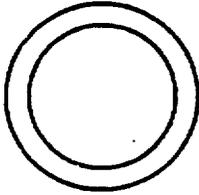


06C

Office of Civilian Radioactive Waste Management

Office of Resource Management



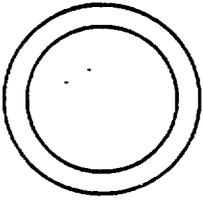
*Licensing Support System
Conceptual Design Analysis*

May 1988

*U.S. Department of Energy
Office of Civilian Radioactive Waste Management*

Office of Civilian Radioactive Waste Management

Office of Resource Management



***Licensing Support System
Conceptual Design Analysis***

May 1988

*U.S. Department of Energy
Office of Civilian Radioactive Waste Management*

TABLE OF CONTENTS

	Page
Preface.....	i
Executive Summary.....	ii
1.0 Introduction.....	1
1.1 Purpose and Scope.....	1
1.2 LSS Subsystems.....	1
1.2.1 Text Storage.....	2
1.2.1.1 Records Access Subsystem.....	2
1.2.1.2 Regulations Access Subsystem.....	2
1.2.2 Tracking.....	2
1.2.2.1 Issues Tracking Subsystem.....	3
1.2.2.2 Commitments Tracking Subsystem.....	3
2.0 Required LSS Functions and Their Conceptual Implementations. 5	
2.1 Capture of Information into the LSS.....	5
2.1.1 Capture of Tracking Information.....	6
2.1.1.1 Issues Tracking.....	6
2.1.1.2 Commitment Tracking.....	6
2.1.2 Capture of Records and Regulations Information.....	7
2.1.2.1 Image Capture (Scanning).....	7
2.1.2.2 Text Capture.....	7
2.1.2.2.1 Direct Keying of Document Text.....	8
2.1.2.2.2 Optical Character Recognition.....	8
2.1.2.2.3 Input of Standard Word Processor Files.....	8
2.1.2.2.4 Electronic Mail Input.....	9
2.1.2.3 Cataloging.....	9
2.1.3 Capture Quality Control.....	11
2.1.4 Preparation, Transfer and Loading.....	12
2.1.4.1 Headers and Full Text.....	13
2.1.4.2 Images.....	13
2.2 Storage of Information in the LSS.....	14
2.2.1 Contents of Tracking Subsystems.....	14
2.2.2 Contents of Records Access Subsystem.....	15
2.2.2.1 Standard Documents.....	15
2.2.2.1.1 Headers.....	15
2.2.2.1.2 Text.....	16
2.2.2.1.3 Images.....	17
2.2.2.2 Special Cases.....	17
2.2.2.2.1 Satellite File Indices (SIMS).....	18
2.2.2.2.2 Document Archive Indices.....	18
2.2.2.2.3 Privileged Information.....	18
2.2.2.2.4 OCRWM QA Records.....	18
2.2.3 Contents of Regulations Access Subsystem.....	19
2.2.4 LSS Storage Topology.....	20
2.2.4.1 Workstations.....	20
2.2.4.2 Node(s).....	21

TABLE OF CONTENTS
(continued)

	Page
2.3 Access to LSS Information.....	21
2.3.1 Access to Information in the LSS Tracking Subsystems..	21
2.3.2 Access to Information in LSS Records and Regulations Data Bases.....	22
2.3.2.1 Structured Index Searching.....	22
2.3.2.2 Full-Text Searching.....	23
2.4 Output of LSS Information.....	24
2.4.1 LSS Output.....	24
2.4.1.1 Workstations.....	24
2.4.1.2 Printers.....	26
2.4.2 Types of LSS Output.....	26
2.4.2.1 ASCII.....	26
2.4.2.2 Image.....	27
2.5 Communication of LSS Information.....	28
2.5.1 Node to Node Communication.....	28
2.5.2 Capture Station to Storage Communications.....	29
2.5.3 Node to Workstation Communications.....	29
2.6 Electronic Mail.....	30
2.6.1 Conventional E-Mail Functions.....	30
2.6.1.1 Uploading Files from a Workstation.....	31
2.6.1.2 Message Transmission and Receipt.....	31
2.6.1.3 Privacy and Authenticity.....	31
2.6.2 Special LSS E-Mail Functions.....	31
2.7 LSS Management Functions.....	31
2.7.1 Performance and Usage Monitoring.....	32
2.7.2 LSS System Administration.....	32
2.7.2.1 LSS Quality Control.....	32
2.7.2.2 LSS Configuration Management.....	32
2.7.3 Data Base Administration.....	33
2.7.3.1 Data Base Maintenance.....	33
2.7.3.2 Loss Protection.....	34
2.7.3.3 Access Control.....	34
3.0 Concept Of Operation: Data Capture and Data Retrieval.....	35
3.1 Document Preparation and Data Capture.....	35
3.2 Data Retrieval Operations.....	38
3.2.1 Access.....	38
3.2.2 Query Operations.....	39
3.2.3 Retrieval Operations.....	40
3.2.3.1 Headers.....	40
3.2.3.2 Full Text.....	40
3.2.3.3 Images.....	41
3.2.4 User Session.....	41

TABLE OF CONTENTS
(continued)

	Page
4.0 Conceptual Design.....	43
4.1 Base Conceptual Design.....	44
4.1.1 Base Conceptual Design Hardware.....	44
4.1.1.1 Capture System.....	44
4.1.1.2 Search System.....	47
4.1.1.3 Image System.....	48
4.1.1.4 Communications System.....	49
4.1.1.5 Workstations.....	52
4.1.2 Base Conceptual Design Software.....	53
4.1.2.1 Capture Software.....	53
4.1.2.2 Search Software.....	54
4.1.2.3 Image System Software.....	56
4.1.2.4 Communications Software.....	57
4.1.2.5 Workstation Software.....	58
4.2 Variants to Base Conceptual Design.....	60
4.2.1 Variant I - Full Replicated Nodes.....	60
4.2.1.1 Description.....	60
4.2.1.2 Impact On The Capture System.....	60
4.2.1.3 Impact On The Search System.....	60
4.2.1.4 Impact On The Image System.....	60
4.2.1.5 Impact On The Communications System.....	62
4.2.1.6 Impact On The Workstations.....	62
4.2.2 Variant II - Hardware Full Text Search.....	62
4.2.2.1 Description.....	62
4.2.2.2 Impact On The Capture System.....	64
4.2.2.3 Impact On The Search System.....	64
4.2.2.4 Other Impacts.....	64
4.2.3 Variant III - Images Are Not Supported At Workstations.....	64
4.2.3.1 Description.....	64
4.2.3.2 Impact On The Image System.....	64
4.2.3.3 Impact On The Communications System.....	64
4.2.3.4 Impact On The Workstations.....	66
4.2.3.5 Other Impacts.....	66
4.2.4 Variant IV - Microform Digitizers in Capture and Image Systems.....	66
4.2.4.1 Description.....	66
4.2.4.2 Impact On The Capture System.....	66
4.2.4.3 Impact On The Search System.....	66
4.2.4.4 Impact On The Image System.....	66
4.2.4.5 Impact On The Communications System.....	68
4.2.4.6 Impact On The Workstations.....	68
4.2.5 Variant V - Microform Off-Line Image Storage and Retrieval.....	68
4.2.5.1 Description.....	68
4.2.5.2 Impact On The Capture System.....	68
4.2.5.3 Impact On The Search System.....	70

TABLE OF CONTENTS
(continued)

	Page
4.2.5.4 Impact On The Image System.....	70
4.2.5.5 Impact On The Communications System.....	70
4.2.5.6 Impact On The Workstations.....	70
4.2.6 Variant VI - Full Text <u>via</u> Re-keying.....	70
4.2.6.1 Description.....	70
4.2.6.2 Impact On The Capture System.....	72
4.2.6.3 Other Impacts.....	72
4.2.7 Variant VII - Combined Variants III, V, and VI.....	72
4.2.7.1 Description.....	72
4.2.7.2 Other Impacts.....	72
5.0 Conclusions.....	74
REFERENCES.....	76
APPENDIX A - Example of LSS User Session Scenario.....	77
APPENDIX B - Revised Projection of the Size of the LSS Data Base, 1990-2009.....	79
APPENDIX C - Abbreviations Used.....	81

Preface

This is the third in a series of four reports on the Licensing Support System (LSS) prepared by the DOE Office of Civilian Radioactive Waste Management (OCRWM) for the Office of Management and Budget (OMB). The LSS is an information management system intended to support the needs of all the parties involved in repository licensing, including the Department of Energy (DOE) and the Nuclear Regulatory Commission (NRC). These reports are:

Preliminary Needs Analysis

Preliminary Data Scope Analysis

Conceptual Design Analysis

Benefit-Cost Analysis

The Preliminary Needs Analysis, issued in February 1988, and the Preliminary Data Scope Analysis, issued in March 1988, constitute the system requirements basis for developing a Conceptual Design, presented in this report. The Benefit-Cost Analysis evaluates alternatives within this conceptual design. These four reports, and subsequent refinements, are intended to provide the basis for determining the LSS design specifications.

Note that Appendix B contains a revised version of Table 8 (Projected size of the LSS data base, 1990 - 2009) of the Preliminary Data Scope Analysis. Revisions have been made for the estimates of pages added during 1993 and 1998, based on a re-evaluation of the expected levels of activity consistent with the methods described in that report.

Executive Summary

This report analyzes the design requirements for the Licensing Support System (LSS) and presents viable design concepts and implementation approaches that satisfy the expected functional requirements of LSS users. It presents a detailed conceptual design base that is highly responsive to these requirements and has a low associated development risk.

The Preliminary Needs Analysis and Preliminary Data Scope Analysis have defined the requirements of an automated computer-based information storage and retrieval system that must accommodate millions of documents. Based on these studies and the directions perceived from the LSS negotiated rulemaking process, a conceptual design has been formulated which meets the rulemaking requirements. The conceptual design has the following major features:

- 1) Headers and searchable full text of all documents suitable for inclusion in LSS
- 2) Bit-map images of all documents in LSS
 - reproduction of documents for quick distribution from central location
 - on-line display and local printing at special workstations
- 3) Centralized search system and on-line optical disk image system in Washington, DC or Las Vegas, NV
- 4) Multiple capture systems for:
 - scanning
 - text conversion
 - correction
 - cataloging
- 5) Workstations capable of displaying readers, ASCII text and images
- 6) Support for workstations displaying headers and ASCII text only
- 7) Retrieval through structured index searching of cataloging information and software full text searching of documents
- 8) Electronic mail

In analyzing the LSS requirements and formulating the conceptual design presented, a number of alternatives were examined. Some were rejected due to a low probability of success. Others not only offered a reasonable risk but also were potentially more cost effective, although they may not meet all of the "non-firm" requirements identified in the Preliminary Needs Analysis. These alternatives were identified as variants to the base and will be subject to further technical and economic investigation.

The variants examined differed from the base conceptual design in the following ways:

- I. Two search and image systems replicating the data base, rather than one, located in Washington, DC and Las Vegas
- II. Hardware full text search, rather than software
- III. No workstations capable of displaying images
- IV. Microform digitization rather than optical disk storage of images
- V. Off-line microform printing rather than on-line bit-mapped image system
- VI. Re-keying text rather than text conversion from scanned bit-map.
- VII. Combination of 3, 5 and 6 above.

The base conceptual design and variants are consistent with the requirements identified to date, including the deliberations of the Negotiated Rulemaking Advisory Committee. Indeed, the rulemaking activities have not yet imposed any requirements on the design which were not anticipated in the Preliminary Needs Analysis. It is further not expected that the rulemaking activities will result in any requirements which cannot be met by the base design or one of the variants; however, the possibility still exists that design refinements may be required to reflect changing requirements. It is important to note that the need for the Tracking Subsystems has not been identified to date by the Negotiated Rulemaking Advisory Committee, but rather arises from the Preliminary Needs Analysis discussions with potential DOE (and contractor) users and from the LSS Design and Implementation Contract (DOE, 1987) requirements. For this reason, it is probable that the Tracking Subsystems described here shall be removed from the LSS design and implemented for DOE only.

Significant to the refinement of the design is the feedback from potential users and the benefit-cost analysis. All of the dialogue conducted to this point with potential users of the LSS has been in the absence of a common reference design. This has caused much difficulty in communication, because while one must speak from a certain frame of reference, it is almost assuredly not the same as that of the other party. With this study, a reference point is established for the LSS conceptual design that will facilitate future deliberations. Refinement of the design can benefit from additional experiences of and feedback from potential users, especially in the area of features available to the user. With this base frame of reference, the potential user can now envision certain scenarios of LSS applications and can fill in the various details necessary to more accurately predict system sizing and response.

The Benefit-Cost Analysis, the next report in the series, will permit the investigation of the variants from the standpoint of cost (both savings and additions), and the associated benefits (or lack thereof) of each variant. Such information will provide an economic basis for decisions on how best to meet the various requirements on the system.

The opportunity is now available for substantive input to the design process, which is both encouraged and necessary to refine the process further.

1.0 INTRODUCTION

This report analyzes the design requirements for the Licensing Support System (LSS) and presents viable design concepts and implementation approaches that satisfy the expected functional requirements of LSS users. It presents a detailed conceptual design base that is highly responsive to these requirements and has a low associated development risk.

1.1 Purpose and Scope

The first two reports in this series (Preliminary Needs Analysis and Preliminary Data Scope Analysis) have been a first effort under the LSS Design and Implementation Contract toward developing a sound requirements foundation for the LSS design. It is clear that there is not one unique design which can meet these requirements and the detailed specifications to be refined from them. Rather, a variety of basic concepts and myriad variations on those concepts can be acceptable.

The purpose of this analysis is to develop one such concept, for which a benefit-cost analysis will be performed in the last report in this series. Although a number of general LSS concepts have been proposed, the conceptual design presented here is the most detailed to date. In order to present this level of detail, the basic concepts examined needed to be limited. The basic concept presented here (the base conceptual design presented in Section 4.1) was selected because it is highly responsive to the requirements identified and has a low associated development risk.

Section 2.0 specifies the major functions required of LSS and discusses how these functions can be implemented. The basic concept of LSS operation is presented in Section 3.0 and covers both information capture and information retrieval. Section 4.1 describes the proposed base conceptual design in detail, and Section 4.2 describes seven major variants on this design. These variants have been selected to examine significant technological or operational options in implementing LSS functions. These variants will be examined along with the Base Conceptual Design in the forthcoming Benefit-Cost Analysis.

1.2 LSS Subsystems

The Licensing Support System (as described in the LSS Design and Implementation contract RFP) consists of four subsystems, grouped into two categories - Text Storage and Tracking. Each component provides on-line access to a particular type of information.

1.2.1 Text Storage

The Text Storage category has two subsystems: the Records Access Subsystem and the Regulation Access Subsystem. Each subsystem is a library of documents generated by or related to primarily the OCRWM mined geologic repository program and its license application process. The full text of the documents in each will be stored, indexed and made available for on-line search and retrieval. A set of catalog data containing information about each document (such as author(s), title, date, document type, and accession number) will be maintained, indexed and made available for on-line access.

1.2.1.1 Records Access Subsystem

The Records Access Subsystem will primarily contain documents generated by OCRWM program participants. Users of the subsystem will include attorneys, licensing staff, project engineers, and designers. Hard copies or microform copies of all documents in the subsystem will be kept in the LSS Document Archives. The Records Access Subsystem will also contain cataloging data for non-text programmatic records (such as material samples, field instrument strip charts and design drawings). The non-text records will be kept in the LSS Satellite Archives. The Records Access Subsystem cataloging data for each of these items will contain its location in the LSS Archive storage facilities.

1.2.1.2 Regulations Access Subsystem

The Regulations Access Subsystem will contain documents that impose legal requirements on the program, such as, the Nuclear Waste Policy Act (NWPA) of 1982 and its Amendments Act, 10 CFR Part 60, other Federal and state regulations, regulatory guides, and DOE agreements with regulatory agencies. Both structured cataloging data and full text will be stored and available for search and retrieval. Each of the user groups identified in the Preliminary Needs Analysis are expected to use this subsystem.

1.2.2 Tracking

The LSS will also include two Tracking subsystems: the Issues Tracking Subsystem and the Commitments Tracking Subsystem. These subsystems will be used primarily by OCRWM and NRC licensing staff, project management personnel and technical staff to record and review identified and agreed to issues and commitments and the status of progress towards their compliance or completion. In both subsystems the data will be stored in a structured form, similar to the catalog data in the Records Access and Regulations Access Subsystems.

It is important to note that the need for the Tracking Subsystems has not been identified to date by the Negotiated Rulemaking Advisory Committee, but rather arises from the Preliminary Needs Analysis discussions with potential DOE (and contractor) users and from the LSS Design and Implementation Contract (DOE, 1987) requirements. For this reason, it is

probable that the Tracking Subsystems described here shall be removed from the LSS design and implemented for DOE only.

1.2.2.1 Issues Tracking Subsystem

Issues are identified efforts that clarify what part of a regulation applies to the program and how compliance is to be achieved. Issues can be raised by DOE or a regulatory agency and resolution is accomplished through research, investigation and negotiation, the results of which are documented and stored in the Records Access Subsystem. The Issues Tracking Subsystem will provide the capability to record and track the resolution of issues in a controlled, automated fashion. Some uses of the subsystem will be by:

- 1) OCRWM licensing staff who need to track licensing and regulatory related issues for the geologic repository program
- 2) NRC and other non-DOE parties identified in the Nuclear Waste Policy Act, who are sources of issues
- 3) OCRWM Site Characterization staff and NNWSI project staff who will follow progress towards resolving the issues in the NNWSI Site Characterization Plan.

1.2.2.2 Commitments Tracking Subsystem

The Commitments Tracking Subsystem is similar to the system described above, but records and tracks the progress of:

- 1) Commitments made by OCRWM to other organizations
- 2) Commitments made to OCRWM by other organizations
- 3) Commitments or action items internal to the OCRWM program.

2.0 REQUIRED LSS FUNCTIONS AND THEIR CONCEPTUAL IMPLEMENTATIONS

The LSS will be used to capture, store, access, and present (output) all records, regulations and tracking information that is deemed relevant to obtaining the necessary licenses and permits for the siting, design, construction, operation, and closure of a geologic repository for the disposal of spent nuclear fuel and high-level nuclear waste as authorized by the Nuclear Waste Policy Act of 1982 and the Nuclear Waste Policy Amendments Act of 1987. This section describes needed LSS functions and features in these areas that were identified in the Preliminary Needs and the Preliminary Data Scope Analyses, and presents possible implementations. Additional communications, electronic mail and management functions necessary to LSS are also described and evaluated here.

2.1 Capture of Information into the LSS

The LSS must be able to accept a variety of inputs from various sources. Information in the Tracking Subsystems and header information in the Records and Regulations Access Subsystems is created specifically for the LSS and must be keyed into the system. According to the Preliminary Data Scope Analysis, the vast majority of the backlog material appropriate for inclusion in the Records and the Regulations Access Subsystems exists in the form of hardcopy and microform documents. The text of hardcopy documents, needed for the full-text search capability which is expected to be required by the Negotiated Rulemaking Advisory Committee (NRAC), may be entered by re-keying or by the conversion of a scanned image of the document into a text file. Microform information must be converted to and entered as hardcopy or must be directly digitized and similarly converted. The NRAC is also expected to require images of all material in the LSS. The system is also expected to support electronic mail among users (Preliminary Needs Analysis) and will probably be required to accept material from the electronic mail system as input to the Records Access Subsystem. The following sections present the capture functions and features required to support data entry into these subsystems.

