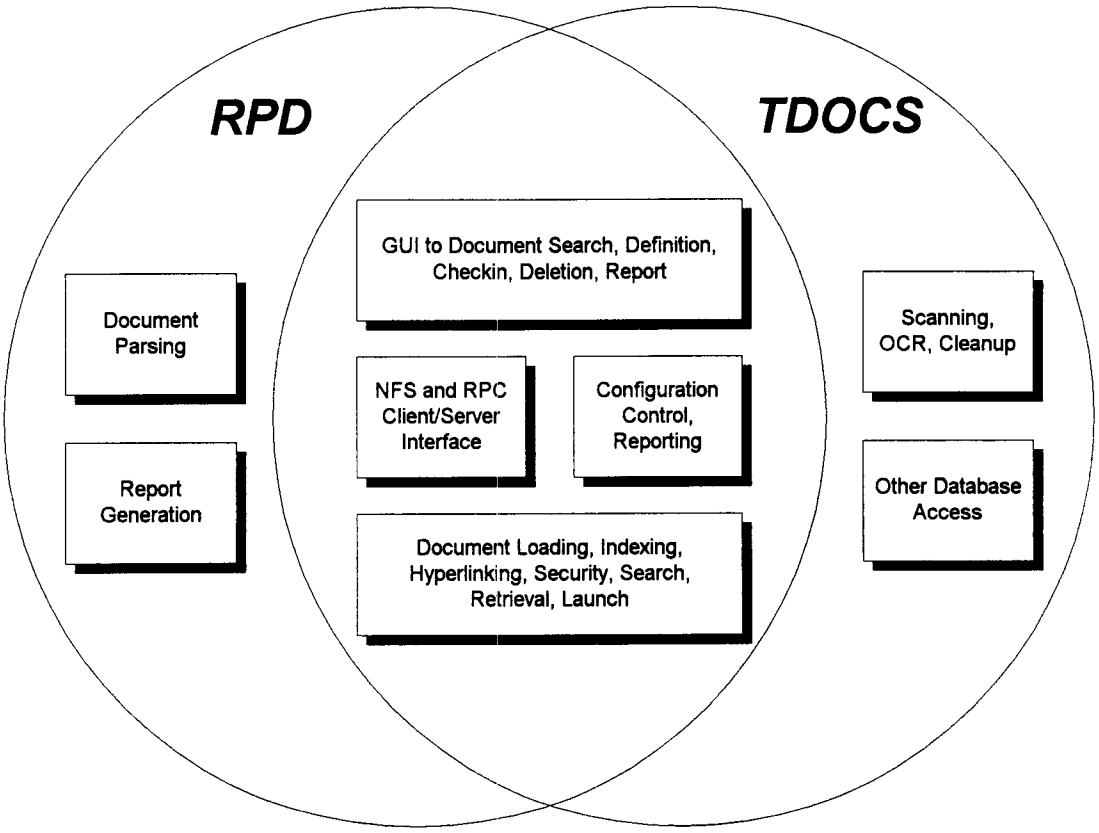

Figure 4-1. Comparison of TDOCS and RPD functions

Figure 4-1. Comparison of TDOCS and RPD functions

and retrieval, and data/system maintenance. The strategic software products selected for RPD implementation have direct applicability to the implementation of TDOCS in that both systems have common core requirements for GUIs across multiple platforms, full-text search and retrieval, and system management (security, configuration control, reporting, etc.). These requirements are discussed in the following sections in terms of the strategic software application packages selected for RPD implementation, and their associated performance evaluation. It is a summary of CNWRA requirement and evaluation (DeWispelare et al., 1993a, 1993b) for RPD, focusing on areas it has in common with TDOCS.

4.2.1 Graphical User Interfaces

RPD development included plans for the design and implementation of GUIs for enhanced system usability. Enhancing system usability is highly supportive of the overlapping RPD and TDOCS requirements for support of a textual data repository. These requirements include:

- Full-text search and retrieval
- Data management
- Word processor access/compatibility
- System response time performance
- Minimized impact on computer support plans
- Compatibility with upgrade initiatives
- Growth potential

The Galaxy software package was selected for this purpose over other candidate software packages. Galaxy is an off-the-shelf package that supports the development of GUI applications that can run on multiple computers and operating systems such as Apple System 7, Sun Openlook and Motif, IBM OS/2, and Microsoft Windows, while providing users with common interface displays.

As part of the evaluation and selection process for candidate GUI software packages, various products were reviewed for applicability. A prototype GUI was implemented to test and verify intended functionality among the multiple RPD platforms and other strategic software packages. The main functions tested were:

- Database table access and control
- Control of colors and fonts
- Access to operating system functions, such as memory allocation

- Platform stability

Galaxy was judged to be better than other candidate GUI software packages for allowing more operating system functionality and more portability for cross-platform development.

The performance of Galaxy to date, even though the system is not fully implemented, has met or exceeded expectations for its ability to integrate with other applications, its portability across supported platforms, and the seamless user interface it provides. Extensive Galaxy knowledge has been gained by IMS staff members as a result of the prototype effort and associated Galaxy training classes.

4.2.2 Full-Text Search and Retrieval

RPD requirements called for a powerful and efficient full-text search capability that would provide a user-friendly method for locating and viewing documents. Considerations for selecting a full-text search and retrieval system included:

- Support for all present and anticipated hardware platforms
- A client/server design
- Fast response time
- Use of logical operators and multiple search terms
- Highlighting of search arguments in the text
- Ability to move between search terms and related documents
- Support for plain text and WordPerfect data formats
- Cut-and-paste operations
- Concurrent viewing of multiple documents
- A user-friendly GUI interface
- Ability to store full-text queries
- Ability to launch database application code
- Hyperlinking
- Concept-based searches

The design and implementation of RPD and TDOCS place great emphasis on the retrieval of documents. Since the ultimate purpose of a text management system is to retrieve information, the retrieval mechanism is critical to the overall effectiveness of the text management system. The best solutions offered by current software technology for effective information retrieval combine full-text and header-based searches with the capability to conduct concept-based searches. A user may define a concept by combining a wide range of queries and search terms based on headers and full-text which can then be saved. The definition of concepts thus permits users to accumulate and share knowledge about how best to retrieve information on specific subjects from the text management system.

The Topic full-text search and retrieval software package was selected for the RPD system after evaluating it against other candidate software packages. As part of the selection process, a survey of products was conducted to identify full-text search software packages suitable for consideration. Evaluations were then performed using test documents to check performance and functionality. Topic was selected due to its robust performance, multiple platform capabilities, concept-based search capabilities, and ability to satisfy the system support requirements addressed previously.

Testing of the Topic software provided extensive hands-on use of the package by IMS staff members. It was determined that the product was well-adapted for handling both general and highly focused searches against large data aggregates and integrated well with the other strategic software packages. Topic performance to date has met or exceeded expectations. As with Galaxy, the IMS staff gained valuable experience with Topic during the evaluation process and attended Topic training classes.

4.2.3 System Management

RPD required a database capability to support the storage of regulatory program records and fields in a relational format. Additional criteria for the database were that it had to interface with standard programming languages (C and FORTRAN), provide *ad hoc* query language capabilities, and support a well-defined application interface to the database software. Testing and evaluation resulted in the selection of Oracle RDBMS Version 6 as the relational database software for RPD implementation. The software was fully compliant with the standard relational database language called Structured Query Language (SQL); supported the open system standard (POSIX), operated on a Unix-based platform; and had superior performance and features that were directly applicable to many required RPD system management functions, such as:

- Storage and retrieval of textual materials
  - Retrieval of materials in multiple sequences
  - Selection and implementation appropriate basic units of storage (document, chapter, paragraph, etc.)

- Maintenance of textual materials
  - Support of unique identifiers for selection of materials (i.e., header-based searches)
  - Ability to insert, change, and delete materials in the text repository
  - History of changes maintenance
  - Access to current versions of a document while a new version is being prepared
  - Support of checkin, checkout, content verification, and configuration control procedures
- Reporting of textual materials
  - Identification and storage of external formatting information to support multiple report formats
  - Parsing, identifying, storing, and retrieval of lower-level textual entities to support variable content reporting
  - Storage and maintenance of intrinsic formatting information as part of textual information (superscripts, subscripts)
- Control of textual materials
  - Access control
  - Update control (insert, modify, delete materials)
  - Version tracking and control

During the relational database evaluation and selection process, Oracle was determined to provide the best range of functionality to support RPD. In addition, it satisfied other features associated with RPD, such as: (i) cost-based optimized implementation; (ii) Federal Information Processing Standards (FIPS) compliancy; (iii) IBM SQL compliant data types; and (iv) row, page, and table locking. Experience to date has shown that Oracle satisfies the required feature set, and has the necessary performance capabilities. A new release of the Oracle RDBMS (Version 7), with additional features and enhancements will be incorporated into RPD in the future.