The Preliminary Needs Analysis also indicates that rigorous quality control (QC) in the capture process is essential to the usefulness of the LSS. Such QC would not be limited to assuring the conformity of captured material with the original, but must also include the elimination of duplicate input and the verification of cataloging information.

2.1.1 Capture of Tracking Information

2.1.1.1 Issues Tracking

The Issues Tracking function requires the following types of input:

- 1) Input Issue-Related Information
LSS must provide for the input of issues which relate to NRC, DOE, other federal, state, and local rules and regulations.
- 2) Input Information Related to the Resolution of Issues
LSS must provide for the input of the description of the resolution of issues. Each resolution must be related to the relevant issue record.
- 3) Input Information Related to the Status of Issues
LSS must provide for the input of milestones and events related to the handling of an issue.
- 4) Input Issue-Related Work Plans
LSS must provide for the input of the action plans for issue resolution and responsibilities for attaining resolution of each issue.
- 5) Input References to Information Related to the Issue
LSS must provide for the input of references to the source documents that initiated or identified the issue, regulations that impact the issue, and information related to the resolution of the issue.

The input of Issues Tracking related textual information will be from LSS terminals (local and remote). The data will be entered from terminal keyboards via on-line, fill-in-the-blank type of screens. The procedures that will govern the selection and specification (format) of the data to be entered, as defined above, will contain data entry forms for the transcription of the data to be entered. Each Issues record will contain a unique accession number for identification and retrieval purposes. The data entry procedures will also provide for the edit/changes of text in existing Issues records in the data base. Markups of text in existing documents may be attached to the data entry form to minimize the amount of transcription required.

2.1.1.2 Commitment Tracking

The Commitment Tracking function requires the following types of input:

- 1) Input Information Related to Commitments
LSS must provide for the input of commitments made to federal agencies, regulatory agencies, states, tribes, and legal parties identified in the Nuclear Waste Policy Act.
- 2) Input the Resolution of Commitments
LSS must provide for the input of the description of the resolution

of the commitment. Each resolution must be related to the relevant commitment record.

3) Input Information on the Status of Commitments

LSS must provide for the input of milestones and events related information on the status of a commitment.

The input of Commitment Tracking related textual information will be from LSS terminals (local and remote). The data will be entered from terminal keyboards via on-line, fill-in-the-blank type of screens. The procedures which will govern the selection and specification (format) of the data to be entered, as defined above, will contain data entry forms for the transcription of the data to be entered. Each Commitment record will contain an unique accession number for identification and retrieval purposes. The data entry procedures will also provide for the edit/change of text in existing Commitment records in the data base.

2.1.2 Capture of Records and Regulations Information

Capture of the LSS records and regulations consists of the processes of scanning documents to obtain the electronic (bit-mapped) image, obtaining text in electronic form for the purpose of preparing for full-text search, cataloging the document for retrieval, and the associated quality control. Due to the large amount of data to be processed, capture of the LSS documents and regulations, and their corresponding cataloging data, is expected to be accomplished at document capture centers located in the vicinity of sites which contain significant backlog or generate significant new documents.

2.1.2.1 Image Capture (Scanning)

Scanning of hardcopy pages to capture the electronic image is required for pages which are to be submitted to the character recognition process and for pages which are not in text form such as graphics, maps, and figures. In addition, since it is required to store the image of the document for the purpose of providing hardcopy on request, it may be reasonable to store those images in electronic form.

2.1.2.2 Text Capture

ASCII text of documents may be captured by the system by one of the following three methods:

- 1) Direct keying of document text from terminal keyboards
- 2) Optical character recognition (OCR)
- 3) Input of ASCII word processor files.

2.1.2.2.1 Direct Keying of Document Text

The LSS should provide the capability to capture document text in ASCII format by direct keying of document text from terminal keyboards. This capability is particularly important for documents that cannot be efficiently converted through the Optical Character Recognition (OCR) process, for example, barely legible copies.

2.1.2.2.2 Optical Character Recognition

The OCR process converts an electronic (bit-mapped) image of a page into ASCII text (a standard pattern for each character and punctuation). The quality of the text produced is highly dependent on the quality of the image that is submitted to the process, i.e. an original printed page with uniform type will produce better results than a fourth generation photocopy with smudges and extraneous markings. Current generation OCR devices can produce text with 99.5% to 99.9% accuracy under optimum conditions. Note that this would still result in 3 to 15 errors in a 3000 character page.

Correction of errors is a manual process, although software tools such as spelling checkers can assist. (A nontrivial consideration is whether or not to correct spelling errors in the original text.) The necessity to correct the errors is dependent on their magnitude and other factors such as:

- 1) The effect of the errors on full-text retrieval
- 2) The use of the ASCII text in reading or browsing the document
- 3) The use of the ASCII text for downloading and file transfer.

The advantage of the OCR process is that it is relatively automated and can be performed without much human intervention up to the point of review and correction. Continuous improvements are being made in OCR technology that will increase speed of production and reduce the error rate. Presently OCR of an image made from scanning a good quality paper copy can be reasonably performed, however OCR from an image produced by blow-back of a microfiche or microfilm is not considered feasible.

2.1.2.2.3 Input of Standard Word Processor Files

It is expected that much of the future documentation for the repository design and licensing will be prepared with word processing equipment. Therefore it is important that the LSS should provide the capability to capture document text directly in electronic form. In order to accommodate the various software programs, a standard input format utilizing the ASCII form (no special codes or printer control characters) will be required. The permissible media include floppy disks and magnetic tape. In order that the image may be captured, a good quality hardcopy must also be provided with the ASCII file.

The major problem with receiving data in machine readable format is the quality assurance of such inputs. It is necessary that the machine readable version of the document be verified as a true representation of the

hardcopy. (In many cases last minute changes to a document are made on a typewriter.) Therefore, the quality assurance procedures for the input and verification of word processor inputs to the LSS must be clearly defined and enforced to ensure data base integrity.

2.1.2.2.4 Electronic Mail Input

The LSS is required to serve as a mechanism for the electronic transmission of filings by the parties during the high-level waste proceeding by means of an Electronic mail (E-Mail) capability. These filings must also be captured as records at the request of the sender. This can be accomplished by having the sender transmit a copy of the message along with "fill-in-the-blanks" header information to an LSS mailbox at a document capture workstation. The E-mail message can then be processed through the same input, cataloging, and quality assurance procedures as any other document, except that an image of the message will not exist.

2.1.2.3 Cataloging

The LSS should provide for the capture of catalog information, input by terminal keyboards, that will contain bibliographic data and keywords to support the user access to the documents stored in the Records or Regulations Access Subsystem. The catalog information will be stored in a "header" for each record. The header is created by extracting some of the information in the record and placing it in appropriate fields stored at the beginning or "head" of the record (descriptive cataloging). In addition, the header can include information about the subject content of the record (subject cataloging). A header describing a document can also be entered into a system without any additional information (such as the text of a document or an abstract). Examples of kinds of information obtained during cataloging are:

- 1) Descriptive information (extracted from a document)
 - Author's name and affiliation
 - Date of the document
 - Title of the document
 - Publication information (journal name, publisher, etc.)
 - Recipient
 - Abstract (if part of the document)
- 2) Subjective information (assigned to a document)
 - Subject category
 - Degree of technical detail
 - Descriptors or keywords
 - Abstract (if written by cataloger)
 - Pointer to relevant LSS Issues and Commitments entries.

Cataloging can also be computer assisted. Examples of computer-assisted cataloging are assigning an accession number to the document, extracting heavily used words for use as keywords, etc. Cataloging results are not necessarily limited to the header, as in the case where markers of special information are embedded in the text of a document file to facilitate

retrieval. Since the primary purpose of cataloging is to provide access points for retrieving (finding) the document in the data base without having to examine sequentially each document in the collection, the development of the header format, i.e., determining which fields are to be included in the header, is a function of both the cataloging process and the retrieval requirements.

The primary tool for the cataloging process is the cataloging manual. The cataloging manual consists of rules or conventions for entering data into each field. These rules or conventions govern either the form or the content of each entry. Generally, rules of form standardize the format in which data is entered, e.g., dates are to be entered in the format yy/mm/dd. Each field in the LSS header will have rules of form. These rules will be part of the cataloging manual. Rules governing the content of the entry are usually contained in controlled vocabularies. A controlled vocabulary or authority list is simply a list of those words available for entry into a field. Words or terms not contained in the list are unacceptable. Controlled vocabularies carry standardization into the content of the field entry by specifying the correct word form(s) of names, subjects, etc., to be used in the data base. Fields such as agency name, corporate affiliation, or document type will have a controlled vocabulary as part of the cataloging manual. These lists will remove any doubt whether to use DOE, Dept. of Energy, Department of Energy, Energy Dept., U.S. Department of Energy, etc. for that government agency. Similarly, the decision to use the term letter, memo, correspondence, or mail for a particular document type will be resolved by consulting the controlled vocabulary for that field.

Controlled vocabularies are also used for subject cataloging. The LSS would use a thesaurus as the controlled vocabulary for terms describing the content of each record. A thesaurus is a controlled vocabulary list that shows the relationships between words or terms in the vocabulary. The relationships are shown as broader, narrower, or related terms. In addition, "use" and "used for" terms are also included. As an example, the broader term for "automobile" is "vehicle", a narrower term is "four-door sedan", and a related term is "motorcycle". A "used for" term would be "car", while the entry for "car" would state "use automobile". Thesaurus entries also contain notes or definitions to explain the application of the term, e.g. "hearing - a legal procedure" (as opposed to "hearing - a physiological process"). Thus a thesaurus would eliminate uncertainties on the part of the cataloger and the user as to which terms to use for cataloging and retrieval. The thesaurus would be arranged both alphabetically and hierarchically to assist the cataloger and the user in describing documents or non-documents consistently, logically, and at an appropriate level of detail. Instructions on how to use the thesaurus and how to submit recommended terms would be included in the thesaurus. Procedures for maintaining, updating, and revising all controlled vocabularies will be included in the cataloging manual.

One of the most significant conclusions of the Preliminary Needs Analysis is that appropriate and extremely high quality cataloging is essential to guarantee the usefulness of the Records and Regulations Access Subsystems. The quality of the cataloging directly determines the usefulness of structured index access to the LSS, which, in turn, is a critical (if not the most critical) technique for identifying material in

the system. Much of this cataloging must be performed by highly trained and experienced personnel, particularly the subject cataloging. The development of an appropriate header format and the design of rigorous and reliable cataloging procedures are therefore key aspects of the detailed LSS design.

2.1.3 Capture Quality Control

The capture quality control process should include not only verification of correctness and of fidelity of free text and images to the original material, but also explanation of anomalies or relationships that might surprise an LSS user.

Quality control is accomplished by:

- 1) Use of an integrated production process that minimizes opportunity for introduction of error
- 2) A production process that focuses human attention and uses human time efficiently
- 3) Human follow-up and spot checks
- 4) Use of automated tools to verify data, enforce consistency, and identify anomalies.

Quality control of imaged data includes:

- 1) Verification that all pages are present and correctly sequenced
- 2) Verification that imaged pages are legible, complete, and unskewed
- 3) Entry of a standard indicator for missing pages or oversized pages.

Software tools to support the image quality control process might include:

- 1) Tools to re-sequence pages
- 2) Tools to delete pages.

General policies regarding trade-offs of fidelity to source material versus utility of the data must be articulated and agreed to before the quality control process is implemented. One policy might be to ensure that all data stored in headers complies with standards for spelling and abbreviation, but document text is to reflect spelling, including errors, as they appear in the source document.

Consistency of assignment and of data representation is critical to the usability of catalog data. Catalogers will be highly trained and will have specific, written guidelines available for all fields that require a specific format or that require evaluation and judgment. A variety of software tools support entry and quality control of catalog data. They include:

- 1) Menus accessing authority lists legal values for fields having controlled vocabularies (e.g., journal title, document type, author affiliation) to provide on-line tools to help in validity checking

- 2) Range checks, date conversion checks, and format checks for applicable fields
- 3) Cut-and-paste capabilities for data such as author names and document titles and importing from windows for data such as descriptors and major subjects.

As the data base grows, far more sophisticated computer assistance can be developed. For example, author names can be checked for similarity to previously-stored data. Names that differ by only a few characters may be presented for human verification. Similarly, publication dates might be verified for feasibility with respect to other publications by the same author. Computer-assisted duplicate detection is required in the early phases of population. Later, automated duplicate detection, perhaps with human verification, may be more practical.

Quality control of document text that originates in hardcopy includes use of automated spelling checks, with editing tools for human correction of errors, along with tools and controls for maintaining the dictionary of valid spellings. The dictionary will grow rapidly during early phases of data capture as numerous technical terms and proper names are added. Periodic reviews of the dictionary, along with the careful control of additions, will ensure that few errors are introduced.

Any omissions of material must be explained in a revealing and consistent way. Formulas or other text containing non-roman characters, illustrations, graphics, and tables that cannot be rendered as ASCII text must be identified. Missing portions of text in the original must be so indicated.

As the data base grows, it may be worthwhile to develop software to verify citations in document text against catalog data. Such checks would be useful to detect possible errors in the data and to identify material for inclusion in the LSS data base.

Random checking of the data by staff knowledgeable in the subject matter must be performed on a regular basis. In a data conversion effort for another government organization, such checks revealed the consistent interpretation of a peculiar typographic form of "waves" as "wives". Software can be developed to detect, and perhaps perform context-sensitive filtering of, recurring errors.

2.1.4 Preparation, Transfer and Loading

Following release from the capture quality control process, the document and regulations full text, associated catalog data, and bit-mapped images will require preparation, transfer, and loading into the LSS data base. Differing requirements for access and storage imply that the data will be partitioned into various major data bases such as:

- 1) A records data base containing the full text, cataloging data, and indices for documents

- 2) A regulations data base containing the full text, cataloging data, and indices for relevant Federal, State, and local regulations
- 3) An issues tracking data base that contains the issues information and associated structured indices
- 4) A commitment tracking data base that contains the commitments information and associated structured indices
- 5) An image data base that contains the bit-mapped image data for all page images and indices for retrieval of such data.

Since the tracking data can be entered directly into the data bases as described in Sections 2.1.1.1 and 2.1.1.2, no further processing is required.

2.1.4.1 Headers and Full Text

Because of the size of the records full-text data base, it and possibly its associated catalog data are likely to be partitioned into several data base segments. The first step in preparation for storage is to sort the records according to segment. The separate files are then transferred to the computer(s) on which the data base loading will be processed. The loading will store the header data and full text as well as construct or update the structured indices for the header data and the inverted index, which contains pointers to each unique word and its location within the full text. Both the structured and inverted indices should provide quick data access in response to user queries.

For regulations, no segmentation will be required; thus the files prepared during capture can be directly transferred to the loading process, which is identical to records loading.

If full-text searching is implemented via special full-text search hardware then the full-text of the records and regulations will be separately loaded and no inverted index will be built.

If there is more than one node, then duplicate copies of the prepared files will be made and distributed to each node for loading.

2.1.4.2 Images

The bit-mapped images stored during the capture process should include unique accession numbers and page numbers which link them to the catalog data and full-text data bases for records and regulations and support retrieval. The captured images will be transferred to the image storage system for loading.

An alternative method for storing images for display and printing is microform. In that case, the microform is produced during or prior to the capture processing. For the case in which the images are to be available

on-line, in preparation for loading into the retrieval storage device, a link tying accession and page number to microform will be constructed.

2.2 Storage of Information in the LSS

The Preliminary Data Scope Analysis concludes that the LSS should support all the information needs of all parties involved in repository licensing, serving as the sole basis for expedited document discovery. The report estimates that this will constitute from 9 to 11 million pages of relevant information in August 1990 (when the LSS is scheduled to be operational), increasing to between 32 to 42 million pages in the next twenty years. NRAC interprets this to mean that all (technologically suitable) material should be in searchable full-text. If this capability is to be provided by software, the associated inverted indices (see Sections 2.1.4.1 and 2.3.2.2) would also be stored by the system.

That report also concluded that for a data base as large and diverse as that envisioned for the LSS to be useful, the material it contains should be extensively, accurately and consistently cataloged. This cataloging information, stored as headers associated with each document in the Records and Regulations Access Subsystems, would also be stored along with their structured indices. Such headers are also likely to be required by the NRAC.

The Preliminary Needs Analysis found that some records may be needed in the form of bit-mapped images while the NRAC is expected to require that images of most pages (in some form) be available through the system. This can require significant additional storage in some designs.

In addition to the storage of standard documents, the LSS is required to store several other types of information. The OCRWM Quality Assurance Plan (DOE, 1986) requires the LSS to be the repository of all Program QA records, which constitute a special class of documents requiring special treatment. Other information, in the form of headers only, must also be stored in the LSS. The statement of work of the LSS Design and Implementation Contract (DOE, 1987) requires the LSS to be able to store indices of the LSS archives. As discussed in Section 1.2, these archives are made up of Document Files and Satellite Files. The Satellite File index should effectively be linked to a Sample Inventory and Management System (SIMS). The NRAC also intends to require the system to store information identifying (but not including the contents of) privileged documents.

Although not large in comparison to the amount of material relevant to the Records Access Subsystem, information in the Tracking Subsystems and the Regulations Access Subsystem will also require some storage capacity.

2.2.1 Contents of Tracking Subsystems

The Tracking Subsystem's data consist of both definition information and status data, in the form of ASCII text. Most of the data elements that describe both types are also stored as structured indices. The definition information is relatively static, while the status data is fairly dynamic. Thus the data should be stored on media that can be easily modified.

For both issues and commitments, once a definition has been entered, an on-line record would likely be maintained for the duration of the LSS operational period. The estimates for the amount of storage required are:

Issues - for 5000 issues, investigations, studies, and activities: the total data storage would be about 13 million bytes and the structured indices would be about 10 million bytes.

Commitments - for 3000 commitments/action items: the data storage would be about 9 million bytes. The structured indices would be an additional 4 million bytes.

2.2.2 Contents of Records Access Subsystem

The Records Access Subsystem storage requirements derive from the need to store the full text of documents subject to discovery during licensing hearings and appeals as defined by the NRAC. In addition to the full text of documents, the Records Access Subsystem document headers, images of non-textual pages, structured indices and inverted indices, which support the full-text retrieval of the document text, also contribute to the storage requirements.

2.2.2.1 Standard Documents

Standard documents stored in the Records Access and Regulations Access Subsystems should consist of header records, ASCII text records, and bit-mapped images records for non-textual document pages. Headers only will be stored for special cases, such as non-document information, documents not included in full text, and privileged information as discussed in Section 2.2.2.2 below. In accordance with ANSI/ASME NQA-1 (1983), Quality Assurance Program Requirements, OCRWM Quality Assurance records should be included in the data base to indicate the level and status of required quality assurance processing.

2.2.2.1.1 Headers

Headers, which are manually input from terminal keyboards, contain cataloging information consisting of such fields as author, corporate affiliation, document title, number of pages, etc. As indicated in the Preliminary Needs Analysis, the following field descriptions were found desirable by a majority of users:

- o Originating and Recipient Organization(s)
- o Date the Document Was Created
- o Author(s)
- o Document Type
- o Baseline Data Flag
- o Submitter Reference Number

Based on the document counts estimated in the Preliminary Data Scope Analysis (see Appendix B), the amount of storage space required to accommodate the header data for all documents and non-documents is:

Time	Storage in megabytes		
	Data	Structured Indices	Total
1990	1,500 - 3,800	300 - 800	1,800 - 4,600
1994	4,600 - 15,000	900 - 3,000	5,500 - 18,000
1998	6,400 - 21,000	1,300 - 4,200	7,700 - 25,000

2.2.2.1.2 Text

The full text of a document should be entered and stored in the LSS and be available for full-text search and retrieval by software or hardware. As a part of the document selection process, the user should be able to browse or read all or parts of the retrieved document.