4.2.4 Implementation

The Phase 1 implementation of RPD has been completed (DeWispelare et al., 1993b). The testing took the form of initial functional evaluations, using specially prepared test data. This was

followed by a full system test using actual regulatory program documents. The basic integration of the software packages has been accomplished. Utilization and integration of these products is viewed by the CNWRA as a successful effort. The in-house experience, training, resources, and application knowledge gained during the process can also be applied to other complex applications such as TDOCS.

4.3 REVIEW AND REVISION

Following exploratory work, a Verity Corp. consultant was contracted to review the current customization of Topic and make recommendations concerning refinements and extensions for TDOCS, particularly the integration of document scanning, OCR, and cleanup with document loading; the hyperlink launching of document and image viewers; and end-user training. Another consultant was contracted to provide additional training in Galaxy and review of the current user interface.

4.3.1 Document Loading Procedures

Document loading procedures have been revised. Exploratory work loaded documents on demand; when a document was defined or checked in, it was immediately loaded into Topic and made available for search and retrieval. This approach has been found to be exceedingly slow and requires that users restart the Topic search-and-retrieval engine each time a document is loaded to synchronize client and server document indexes. Verity has suggested a solution that greatly reduces load times and mitigates but does not eliminate, the synchronization problem (the end-user would still need to restart the client search-and-retrieval engine to see the new document). It is, however, a complicated solution to a complex problem.

It is recommended, instead, that the design of TDOCS take a simpler batch loading approach. For TDOCS, the requirement for on-demand loading involves only paper documents. The output of the scanning, OCR, and cleanup process is an electronic document that could be downloaded directly to the end user requesting it and queued for loading later. At the end of each day, all on-demand and routinely processed documents would be loaded in batch. End users would still have to restart the Topic search-and-retrieval engine daily to synchronize with the database index. This solution is recommended over Verity's because it alleviates the synchronization problem, further reduces document load times, and, more importantly, simplifies rather than complicates design and implementation.

4.3.2 Search and Retrieval Configuration

Most of the exploratory work concentrated on the interfaces to document definition, checkin, and deletion. Configuration of the Topic search-and-retrieval engine was limited to a security, search result list format, and a single-form query interface. Because this is where a majority of the TDOCS users will spend their time, it will be quite productive if a greater effort is expended in customizing the search-and-retrieval engine.

Topic will be configured so that users can search across the entire database or restrict searching within certain document types. For example, users will be able to search for DOE reports, National Institute of Science and Technology (NIST) analyses, or TDI correspondence. In addition, Topic will be

configured so that it can be started with a form query interface appropriate to the document type being searched.

Other configurations are also possible. One of the most powerful aspects of Topic is its ability to perform concept-based searches. Concepts appropriate to HLW and the regulatory program can be added to the Topic database. This type of configuration demands technical and linguistic knowledge. Building relevant concepts will require the combined efforts of the IMS team and the Advisory Groups, as well as experts in the various fields related to the HLW program.

4.3.3 Technical Issues and Policy Matters

Not all technical issues and policy matters can be resolved in this report. Some require further investigation by the CNWRA IMS team before they can be designed and implemented. Policies require resolution and formulation by the Advisory Groups in conjunction with the CNWRA IMS team.

Exploratory work has laid the foundation for TDOCS, but technical issues remain. These include:

- Scanning, OCR, cleanup and hardware and software, and integration with Topic
- Access to other databases, particularly NUDOCS
- Image manipulation and enhancement

In addition, the following policy matters remain to be resolved:

- Selection of document sets
- Specification of required and optional header fields
- Formulation of procedures, policies, and training needs
- Decision on remote database access at the DHLWM or dual operations at the CNWRA and the DHLWM

These technical issues and policy matters are discussed in Section 6.

5 PROTOTYPE SYSTEM DESIGN

Review and revision of exploratory work has led to the prototype system design presented in this section. The design, in accordance with the approach adopted as a design plan, is open-ended, flexible and expandable. Document scanning, OCR, and cleanup, for example, have been given only a general place in the design; precisely how these processes will fit into the design is a matter for investigation during the prototype phase. In other words, it is anticipated that this design will change, and that another cycle of implementation, review, and revision will take place during the prototype phase in order to arrive at a production system design. This section, then, provides an overview of the prototype system and discusses each of its major components.

5.1 DESIGN OVERVIEW

The TDOCS prototype system, illustrated in Figure 5-1, will consist of three major modules:

- Document Management Server
- Document Processing Clients
- Document Search-and-Retrieval Clients

The Document Management Server will consist of two submodules—Database and Services. The Database will serve as the repository for document headers, text, images, and indexes. Even though three software tools—Oracle, Unix, and Topic—will be used to implement this submodule, it will be treated as a single database. This treatment is possible because clients will have access to the Database only through the services of the Document Management Server. Document Processing Clients will submit processed documents to the server for storage and later loading, and Document Search-and-Retrieval Clients will access those documents through the server. In addition, the Services submodule will be used for batch document loading, as well as database administration and maintenance, limited to an authorized database administrator (DBA).

Document Processing Clients will be used to define, submit, and delete documents and to associate documents with images. One of these clients will be connected to a scanner; this client will be used to perform scanning, OCR, and cleanup operations. Document Search-and-Retrieval Clients will be used to perform full-text and header search; document viewing, cut and paste, hyperlinking, and launching; document downloading and editing; on-demand document requests; database reporting; and other database and services access. Both of these clients will be available on multiple platforms, except for those connected to the scanner. Password-protected access to the Document Processing Clients will be limited to authorized custodians, and the Document Search-and-Retrieval Clients will be limited to authorized users.

The design employs a client/server architecture. The Document Management Server will reside on the TDOCS Sun Server. Document Processing and Document Search-and-Retrieval Clients will reside

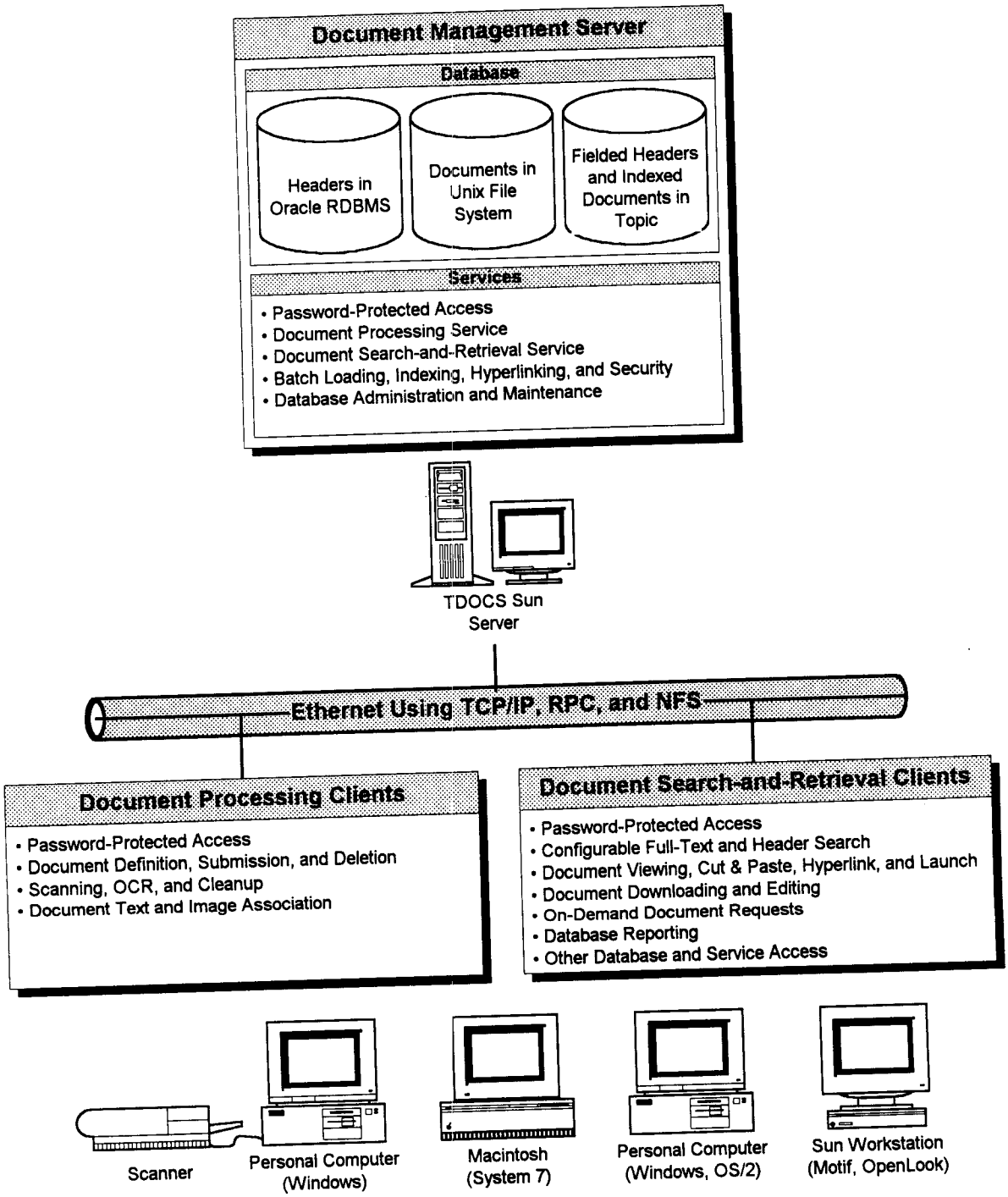

Figure 5-1. TDOCS Prototype System Design

on Sun Workstations using OpenLook or Motif, personal computers using Windows or OS/2, or Macintoshes using System 7. Clients will use Transmission Control Protocol/Internet Protocol (TCP/IP), RPC, and NFS to communicate over the network with the server. RPC will be integrated with TCP/IP and allow a client application to make calls to functions that are part of a server application. NFS allows a client system to access remote file systems. These communication services are available on the AUTOS and ACRS systems.

It should be noted that the logical design of the system presented 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, user privileges determining which portions of it are accessible. Design dictates what must be implemented, not how implementation will be achieved.

5.2 DOCUMENT MANAGEMENT SERVER

The Document Management Server will consist of Database and Services modules. Its Database will serve as the repository for document headers, text, images, and indexes. Its Services will provide all access to the Database, whether that be client-requested document submittal, search and retrieval, server-based document loading; or administration, and maintenance. The Document Management Server will reside on the TDOCS Sun Server. Direct access for administrative purposes will be limited to an authorized DBA. All other access to the Database will be through the Services of the Document Management Server.

5.2.1 Database

The Document Management Server Database will serve as the repository for document headers, text, images, and indexes. Although, three software tools will be used to implement this module; it will be accessed and maintained as a single database. Headers will be stored in Oracle and captured in Topic. Document text and images will be stored in the Unix file system, and the text will be indexed in Topic.

5.2.1.1 The Role of Oracle

The Oracle RDBMS will be used as one of the tools for the Document Management Server Database. Oracle provides data integrity, an interface through the C programming language, and password-protected access. The primary role for Oracle will be to store document headers. These headers will support document loading, configuration control, and database reporting. Oracle may also be used in a secondary role to support downloads from other databases (e.g., from NUDOCS, which also uses the Oracle RDBMS).

Document headers will consist of fields. Some fields will be used internally by the system to uniquely identify documents and to reference their storage in the Unix file system and their indexes in Topic. These fields are listed and explained in Table 5-1.

Table 5-1. System-internal document header fields

Field	Explanation
Document Type	Specifies type of document, e.g., analysis, correspondence, regulation
Document ID	Uniquely identifies a document within a document type
Instance ID	Uniquely identifies a document version for tracking document revisions
Document Status	Specifies a document's status in the system, e.g., defined, submitted, deleted
Submitter	Specifies the name of the person who submitted the document to the system
Submittal Date	Specifies the date the document was submitted to the system
Partition	Specifies the location of the document in the Unix file system and Topic
File Name	Names of the file used to store the document in the Unix file system

Together, the Document Type, Document ID, and Instance ID fields will form the unique primary key to identify a document. All internal header fields will be supplied automatically by the system and kept entirely transparent to the user except for Document Type, which the user will be able to select from a system-generated list of specified document types.

Other header fields will be bibliographic in nature. These fields will be visible, meaningful, and useful to users in identifying and searching for documents. The specification of document types and their associated header fields, a collaborative effort for the IMS team and Advisory Groups, is discussed in Section 6.

Oracle will also serve a role in the administration and maintenance of user accounts. All users will be provided accounts stored in Oracle. These will consist of user name, password, and privileges. A user name must be unique but may be the same as that used on other systems; and users will be able to change passwords. The specification of privileges, however, will be a matter of policy, and is discussed in Section 6.

5.2.1.2 The Role of the Unix File System

The Unix file system will be used as another Document Management Server Database tool. It provides for password-protected access as well as group- and user-level read/write privileges on directories and files. That is, access can be controlled according to privileges assigned to an individual user or the group of users to which he or she belongs. Its primary use will be for the storage of documents. Both Oracle and Topic make extensive use of this system for storage. Password-protected access and read/write privileges will be used to protect documents, headers, and their indexes from unauthorized modification or deletion. The Unix system will also provide many of the utilities necessary for system administration and maintenance.

Text and images will be stored in partitions. These partitions will be formed as a directory structure that parallels the partition structure used to store indexes in Topic. The division of documents into partitions will be based on the various document types stored in TDOCS.

5.2.1.3 The Role of Topic

The Topic database and full-text search-and-retrieval engine will also be used as a Document Management Server Database tool. While Oracle provides for storage of headers and the Unix file system for documents, Topic provides for password-protected indexing, search, and retrieval of those headers and documents. Protected access can be applied at the group and user levels.

The role of Topic in the database will be to store captured headers and indexed text. Documents will be partitioned primarily according to document type and secondarily according to the physical storage size of files in each partition. Partitioning documents in this way will support maintaining security levels and providing header query forms for each type of document.

Topic provides the utilities for fielding headers and indexing text. It also provides the GUI-based software engine 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. These search-and-retrieval capabilities are discussed in Section 5.4.

5.2.2 Services

Document Management Services will provide password-protected access to the Document Management Database. Some of these services will be responsible for handling client requests for document processing and for search and retrieval. Other services include batch document loading, as well as database administration and maintenance.

5.2.2.1 Document Processing Service

Custodians will input document headers, text, and/or images through Document Processing Clients (see Section 5.3). These clients will request document loading or deletion through the Document Processing Service of the server. For each document processing service request, the server will receive an RPC data package or packages. Depending on the nature of the submittal, the package or packages may contain a document header; a header and an image; or a header, text, and images. In other words, it will be possible to submit headers without text or images, headers for images, and headers along with text and images.

This service will process requests in several steps. First, the RPC data package or packages will be checked to see that header, text, and images have been properly received. Second, the header will be checked for correctness and completeness and, against the Oracle database, for nonduplication. Third, using user-supplied header information, internal header fields will be completed, and the status will be marked for later batch loading of the document. Fourth, the header will be stored in Oracle. Finally, using internal header information, if a document is submitted, text and/or images will be stored in the Unix file system. In the case of requests for document deletion, the current header status will be so

marked. If any errors occur, an error flag and message will be returned to the client requesting the service; otherwise, a success flag will be returned.

The Document Processing Service does not immediately load documents in Topic. This is another service performed by the Document Management Server in a batch process (see Section 5.2.2.3).

5.2.2.2 Document Search-and-Retrieval Services

Most users will use Document Search-and-Retrieval Clients (see Section 5.4). These clients will request the Document Search-and-Retrieval Services of the server. These services will involve, for the most part, customization of the tools being used to implement the server.

Requests for full-text search, headers queries, document viewing, cut and paste, hyperlinking, and WordPerfect and image viewer launching will be made through the Topic search engine. Document downloading and editing will be made through RPC and/or NFS. On-demand requests will be made through e-mail. Database reports will be drawn from Oracle, and transfered via RPC. These search-and-retrieval processes are discussed in Section 5.4.

5.2.2.3 Batch Loading, Indexing, Hyperlinking, and Security

Documents processed and submitted during the day for loading or deleting will be loaded or deleted overnight in a batch process by the Document Management Server. As discussed in Section 4.3, batch processing will significantly reduce load times for Document Processing Clients and alleviate the need for Document Search-and-Retrieval Clients to restart the Topic other than once each morning.