In order for all the words in documents to be searched by software, the text must be indexed. Software full-text search programs include the tools to accomplish this process; thus it is a relatively automated process and does not require skilled information management personnel. The resulting index, sometimes referred to as an inverted index, contains a sorted list of all words in the documents (except specified common words such as a, an, the, etc.) and a pointer to the storage location(s), i.e. disk volume, file, record, of the words in the documents. The size of the inverted index is a function of the program which is used for the indexing, but it can vary from 50% to 200% of the size of the original ASCII text file.

Full-text inversion, although not labor intensive, requires major computer resources and time to process large files. The files may require segmentation, although this may be invisible to the user.

Based on the page counts estimated in the Preliminary Data Scope Analysis (see Appendix B), the amount of storage space required to accommodate the full text ASCII and its inverted indices is:

Time	Storage in megabytes		
	ASCII Text	Inverted Indices	Total
1990	9,000 - 15,000	8,000 - 17,000	17,000 - 32,000
1994	29,000 - 57,000	25,000 - 69,000	54,000 - 126,000
1998	39,000 - 80,000	35,000 - 95,000	74,000 - 175,000

2.2.2.1.3 Images

Images of documents should be stored in the LSS for the distribution of copies to users on request. Images may be stored in the following forms:

- 1) Hardcopy (paper)
- 2) Microform (microfilm, microfiche)
- 3) Electronic format (compressed bit-mapped images).

The form of image storage depends on two factors: (1) the requirements for the time of receipt of hardcopy as an output of LSS, and (2) the location of the user who makes the request for the hardcopy of a document. The Preliminary Needs Analysis discovered that on the average managers would be willing to wait 3 hours on the average for documents of 100 pages or less. Technical users would wait 7 hours and regulatory staff over 22 hours for the same materials. For hardcopy more than 100 pages, most users would wait overnight.

If a maximum waiting time of 3 hours for a hardcopy of a document of 100 pages or less is considered an LSS requirement, the location of the user who made the request must be considered. If the requesting user's location is local to the LSS node that contains the image, then any of the three possible image forms is acceptable. The hardcopy could be produced by photocopy of the paper image, blow-back from the microform image, or laser printout from the bit-mapped image and hand delivered to the user. However, if the requesting user's location is remote (in another city or state) to the LSS node which contains the image, then electronic transfer (via telecommunications) of the bit-mapped image is the only option available to meet the receipt time requirements.

For the overnight delivery of hardcopies of more than 100 pages, it is possible to meet the requirement with any of the three possible forms of image storage.

Based on the page counts estimated in the Preliminary Data Scope Analysis (see Appendix B), and assuming all images are in electronic form, the amount of storage required to accommodate the bit-mapped images is:

<u>Time</u>	<u>Storage in gigabytes</u>
	Bit-mapped images
1990	170 - 260
1994	540 - 1,000
1998	750 - 1,400

2.2.2.2 Special Cases

The storage and access to some LSS information is considered as special cases. This information includes documents and non-documents in the LSS Archives and privileged information or documents that will not be stored in the LSS.

The LSS Archives should be an integral part of the LSS and should be places to physically deposit those products of the OCRWM program that must be retained and accessible throughout the life of the OCRWM program. The LSS Archives should be able to store documents (including microform) and other physical forms of materials that might include samples of soil, rock, water, plant, etc., and data from tests or explorations that are in non-reproducible form, or in electronic form which, in its unprocessed state, is not useful for inclusion in the on-line portion of the LSS.

2.2.2.2.1 Satellite File Indices (SIMS)

Satellite Files are defined as that portion of the LSS Archives which contains (in boxes, cabinets, shelves) physical forms of materials (such as core) and non-reproducible test or exploration data (such as recorder charts). The on-line portion of the LSS data base will contain only header information concerning the material, indicating where in the Sample Inventory Management System (SIMS) further detail can be found. The SIMS will provide, for each unique item, a description of the item and the physical location of the item for access purposes.

2.2.2.2.2 Document Archive Indices

Document Archives are defined as that portion of the LSS Archives that contains the images (hardcopy, microform, or electronic form) of all LSS documents, with the exception of privileged information documents. As stated in Section 2.2.2.1.1 above, headers for all documents stored in full text in the LSS will be included in the LSS data base. Also headers only will be included in the LSS data base for all document images contained in the Document Archives that have not been stored in full text in the LSS data base (for what ever reasons). These "header only" entries will contain the same information as full-text document headers and in addition will contain the location within the Archive of the document image.

2.2.2.2.3 Privileged Information

Information relevant to the repository licensing that is deemed privileged by any party to the process will be identified in the LSS through the entry of a header with a limited number of fields, which includes a brief description of the document and the privilege being claimed. No entry of text or image will be made until and unless the claim of privilege is overruled. The document will then receive expedited process for entry into the LSS data base in searchable full text.

2.2.2.2.4 OCRWM QA Records

Quality records play a particularly important role in the OCRWM program and the repository licensing process. As such, they require special identification. Quality assurance audits, quality policy manuals, implementing procedures, activity plans, inspection records and test results

are examples of quality related records. Codes in the header will indicate that records are quality related. Another header field will indicate the quality level associated with each OCRWM program-generated document.

Often quality records are in sets or packages, such as all of the records related to one QA audit. To associate or tie such a set together, a package header will be used. It will contain the package name, date, source, and identifier. The package identifier will also be in the header of each record in the set. Structured indexes will be created for the package identifier in both the package header and the record header. Other fields in the package header will also have structured indices.

2.2.3 Contents of Regulations Access Subsystem

The Regulations Access Subsystem, like the Records Access Subsystem, has headers, full text, bit-mapped images and the associated indices for each which are stored to support on-line access. A unique aspect of the regulations data base is the requirement to maintain the latest contiguous version of the document even if the regulation is amended in parts.

The amount of storage required, based on the Preliminary Data Scope Analysis (see Appendix B), is:

<u>Time</u>	<u>Header storage in megabytes</u>		
	Headers	Structured Indices	Total
1990	1.1 - 2.1	0.2 - 0.4	1.3 - 2.5
1994	1.3 - 2.6	0.3 - 0.5	1.6 - 3.1
1998	1.6 - 3.2	0.3 - 0.6	1.9 - 3.8

<u>Time</u>	<u>Full Text storage in megabytes</u>		
	ASCII Text	Inverted Indices	Total
1990	43	39 - 52	82 - 95
1994	54	49 - 65	100 - 120
1998	65	58 - 78	120 - 140

<u>Time</u>	<u>Image storage in megabytes</u>
	Bit-mapped Images
1990	350 - 500
1994	440 - 630
1998	530 - 750

2.2.4 LSS Storage Topology

The key objectives for the LSS are the capability for:

- 1) Full-text storage/retrieval of a large number of documents
- 2) Rapid, full-text search
- 3) Full-text access from diverse geographic locations
- 4) Hardcopy production at terminal locations.

Each objective has a direct bearing on the LSS storage topology, that is, where the LSS data base is physically stored. Similarly, the LSS storage topology has a direct bearing on the total life-cycle cost of the LSS, which includes the development, user/operator training, operations, configuration management, facility preparation, equipment maintenance, telecommunication usage, and data base maintenance.

If it were not for the objective of full-text access and responsive hardcopy production at user terminals at dispersed geographic locations, then a single node (central computer system and data base storage location) would be the most cost effective LSS storage topology. The Preliminary Needs Analysis indicates that there are three major and ten other potential locations across the country for LSS users. The Preliminary Needs Analysis further suggests that in the 1990 to 1992 time frame, the distribution of usage among these locations is approximately:

Washington, DC Area	50%
Las Vegas, NV	30%
San Antonio, TX	10%
Other Locations	10%

Based upon this distribution of potential LSS users, and with 80% of the users located in either Washington, DC or Las Vegas, NV, it is clear that no more than two LSS storage topologies need to be considered as follows:

- 1) A single-node configuration (central computer system and data base storage) in the Washington, DC or Las Vegas area, and
- 2) A two-node configuration (two computer systems and distributed data base storage) in both the Washington, DC and Las Vegas, NV locations.

2.2.4.1 Workstations

Because LSS user workstations may differ widely in type and configuration and because of the quality assurance and data base integrity requirements of LSS data, it is nearly operationally impossible to store the LSS data at the workstations. However, the Preliminary Needs Analysis states that about half of the potential users interviewed stated requirements for downloading LSS information to other systems (which include PC-based workstations). The average and peak downloads generally ranged from two to three pages to 20 to 30 pages. However, some regulatory staff indicated peak download requirements in excess of 1,000 pages. Based on these requirements,

workstations requesting the download of LSS data would require from 9,000 to 3 million bytes of disk storage.

2.2.4.2 Node(s)

The system design options for the LSS storage topology (*i.e.* the number and locations of LSS nodes) must address not only the computational, storage, and input/output rates required to meet the functional and operational requirements of the LSS, but additionally must consider the total life-cycle cost of the LSS. Experience has shown that the operational cost of a system, such as the LSS, far exceeds the initial developmental cost, and therefore the recurring operational cost will be a determining factor in the optimal design of the system.

Based upon the geographic distribution and temporal distribution of usage demand contained in the Preliminary Needs Analysis, a central one-node configuration located in the Washington, DC area or Las Vegas and a distributed two-node configuration located in the Washington DC and Las Vegas, NV areas need to be considered as LSS system design options. These two options provide access (*via* direct terminal connection or low cost Local Area Networks) to the LSS for approximately 80% of the users and support 80% of the usage demand.

2.3 Access to LSS Information

The Preliminary Needs Analysis concluded that structure index searching *via* detailed and extensive headers should be available, involving subject terms and keywords assigned with the aid of a controlled vocabulary. Full-text search capability on both document text and headers should be also available. These methods should be supported by extensive search aids. Both access methodologies are expected to be required by the NRAC. Further, the system should be easy to use with a minimum of training, should contain built-in help functions, and should provide assistance when needed, either through an expert system or on-call assistance.

2.3.1 Access to Information in the LSS Tracking Subsystems

Each of the user groups identified in the Preliminary Needs Analysis report will also be users of the Tracking Subsystems. Some of the uses will be to:

- 1) Track licensing and regulatory related issues for the geologic repository program
- 2) Follow the progress on satisfaction of the Site Characterization Plan (SCP) issues
- 3) Follow progress on the completion of commitments and action items.

Both the Issues Tracking and the Commitments Tracking Subsystems consist of records, similar to the LSS header, which contain fields such as titles, identification codes, keywords, contact persons and organizations.

All of these fields can be used to search and to specify which records are to be contained in requested reports. The ITS will also have a field which is the full text of the issue. This text in this field will be searchable in a manner similar to the full-text searching of LSS document text.

Both prompting menus and fill-in-the-blanks screens will be available for specifying search criteria.

2.3.2 Access to Information in LSS Records and Regulations Data Bases

Because of the very large and complex data base in the Records and Regulations Access Subsystems, the capability of efficiently and accurately identifying and accessing the information sought is crucial to the usefulness of the system. The Preliminary Needs Analysis has shown that LSS users will require access to LSS data through both structured index searching (on header fields) and through full-text searching. These access capabilities are also explicitly required by the Negotiated Rulemaking Advisory Committee.

The ultimate purpose of any computer-based text storage system is, in response to a specific query, to retrieve the largest possible fraction of documents that are appropriate to the query (high recall) while retrieving as few as possible documents that are inappropriate (high precision), all within the time and complexity constraints of the user. Rather than attempting to identify all "hits" and include no "misses" in a single pass, a very large system such as the LSS uses the concept of set refinement. Here, a retrieved set of results, which has been conservatively defined by an initial query to contain a very large number of possible hits (hopefully all) but also contains much irrelevant material, is iteratively refined by applying additional constraints until most (hopefully all) irrelevant material has been removed with little (hopefully no) relevant material lost. The user is interactively guided during the refinement process by being given the number of hits in the current retrieval set and by being able to examine members of the set to assess the set's degree of relevance (and therefore the degree to which the query has been completed). The results of multiple searches can be logically combined to create hybrid sets (sets that contain the results of multiple queries). Both the results of the queries and the queries themselves can be stored by the user, for later use. Thus, complex routine queries can be easily re-run.

In the LSS concept presented here, the refinement of the retrieval set of a query can be achieved through structured-index searching (Section 2.3.2.1) or full-text searching (Section 2.3.2.2) or a combination of both search techniques.

2.3.2.1 Structured-Index Searching

Structured-index searching, the conventional method used by data base management software to access data, searches indices constructed to support the specific type of queries. These indices are constructed on one field or a combination of fields in the header of a document, such as author, date published or keyword. They can be considered as a surrogate data base that

has been specially sorted to support rapid retrieval and response to a specific type of anticipated query on a specific field (or set of fields) in the header. (See Section 2.1.2.3 for the method by which header information is obtained. See Section 2.2.2.1.1 for header and index storage size requirements.)

The completeness of the header design and content (obtained by reliably anticipating the header information which will be needed by users and the ways in which they will search that information) is key to efficient structured-index searching. Significant effort in these areas will be required during detailed LSS design, since structured-index searching takes advantage of a certain amount of pre-processing of the query, and since full-text searching (discussed in the following section) can yield an unmanageably large (and therefore not useful) number of hits when operating on a very large data base. One approach would be to limit the initial steps in the identification and refinement of a retrieval set to structured index searching. Once the set has been reduced to a manageable size (and full-text searches are useful) both methods may be used to resolve the set. The threshold for when both access methods become available must be designed based on the expected response time of a full-text search request.

The ability to formulate Boolean logic expressions among header fields is necessary for useful full-text searching. These expressions need not be directly created by the user, but in order to maximize ease of use, they may be constructed by the system from information prompted from the user. Structured-index searching also requires the creation and storage of index files. The tradeoff between index storage requirements and availability of indices must be examined during the detailed design, but structured indices generally require much less storage than full-text search inverted indices (see next section) because they are based on the less voluminous header information, rather than on the entire text of the documents.

2.3.2.2 Full-Text Searching

Full-text searching is a computerized text processing technique that locates the occurrence of specific words (or groups of words) within text files. In addition to simply finding words, full-text search systems can locate strings in logical or physical relation to each other in a file.

The logical relationships can be specified by Boolean logic expressions when formulating the search condition (e.g. "Find places in the text where 'hot' and 'cold' occur within the same paragraph"). These expressions need not be directly created by the user, but in order to maximize ease of use, they may be constructed by the system from information prompted from the user.

Full-text search capabilities may be implemented in a system such as LSS via either software or hardware techniques. Software techniques create a special file of sorted pointers (called an inverted index) that contain each word in the data base (excluding a set of pre-defined "stop" or trivial words) and their location in the text. In responding to a search, the full-text software uses these locations to respond to the user, performing logical operations on information in the inverted index to respond to

queries containing logical constraints. Response time in locating a word can therefore be appreciably shorter than finding the response set to a complex Boolean search. (See Section 2.1.4.1 for a discussion of the creation of the inverted index and Section 2.2.2.1.2 for inverted index storage requirements).

Hardware implementations of full-text searching do not use inverted indices, but rather operate directly on the text data base. During a query, the portion of the data base being searched is streamed past a series of comparators that evaluates it with respect to the query. Matches are recorded and reported to the user. Hardware implementations are also able to perform a Boolean search, and are similarly slowed by logical evaluation over simple location queries.

Hardware full-text searching requires specialized hardware, is a somewhat newer technology and therefore has associated with it a greater development risk. A detailed evaluation of the relative merits and the selection of either a hardware or a software implementation of the required LSS full-text search capability is required during the detailed design. These alternatives are considered significant variants in the conceptual design presented in this report and will also be evaluated in the forthcoming Benefit-Cost Analysis.

2.4 Output of LSS Information

The Preliminary Needs Analysis established the requirements for information to be provided to the user in two basic forms. First is information that is made available to the user in electronic form through an interactive workstation including video displays, printers, local storage, and other capabilities. Second, users require a hardcopy of the document that is faithful in appearance to the original. Perceived response time requirements for both forms are given.

The Preliminary Needs Analysis established several requirements for the ability to interactively query the LSS for information and for the supply of information to the user to be in a form that can best aid the users in performing their job functions. These include, for example, a list of documents that match a query, the text of a specific document, summary reports, and copies of documents representing the "original" format. Most of these requirements will be met through the design of the user workstation, the general term used for the user/machine interface with the LSS.

2.4.1 LSS Output

2.4.1.1 Workstations

The ability to communicate effectively with the LSS from a local workstation, incorporating the major functions that have been identified in the Preliminary Needs Analysis, requires the minimum hardware configuration of a basic personal computer. This level of capability has therefore been defined as the basic or Level 1 workstation configuration. The high

availability, relatively low cost, and the fact that many users may already have this hardware in use are factors which lead to this conclusion.

The functional capability to view electronic (bit-mapped) images on the screen at the workstation requires enhanced capabilities to decompress image files, display high resolution images on a screen, temporarily store images, print images, and provide higher throughput communications with the host. Since this capability was not identified as necessary to all potential users, the enhanced workstation (designated Level 2) is not required for LSS access. The summary descriptions of these workstations and their associated capabilities are compared in Table 1.

Since the user will require interactive communication with the LSS, an alphanumeric video screen with keyboard is assumed to be the basic communication device, due to its wide acceptance in the computer industry. This does not rule out the assistance of other devices such as a "mouse", however, since the LSS is a textual data base, a keyboard is a necessity. This will meet most of the functional requirements of entering queries, receiving responses, and viewing document text that are required to determine the relevance of a document to the user's needs. A minimum definition of a keyboard for the workstation would be a standard typewriter QWERTY key arrangement with upper and lower case, shift, return or enter key, and cursor control arrows.

TABLE 1. COMPARISON OF LEVEL 1 AND LEVEL 2 USER WORKSTATIONS

	Level 1	Level 2
Description	Personal Computer Floppy disk and/or hard disk* Monochrome text monitor Keyboard Dot-Matrix printer* Modem or Network connection	Personal Computer Hard disk High-resolution full-page monitor Keyboard Laser printer* Network connection Decompression board Mouse
Functions	E-Mail: send and receive, upload/download Data: queries and reports, view ASCII text, download data, request hardcopy.	E-Mail: send and receive, upload/download Data: same as Level 1, plus ability to view, download and print images

* Optional

For Level 1 workstations (without the capability of displaying electronic images), the minimum video screen is an alphanumeric (text mode) monochrome display capable of providing an 80 column by 25 line display. For the Level 2 workstations (with image and full page ASCII text display capability), the preferred system would be a high resolution graphical display in a landscape format, capable of displaying a text display along side of an image display or two image displays simultaneously. Since the image will be stored in a compressed form, the decompression board (assuming decompression is part of the workstation function) must be compatible with the compression techniques used.

Print capability at a workstation will be similarly dependent on the image display (and thus, image reception) capability. For the text-only workstation, a local print capability of a dot matrix (or impact) printer will be sufficient. For those workstations with image display capability, a laser printer would be required, although it would be sufficient to share a printer among several terminals located in the same work area.

The capability to download data from the primary LSS data storage to the LSS local workstation implies an intelligent terminal with capabilities similar to a personal computer, *i.e.* the workstation must have local storage sufficient to receive the data and file transfer capability to put the data into a transportable format (electronic re-transmission or floppy disk for example) or be able to process the data locally (word processing, for example). This function therefore could be met by defining the LSS workstation to have the basic capabilities of a personal computer. Such a definition is not inconsistent with the keyboard, display, and print capabilities noted above.