Batch processing will take several steps. First, on a partition-by-partition basis, the batch process will fetch from Oracle every header whose status is marked for loading. Second, using the headers to reference document text stored in the Unix file system, headers will be fielded and text will be indexed in Topic. Third, again using header references, document text will be hyperlinked to document images. Fourth, the documents will be secured for authorized access. Finally, the status in document headers will be marked active.

Batch processing will also fetch from Oracle all headers marked for deletion. Whether these documents will be actually deleted from Oracle, the Unix file system, and Topic or archived will depend on policy decisions about document types (see Section 6).

Logs will be generated during these batch processes and made available in the form of reports. In this way, custodians using Document Processing Clients will be able to check on the success of document loading and take corrective action in the case of failure.

5.2.2.4 Database Administration and Maintenance

The Document Management Server will also provide Database Administration and Maintenance Services. These services will be accessible only to an authorized DBA. One set of services will support the DBA in adding and dropping custodians and users and modifying their privileges and accounts. The other set of services will support the DBA in periodic backup of the Document Management Database.

5.3 DOCUMENT PROCESSING CLIENTS

Document Processing Clients will be used to define, submit, and delete documents, as well as to perform related operations involving the scanning, OCR, and cleanup of documents and the association of documents with images. At least one of these clients will be connected to a scanner and will be used to perform scanning, OCR, and cleanup operations. Document Processing Clients will have easy-to-use GUIs available on multiple platforms except, possibly, any client connected to the scanner.

5.3.1 Password-Protected Access

Access to the Document Processing Clients will be password protected and limited to authorized custodians. Custodians will be those who are charged with the responsibility of managing documents in TDOCS. They will be drawn from the clerical staff and interested technical staff at the CNWRA and the DHLWM. Each custodian will likely be responsible for a type or set of documents. The selection of custodians and the delineation of their responsibilities will be matters for policy decision (discussed in Section 6).

5.3.2 Document Definition, Submission, and Deletion

The main operations performed by Document Processing Clients will be document definition, submission, and deletion. These operations will be geared toward the processing of electronic documents with embedded images in WordPerfect format or plain text formats. In the case of paper documents, these operations will be supported by scanning, OCR, and cleanup operations; and, in the case of documents with referenced images stored in separate files, by document text and image association operations. All of these operations will be performed by custodians.

Document definition will be a prerequisite to the submission of a document to the system. This will be true whether only a header is submitted; a header and text; or a header, text, and images. Document definition will provide a means for user identification of documents and the data for structured header queries as well as database reports. Thus, documents will be defined by header information that is meaningful and useful to users, such as title, author, date, accession number, and other identifying information. The specific fields required for document definition will depend on the type of document submitted. The selection of document types and the specification of header fields are matters for policy decision (discussed in Section 6).

During the document definition step, the custodian will select a document type and then fill in those fields appropriate to the selected document type. The custodian will then use a file chooser to identify the text file to be submitted. In those cases where only a header will be submitted, file selection will not be necessary. In the case of paper documents scanning, OCR, and cleanup will need to be performed to generate files for selection. In the case of documents with associated image files, document text and image association will need to be performed to select the image files.

Document submission will follow document definition. In this step, when the custodian submits a document, the Document Processing Client will transfer its header, text, and images in a package or packages by means of RPC to the Document Management Server. As described earlier, the server's Document Processing Service will attempt to store the document in Oracle and the Unix file system and queue it for later batch loading into Topic. If it is successful, it will return a success flag; otherwise, it will return an error flag and message, in which case the custodian must take corrective action.

Document deletion will also be performed by custodians. They will be able to select and identify documents for deletion by header fields (such as title). The Document Processing Client will then transfer the document-identifying information in a package by means of RPC to the Document Management Server. As described earlier, the server's Document Processing Service will attempt to locate the document header and mark it for later batch deletion. If it is successful, it will return a success flag; otherwise, it will return an error flag and message, in which case the custodian must take corrective action.

The GUI implemented to facilitate the interactive document definition, submission, and deletion will also facilitate the batch processing of sets of documents. Batch processing will be particularly useful in the case of database downloads or in the case of work contracted out for scanning, OCR, and cleanup. What will be required the preparation of a plain text file listing the same sort of document headers and file references that would normally be defined interactively. The list could be prepared by programatically extracting fields from the documents, provided they are consistently formatted; exporting records from another database; or including it as a requirement in a contract for batch scanning, OCR, and cleanup. The custodian will simply use a file chooser to select the listing. Once selected, the Document Processing Client will read the listing and submit the documents one by one to the Document Processing Service of the server.

5.3.3 Scanning, Optical Character Recognition, and Cleanup

Scanning, OCR, and cleanup operations will support the processing of paper documents. These operations will be performed at Document Processing Clients that have access to a scanner. These clients will be shared by custodians who process paper documents.

Paper documents must, of course, be defined by entering header information. Prior to submission to the server, however, they must be converted to electronic format. The conversion process will involve scanning the paper document into an electronic image format; performing OCR on the image, converting it to text and extracting images; and correcting errors, formatting, and, in general, cleaning up the resulting document. It is anticipated that the resulting electronic document will consist of text in WordPerfect format and accompanying images. If this is the case, then the images will next be associated with the text and the document submitted.

While scanning is largely an automatic process, it must be set up by an operator; OCR and cleanup will also require operator assistance. The incorporation of these operations in the document processing scheme, and the selection of hardware and software to support them, are issues scheduled for technical investigation during the prototype phase of TDOCS design and implementation. These issues are discussed further in Section 6.

5.3.4 Document Text and Image Association

Document text and image association will support the processing of those documents whose images are not embedded in the text. Such association will utilize hyperlinking from document text to referenced images. Thus, this operation will be performed by custodians through Document Processing Clients.

The process will require associating references to images in the text of documents and marking them as "launch pads" with references to image file "landing pads." When the document is submitted for loading, Topic utilities will be used to set up hyperlinks between launch and landing pads. Then, when the document is viewed in Topic, the hyperlinks will appear as icons that can be selected to launch the image files into an image viewer.

Like scanning, OCR, and cleanup, text and image association will require operator assistance. Some of this process, specifically the selection of image file formats and image viewers, will be the subject of technical investigation during the prototype phase (see discussion in Section 6).

5.4 DOCUMENT SEARCH-AND-RETRIEVAL CLIENTS

Document Search-and-Retrieval Clients will be used to perform full-text and header search; document viewing, cut and paste, hyperlinking, and launching; document downloading and editing; on-demand document requests; database reporting; and other database and services access. These clients will be available on multiple platforms.

5.4.1 Password-Protected Access

Access to the Document Search-and-Retrieval Clients will be password protected and thus limited to authorized users. Users will be those authorized to search for and retrieve documents in TDOCS. Most likely, users will be restricted in terms of what document sets they are privileged to access. They will be members of the technical staff at the DHLWM and the CNWRA. The selection of users and the delineation of their privileges will be matters for policy decision (discussed in Section 6).

5.4.2 Configurable Full-Text and Header Search

Users will be able to perform full-text and header searches through the GUI of the Topic search-and-retrieval engine. Topic supports:

- Wildcards and Boolean operators: the capability to combine search criteria with logical operators such as AND, OR, and NOT to create more specific queries

- Near spell search: the capability to find words regardless of syntactic variety (number, tense, and other affixes)

- Fuzzy search: the capability to find words even though they may be misspelled in search criteria or in the documents themselves

- Phrase search: the capability to find not just words but phrases
- Proximity search: the capability to find combinations of words and to specify the distance between those words in a document (e.g., phrase, sentence, paragraph, etc.)
- Cross-partition search: the capability to search for documents of specified type or category
- Search result ranking: the capability to rank search results in display lists according to closeness of match between search criteria and documents

It also supports the use of topics or concepts in searching. A concept-based search can be constructed from various combinations of wild card and Boolean, near spell, fuzzy, phrase, and proximity operators. For example:

- A paragraph containing "operations" and ("environment," "closure," or "container")
- A phrase containing "60.111(b)"
- A paragraph containing "retrieve" and ("waste," "container," "canister," or "HLW")
- A paragraph containing "underground facilities" and ("fracture," "opening," or "stability")
- A paragraph containing "emplacement" and ("operations" or "container")
- A paragraph containing "backfill" and ("EBS," "Engineered Barrier System," or "Waste Package")
- A paragraph containing "thermal load" and ("stability" or "mechanical")
- A paragraph containing ("EBS" or "Engineered Barrier System") and ("Canister," "Container," or "Waste Package")
- A paragraph containing "recovery" and ("spent fuel," "value," or "resource")

Queries and concepts may be saved, recalled, and edited. On execution, they are checked for syntactic accuracy, and progress is reported. The GUI also provides a labeled list of fields that the user can fill in for header searches. Full-text and header search can also be combined in queries. The results of queries generally are displayed in a list of document titles scored by closeness of match. If the list is too long, the query can be edited in place to narrow the search conditions and reduce the number of finds.

Topic will be customized with privileges, preferences, and sources. Each user will have a password-protected account. This account will specify privileges in terms of what documents the user can and cannot see. From the Document Search-and-Retrieval Client, the user will be able to select from among a set of preferences with which to start Topic. These preferences will determine what fields are used for header queries and search results and will be configured for individual document types. In

addition, Topic will be configured with sources that limit searches to a specific document type or across document types. Users will be able to select sources in Topic when editing queries.

Topic is a highly sophisticated search-and-retrieval engine. Because its GUI is user friendly some users will quickly tap its full power. However, because it is sophisticated, most users will need training on it (see discussion in Section 6).

5.4.3 Document Viewing, Cut and Paste, Hyperlink, and Launch

The results of queries, as stated previously, are displayed in a list of document titles scored by closeness of match. In order to view a document, the user simply selects a document from the results list. Topic will then display the document. What will be viewed is a copy of the document in plain text. In the Topic viewer, text may be cut from the document viewer and pasted into another document, saved to a local file, and printed.

In the viewer, if the document has images associated with it, icons will be embedded in the text where those images would normally appear. The user may select an icon to view the associated image. Topic will then launch the image file into an image viewer.

Also in the viewer, if the document has been submitted in WordPerfect format, an icon will appear at the top of the document labeled "WordPerfect." The user may select the icon to launch the document into WordPerfect and view it in its original format. Within WordPerfect, formatted text may be cut and pasted into other documents, saved to a local file, and/or printed.

It should be noted that what the user may edit, save, and print are unofficial copies and not the original documents—users will not be able to modify documents in the TDOCS repository.

5.4.4 Document Downloading and Editing

Quick access to documents will also be possible. From the Document Search-and-Retrieval Clients, users will be able to select a document from a list by type, title, and other header information, and then elect either to copy the document to a local file or edit it in WordPerfect. Again, what the user may copy and edit are unofficial copies and not the original documents—users will not be able to modify documents in the TDOCS repository.

5.4.5 On-Demand Document Requests

Occasionally, a document will not be found in TDOCS. This may be because it has not been submitted or only a header has been submitted. Since access to the document may be essential to technical analysis and review, the user will be able to request it on demand.

Requesting a document on demand will involve a number of steps. The user will fill in an electronic form identifying the document much the same way custodians fill in headers for document submittal. For example, the user may have located a document's header in TDOCS and, thus, its hardcopy or microfilm accession number. This information is entered into the electronic form. The form

will then be e-mailed, automatically if possible, to the appropriate TDOCS custodian for document processing. As soon as the document has been processed through scanning, OCR, and cleanup operations, the resulting electronic document will be e-mailed back to the user and, if it is deemed significant enough, queued for loading into TDOCS. Alternatively, if the user is also an authorized custodian, he or she may use a scanning workstation to process and obtain the document.

5.4.6 Database Reporting

From Document Search-and-Retrieval Clients, users will be able to view and print database content reports. A number of reports will be possible, including listings of documents and their status, listings of headers, and listings of new additions to TDOCS.

5.4.7 Other Database and Service Access

Access to other databases and services will be possible through the Document Search-and-Retrieval Clients. In some cases, this will be a simple matter. Because TDOCS and RPD share common functionality and tools, for example, TDOCS users with proper authority will be able to access RPD directly. Other databases, such as the DHLWM Hydrology, Site Characterization Plan, and NIST Analyses databases, and the CNWRA TDI database, will be downloaded and incorporated into TDOCS. In other cases, there are a number of technical issues and policy matters that remain to be resolved. Requirements call for access to at least the NUDOCS database and possibly the on-line library service, Dialog. These issues and matters are discussed in Section 6.

6 PRODUCTION SYSTEM ISSUES

Exploratory work, review, and revision have resolved many issues and have been highly productive concerning GUIs, search and retrieval, document launching, security and control, and reports. Galaxy has proven itself to be a portable GUI application development environment; a common library of code can be transferred. Topic is a powerful search-and-retrieval engine; procedures for loading documents into it need only minimal refinement. Oracle is effective for configuration control and reporting; code for database access needs only minimal modification.

However, many unresolved technical issues and policy matters remain. Document sets to be loaded into TDOCS must be selected. Scanning, OCR, and cleanup must be investigated for integration with Topic. Header fields, both required and optional, must be specified. Access to NUDOCS, and possibly a selection of other databases and on-line services, must be investigated. Procedures, policies, and training needs for TDOCS must be formulated. Access to TDOCS at the DHLWM during implementation must be examined. Hardware and software for image manipulation and enhancement must be recommended. This section discusses these unresolved technical issues and policy matters.

6.1 DOCUMENT SETS

Document sets to be loaded into TDOCS must be selected. This section makes recommendations and states system-related restrictions for this selection. However, it will be the responsibility of the Advisory Groups to select document sets and prioritize their loading. For the immediate benefit of the DHLWM, selections should be made for loading TDOCS during prototype and production phases. Remaining document sets can then be prioritized for loading into TDOCS once it has been delivered.

Initial emphasis for document selection should be placed on materials generated by and currently in use at the DHLWM and the CNWRA. These include:

- The DHLWM in-house databases: Hydrology, Site Characterization Plan, and NIST/Materials
- The CNWRA in-house databases: TDI, Correspondence, and QA
- Regulations, such as 10 CFR Part 60
- DOE reports received by the DHLWM
- United States Geological Survey (USGS), National Laboratory, and DOE contractor reports pertaining to the Yucca Mountain Site
- Technical journal articles

Selection and prioritization of document sets should also be made in conjunction with the selection and prioritization of access to other databases and on-line services.

In selecting document sets, the Advisory Groups should consider that TDOCS will be a limited system. The planned phases of design and implementation place restrictions on what documents can be processed immediately. The processing of paper documents will be limited in volume, and the processing of electronic documents will be limited to certain formats.

The Advisory Groups should consider the proposed phases of design and implementation. The TDOCS prototype system of Phase 1 will be limited to loading headers and documents from electronic sources. The TDOCS production system of Phase 2, where scanning, OCR, and cleanup processes will be implemented, will be limited to a small set of paper documents. Delivery of TDOCS will not remove limitations on the volume of documents that can be processed.

Based on experience at the CNWRA, approximately 20 documents are indexed each day. This includes documents that would be loaded completely electronically as full text, input of headers only, electronically and as keyed input, and scanned documents. Given an average of 20 pages per document and the possible 5–15 minutes per page for cleanup, perhaps 2 to 4 of the estimated 20 documents at most could be processed currently by means of scanning, OCR, and cleanup. TDOCS scanning will be done by the DHLWM technical and clerical staff and not a dedicated operator as suggested by the DHLWM staff. If larger volumes are scanned, the cleanup operations will require staff that neither the DHLWM or CNWRA currently have available or in place. For this reason, the Advisory Groups might consider for large volumes, the possibility of contracting companies that specialize in batch scanning, OCR, and cleanup.

TDOCS will provide for the deletion of documents. Thus, in selecting document sets, the Advisory Groups should also give consideration to policies for deletion. It may be that certain types of documents should never be removed from TDOCS, while others might be removed and archived.

The DHLWM and the CNWRA have standardized on WordPerfect for word processing. Therefore, document formats should be limited to plain text and WordPerfect, though allowances could be made for other formats provided they can be filtered by Topic and converted by WordPerfect. Image formats must be restricted as well to a few like encapsulated postscript (EPS) or Targa Information File Format (TIFF). The decision of which image file format to support will, in part, determine the type of image viewing software required for the system.