2.4.1.2 Printers

The LSS user requires that a hardcopy of selected documents be made available to them for personal use. If the document has existed in paper form (*i.e.* not originally submitted in electronic form), then the hardcopy should be made from an image of the original paper form. This function can be accomplished in one of two ways. Each (Level 2) workstation could include a laser printer, and the electronic images of the document could be transmitted from the LSS image storage system to the workstation for printing. Alternatively, the hardcopy could be made at the locations where the image is stored and transmitted to the user by express delivery. Considering the difficulties presented by the transmission of images to all workstations (see Section 2.5.3), some combination of these alternatives appears reasonable (*e.g.*, the use of express delivery for large documents and for all documents to locations without image/laser printer capabilities).

2.4.2 Types of LSS Output

2.4.2.1 ASCII

It is clear that for an alternative workstation to permit user queries to a remote textual data base, a minimum capability of transmitting and

receiving ASCII format data is required. With the additional requirement for downloading of ASCII files, a workstation based on a personal computer would provide these capabilities.

2.4.2.2 Image

The requirement for electronic images of documents was identified in the Preliminary Needs Analysis, at least for those pages of information that are not capable of being translated into ASCII such as maps, figures and other such graphic information. Similarly, if the format of the ASCII version of a page is sufficiently different from the original (because of the displacement of footnotes or page numbers, for example) for a comparison to be useful, an image of the text pages can also be needed. Since the requirement has been expressed by potential users as well as by the NRAC for storage of images by the LSS (to produce hardcopy), the question remains only to determine the method for making these images available at the workstation. Alternatives are:

- 1) Store the image in electronic form (or a form readily converted to electronic form) and transmit the image to the workstation as requested.
- 2) Store the image in electronic form on a reproducible media such as CD-ROM, distribute the media, and provide the capability at the workstation to locate and display the image.
- 3) Store the image on a reproducible media such as microform and distribute the media along with a location index and a reader.

In any event, it is clear that not all users of the LSS require that images be available to them at their workstation, thus enabling the definition of workstations with and without image capability, and providing only a fraction of the workstations at a facility with image capability.

With the anticipated data base of the LSS records extending to tens of millions of pages, the distribution and local storage of all images at the workstation either in electronic form or microform, becomes a major operational problem. For example, at current densities of storage the LSS administrator would be distributing 2 to 3 CD-ROM disks per day to each user. Therefore, local distribution of images is viable only if those images are limited to pages or portions of pages that are unsuitable for conversion into ASCII format. This solution presents the difficult configuration management problem of separating images of graphics from the parent document and ensuring that the dislocated pages are correctly filed and distributed.

The magnitude of the problem is a function of the percentage of documents which contain graphics and the percentage of pages within those documents which are graphic; neither of these two percentages have yet been defined to an accuracy that one can assess the magnitude of the problem in detail. The conservative assumption for the conceptual design was therefore made that images at all of the workstation will be provided over a communications link from a central image storage location. Since all images

will be available at this location, it is not necessary to differentiate between images of graphics and images of text pages. An alternative is also considered where there would be no images provided directly to workstations. The actual design will probably fall between these extremes, i.e. images may be provided to some but not to all workstations.

2.5 Communication of LSS Information

The Preliminary Needs Analysis estimates that the LSS will need to support access to the LSS from multiple sites, in over a dozen cities throughout the U.S. The majority of LSS workstations are expected to be in the Washington, DC and Las Vegas areas.

In addition to the explicit communications requirements associated with connecting workstations to the LSS, several implicit communications requirements also arise from potential system designs.

2.5.1 Node to Node Communication

If the LSS computer system is not centralized at a single site, there would be a requirement for communication between all of the LSS computer processing subsystems (or nodes). There is a wide range of possibilities for communicating information from node to node. This range includes:

- 1) A dedicated, real-time communication facility
- 2) A shared, batch communication facility
- 3) An overnight, courier delivery service.

The main determinants on which communication environment would best meet the LSS requirements are the time-sensitivity of LSS information, geographic proximity of LSS processing nodes, and the LSS data base structure and requirements.

If either of the electronic communication environments are required, each node site would require some communication equipment. Each node would have to dedicate physical ports on their data processing hardware for communication. These ports would need to be locally connected to a communication signaling device (such as a modem or multiplexer). The signaling devices require a communication circuit (such as a telephone line) which extends between the two node locations.

In order to reliably communicate electronic information, each LSS node must be capable of supporting a defined set of rules, i.e. a protocol. The protocol establishes the vehicle for the flow of information between LSS node sites. Various protocols exist for various data processing applications. The appropriate protocol must be supported by the various vendors' hardware and software that comprise an LSS node site.

2.5.2 Capture Station to Storage Communications

If the LSS capture station is not physically co-located with the LSS data storage facility, then there will be a need to transmit the captured data to the storage facility. Like the node-to-node communication environment, the capture-to-storage process also has three alternate configurations:

- 1) A dedicated, real-time communication facility
- 2) A shared, batch communication facility
- 3) An overnight, courier delivery service.

If either of the electronic alternatives are chosen, there are certain generic requirements for communication. The capture workstation must have an interim data storage device that serves as a buffer and back-up mechanism. This data storage device could either be connected to a communication device directly or it could transmit its information through a communication device associated with the capture station. These local communication requirements (capture device to interim storage device) would be dictated by the vendor of the capture device.

To move the information from the interim to the permanent storage location would require a physical port on each device. These ports would have to be connected to a communication signaling device (such as a modem or multiplexer). The signaling devices require a communication circuit (such as a telephone line) that extends between each capture node to the data processing node. The speed and type of these connections depend on the time-sensitivity of the data.

Overnight mail of high density storage media (such as magnetic tape for ASCII files or optical disks for images) can well be a viable communications mode for sending information from a capture station that is not near an LSS node. This requires compatible capability for reading and writing such disks at both systems.

2.5.3 Node to Workstation Communications

The function of this communication is to provide reliable access to the LSS data base with the added features of downloading data to local workstations and possibly displaying or outputting bit-mapped images locally. Because of the concentration of LSS users in three major geographic areas (Washington, DC, Las Vegas, NV, San Antonio, TX, according to the Preliminary Needs Analysis) the communication requirements should be modeled to support intensive communication from these sites to the processing or host site or sites.

In an effort to pool LSS resources at the seven most concentrated usage locations (White Flint, Forrestal, Weston, M&O Contractor, San Antonio, Las Vegas and Carson City), local area networks (LAN) are recommended. LANs allow users to share common equipment such as laser printers, personal computer (PC) storage devices and communication facilities. Each PC or associated device at each site would require access to the LAN. The access usually entails a physical communication port (or card) and a communication

circuit (or cable). The LAN would benefit the user most by having bridged access to the remote LSS host location. This form of remote connectivity allows a LAN user to request large LSS files to be printed or stored at any device attached to their LAN. This offloading of data from the PC maximizes the communication between sites, provides timely/efficient print and storage functions and minimizes the investments in individual workstations while enhancing the capabilities of the total LAN user group.

Once LANs are established at the major usage locations, the shared communication circuit (or telephone line) to the host sites would be sized according to the anticipated traffic of the entire LAN user group. The rate or line speed of the communication circuit is variable and could be sized according to the unique demands of each site. If LSS users and a capture station were co-located, and the capture station was electronically connected to an LSS node, then each application (access and capture) could share a common communication facility to the host site.

For LSS users not located at the seven major usage locations, basic workstation access needs to provide terminal to host communication. Each workstation would need access to the LSS data base and some workstations would need the capability to download files or view images locally. Each workstation would require a physical communication port, communication device, communication software and communication circuit. The specific type of each of these four functions would be workstation dependent. The major variant is the rate at which communication will occur which provides the necessary response time for the desired application (such as viewing images). The most reliable way to provide terminal access to an LSS user would be to extend a dedicated communication circuit from the remote location (e.g. Lawrence Livermore National Laboratory) to the nearest LSS LAN site (e.g. Las Vegas, NV). This dedicated circuit would allow the remote site to share the larger, more economic communication circuit between the LAN site and the host site. For users only requiring limited access to the LSS data base (such as those needing ASCII output only), a dial-up communication circuit at a reasonably high data rate (such as 9.6 Kbps) should suffice most applications.

2.6 Electronic Mail

The Preliminary Needs Analysis found that an electronic mail capability within LSS may be needed. It appears that the NRAC will find that the LSS should be able to accept input from such a system. When this feature is used to electronically file briefs, for example, the submitter should be able to certify that the material captured by the system conforms with what was sent.

2.6.1 Conventional E-Mail Functions

The following LSS functional capabilities can be met with currently existing electronic mail software and therefore require little or no customization.

2.6.1.1 Uploading Files from a Workstation

Since many of the hearing related documents will be multipage messages, it will be more efficient to provide the capability of uploading message text files to the LSS which can then be transferred to the E-mail function (rather than requiring the message to be keyed in on-line). This will allow the user to prepare the text of the message on office word processing equipment "off-line" to the LSS. For consistency, the format for acceptance of the message file into the LSS E-mail system should be the same as the format for word processing files submitted for the LSS records.

2.6.1.2 Message Transmission and Receipt

The filing of hearing related documents will require certain common E-mail functions to be available. Since messages will normally be sent to multiple parties, the system should provide for the ability to send messages to a distribution list. To assure completion of the process, both delivery receipt (notification to sender that the message is in the receiver's "mailbox") and read receipt (notification to sender that someone with the receiver's access code has read the message) functions are required. To provide a more positive indication, a receipt acknowledgment message capability initiated by the receiver to the sender may be provided. Date and time "stamping" of the message will automatically be provided so that the receiver will know how long the message has been available in his mailbox.

2.6.1.3 Privacy and Authenticity

The E-mail system (as well as the LSS in general) can be provided with password security access and identification which will provide privacy on the contents of a mailbox as well as an authentication of the sender's identity.

2.6.2 Special LSS E-Mail Functions

Hearing related documents such as motions and pleadings which are sent out on the E-mail system must be sent as well to the LSS records data base for input. While inclusion of the document could be made automatically (assuming the E-mail function is integral to the LSS), for reasons of quality assurance and management it is preferable to designate one of the capture workstations as an "LSS mailbox" to receive copies of the messages. These can then be processed through the cataloging and capture process in the same manner as hardcopy documents or word processing files. The submitter can be allowed to verify the correctness of the captured file and certify it if required.

2.7 LSS Management Functions

The need for a rigorous LSS quality control process for both data in the LSS and the system itself is concluded in the Preliminary Needs Analysis

and is expected to be required by the NRAC. Performance and usage monitoring (required by the statement of work of the LSS Design and Implementation contract, DOE, 1987) is also identified as necessary for providing the data for system optimization.

2.7.1 Performance and Usage Monitoring

To obtain the best performance possible, the LSS includes hardware and software which enable the LSS administrator to monitor its performance. The performance information provided will assist the LSS operators in determining what immediate and future adjustments would be needed to improve performance. The following performance monitoring tools will be included in the system:

Telecommunications - equipment and software which measures and records the transmission loads and data errors on each link in the network as well as equipment malfunctions.

Hardware - software that measures and records the input/output and computation loads on all computers, the loads placed on the magnetic and optical disk storage devices, and the amount of storage fragmentation of the data base.

Software - special software will sample and record the data entry loads and the types of user queries being made.

Analysis of the monitoring data obtained will assist in the identification of hardware needing maintenance, planning maintenance schedules, changing communication routings, relocation of cluster hardware and possible software or data base architecture changes to better support the types of queries being experienced.

2.7.2 LSS System Administration

2.7.2.1 LSS Quality Control

The quality of the LSS will be determined by the accuracy and completeness of its contents, the quality of the cataloging processes, the richness and accuracy of the thesaurus and other retrieval aids, and the reliability of the hardware and software components. To ensure a quality system is maintained, a quality control process should be established that sets standards, develops policies and audits the system's quality. The performance monitoring tools described in Section 2.7.1, the capture quality control tools described in Section 2.1.3, and the configuration management tools described in Section 2.7.2.2 can also assist the quality monitoring function.

2.7.2.2 LSS Configuration Management

Knowing precisely what hardware and software comprise the LSS is required to efficiently manage its operation. To support this need, the LSS

design must include configuration management software as tools for use by the LSS administrator. A LSS hardware configuration management software package should include the following:

- 1) A current listing of the LSS system's hardware configuration, including component vendor and part number
- 2) A listing of the spare parts inventory of each hardware component
- 3) A listing of and status of each hardware trouble report, which documents any hardware malfunction requiring maintenance.

Software configuration management tools should be provided or developed for the following:

- 1) Operating system software generation and update capability to generate and maintain the operating system and utilities to the current, approved version level
- 2) Software for generating, updating, and controlling the LSS application software consistent with the requirements of the configuration management program
- 3) A listing of and current status of any software Trouble Reports, which document any software system error that requires maintenance
- 4) Software for data base management and control including data base backup and recovery and for the periodic generation of data base copies to be stored off-site to prevent loss of the data base due to fire or some other catastrophic occurrence.

2.7.3 Data Base Administration

2.7.3.1 Data Base Maintenance

Data base maintenance for LSS encompasses numerous activities that are critical to system reliability, the usability of the LSS data, and to user confidence in the LSS. Data base maintenance includes activities such as:

- 1) Appending new data to the LSS
- 2) Correcting erroneous entries
- 3) Restructuring to consolidate fragmented space
- 4) Maintaining the LSS data dictionary containing
 - Descriptions of the data structure and of schema
 - A glossary defining all terms in the controlled vocabulary used for cataloging entries.

The above activities require both supporting software and organizational procedures for executing the actions at routine intervals, defined according to calendar time or specific data base activities (e.g., before and after

new data is appended to a data base). The data base administrator should keep a log of all maintenance activities performed on the LSS.

2.7.3.2 Loss Protection

The data base administrator will perform the following activities on a regular basis:

- 1) Routine backups of the data
- 2) Execute recovery procedures when required
- 3) Storage and maintenance of backups, both on-site and off-site
- 4) Execute various check routines to ensure the physical integrity of the data base.

2.7.3.3 Access Control

LSS users will have one more of the following privileges:

- 1) Search and retrieve all data in the LSS
- 2) Search and retrieve all header data in the LSS
- 3) Search and retrieve data in the Tracking Subsystems
- 4) Update authorization for the Tracking Subsystems
- 5) Update authorization to the Records and Regulations Access Subsystems.

User names and passwords will be used to enforce privileges; the system administrator will add and delete passwords for individual users.

3.0 CONCEPT OF OPERATION: DATA CAPTURE AND DATA RETRIEVAL

The preceding section has given a detailed description of the functions needed by LSS users. The various ways in which these functions can be implemented cannot be indiscriminantly combined into a single system. This section presents an operational concept of LSS that integrates these functions effectively to satisfy the requirements identified. The concept of operation presented here is subdivided into the two main operational areas of the LSS: information capture (Section 3.1) and retrieval (Section 3.2). The base conceptual design presented in Section 4.1 represents a possible hardware and software environment that supports the operations described in this section.

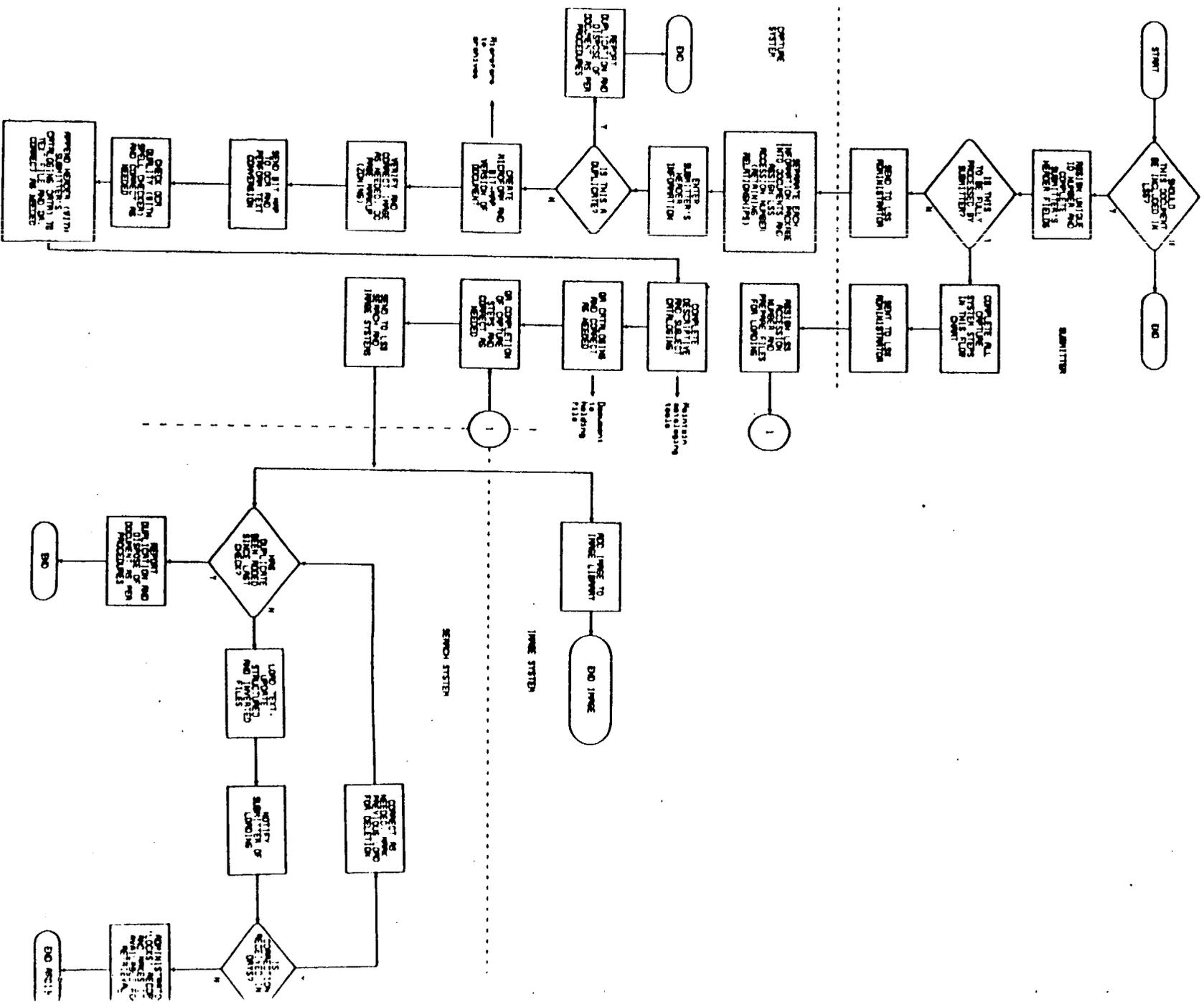
3.1 Document Preparation and Data Capture

The document flow described here is tied to the Base Conceptual Design. In this concept of operation, items will be processed on a single pass at a capture site. Moreover, this operation is based on the assumption that technology advances anticipated in the next one to two years will be incorporated into the document preparation system. Such advances include the use of hand-held scanners to capture descriptive cataloging data and the development of software to assist in the document cataloging process. In the event that a different document capture or preparation process is adopted (for example, a plan to defer conversion from image to text until some later time), the integrated process described below would change significantly.

The LSS administrator's responsibility for documents begins when the document is received at the capture site. Note that the submitter has the option of fully processing the document. In this case all the capture steps described below, are performed by the submitter. The administrator needs only QA check the material and add the submitted material to the normal capture system information flow. Several stages of document processing must be completed before the document is available to LSS users. Figure 1 depicts this process and associate flows. The processing tasks require both manual and automated support, as described in the following paragraphs.