6.2 SCANNING, OPTICAL CHARACTER RECOGNITION, AND CLEANUP

A process of scanning materials and converting them to full text through OCR will be required for both routine and on-demand loading of technical references. This is typically a labor-intensive process that involves the following steps:

- Taking an image scan of each document page
- Identifying graphical image zones to ignore during OCR
- Performing the OCR conversion

- Resolving ambiguities identified during the OCR operations
- Taking a scan of graphic images or "clipping" them from original page scan
- Providing a file name for each previously identified graphical image that uniquely associates the image with a location in the document
- Comparing original and new document pages to ensure textual integrity
- Embedding hyperlinks for all document graphics to support image view launching during document viewing
- Performing a spell check on the document

The labor required to perform these steps depends not only on the length of the document but also on the scanning and OCR effectiveness, and the degree of cleanup accuracy desired.

While scanning and OCR can achieve a less than 1 percent error rate with good quality source documents, this level of reliability implies that a manual document cleanup procedure will be required as part of the conversion process. Therefore, it is an important design consideration that the scanning and OCR process be as effective as possible.

In terms of scanning and OCR, graphical images within a document consist of figures, tables, and equations. These nontext objects must be recognized, identified, stored in an appropriate format, and properly associated with the textual materials through both an appropriate file naming scheme and embedded hyperlinks. These functions are accomplished interactively by an operator during document conversion through interaction with the scanning, OCR, and hyperlink software. This can be very labor-intensive if a large number of images are contained in the document.

Since the cleanup process involves human intervention, the training and experience level of the person performing the cleanup procedure can have a significant effect on the accuracy and timeliness of the derived document. The scanning and OCR process must, therefore, be coupled with appropriate document cleanup procedures to ensure confidence in the converted material.

The TDOCS design approach for development of a scanning, OCR, and cleanup capability will include the following considerations:

- Selection of the highest quality scanning and OCR hardware and software to minimize conversion errors
- Selection of software that minimizes operator intervention during the cleanup process
- Selection of software that supports efficient hyperlink embedding as part of the cleanup process

- Development of a concise scanning, OCR, and cleanup methodology that maximizes operator efficiency, enhances cleanup accuracy, and minimizes operator training requirements

In addition, scanning and OCR software that supports appropriate text and image output formats will be selected to preserve the original document formatting information and provide image viewing compatibility.

6.3 DOCUMENT HEADERS

Headers will be required for all documents loaded into TDOCS. They will be used to support configuration control, database reporting, and the structured queries with which users of such systems as NUDOCS and TDI are accustomed. In some cases, all that will be stored in TDOCS will be document headers; these will be considered slots that could be filled on demand. The issue is not whether headers are required but where headers should be stored and what fields need to be included.

There are two places where headers could be stored, Topic and Oracle. It is possible to store headers in Topic alone. This would support configuration control, database reporting, and structured queries. However, Topic-supported configuration control would be complicated, because Topic utility output must be redirected to text files which must then be read and parsed. Topic-supported database reporting would be limited to database dumps of simple format. Oracle overcomes these problems by providing more sophisticated, direct access to data. Thus, it is recommended that headers be stored in Topic to support structured queries and in Oracle to support configuration control and content reporting.

As a further technical issue, the system will require some internal fields to support configuration control. These include such fields as document ID, instance ID, partition, filename, etc., as discussed in Section 5. These internal fields will be system supplied and transparent to end users.

All other fields should be selected on the basis of ease of entry and usefulness for search and retrieval. Header fields can be categorized as those that are required and those that are optional. Required fields should be limited to those that uniquely identify a document to an end user, such as document type, title, author, and date. These required fields will be used across document sets. Optional fields could include any others considered useful for structured queries and reports. These optional fields could vary across document sets. Appendix A lists a set of possible fields (Acree et al., 1992) supported by the LSS Advisory Board in 1992. For additional background on the LSS and header fields it is recommended that the referenced report (Acree et al., 1992) be reviewed.

The IMS team and the Advisory Groups will work together to select fields to be included in document headers. There is an immediate need for required fields. Optional fields can be selected as document sets are selected.

6.4 OTHER DATABASES AND SERVICES

TDOCS will support access to other databases and services. As stated in the TDOCS requirements report (Johnson, et. al., 1993), these include:

- NUDOCS
- LSS/INFOStreams
- The DOE Improved Records Information System (IRIS)
- The CNWRA RPD
- DIALOG, an on-line library service

The most important of these databases is NUDOCS. It represents the official repository for the NRC documents and contains approximately 2 million records. Of these records approximately 35,000 documents have been scanned full-text, and another 35,000 have been abstracted for the DHLWM. Even though dial-in access is currently supported, it would be premature to begin design and implementation with regard to NUDOCS, since the Office of Information Resources Management (IRM) is revising it with TCP/IP access expected to be available sometime in 1994. Progress on the NUDOCS revision will continue to be monitored. TDOCS design and implementation must be open and flexible enough to accommodate such access; this is one argument supporting the use of Oracle.

The same approach will be adopted for the LSS/INFOStreams. This is important because it will replace NUDOCS as the NRC official document repository. This is not likely to happen prior to completion of TDOCS.

Access to RPD will be possible through TDOCS. This is possible because the two systems have similar design and employ the same software tools.

The IMS team working with the Advisory Groups will select which other databases and on-line services TDOCS will access.

6.5 PROCEDURES, POLICIES, AND TRAINING NEEDS

It is expected that procedures for loading and managing the document database will be responsibilities shared among the DHLWM technical and clerical staff. These procedures will require training. It is further expected that procedures for administering user accounts and maintaining the system (i.e., performing backups) will be the responsibility of the IRM system administrator who now performs similar activities associated with maintaining network and computer system operations at NRC.

The development of procedures and policies for the use of TDOCS will be the responsibility of the CNWRA IMS team working with the Advisory Groups. Procedures will involve loading and managing the document database, administering user accounts, and maintaining the system. Policies will make procedural responsibilities and rules explicit. An important aspect of these policies will be deciding who the users and custodians of the system will be and what privileges should be assigned to them.

Procedures and policies will be developed during design and development and be submitted for review in the form of a users' guide and training for TDOCS, both of which are required deliverables. The development of procedures and policies will require the close cooperation of the Advisory Groups.

It is suggested that training in the use of the Topic search-and-retrieval engine be contracted through Verity, Inc. A variety of courses are offered, from half-day group quick starts to full-day individual training, either on-site or at a Verity Education Center in Mountain View, California, or McLean, Virginia.

6.6 ACCESS TO TDOCS

TDOCS is available for single-user demonstration over the 56 KBPS line between the CNWRA, where the application executes, and the DHLWM, where it displays. It is planned that, during the TDOCS production phase, clients will be available for executing document searches and viewing locally at the DHLWM, and retrieving indexes and documents over the 56 KBPS line from the database server at the CNWRA.

Most likely, the 56 KBPS line will not provide reasonable response times. The situation will be monitored closely during the production phase. This issue can be resolved quickly by upgrading the line to a T1 (1.544 MBPS) capacity.

6.7 IMAGE MANIPULATION AND ENHANCEMENT

The requirements for TDOCS include image manipulation and enhancement. Specific capabilities that have been discussed with the DHLWM include video frame capture and large size image viewing, panning, and zoom in and out. Two recommendations are made concerning these capabilities for review by the Advisory Groups.

First, it is recommended that these capabilities be made available on only a single, shared workstation located in a common work area. This will reduce the potential cost of hardware and software upgrades needed to support these specialized capabilities. The DHLWM staff has agreed that this arrangement would be adequate.

Second, since this consideration is somewhat peripheral to the document management, search, and retrieval requirements for TDOCS, it is recommended that only investigation and specification development be completed for the hardware and software of such a workstation.

7 CONCLUSIONS

7.1 SUMMARY OF DESIGN PLAN

The following major conclusions can be drawn from the TDOCS design plan:

- TDOCS will be designed and implemented in at least three phases during FY94 and FY95. Based on direction and feedback from Advisory Groups at the DHLWM and the CNWRA, the design will be finalized using a prototype system derived from exploratory work already done on the RPD.

- Exploratory work, review, and revision have resolved many issues and have been highly productive concerning GUIs, search and retrieval, document launching, security and control, and reports. Galaxy has proven itself to be a portable GUI application development environment; a common library of code can be transferred. Topic is a powerful search-and-retrieval engine; procedures for loading documents into it need only minimal refinement. Oracle is effective for configuration control and reporting; code for database access needs only minimal modification.

- The design, in accordance with the iterative approach adopted in the design plan, is open-ended, flexible, and expandable. Document scanning, OCR, and cleanup, for example, have been given only a general place in the design; precisely how these processes will fit into the design is a matter for investigation during the prototype phase. In other words, it is anticipated that this design will change, and that another cycle of implementation, review, and revision will take place during the prototype phase in order to arrive at a production system design.

7.2 ACTION ITEMS FOR THE INFORMATION MANAGEMENT SYSTEM TEAM AND ADVISORY GROUPS

As expected, many unresolved technical issues and policy matters remain. Document sets to be loaded into TDOCS must be selected. Scanning, OCR, and cleanup must be investigated for integration with Topic. Header fields, both required and optional, must be specified. Access to NUDOCS and possibly a selection of other databases and on-line services must be investigated. Procedures, policies, and training needs for TDOCS must be formulated. Access to TDOCS at the DHLWM during implementation must be examined. Hardware and software for image manipulation and enhancement must be recommended. Meetings are planned for February and April 1994 between the IMS team and the DHLWM Advisory Group to resolve technical issues and policy matters.

8 REFERENCES

Acree, C., R.D. Johnson, B. Brient, and S. Spector, 1992. Classification and Attributes of Non-Text-Searchable Documentary Material and Its Treatment in the Licensing Support System (LSS). San Antonio, TX: Center for Nuclear Waste Regulatory Analyses.

Chilk, S.J. 1993. SECY-93-107 — Licensing Support System Program and Budget Responsibilities. Memorandum for James M. Taylor, Executive Director of Operations. Washington, DC: U.S. Nuclear Regulatory Commission.

Center for Nuclear Waste Regulatory Analyses. 1992. *(DHLWM) Advanced Computer Review System for Technical License Review — Support Tasks.* San Antonio, TX: Center for Nuclear Waste Regulatory Analyses.

Center for Nuclear Waste Regulatory Analyses. 1993. *CNWRA FY94-95 Operations Plan for the Division of High-Level Waste Management.* San Antonio, TX: Center for Nuclear Waste Regulatory Analyses.

DeWispelare, A.R., R.D. Johnson, R.L. Marshall, and J.H. Cooper. 1993a. *Development Plan for PASS/PADB System Design Version 3.0.* San Antonio, TX: Center for Nuclear Waste Regulatory Analyses.

DeWispelare, A.R., P.C. Mackin, J.H. Cooper, and R.L. Marshall. 1993b. *Regulatory Program Database Version 1.0 Development Status FY93.* 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.

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: U.S. Nuclear Regulatory Commission.

Youngblood, B.J. 1993. Overall Review Strategy for the Nuclear Regulatory Commission's High-Level Waste Repository Program. Note to Division of High-Level Waste Management Staff. Washington, DC: U.S. Nuclear Regulatory Commission.

APPENDIX A

LICENSING SUPPORT SYSTEM (LSS) PROPOSED BIBLIOGRAPHIC HEADER FIELDS

LICENSING SUPPORT SYSTEM (LSS) PROPOSED BIBLIOGRAPHIC HEADER FIELDS

The LSS Advisory Review Panel-approved LSS bibliographic header fields[1,2] should apply to Technical Investigation Package (TIP) units and to individually indexed (packaged and unpackaged) non-text-searchable documentary materials as follows. The general instructions contained in the draft LSS Cataloging Manual[3] will apply unless otherwise noted.

PARTICIPANT ACCESSION NUMBER: This unique identifier will be assigned by the submitter for each package-unit and to each individually indexed unit (packaged or not).

SUBMITTER CENTER: This field will be used in the same way that is used for all LSS units, subject to a controlled vocabulary, to identify the responsible, originating organization for each package or individually indexed item.

SUBMITTER PAGE COUNT: In the case of package-units, this field will include the total count indicated on the table of contents plus the number of pages on which that listing itself is printed. The number of pages in a package will often exceed 1,000, requiring modification of the tentative 999 maximum permitted by the draft LSS Cataloging Manual. In the case of items residing on machine-dependent media, most of them are expected to be resident on electronic media, such as magnetic tape, where page counts are unlikely to apply; however, photographic materials (microforms, slides, etc.) that cannot be converted to electronic images without loss of content will have page (frame) counts that will be entered.

TITLE/DESCRIPTION: In the case of a package-unit, the entry for this field should be taken directly from the title on the table of contents. It is very important that an LSS requestor be immediately alerted to the fact that a TIP has been found. For this reason, the CNWRA previously suggested that the word "package," if applicable, should precede the actual title in this field. That suggestion was based on the assumption that the CNWRA could not recommend a rearrangement of the LSS header fields. A better way to alert a requestor to this circumstance would be to display the Document Type field (showing "package") next to the Title field on the LSS screen when a header is displayed. For the same reason, the Package ID field should be displayed near these two fields, so that individually indexed materials within a TIP will be prominently related to their packaged collection. In the case of individually indexed materials, their titles should be taken from the descriptions on the related slip sheets (input forms) or created in the same way that LSS titles are generally created when there is no special input form. A title

---

[1]Header Working Group, *Recommended Fields for LSS Header Records*, Marshall, VA: Header Working Group, 1990.

[2]Header Working Group, *Additional Fields for Headers*, Marshall, VA: Header Working Group, 1991.

[3]Science Application International Corp., *Licensing Support System Cataloging Manual (draft)*. McLean, V: Science Application International Corp., 1990.

for an individually indexed packaged item should include the TIP title, as well as the item's specific subject area (e.g., "Injection Test Data included within Geological and Drill-Hole Data for Test Well USW H-5, Yucca Mountain, Nye County, Nevada"). The Document Type (e.g., magnetic tape), and the Package ID (showing that is part of an identified TIP) would be simultaneously viewable on the LSS screen.

AUTHOR: In the case of a package-unit, the name on the TIP's signature line should be used, plus any other names found on the table of contents, description of a finished product, circulated draft, or raw-data item (when exceptionally noted). The CNWRA recommends that the full first name and middle initial be entered, not merely the person's initials as mandated by the draft LSS Cataloging Manual. (The individuals who were involved in the TIP approval/review process are not, considered authors and are therefore, excluded from this field.) In the case of an individually indexed package item, the same guideline should apply (all authors of the package should be mentioned) unless it is exceptionally identified as the creation of another person on the table of contents.

AUTHOR ORGANIZATION: In the case of a package-unit, this field will take the name of the responsible organization from the table of contents. It will also include any other names noted, on an exceptional basis, next to raw-data items that have been produced by other organizations (including subcontractors) and next to authors of finished products. In the case of individually indexed packaged items, the name of the responsible organization will be entered unless it is a packaged item identified as originating elsewhere. Names on the table of contents will, in all instances, conform with the controlled Organization Name Authority List. As stipulated by the draft LSS Cataloging Manual, the correlation between each author and that author's organizational affiliation will be maintained.

ADDRESSEE: This field is not applicable to packages, but it is applicable to individually indexed correspondence that a TIP may contain.

ADDRESSEE ORGANIZATION: This field is not applicable to packages but it is applicable to individually indexed correspondence that a TIP may contain, using the Organization Name Authority List.

DOCUMENT DATE: In the case of package-units, the date of its completion located at the bottom of the table of contents should be used. In the case of individually indexed packaged items, the particular date provided on the table of contents should be entered.

DOCUMENT/REPORT NUMBER: This number or numbers assigned to a unit by the submitting organization for identification or control purposes should be applied to package-units and to other items in accordance with the organization's internal procedures. Technical reports and approval/review documents will be numbered in accordance with the format rules of the Document Number Authority List if they apply.

DOCUMENT CONDITION: Packages may have missing pages, illegible pages, or pages bearing marginalia, in which case these characteristics should be entered into this field in accordance with the controlled vocabulary in the Document Condition List as stipulated by the draft LSS Cataloging Manual. In the case of nonimageable items, their nonimageable status should be indicated.

EDITION/VERSION: This field would have applicability to circulated drafts, some finished products, and some nonimageable materials, computer programs in particular. The draft LSS Cataloging Manual excludes this field and would apparently use the Pointers field for this purpose.

EVENT DATE, CODE: If TIPs or individually indexed materials have a particular dated event (audit, hearing, inspection, or meeting) as their primary topic, this date would be entered using the same controlled vocabulary that will be used for all materials. Otherwise, the Document Date, showing a TIP's completion date will be sufficient, requiring nothing to be entered here.

PROTECTED STATUS: Privileges or exceptions claimed for a TIP or for individually indexed materials should be explained using the controlled vocabulary (The draft LSS Cataloging Manual excludes this field).