In this concept of operation, items from authorized submitters arrive at LSS administration as "information packages" that contain one or more related items of information such as printed matter, maps, and computer printouts. Each information package is assigned a unique identification code number, as are the components of the package. This number will also reflect the relationships among the components of the package. The date of receipt and the data supplied by the submitter for each component, along with the information package association, is recorded. The unit of processing from this point on is the individual numbered item.

FIGURE 1
 PROCESS FLOW IN THE CONCEPT OF OPERATION FOR LSS DATA CAPTURE



The item is first checked against the data base to determine whether it is a possible duplicate of a previously-received item. If a candidate duplicate is found, a cataloger checks both items to determine whether they are, in fact, duplicates. If they are identical, processing of the newly-received item is terminated. Disposition of the item is recorded in the data base and reported to the submitter. If the item is not a duplicate, processing continues. However, if the item has an important physical or logical relationship to another item in the data base (such as a copy of a document with extensive marginal notes added), the association will be recorded in the catalog data for the item.

Items next proceed to microform conversion and scanning (to create the required microform archive copy and the bit-map copy needed for OCR conversion). The images are verified for skew, inversion, page order and completeness. If a page is missing from a document, a substitute page containing a message to that effect will be inserted. Tools will be available for scanning operators to insert, delete and correct page order.

The next step is to mark up or (zone) the image for text conversion. The mark-up, performed with a mouse, will delineate formulas, figures and graphs not suitable for conversion. For example, if captions are converted to text, captions will be marked so that they can be readily identified after conversion has occurred. In the same manner, abstracts, footnotes, and any other portions of text that need to be explicitly identifiable to text users, will be marked.

If the item is a candidate for automatic text conversion, the image files are dispatched for batch OCR processing. If the item is not suitable for automatic text conversion (e.g., if it is a handwritten document), it is so marked and sent directly to cataloging. The decision to key in the contents of the document is made at this point.

After text conversion the text file is spell checked. Spell checking provides a rapid way to assess the accuracy of the OCR process and to help facilitate correction. Spelling correction will be done by an editor.

The next step is to complete the catalog information. At the cataloging workstation, the cataloger will have access to both the electronic and the hardcopy version of each item. Menus, on-line thesauri, and other manual and automated tools will be available for catalogers to select appropriate values for each header field. The software will check spelling and legal values for catalog data and will verify that all required fields are populated. Any anomalies are marked, and alternative spellings or values will be presented where feasible.

The file is then presented to the quality control (QC) cataloger who will review the correctness of the cataloging data and change any that is incomplete or incorrect. The QC cataloger will also have both electronic and hardcopy versions of the item for these activities. As terms are identified that need to be added to the lists of valid spellings and legal values (e.g., a new term for the thesaurus), the QC cataloger will propose such changes to a technical overview group responsible for approving changes. Once approved, changes will be implemented by a data administrator.

A software verification of the new catalog data is performed. If the file is verified, it is ready for loading. If errors are found, the file is returned to the QC cataloger for correction and re-verification. Since image files, header files, and text files associated with a single item will be added to the LSS data bases as one task, it is also the responsibility of the QC cataloger to ensure that these linked files are consistent, i.e., that they are all identified with the same item. Index creation for text and header data occurs as part of the loading process.

Prior to loading, the system will again be checked for duplicates to ensure that a duplicate item has not been entered during the time the current item was in process. If a candidate duplicate is found, a cataloger checks both items to determine if they are identical. If the items are indeed identical, processing is terminated and disposition of the newly processed item is recorded in the data base and reported to the submitter.

After the files have been loaded, the submitter will be notified. The submitter has five days in which to identify any corrections to the LSS administrator. If corrections are submitted, they will be made in accordance with the appropriate procedures. If no corrections have been received, the LSS administrator locks the record against any changes. The record is then available for full access.

Backups of items in preparation will be performed on a regular basis. Backups of image files will be retained until the file has been backed up as part of the permanent LSS data base. Backups of fully-processed versions of text and header files will be retained until the document has been backed up as part of the permanent LSS data base.

3.2 Data Retrieval Operations

For the LSS to be successful, the operational system that is implemented from the design must meet the needs of each of the user classes identified in the Preliminary Needs Analysis. This section discusses the retrieval operations characteristics of the Base Conceptual Design.

3.2.1 Access

The LSS administrator issues users' names and passwords and assigns privileges such as which LSS subsystems each user may access.

To obtain access for a particular session, a user goes to the location of the nearest workstation. For those in one of the primary LSS user locations (White Flint, NRC; Forrestal, DOE Headquarters; M&O contractor; Las Vegas, NNWSI; Carson City, State of Nevada; or San Antonio, FFRDC), Level 2 workstations, tied directly to the dedicated LSS communication network, will be available in the offices of those whose jobs involve frequent use of the LSS. For users who have less frequent need to use the LSS, areas with both Level 1 and Level 2 workstations, also tied to the dedicated LSS communications network, will be established. Users not in

those locations will access the LSS using Level 1 workstations and dial up communications.

The Issues and Commitment Tracking Subsystems will be updated on-line as new issues or commitments are defined and as status information is entered. For the Records and Regulations Access Subsystems, new documents are processed daily as described in Section 3.1, however, access to the recently processed records would be provided on a periodic basis such as every two weeks or once a month. This will ensure that searches and retrievals performed during these periods work with the same data base contents. Upon entry into the subsystem a summary of the newly available data will be displayed.

The LSS would be available for access from 7:00 am to 11:59 p.m. Eastern Standard Time every day except holidays. This will support the vast majority of the access periods identified in the Preliminary Needs Analysis.

3.2.2 Query Operations

Once access been obtained to the LSS and either the Records or Regulations Access Subsystem has been chosen, a user has a choice of four styles of user interfaces for formulating queries:

- 1) Menus - which allow the selection of specific, predetermined queries from a list
- 2) Query screens - which provide fill-in-the-blank type of screens to specify which fields should be searched for what terms. Windows which can be used to display the legal values for a field or expand the contents will likely be incorporated.
- 3) Prompting dialog - in which the search software "converses" with the user via a series of questions to construct a query
- 4) Query language of the data base software - which provides the most flexibility, but is only for trained, experienced and frequent users, such as intermediaries.

Both the headers and full-text data can be searched using any of the four interface styles. Each query may involve specifying one or more header fields such as subject, title or date; one or more phrases, or two terms within a certain proximity, to be searched in the full text. Results of a query may be combined with earlier query results of either headers or full text using the Boolean operators of OR, AND, and NOT.

The Tracking subsystems' data would be stored as structured data, similar to the headers in the Records and Regulations Access Subsystems, thus full-text searches would not be applicable. Also, menus, query screens and query language of the data base software will be part of the Tracking subsystems.

Upon the completion of each query, the results would be displayed showing the number of instances in which the search criteria have been met and the number of documents involved.

Assistance for the users would be available via an on-line facility as well as through user guides and manuals.

3.2.3 Retrieval Operations

The Base Conceptual Design supports retrieving technical data either as on-line displays or in the form of hardcopy. Special policies and procedures would govern retrieval of non-documents in the archives.

3.2.3.1 Headers

The on-line display of headers would be via formatted screens with each file labeled and its values shown. For some large fields, only a portion of the contents might be displayed. To view all of the contents, the user may scroll through the field's "window", or may "zoom" the field to cover most of the space on the screen.

Headers would be displayed for viewing one at a time, but a user could easily browse through a result set, either forwards or backwards. The user may specify how the result set is to be ordered, before it is viewed. Header information relating to a result set could also be viewed in a tabular listing, with the columns being selected header fields and each row the field contents for headers in the result set.

3.2.3.2 Full Text

In the Records and Regulations Access Subsystems, full text could be displayed. On Level 2 workstations, the text would be viewed in full page mode, formatted similar to its appearance in the original document. Exceptions would be that all text would be displayed in a single size and style, and figures, graphics and formulas would be replaced with a note. To see an exact copy, an image display or hardcopy would be requested. The display of text on Level 1 workstations would be a partial page of up to 24 lines. The ability to scan up and down a page would be provided.

Text, like the headers, could also be browsed from page to page or possibly from document to document. After a result set of documents and their associated full text has been established by a search, the documents in the results could be ordered before browsing. Header fields could be specified to determine the order.

The Base Conceptual Design provides for downloading an ASCII file, containing the text associated with documents in a result set for processing in local workstation mode.

3.2.3.3 Images

The Base Conceptual Design provides for the displaying of Records and Regulations Access Subsystem pages in the form of images on Level 2 workstations. The images associated with a result set could be browsed similar to the manner in which full text is browsed. Images could also be displayed one at a time by specifying the document and page number or by requesting the image associated with the full text ASCII page currently displayed. Level 1 workstations will not have an image display capability.

All LSS users would be able to obtain images in the form of hardcopy. Users at Level 2 workstations would be able to print images at the workstation. There will likely be a per request limit on the number of images that may be printed at the workstation. Large volume and all Level 1 workstation print requests would be routed to the image system for printing and priority shipping.

3.2.4 User Session

In any one use of the LSS, users would combine the query and retrieval operations, described above, to best obtain the information they desire. To estimate the communication, search and retrieval loads that are likely to be placed on the LSS, a series of user session scenarios specifying how search and retrieval activities like those presented in Table 2 will be developed and quantified. An example scenario is presented in Appendix A.

TABLE 2. USER SEARCH & RETRIEVAL ACTIVITIES

Header searches:	Searches on the header or catalog descriptions of documents and non-document material
Return number of hits:	The results displayed are only a report of the number of instances the search criteria were met and of the number of documents in which hits were found
Return header:	The user chooses to have header data displayed (as opposed to a report of the number of hits or to seeing the document text)
Return text:	The user asks to see the text of the document, rather than just the header data.
Image browsing:	The user asks to see page images of the document, in addition to the document text.
Full text browsing:	The universe in which searches are performed includes the full text of documents, rather than just the header or catalog data
Local print requests:	Requests to have material printed at the local site
Printed page images:	Printing image versions of document pages
Print Header information:	Printing header or catalog data
E-mail messages:	Electronic mail messages sent during the session

4.0 CONCEPTUAL DESIGN

Now that a preliminary requirements analysis for the LSS has been completed, consisting of the first two reports in this series, it is appropriate to focus on possible design concepts to implement the requirements identified. No single optimum design exists. Rather, a variety of basic concepts and myriad variations on those concepts can be acceptable. The object of this analysis is to focus on one family of high-level designs, with low development risk (i.e., very likely to meet all the requirements and very likely to be technically achievable). The forthcoming Benefit-Cost Analysis will evaluate the financial feasibility of the family of design developed here.

The conceptual design outlined here is a high-level system design, identifying the major LSS functions, hardware and software subsystems, and subsystem interfaces. It represents the consensus of a team of specialists in information management, software design and development, systems integration, communications, image storage and retrieval, nuclear regulatory development and compliance, and nuclear waste management. The high-level design developed by this team presents the best system (i.e., lowest development risk) to implement the identified LSS requirements and is presented as a first step in evolving a detailed design for the LSS. It provides a basis for further review and discussion.

The basic design concept that has been developed was not selected from an exhaustive list of potentially viable alternatives. Rather, after an examination of previous work on LSS and of similar systems by the design team, a concept perceived to have very low development risk was selected. Other alternatives were not considered in detail. For example, in the selection processes one potentially viable alternative that was considered but not developed involves replicating the LSS data base (in totality, or in part - such as text only, or images only) at each user workstation. This concept requires the regular copying and distribution of new material to maintain and keep consistent over one hundred copies of the LSS data base. Because of the very large size of the data base and the extremely difficult data configuration management problem that such a distributed system would pose, the design team preferred a base design with lower development and operational risk.

The conceptual design outlined here is a concept, having features which can be implemented in a number of viable ways. Not all of the options described are compatible or have the same functionality. The approach taken during this analysis was to develop the base conceptual design and a number of internally consistent variants. Each variant considered introduces a different way to implement a feature of the system and identifies the repercussions of this choice on the other system features and functions. The forthcoming Benefit-Cost Analysis will evaluate the financial feasibility of the family of designs developed here and provide a framework for evaluating the trade-offs between the costs and technical risks versus the amount of functionality offered by each design option.

4.1 Base Conceptual Design

4.1.1 Base Conceptual Design Hardware

The Base Conceptual Design hardware architecture, is illustrated in Figure 2, and consists of several replicated capture systems, the main computing and data storage capacity (called the search system), the image system, communications system and workstations. The following sections describe the hardware and software that constitute this Base Conceptual Design.

4.1.1.1 Capture System

A document capture system consists of a computer system and attached (local) terminals for control of the document capture process, cataloging and correction (if not done off-line). The computer system interfaces and controls, via software, the peripheral devices required for the scanning, and optical character recognition processes needed for the capture and quality control of the document images and ASCII text.

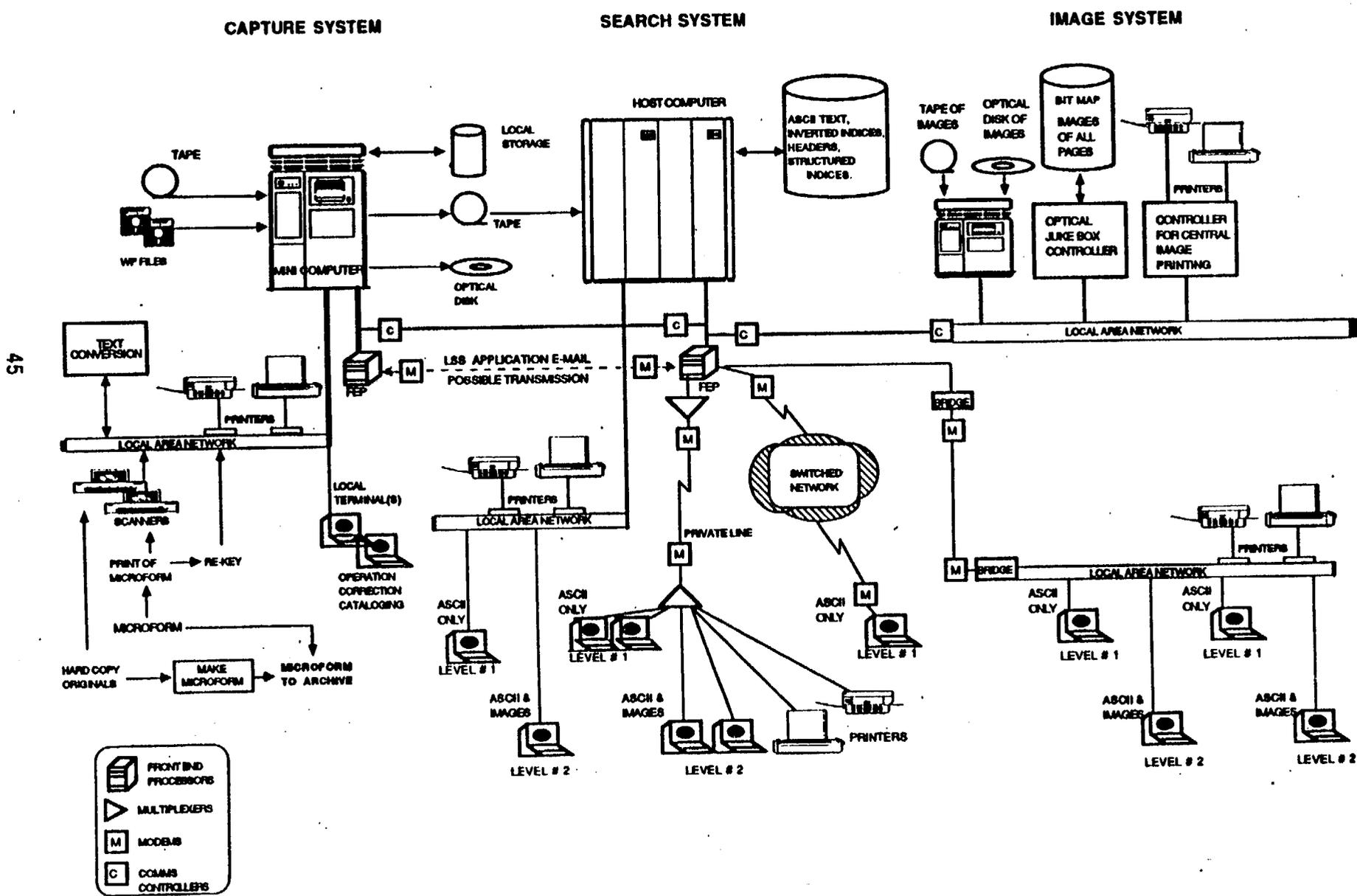
Local terminals attached to the computer system, with appropriate software (see Section 4.1.2.1), are used for:

- 1) Validation and quality control of the digital scanning process, which will capture and store bit-mapped images
- 2) Validation and quality control of the microform conversion process, which will read the microform media and either produce a hardcopy of the image page or produce an ASCII text record of the image page
- 3) Validation and quality control of the OCR process, which will produce an ASCII text record of a bit-mapped image
- 4) Input and validation of documents by re-keying the text of poor quality documents such as those produced from microform or deteriorated hardcopy
- 5) Input and validation of cataloging information
- 6) Input, validation, and cataloging of messages from the LSS electronic mail system.

The document capture center's peripheral devices include:

- 1) A high speed digital scanner
- 2) A text image recognition and conversion processor

FIGURE 2
BASE CONCEPTUAL DESIGN HARDWARE ARCHITECTURE



45

- 3) A microform reader and conversion device for producing hardcopy and/or ASCII text records of microform images
- 4) Magnetic disk drives for temporary storage
- 5) Optical disk drives
- 6) Magnetic tape drives for the storage and distribution of captured data and for input of documents in machine readable format, which have been prepared by word processing software
- 7) Magnetic floppy diskette drives for input of documents in machine readable format, which have been prepared by word processing software.

The digital scanning process should provide the following minimum capabilities:

- 1) An automatic document feed capability of 30 to 50 document pages of up to 8.5 x 14 inches in size.
- 2) Capture each document page as a bit-mapped image with a minimum of 300 dpi (dots per inch) resolution.
- 3) Perform CCITT Group IV image compression with operator selectable compression ratios.
- 4) Store the bit-mapped image on magnetic disk coded with an accession number for later retrieval for quality control processing and optical character recognition processing.

As discussed in Section 2.1.2, the capture of material into LSS is to be accomplished at several capture centers, some of which may be co-located with the LSS node.

An electronic image of a document page from hardcopy is captured by a process of feeding document pages into a digital scanning device, checking the resultant image, and entering an accession number from a terminal keyboard for identification of the document. As the document image is captured, the digitized bit-mapped image is displayed on a high resolution page-oriented monitor capable of at least 150 dpi resolution used to verify the quality of the scanned image and allow for re-scanning if the image does not meet the quality control requirements. The image is a replica of the original, including margin notes, signatures, graphics, date stamp, etc. which can not be captured in ASCII form. Images are the only reasonable method of capturing graphic oriented document pages.

Although images are electronic, the characters or words on the page cannot be recognized by a computer until the image is processed by optical character recognition. Electronic images require relatively large amounts of storage, typically 50,000 to 70,000 bytes (compressed mode) per 8 1/2 x 11 inch page, as compared to ASCII at 2500 to 3000 bytes per page. Thus the use of images require high density storage devices such as optical disks.

Documents which have been prepared on a computer by word processing software are already in machine readable format. However, because most full-text retrieval programs require that files be entered in ASCII form and because the variety of word processors systems in use by Government and industry are not standardized for computer communications of ASCII text, some conversion is required. Generally speaking, software tools are available to support this conversion. Most word processor systems provide the option for producing an ASCII file of the document text as an output. This file does not contain the special codes, for formatting and printing, that the word processing document file contains. However, even this ASCII file format is not standardized, and therefore a standard file format must be defined for input to the LSS. The Technical Staff of the NRAC is preparing such a standard, and the LSS should accept and store the ASCII text provided in the approved format.