RELATED DOCUMENTS: The original approved description for this field[4] calls for a package and the cataloging units it contains to be identified here, by the submitter, as a "whole/part" type of relationship to be translated at LSS input time into standardized form for the Pointers field. The draft LSS Cataloging manual, which excludes the Related Documents field, suggests that the prefix "PA" from the Pointer Code List be used in the two-way Pointers field to communicate that a unit that "has parts.. or is part of...". This could logically be used for packages and for package parts. However, a Package ID field was adopted by the LSSARP in 1991[5] making such use redundant for TIPs. The CNWRA recommends that, for TIPs, these fields (Related Documents and Pointers) be used instead to link packages that concern the same prolonged investigation, using a special prefix.

SPECIAL CLASS: This field, intended to further clarify units, will employ a controlled vocabulary from a Special Class List according to the draft LSS Cataloging Manual. While that Manual requires this field to be used for packages headers, its Special Class List does not include the term "package" as an option. Given the fact that a package-unit will be identified as such in the Document Type field, there seems to be no reason to include a redundancy here. Nonimageable items should be identified in this field as having a "Header Only," which is an option on the List. The listed option, "Header & Image only," will not ordinarily be used for packaged material of that type, since imageable raw data will not be individually indexed according to the CNWRA's recommendation; but a finished product, such as a map or a design drawing, would be so identified.

DOCUMENT TYPE: This field will be governed by a controlled Document Type List, which was mentioned but not included in the draft LSS Cataloging Manual. The intent here is to describe the kind of unit that the header represents, not in terms of information content, but in terms of form (e.g., correspondence, reports, etc.). A "package" should certainly be an option on this list when it is created. In the case of individually indexed non-text-searchable materials, the classification should be developed into a set of additions to this list. More than one entry should be made if more than one document type applies to particular unit.

---

[4]Header Working Group, 1990.

[5]Header Working Group, 1991.

SPONSORING ORGANIZATION: This is the name of the agency or agencies responsible for funding or otherwise sponsoring the investigative work undertaken by the Submitter Center. It, too, will use the controlled Organization Name Authority List.

PUBLICATION DATA: This field, which is intended to capture "bibliographic information that is not covered in other fields but is important in identifying or citing the unit" has no perceived applicability to package-units. However, as suggested by examples in the draft LSS Cataloging Manual, it is a convenient location to record publication data on indexed maps and computer software programs, which would include characteristics that have been defined for those categories of material. Design drawings and commercially available photographs are other items whose publication data could usefully be placed in this field, when those items are indexed.

DESCRIPTORS: These subject terms, selected from the controlled vocabulary in the LSS Thesaurus to identify the information content of a unit, will be particularly important for the retrieval of TIPs. They should be gleaned not only from the title of a package, but also from the package table of contents, the package abstract, the title of any finished product included in the package, and the abstract, table of contents, index, introduction, and any other summary included in the finished product. In the case of nonimageable packaged items, the subset of the above descriptors that are pertinent to their specific content should be entered.

IDENTIFIERS: These subject terms, which are not yet contained in the LSS Thesaurus but would serve to further identify the information content of a unit, will also be important for the retrieval of TIPs and individually indexed materials. They should be gleaned from the same sources from which descriptors have been gleaned.

COMMENTS: This field is intended to capture "information not covered in (other fields) which... will be necessary to identify or retrieve the unit." Examples provided in the draft LSS Cataloging Manual are specifics about foreign language content, missing pages/attachments (for which only an indicator was used in the Document Condition field), or omitted images when the pages in question were larger than 17 × 22 inches, rendering them "oversize" and thus, presumably nonimageable. Conditions such as these would apply to packaged materials as well as to other LSS documents. However, the CNWRA[6] has recommended that all items, regardless of size (or color), should be scanned, if possible, to produce viewable LSS images, which would negate the Manual's last example. The CNWRA suggests that, in the case of items resident on nonimageable media, this field be used to supplement the proposed controlled-entry Media Type field as described below under that heading.

ABSTRACT/SUMMARY: This narrative description of a unit's content should be included within headers for package-units themselves, for finished products within a package (but not approval/review documents), and for nonimageable items belonging to a package. Headers, for non-text-searchable items having no package association, should also include abstracts. An abstract is most successfully written by the producer of the material by the principal investigator, or by another knowledgeable technician, rather

---

[6]Center for Nuclear Waste Regulatory Analyses, *Alternative Ways of Making Packaged Documentary Materials Accessible within the Licensing Support System*, San Antonio, TX: Center for Nuclear Waste Regulatory Analyses, 1991.

than by records management specialists who will be less acquainted with its technical content. An abstract for a package-unit should be brief and focused entirely on the subject and results of the investigation. It should make no attempt to summarize the table of contents. An abstract of individually indexed materials should mention characteristics such as the geographic area, scale, size, and color of a map should be described in this field.

LSS SYSTEM ACCESSION NUMBER: This unique identification code, assigned to each cataloging unit by the LSS capture station, has no exceptional applicability to packages or to non-text-searchable material in general.

NUMBER OF IMAGES: This number will be based upon the total in the Submitter Page Count field. If materials are scanned as single LSS images regardless of size, as the CNWRA recommends, this field becomes redundant with Submitter Page Count, because an "oversize" image will not be divided into parts. If a printed rendition is ordered by an LSS requestor, it, too, would be delivered whole.

POINTERS: This field is intended to provide references to Related Documents after the entries in that field have been standardized. The CNWRA's recommendations concerning that field apply.

PACKAGE ID: This field was adopted specifically for the benefit of packages, with the aim of relating all parts of a given package to each other. A controlled two-character prefix to the assigned number (obtained from LSS Process Control and typed at the top of a table of contents) is suggested by the draft LSS Cataloging Manual for the purpose of distinguishing "data package" (which this report is calling TIPs) from other types of packaged collections (e.g., training packages or audit packages). The parts of a package that are independently indexed as units (finished products, approval/review documents, and nonimageable items, according to the CNWRA recommendation) should all bear the number of their assigned package. Errata sheets relating to package-units will be linked to those units through this field.

COPYRIGHT: This field has no exceptional applicability to packaged materials.

MEDIA TYPE: This field, already included in the draft LSS Cataloging Manual, would designate the media on which nonimageable materials are stored. It would not apply to headers for TIPs, their finished products, and their approval/review documents. According to the CNWRA's recommendation, nonimageable items will be individually indexed in all instances (in or out of packages) and require this characterization, since they will be unavailable for viewing on LSS screens. These items are currently numerous and are expected to become increasingly so. The draft LSS Cataloging Manual did not provide a controlled Storage Media List, which it proposed. Machine-dependent media types should constitute options for that list, which would permit a search by general media type. However, to enable a requestor to understand what specific equipment will be necessary to examine a given item, a full media description will be necessary, amplifying the listed options to detail the hardware/software dependencies. In view of the fact that a controlled field like this cannot contain such detail, it is suggested that the Comments field be used for this purpose. The detail would be obtained from information provided on the corresponding slip sheets within a package, which were created as input forms. Miscellaneous nonimageable items that have no machine dependencies could be described as "paper," "mylar," "vellum," "acetate," or "other" in the Media Type Field. In the case of imageable items, the field could be left blank or the term "image" could be entered by default.

STORAGE LOCATION: This field will tell where nonimageable materials can be found and how they can be examined. It should contain an acronym/abbreviation from the Organization Name Authority List, which would link to an updatable providing a current mailing address, phone number, FAX number, cognizant administrative authority, and procedure for retrieval. Like the Media Type field, this field will have no applicability to TIP headers themselves but will be essential for nonimageable items, which must be kept apart from the LSS computer system at dispersed storage locations. The draft LSS Cataloging Manual suggests that the Media Type field incorporate a code that could be translated through a table into a procedure for access to the particular item. The CNWRA believes that a separate, searchable field (Storage Location) would be preferable. The field could either be left blank in the case of imageable items, or the term "image" could be entered by default.

QUALITY ASSURANCE STATUS: This would be a simple indicator, saying either "yes" or "no", thereby indicating whether or not the unit was produced by the submitting organization under an approved quality assurance program, in accordance with 10CFR Part 60 Subpart G. The field would have applicability to all LSS documents, including packaged material. In the case of package-units, the entry may be taken from the typed notation on the bottom left corner of the TIP Table of Contents, which identifies its "Quality Assurance Status."