Archive requirements dictate that all LSS documents must be microformed. The Base Conceptual Design meets this requirement by creating a microform copy of all LSS documents as part of the capture processes.

4.1.1.2 Search System

To the LSS user, the search system is the LSS. It is the single point of contact, supporting all of the LSS functions either directly or transparently through directions issued to the image system and the workstations.

The search system supports a myriad of functions including electronic mail, issues and commitment tracking, search and retrieval services for all document types, storage of all the regulations, header, and full text including their associated structured and inverted indices, update of the data bases as new documents are received from the capture system, downloading of ASCII text to the workstations, and routing of commands to the image system for off-line printing or on-line display of document images. All of these functions must be performed within reasonable response requirements such as those listed on Table 1, of the Preliminary Needs Analysis.

The base conceptual design hardware architecture has a single search system, co-located with the image system. The search system receives data from the capture system(s), which will or will not be co-located with the search system. The search system then updates the LSS data bases, including structured and inverted indices. LSS users are supported by two levels of intelligent workstations connected via the communications system.

The search system can consist of a large mainframe or a tightly coupled cluster of super-minicomputers. Either configuration will provide the computational, disk storage, and input/output transfer rates required. The system will be sized to handle the peak loads associated with a LSS user community of 225 to 475 people at the peak point in the licensing process, according to the Preliminary Needs Analysis. Disk requirements for the search system, estimated in Section 2.2.2, include all of the document types plus structured and inverted indices. The capacity and number of

input/output channels are critical for a number of users and the size of the data base that the search system must support. The search system will be designed for incremental growth and for technology insertion of faster processors, larger capacity disks, and higher input/output transfer rates. All configuration changes will be transparent to the users and the software.

Provisions will be made for additional central processors and disk drives beyond those necessary to support the on-line use of the system. These extra computer resources will enable the system operators to:

- 1) Update data bases 24 hours per day. When a data base is to be updated, it will be copied to the extra disks, updated by the extra processor(s) and the system switched to the updated copy. This ability will be crucial during the initial operation of the system when large numbers of documents are entered every day.
- 2) Allow continuous backup of the system without affecting on-line response. This need for backup is also driven by the high rate of change in the data bases.
- 3) Fully develop and test new system and applications software releases prior to going on-line with the LSS users.
- 4) Provide backup for failed components and to deal with unexpected surges in demand.

Beyond quantitative requirements such as the number of users to be supported are consideration of the effects that the operating system, commercial off-the-shelf, applications, and communications software have on the hardware. Software issues will be covered in Section 4.1.2.

The capabilities of the workstations used to access the search system must form a seamless interface with the search system. Workstation hardware and software are covered in Sections 4.1.1.5 and 4.1.2.5 respectively.

4.1.1.3 Image System

The image system stores compressed images of all documents in the LSS on optical disks in optical jukeboxes for on-line retrieval and display on Level 2 workstations or for off-line volume printing of documents via high speed laser printers. It is connected directly to the search system from which it receives commands. Output is routed directly to the workstations via the communications system.

The image system consists of three components that are interconnected by a local area network. The components are: (1) image preprocessor; (2) jukebox controller and jukebox storage unit(s); and (3) printer controller and high speed laser printer. Each of these is covered separately below.

The image preprocessor accepts compressed images and associated retrieval data (document accession number, date of issue, and number of pages) from the capture systems. The data can be supplied on magnetic tape, optical disks, or via the communications system. The image preprocessor

then performs any translations necessary and presents the image and associated retrieval data to the optical jukebox controller for permanent storage on one of the optical disks in the jukebox storage unit(s).

The jukebox controller and its storage unit(s) are the heart of the image system. The controller is a powerful minicomputer with many functions including:

- 1) Acceptance of new images from the the image preprocessor. This is a complex event requiring a check for duplication, calculation of the optimal optical platter and the location of that platter in the jukebox storage unit, transfer of the optical platter to a optical disk drive, writing the image to disk and a transaction log, and update of the controller's internal data base of documents and their storage location.
- 2) Servicing requests for image retrieval and transfer. Based on commands from the search system, the jukebox controller must check its internal data base to verify that it has the document, send a message to the search system if it does not, retrieve the correct optical disk and place it in an optical drive (unless the disk happens to already be in an optical drive), and then transfer the image to the printer controller or to a Level 2 workstation via the communication system. At the workstation the image is decompressed and displayed.
- 3) Maintaining statistics on usage of specific documents - not by which user, but by the number of times an optical disk or document is accessed. The controller will move optical disks within the jukebox(es) and will group documents often accessed in order to minimize retrieval times.

The printer controller and high speed laser printers print whole documents or pages of documents, plus mailing information, based on commands from the search system. After receiving the print command and associated data, the printer requests the document images from the jukebox controller, temporarily stores them at the printer controller, and then decompresses the images and prints them. The temporary storage is necessary to minimize the service time by the jukebox system and to allow special services such as the printing of multiple copies or on-the-fly reordering of printer priorities. For example, a long print request could be suspended while printing, a high priority short document could be printed and the long print request resumed.

4.1.1.4 Communications System

The LSS communication system must provide ASCII and image data transport at a speed consistent with the response time requirements stated in the Preliminary Needs Analysis. These needs are translated into medium (2.4 to 9.6 Kbps) and high speed (≥ 56 Kbps) data transport between the Level 1 and Level 2 workstations and the co-located search and image systems. The data transport speeds will be established based on the results of modeling efforts, which use parameters such as number of users and the

Level 1/Level 2 workstation mix at each site, access needs as a function of time and program schedule, and request service times by the search and image systems.

The key concept in the communications network topology is one of a unified computing resource. The diverse geographic locations can be viewed as different floors of the same large building. Every LSS user, regardless of location, will view the system as though directly connected to the search system. In the Base Conceptual Design, White Flint is taken to be the communication hub only for illustration. The hub will be co-located with the search and image systems. The concept is achievable through the use of five technologies: local area networks, intelligent bridges, high speed multiplexers, high speed modems for voice grade switched circuits, and intelligent communications processors. Figure 3 shows the network topology with Washington, DC as the location of the search and image systems. High concentrations of users in Washington, DC, Nevada and Texas (see Preliminary Needs Analysis) are supported by LANS while other users are supported through dial up services. Each of these technologies, and their role in the communications system of the Base Conceptual Design, are covered below.

Local area networks, specifically those based on Ethernet, are placed at the concentrated usage locations. Ethernet provides a 10 million bit per second pathway between all the devices attached. In this case, the devices are Level 1 and Level 2 workstations, and, optionally, printers. Software on the Level 1 and Level 2 workstations will allow transparent access to the search system if directly connected (co-located) or connected via a bridge.

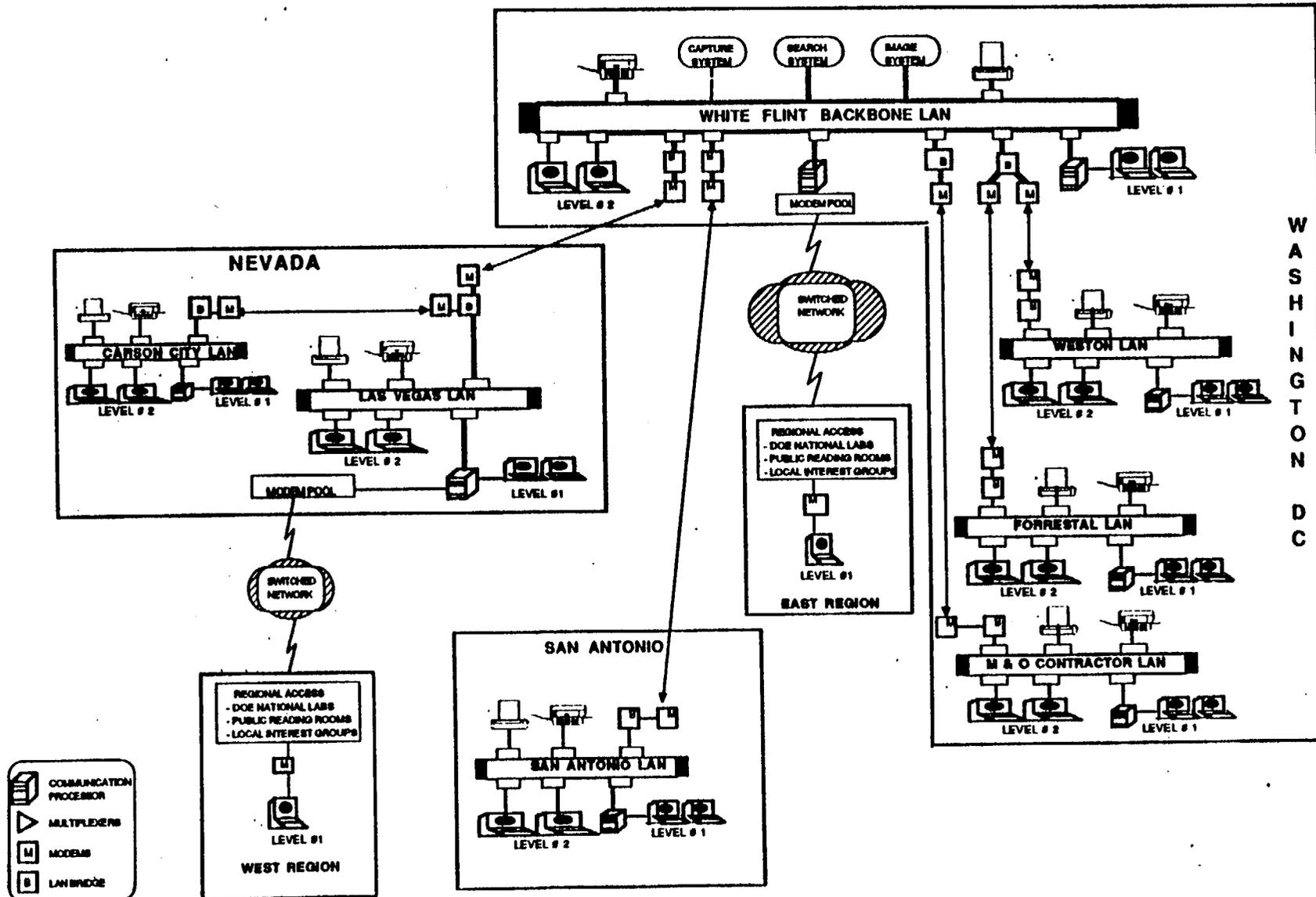
Bridges are intelligent devices that connect two individual local area networks via a wide area network allowing the "bridged" local area networks to appear as a single network. The bridge intelligence is in its ability to route only data from one local area network to another if the data is destined for a particular workstation on the other side of the bridge. This capability becomes important as more local area networks are bridged together. The image system of this conceptual design has its own local area network for communication between the image preprocessor, jukebox controller, printer controller, and the mainframe. A second local area network is supported by the image system for direct and bridged connection to the site local area networks.

High speed multiplexers allow multiple workstations to share a leased communication circuit, each viewing the circuit as though it were dedicated solely to that workstation. High speed modems use new integrated circuit and error correction technologies to send across switched voice grade circuits at speeds such as 9.6 Kbps.

The intelligent communication processors can apply to all the aforementioned technologies simultaneously, providing tailored solutions for each location. In addition, the communication processors monitor the network performance.

FIGURE 3

LSS TOPOLOGICAL COMMUNICATION NETWORK ARCHITECTURE



51

WASHINGTON D.C.

4.1.1.5 Workstations

The workstations are connected via the communications system to the search system. As described in Section 2.4.1.1, two types of workstations will be supported by the search system. The major difference in the workstations is that Level 1 workstations do not support on-line display of document images, and can only display partial full text ASCII pages. Both would have the same query capabilities and allow LSS users to order the document images for subsequent off-line printing.

The Level 1 workstations are micro computers provided by the users. To be a Level 1 workstation the user device must have the following capability:

- 1) Dedicated intelligence equivalent to an IBM PC or compatible.
- 2) Full screen terminal emulation is performed to support the search system software.
- 3) Downloading of ASCII text from the search system is fully supported if the user has compatible file transfer software.
- 4) The workstation can also be used for any standard personal computer software applications.
- 5) A modem is required, which will access the search system via the switched telephone network.
- 6) Text printers can be attached to a Level 1 workstation to print copies of the screen, reproduce all search system responses and user inputs (record a session in hardcopy), and print text downloaded from the search system during a user session.

Images of documents are available to the Level 1 workstation users only through the print capability of the image system. After retrieving and inspecting a document in ASCII, the user can issue an order for hardcopy, which will be sent by express delivery.

The Level 2 workstation, configured specifically for LSS, has all of the capabilities of the Level 1 workstation, plus the ability to transfer, display, and locally print images. Significant features of the Level 2 workstation are:

- 1) The Level 2 workstation would be based on a high speed microprocessor that supports multiple parallel tasks, such as the INTEL 80386.
- 2) Dedicated hardware, including a large format high resolution monitor and image decompression boards and additional software, are used for image display. The large format monitor would also display full page ASCII text.
- 3) A high speed connection to the communications system supports the transfer of images.

- 4) A laser printer is available for the printing of images and ASCII text.

Estimates of the number of workstations that would be required at operational startup and at the first peak usage period are shown below.

Workstation Level	<u>Number of Workstations</u>		January, 1994	
	August, 1990 Min	Max	Min	Max
1	50	100	110	230
2	25	50	65	120
Total	75	150	175	350

These estimates were derived from usage data identified in the Preliminary Data Needs Analysis report and assumptions about which users would need on-line image viewing. The expected number of users during a peak period is 350. The upper limit on the total number of workstations is for each user to have a terminal, with one half of the technical, engineering, regulatory and licensing users requiring on-line images, and thus Level 2 workstations. The minimum estimates would provide one workstation for every two users and only a quarter of the technical, engineering, regulatory and licensing sessions would require on-line images.

4.1.2 Base Conceptual Design Software

Each of the major hardware systems described above require software to perform as an integral part of the total LSS. This software is described on the following sections.

4.1.2.1 Capture Software

There are three major capture functions; image scanning and capture, optical character recognition to obtain full ASCII text, and the capture of cataloging data. The major capture software modules are:

- 1) Image scanning and capture software - produces files containing the bit-mapped images and page order information.
- 2) OCR preparation software - obtains the files produced by the scanning process, selects an OCR station, and presents it a page at a time.
- 3) Check and edit of OCR output - obtains the OCR output, checks the spelling and presents those results and OCR conversion statistics to an operator at a workstation. The operator-corrected ASCII text is written to a file.

- 4) Cataloging software - the keyed cataloging data of the Records and Regulations Access Subsystems are checked against controlled vocabularies and the thesaurus, and are written in a form ready for data base loading. Software tools that automate portions of the cataloging process may be available.
- 5) Duplicate document checking - searches the cataloged data in the operational data base to see if a document has already been entered.
- 6) Cataloging quality control - provides consistency and format checks for cataloging data.
- 7) Controlled vocabulary and thesaurus maintenance - provides for the addition to controlled vocabularies and the thesaurus.
- 8) Electronic text input - handles the receipt of ASCII files for documents prepared by word processors and for electronic filings received via E-mail.

4.1.2.2 Search Software

The search software, runs on the LSS search hardware and implements the main LSS function - computerized access to its data. It includes:

- 1) Applications software to implement the LSS search and retrieval needs that have been and will be identified
- 2) Commercial data base management system software to implement the LSS data base design, and provide many of the query, retrieval, and utility functions
- 3) Operating system software, provided by the manufacturer of the computer hardware, which controls and schedules all of the jobs running on the computer hardware.

To satisfy the LSS search and retrieval needs, the application software will consist of a number of modules, each related to a specific function. The major modules will be:

- 1) Menus - One method a user will be able to choose what they want to do, is to select from a list of choices displayed in menus. The choices presented will likely range from a list of the LSS subsystems to specific search criteria. The menu module will also manage entry to and linking of the other modules. Menu navigation will permit knowledgeable users to go directly to any menu, whether or not it appears as choice on the menu currently displayed.
- 2) Query screens - Fill-in-the-blank type of screen displays which support the formulation of query requests for the Records and Regulation Access headers, and the information in the Tracking Subsystems. These screens will likely also incorporate windows

which can be used to view items such as controlled vocabularies, the thesaurus, indices and previous result reports. Special query screens may also be developed for accessing the full text.

- 3) Prompting dialog - This module would "converse" with a user to construct a query to pass to the data base management system. It would provide the ability to easily store and modify an earlier request. The module accesses both the headers and full text associated with the Records and Regulations Access Subsystems.
- 4) On-line help - Interfaces with the other modules to answer user questions and to display informative and helpful error messages.
- 5) Image display and hardcopy requests - Obtains a request from Level 2 Workstation user for the display or printing at the workstation of one or more images, passes the request to the Image System Link module. Marking of specific pages viewed via full text may be one method of selecting pages to be printed.
- 6) Image system link - Passes request obtained by the image request module to the image system along with the information the communication system needs to route the image to the proper workstation. If the image system also processes the requests for hardcopies that are to be mailed, then this module will also pass those requests to the image system.
- 7) SIMS interface - Provides the software and data base link between the Records Access Subsystem and the Sample Inventory Management System (SIMS).
- 8) Queries and results save and reuse - Will be used by the prompting dialog and query screens modules to enable users to save queries they have constructed for future reuse or as a basis for preparing a new query. It will also save and identify result sets for later display or a restructure for a future query.
- 9) Parallel query processor - The base design assumes that the full text data base can best be implemented by partitioning it into several data bases. This module would make the partitioning as invisible as possible to the user. On the query side, it would invoke the same query on each partition. On the retrieval side, it would combine result sets.
- 10) Data base loading - Controls and schedules the entire data base loading process. The actual loading will be done via the loading utilities provided by the data base management system. Other functions of this module include synchronization with the image loading process and providing information for menu system to display data concerning the contents of the data base.
- 11) Performance monitoring - Collects statistics for the LSS administrator such as the type and complexity of queries generated via other modules, the size of result sets obtained, the time required for classes of queries to be completed and amount of text

and images displayed. This data will be used to optimize system performance.

- 12) Configuration management - Will contain data describing the hardware and software configuration of the LSS and will track changes made to the baseline. This module will be implemented using a commercial software package.

The data base management system software is the key component of the LSS search software. Its ability to handle data bases of the size projected for LSS and to support the following types of query and retrieval functions is crucial for the success of the LSS. The following functions need to be supported by the data base management system:

- 1) Store and invert indices for up to 200 Gbytes of full text (upper estimate for the year 2009)
- 2) Implement both full text and relational data base structures
- 3) Support the following query constructs
 - High level, easy to use full screen query tools
 - Full screen forms
 - Boolean expressions
 - Stem wildcard searches
 - Proximity word searches
 - Near spelling or spelling correction
- 4) Optimization of searches, including segmented refinement of search in order to narrow retrieval sets
- 5) Thesaurus support for retrieval
- 6) Simple and complex user report generator(s)
- 7) On-Line help facility
- 8) Utilities
 - Fast, large volume loading
 - Backup and recovery
 - Integrity checking
 - Usage monitoring
 - Tuning

4.1.2.3 Image System Software

The image system consists of six components - image preprocessor, jukebox controller, jukebox(es), printer controller, printer(s), and local area network. Three of these (the image preprocessor, jukebox controller, and printer controller) have embedded microprocessors that require software. Each is described below.

The image preprocessor has software that:

- 1) Accepts input image data on magnetic tapes and optical disks
- 2) Translates, as necessary, to vendor compressed image format and forward to the jukebox controller.

Software for the jukebox controller:

- 1) Checks for duplication of an existing document by accession number
- 2) Calculates the optimal optical platter for storage of the image and its location
- 3) Writes the image to the chosen optical disk and updates the internal location data base
- 4) Responds to requests from the search system and printer controller.

The printer controller software functions include:

- 1) Receiving requests from the search system for printed images
- 2) Requesting and receiving images from the jukebox controller
- 3) Buffering images to magnetic disks as necessary
- 4) Image decompression and transfer to high speed laser image printers.

Depending on the system chosen, the vendor supplied software could range from primitive subsets to completely integrated packages with high level control languages.

4.1.2.4 Communications Software

Each of the systems in the Base Conceptual Design must communicate with at least one of the other systems. A combination of vendor supplied and custom designed software will be required for integration of the capture, search, image, communication, and workstation systems. The seven layer Open System Interconnection (OSI) standard will be followed wherever possible. The level of custom software required will depend on the final vendors chosen for the systems. Regardless of the selected vendors, there are key functions for each system that must be performed. Each system is addressed separately below.

The capture system minicomputer must be able to communicate with the optical character recognition (OCR) machine, the optical disk drive, the scanners, and the search system. The OCR is sent images and returns ASCII text. The optical disk drive requires the transmission of commands and images from the minicomputer. Software that accepts data from the scanners

and writes the images to local storage is required. Communication with the search system will be via magnetic tape, optical disks, or direct connection. In the case of a direct connection, the capture system minicomputer will have software which performs any protocol conversion for transmission of text and images to the search system.

The search system will accept data from the capture system, transmit commands to the image system, and interface with the workstations. Separate software components will be required for each interface. All of this communication will be simultaneous yet transparent to the LSS user.

The software to support the electronic mail functions described in Section 2.6 would also operate on the search system hardware and be interfaced with the data transfer and data capture software. The integration would enable messages in filings to be composed on a local workstation editor and then sent to the E-mail software for distribution. When appropriate, these items would be forwarded to the data capture software.

The image system components will communicate with each other. The jukebox controller and printer controller will accept commands from the search system, return status information to the search system, and transmit images to the Level 2 workstations.

The workstations will communicate with the search system by emulating terminals and through downloading from the host. Level 2 workstations will also communicate with the image system, accepting compressed images and decompressing them for display and local printing.

Table 3 summarizes the high level communications software functionality required. Some custom software will be required in many cases to interface the vendor supplied software to the LSS applications software. This effort may be complex and, depend on the functionality of vendor equipment and software, especially when multiple vendors are used.

4.1.2.5 Workstation Software

The two levels of workstations for query users each will have some local software operating that enables them to interact with the rest of the LSS retrieval software and hardware.

The Level 1 workstation will require the following software, both of which are commercial products:

- 1) Terminal emulator - allows the workstation to act as a terminal to the search computer. The type of terminal to be emulated depends on what terminals the search computer and software will support.
- 2) Data transfer - for receiving data from the search computer. This needs to be compatible with the LSS communication software. This function is sometimes included in terminal emulation packages. For a Level 1 workstation, this function is optional.

The Level 2 workstation software includes the Level 1 software plus the following:

- 1) Image handler - receives the images from the communications system, controls the display of a image on the workstation's full-page display, and routes images to the local laser printer.

TABLE 3. SOFTWARE REQUIRED FOR LSS COMMUNICATIONS

System	Function	Vendor Supplied Software	Custom Developed Software
Capture:	Communicate between OCR and minicomputer	X	
	Communicate with optical disk drive(s)	X	D
	Communicate with scanners	D	
	Communicate with search system	X	D
Search:	Communicate with capture system	X	D
	Communicate with image system	D	X
	Communicate with workstations	X	D
Image:	Communicate with search system	D	X
	Communicate with workstations	D	D
Workstations:	Communicate with search system	X	
	Communicate with image system	D	D

X - Will be required regardless of vendor supplied software.
D - Depends on the hardware vendors chosen.

- 2) Multi-tasking operating system - such operating system software will allow the image printing to occur in parallel with continuing query activities being conducted via the terminal emulation software.

4.2 Variants to Base Conceptual Design

Several major features of the LSS may be implemented in more than one viable way. This section presents seven variants on the Base Conceptual Design that provide such alternatives. Although all variants satisfy the functional requirements, the efficiency and manner in which they are satisfied will vary.

4.2.1 Variant I - Full Replicated Nodes

4.2.1.1 Description

Variant I (Figure 4) differs from the Base Conceptual Design in that it has two fully replicated search and image system nodes, one located in Washington, DC and one in Las Vegas, NV. The system architecture consists of the identical hardware and software configuration as the Base Conceptual Design. This variant was chosen to exploit the data security that full redundancy offers from a backup and restore perspective. Additional advantages include higher overall availability if the centers are linked plus potentially lower recurring communication charges, since the leased lines required would be shorter in total. Partial replication or distribution of the data base were also considered but not selected as variants because the required complexity of the associated query software and the severe difficulties of data configuration management.

4.2.1.2 Impact On The Capture System

There are no changes in the hardware and/or software configuration of the capture system. The text and images created will be duplicated for input and storage at the two nodes. This additional step will somewhat lengthen the total processing time.

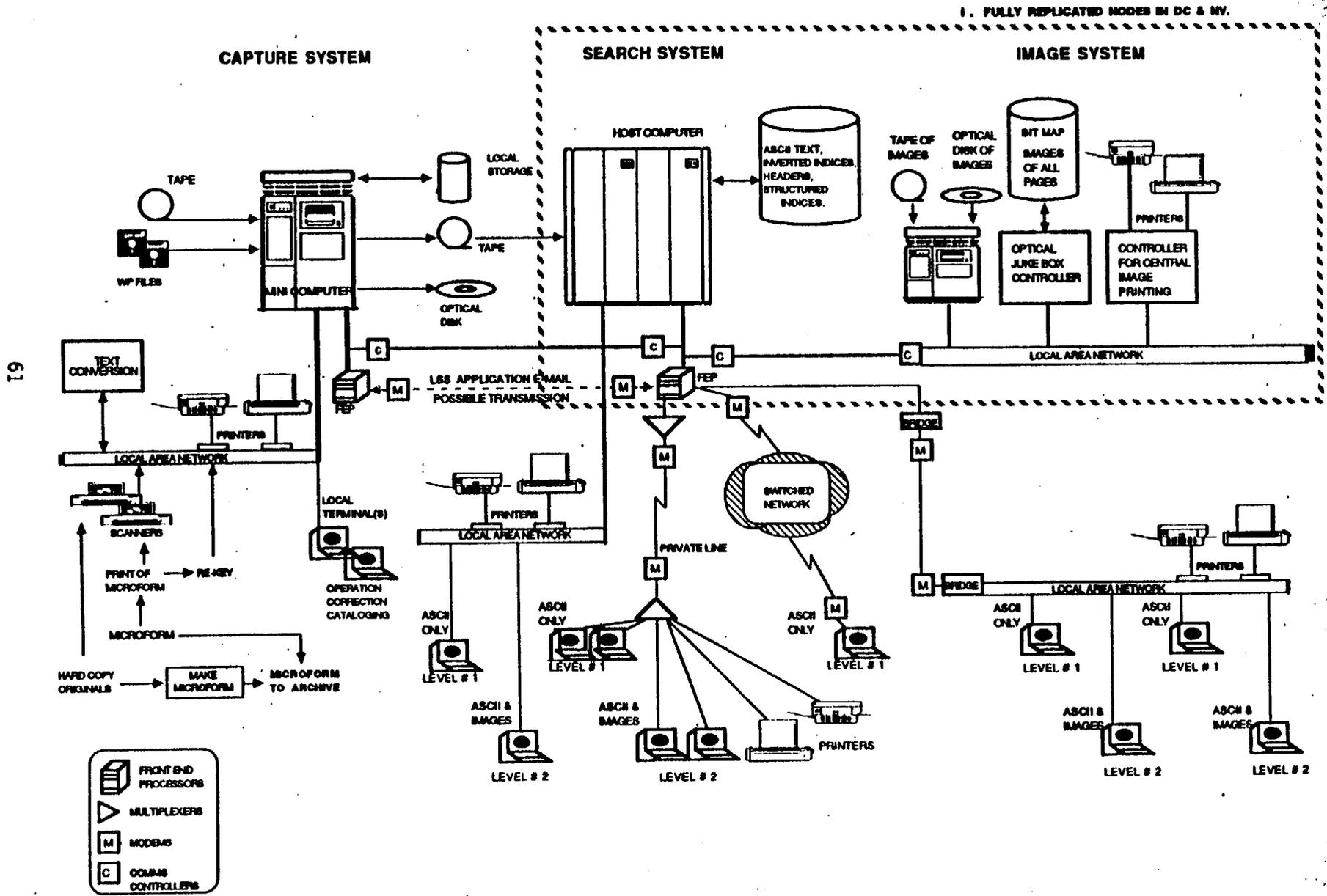
4.2.1.3 Impact On The Search System

The hardware and software configuration for the two-node system will be identical to the one-node Base Conceptual Design architecture. The LSS data bases are replicated and stored at both nodes. There is an increased staffing and configuration management overhead in keeping the two centers synchronized.

4.2.1.4 Impact On The Image System

There will be identical image systems co-located with the search systems at both Washington, DC and Las Vegas. All LSS images are stored

FIGURE 4
VARIANT I



at both centers. The same configuration management problems affecting the search system also apply here.

4.2.1.5 Impact On The Communications System

The network will be geographically split between the Las Vegas and Washington nodes. The communications traffic associated with Las Vegas region (including Carson City) will be supported by Las Vegas. All of the Washington area users will be supported by Washington. San Antonio could be supported by either node. The communications processors at each node would be interconnected to allow transmission of E-mail and failover support if one center should go off-line.

4.2.1.6 Impact On The Workstations

Level 1 and Level 2 workstation configurations are unaltered from the Base Conceptual Design.

4.2.2 Variant II - Hardware Full Text Search

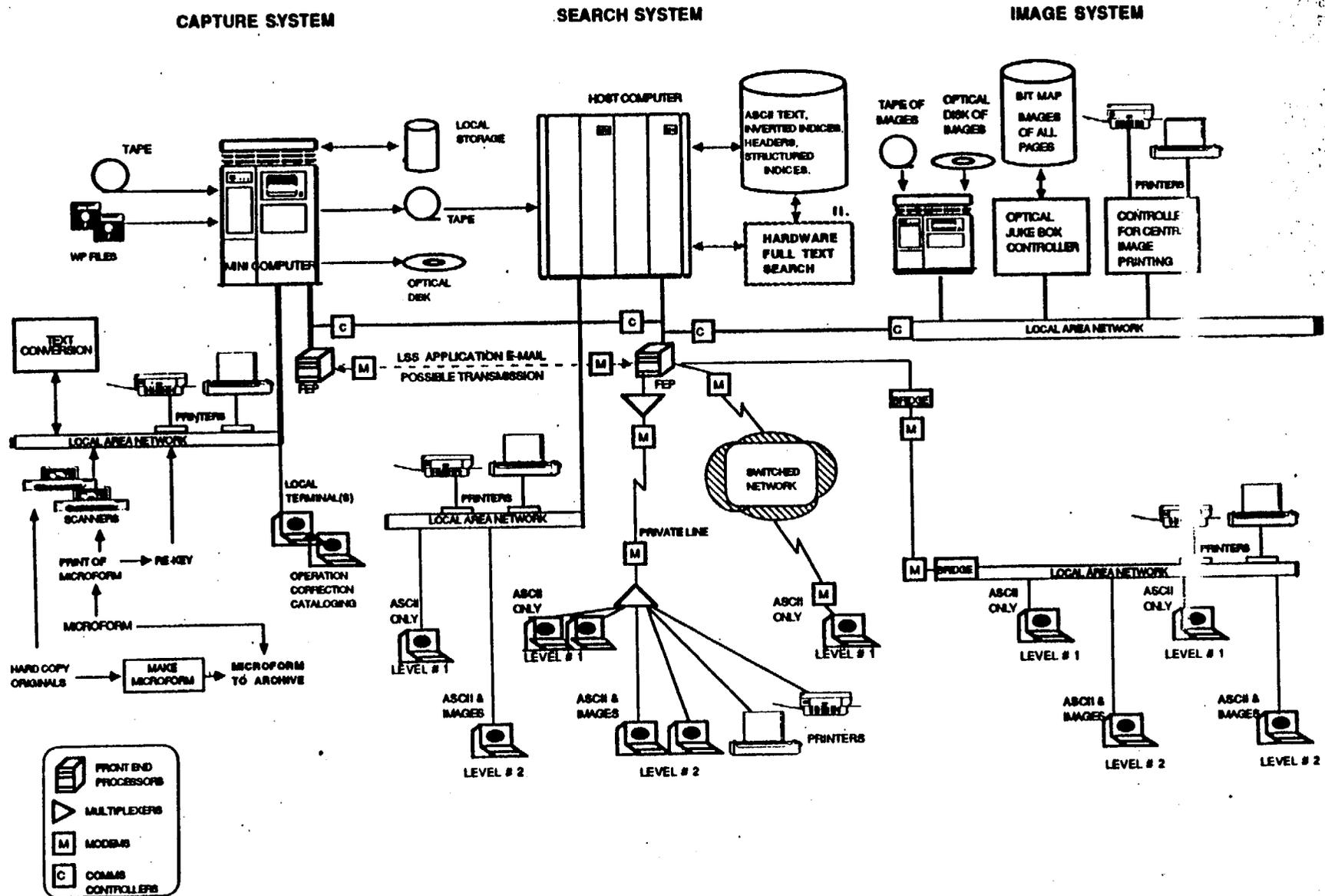
4.2.2.1 Description

As previously described in Section 4.1.2.2, in the Base Conceptual Design full-text search is implemented via storage of the full text plus creation of an inverted index. The full-text data base software uses both to respond to user queries. Variant II (Figure 5) replaces them and the text data base manager with specialized hardware processors. The hardware compresses the full text and stores it on very high transfer rate disk drives for subsequent searching. The search is performed serially through all of the compressed full text. The search speed is a function of data base size, disk transfer speed, and the relative size of the query relative to the width of a comparator. The hardware full-text processor is directly attached to the search host computer from which it receives search requests and returns decompressed full text. Multiple hardware processors can be connected to the host computer. This allows the system to maintain a consistent search time as the data base grows. All other functions (E-mail, header searches, etc.) performed by the search system host computer will remain the same as the Base Conceptual Design.

This variant was chosen in order to capitalize on four advantages offered by hardware full-text search:

- 1) A 4:1 reduction in the amount of required disk storage. This is a result of the average 50% compression achieved by the hardware full-text processor vs. at least a 100% expansion (full text plus inverted index) for the software full-text solution.
- 2) Predictable full-text search response times.
- 3) Update simplicity since new documents are simply added to the end of the data base and no index creation or update is required.

FIGURE 5
VARIANT II



63

- 4) Over the long operational life of the LSS hardware full-text search technology is expected to advance much faster than software technology.

4.2.2.2 Impact On The Capture System

There are no impacts on the capture system, compared to the Base Conceptual Design.

4.2.2.3 Impact On The Search System

Currently available vendor supplied software for the hardware processors is primitive compared with the current state of the art in software data base managers. The result is a significant increase in the software effort and complexity, compared to the Base Conceptual Design, particularly in the software required to coordinate multiple hardware processors.

4.2.2.4 Other Impacts

There are no impacts in Variant II compared to the Base Conceptual Design on the image system, communications and workstations.

4.2.3 Variant III - Images Are Not Supported At Workstations

4.2.3.1 Description

In this variant (Figure 6) from the Base Conceptual Design, the capability to view electronic (bit-mapped) images on the screen at the Level 2 workstations is excluded. This variant was selected since this capability was not identified as necessary to all potential LSS users in the Preliminary Needs Analysis.

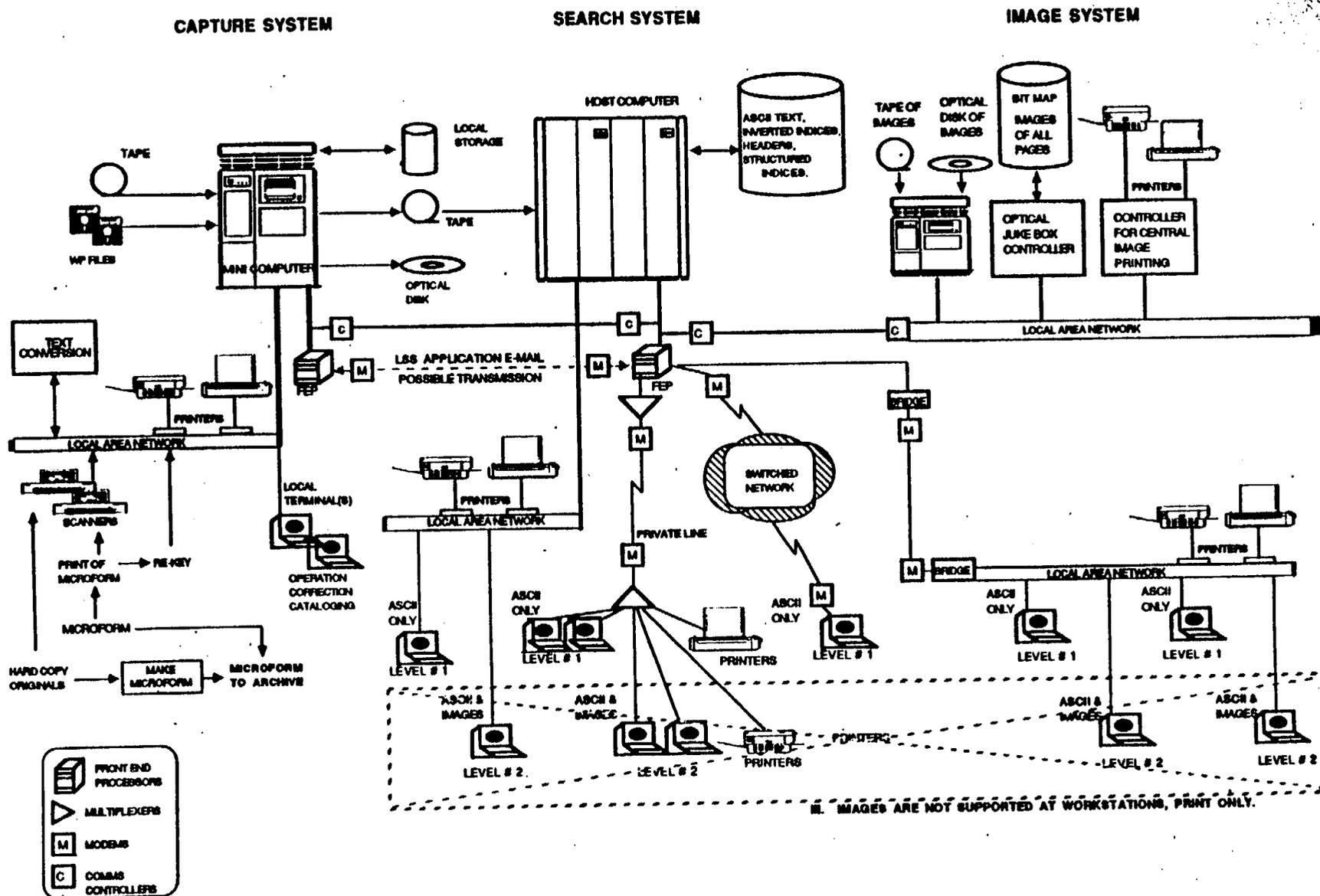
4.2.3.2 Impact On The Image System

The software to transmit images stored in the image system to Level 2 workstations would not be required by this variant. There is a decrease in load on this system resulting from this reduced capability. To compensate for not being able to display images at a workstation, the requests for printing images would increase.

4.2.3.3 Impact On The Communications System

The elimination of the transmission of bit-mapped images to Level 2 workstations reduces the need for high capacity telecommunications links to user locations.

FIGURE 6
VARIANT III



4.2.3.4 Impact On The Workstations

Variant III eliminates the ability to view images on-line. This will result in the elimination of image hardware and software for Level 2 workstations. Level 2 workstations would still be able to display full page ASCII text.

4.2.3.5 Other Impacts

There are no impacts in Variant III compared to the Base Conceptual Design on the capture and search systems.

4.2.4 Variant IV - Microform Digitizers in Capture and Image Systems

4.2.4.1 Description

In this variant (Figure 7) on the Base Conceptual Design, the changes occur in the capture and image systems. Microform digitizers are used to create the OCR input for documents available only on microform. Microform replaces optical disks for the storage and retrieval of images. This variant was chosen to reflect the availability of automated microform systems.

4.2.4.2 Impact On The Capture System

A microform digitizer is added to the capture system to convert documents available only as a microform image to a bit-map similar to the bit maps of hardcopy documents generated by scanners. From this point, conversion to text using the OCR device is the same as in the Base Conceptual Design. Microform images not judged acceptable because of poor film quality will be printed to hardcopy and text re-keyed from the hardcopy. After scanning or re-key, the microform is sent to the microform image retrieval system described in Section 4.2.4.4. Documents that arrive in hardcopy are processed by the capture system as in the Base Conceptual Design.

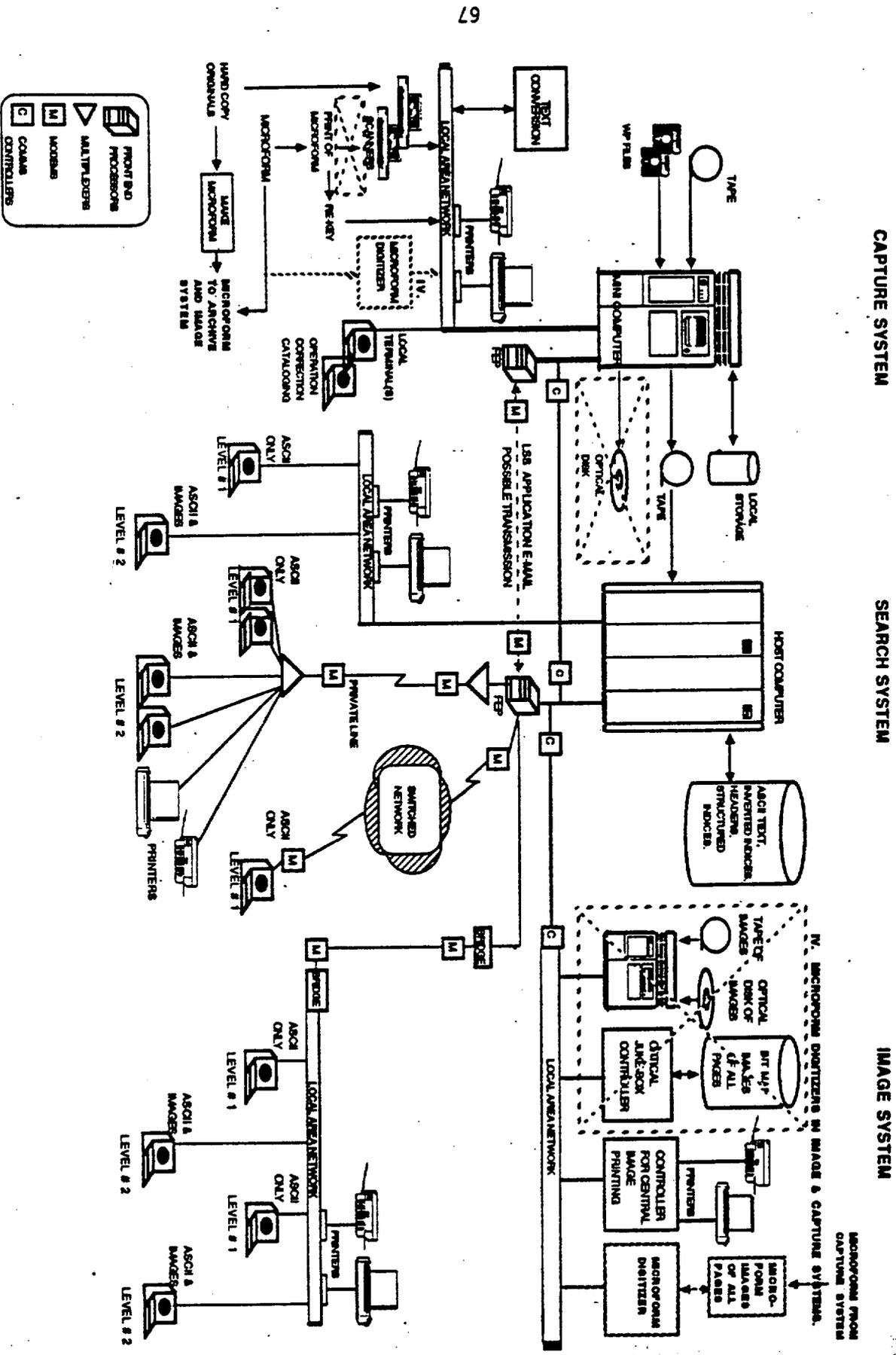
4.2.4.3 Impact On The Search System

The image system link of the search system software must interface with the microform-based image retrieval system in this variant, as opposed to an optical disk system in the Base Conceptual Design. No other impact on the search system results.

4.2.4.4 Impact On The Image System

The optical preprocessor, jukebox controller, and optical jukebox components are replaced in this variant by a microform on-line storage and retrieval system which retrieves, digitizes, compresses, and transmits the microform image to a Level 2 workstation or image printing controller. The

FIGURE 7
VARIANT IV



software required to control the automated microform retrieval and digitizing system will be different from the optical jukebox but does not significantly differ in complexity.

As opposed to the optical disk base conceptual design, in which the storage medium requires no maintenance, a microform-based image storage system can require the replacement of all the storage medium in the system on a frequent basis (depending of specific usage levels).

4.2.4.5 Impact On The Communications System

There should be no impact on the communications system. Compressed images will be of approximately the same size and, depending on the number of microform retrieval units used, will be transmitted at the same frequency.

4.2.4.6 Impact On The Workstations

The Level 1 users will not be affected. Microform stored in such a system deteriorates with age and handling. The Level 2 user may notice minor variations in the quality of the same image if it is retrieved multiple times since the microform image is digitized each time it is retrieved.

4.2.5 Variant V - Microform Off-Line Image Storage and Retrieval

4.2.5.1 Description

Variant V (Figure 8) of the Base Conceptual Design replaces the on-line image system with a off-line service for obtaining hardcopy or microform copies of LSS documents. This is similar to the way commercial and existing DOE bibliographic data base services provide document copies to their users. For example, DIALOG allows users to order documents from NTIS as a command after locating the document.

This variant was developed to present a low-tech solution to meeting the hardcopy receipt time requirements of 2 to 3 days identified in the Preliminary Needs Analysis. The function capability to view electronic (bit-mapped) images on the screen at the Level 2 workstations is excluded, as in Variant III.

4.2.5.2 Impact On The Capture System

The capture system no longer captures the images to optical disks. A duplicate of the microform is sent to the off-line service, which operates similar to the NTIS, described below.

The National Technical Information Service (NTIS) maintains government and contractor documents on microfiche. In addition, some documents are available in hardcopy and on computer tape or diskette. Copies of documents

can be requested by telephone, by mail, or electronically. Documents can be requested as a rush order, in which case the document will be ready for pickup or mailing within 24 hours.

Upon receiving a request the corresponding microfiche is manually retrieved from the archives. Depending upon the request, the microfiche is either duplicated or blown back into paper form. The original is returned to the archive, and the copy is packaged and prepared for delivery. The packaged copy can be either picked up in person at several locations in the Washington, DC area during office hours or shipped to the requester through the postal service or through an express carrier if requested.

4.2.5.3 Impact On the Search System

The search system no longer has to interface with and control the image system. It simply takes orders for documents and forwards them to the off-line service. The elimination of the control software for the image system is significant in terms of personnel and schedule risk reductions.

4.2.5.4 Impact On The Image System

The image system from the Base Conceptual Design is completely eliminated, replaced by the microform off-line system described above.

4.2.5.5 Impact On The Communications System

The elimination of the transmission of images to Level 2 workstations reduces the need for high capacity telecommunication links to the user locations.

4.2.5.6 Impact On The Workstations

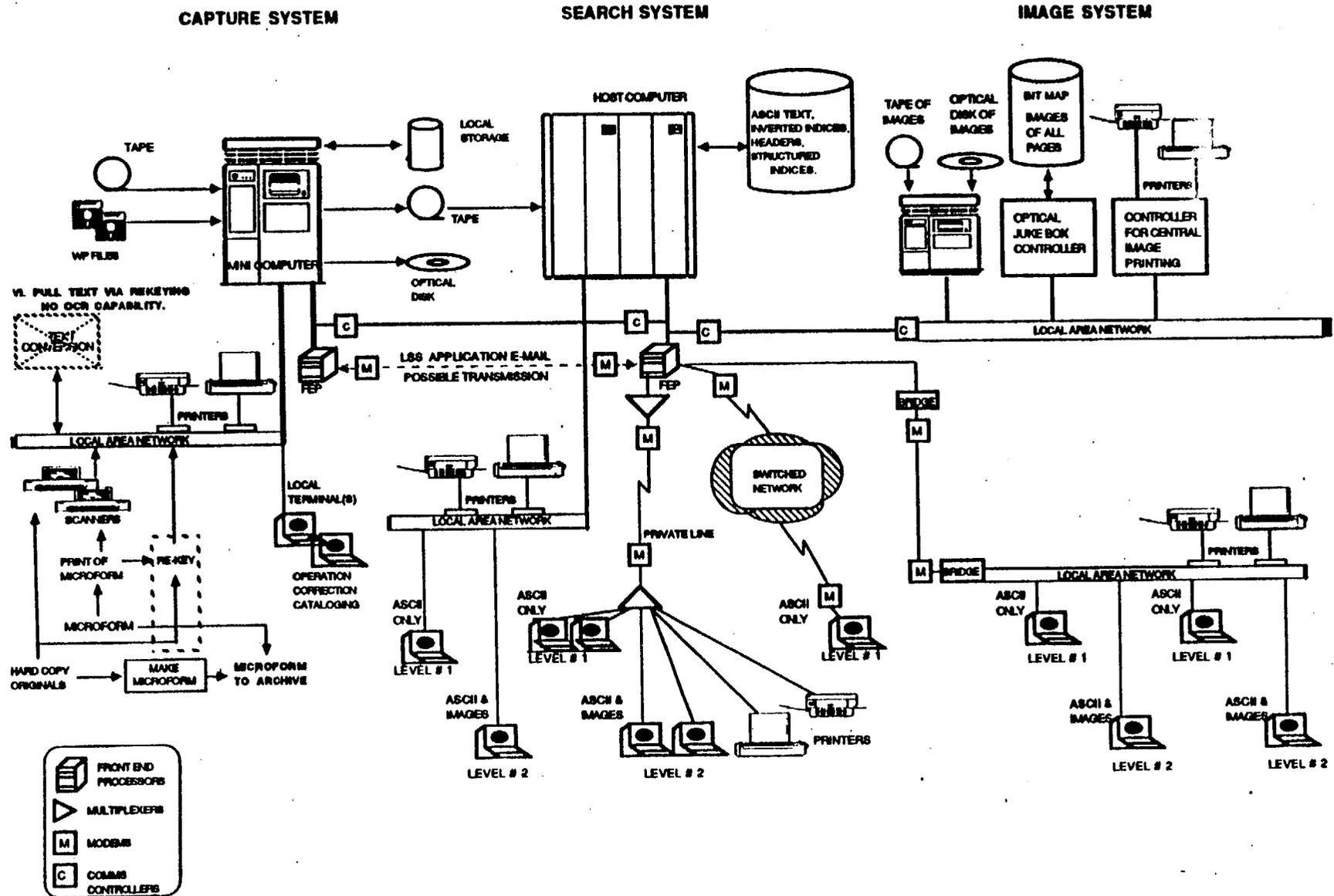
The elimination of the on-line image system results in a similar hardware and software change for the Level 2 workstation, as in Variant III.

4.2.6 Variant VI - Full Text via Re-keying

4.2.6.1 Description

In this variant (Figure 9), there is no automated text conversion (OCR) process. The conversion of hardcopy text to ASCII is accomplished by re-keying the document. An expected 99% accuracy of data via re-keying would be achieved by double keying the original source document. There is a new requirement for software to assist in management and control of the re-keyed documents and bit-mapped image documents. The software should keep track of the documents, location to which they are sent, which have been re-keyed and returned, which have passed the QA process, the status of the image processing and the status of cataloging, so that all of a documents' components are available for loading into the search and image systems.

**FIGURE 9
VARIANT VI**



71

This software will increase the integrity of the process, and reduce the possibility of document duplication.

4.2.6.2 Impact On The Capture System

Since the text conversion will be accomplished via re-keying there will be no requirement for optical character recognition equipment, and associated software. However, the re-keyed documents would require processing through a digital scanning device since bit-mapped image capture and storage is required. (The OCR process allows for the simultaneous conversion of text and digitization of images). There will be a requirement for additional software to assist in the configuration management.

4.2.6.3 Other Impacts

The search system, image system, communication, and workstations in Variant VI are the same as in the Base Conceptual Design.

4.2.7 Variant VII - Combined Variants III, V and VI

4.2.7.1 Description

Variant VII (Figure 10) combines hardware and software changes for the Level 2 workstations (Variant III), the removal of the on-line image system (Variant V) and the re-keying of all documents instead of OCR (Variant VI). This variant was created to present a conceptual design with the lowest schedule risk that minimally meets the requirements presented in the Preliminary Needs Analysis.

4.2.7.2 Other Impacts

The impact on the capture system is the same as Variant VI. The impacts on the search system, image system and communications are the same as Variant V. The impact on workstations is the same as Variant III.

5.0 CONCLUSIONS

The Preliminary Needs Analysis and Preliminary Data Scope Analysis have defined the requirements of an automated computer-based information storage and retrieval system that must accommodate millions of documents. Based on these studies and the directions perceived from the LSS negotiated rulemaking process, a conceptual design has been formulated which meets the stated requirements. While the design was selected because it represents a comprehensive low risk technical solution, it also represents both a large and complex system. As noted in Section 4.0, this design is the formulation of a number of technical experts and is based on the combined experience of the group. In this process a number of alternatives were examined. Some were rejected due to a low probability of success. Others not only offered a reasonable risk but also were potentially more cost effective, although they may not meet all of the "non-firm" requirements identified in the Preliminary Needs Analysis. These alternatives were identified as variants to the base and will be subject to further technical and economic investigation.

The Base Conceptual Design and variants are consistent with the requirements identified to date, including the deliberations of the Negotiated Rulemaking Advisory Committee. Indeed, the rulemaking activities have not yet imposed any requirements on the design which were not anticipated in the Preliminary Needs Analysis. It is further not expected that the rulemaking activities will result in any requirements which cannot be met by the base design or one of the variants; however, the possibility still exists that design refinements may be required to reflect changing requirements. One probable design refinement is for the Tracking Systems needed by DOE to be implemented by DOE only.

Of perhaps more significance to the refinement of the design is the feed-back from potential users and the cost-benefit analysis. All of the dialogue conducted to this point with potential users of the LSS has been in the absence of a common reference design. This has caused much difficulty in communication, because while one must speak from a certain frame of reference, it is almost assuredly not the same as that of the other party. With this study, a reference point is established for the LSS conceptual design that will facilitate future deliberations. Refinement of the design can benefit from additional experiences of and feed-back from potential users, especially in the area of features available to the user. With this base frame of reference, the potential user can now envision certain scenarios of LSS applications and can fill in the various details necessary to more accurately predict system sizing and response.

The Benefit-Cost Analysis, the next report in the series, will permit the investigation of the variants from the standpoint of cost (both savings and additions), and the associated benefits (or lack thereof) of each variant. Such information will provide an economic basis for decisions on how best to meet the various requirements of the system.

The opportunity is now available for substantive input to the design process, which is both encouraged and necessary to refine the process further.

REFERENCES

ANSI/ASME NQA-1, 1983; Quality Assurance Program Requirements for Nuclear Facilities, American National Standards Institute and American Society of Mechanical Engineers, 1 July 1983.

DOE, 1986; Quality Assurance Plan for High-Level Radioactive Waste Repositories, DOE/RW-0095, DOE OCRWM Office of Geologic Repositories, August 1986.

DOE, 1987; Request for Proposal: Design and Implementation of a Licensing Support System, DOE Office of Civilian Radioactive Waste Management, 11 February 1987.

DOE, 1988a; Licensing Support System Preliminary Needs Analysis, OCRWM, Office of Resource Management, February 1988.

DOE, 1988b; Licensing Support System Preliminary Data Scope Analysis, OCRWM, Office of Resource Management, April, 1988.

APPENDIX B
REVISED PROJECTION OF THE SIZE OF THE
LSS DATA BASE, 1990 - 2009

This appendix contains a revised version of Table 8 (Projected Size of the LSS Data Base, 1990-2009) of the Preliminary Data Scope Analysis. Revisions have been made for the estimates of pages added during 1993 and 1998, based on a re-evaluation of the expected levels of activity consistent with the methods described in that report, and include both a low and a high estimate. The revised values are somewhat higher than those published in the Preliminary Data Scope Analysis.

TABLE 8. PROJECTION OF THE SIZE OF THE LSS DATA BASE, 1990 - 2009

<u>Year</u>	LOW ESTIMATE		HIGH ESTIMATE	
	<u>Pages Added During Year</u>	<u>Cumulative Pages At Year-End</u>	<u>Pages Added During Year</u>	<u>Cumulative Pages At Year-End</u>
1990	830,000	9,304,000	1,100,000	11,885,000
1991	1,087,000	10,391,000	1,441,000	13,326,000
1992	1,428,000	11,819,000	1,892,000	15,218,000
1993	1,660,000	13,479,000	2,200,000	17,418,000
1994	2,009,000	15,488,000	2,662,000	20,080,000
1995	1,858,000	17,346,000	2,463,000	22,543,000
1996	1,635,000	18,981,000	2,167,000	24,710,000
1997	1,386,000	20,367,000	1,837,000	26,547,000
1998	1,037,000	21,404,000	1,374,000	27,921,000
1999	1,286,000	22,690,000	1,704,000	29,625,000
2000	1,170,000	23,860,000	1,550,000	31,175,000
2001	1,877,000	25,737,000	2,487,000	33,662,000
2002	1,236,000	26,973,000	1,638,000	35,300,000
2003	1,261,000	28,234,000	1,671,000	36,971,000
2004	1,327,000	29,561,000	1,759,000	38,730,000
2005	1,120,000	30,681,000	1,484,000	40,214,000
2006	415,000	31,096,000	550,000	40,764,000
2007	365,000	31,461,000	484,000	41,248,000
2008	365,000	31,826,000	484,000	41,732,000
2009	365,000	32,191,000	484,000	42,216,000

APPENDIX A

EXAMPLE OF LSS USER SESSION SCENARIO

Technical & Engineering Initial Background Research Scenario

Session Duration: 2 to 3 hours
Session Frequency: 3 to 5 sessions per study
1 to 2 studies per year per scientist/engineer
Data Accessed: Technical reports in the Records Access Subsystem

In this scenario the scientist or engineer is using an LSS workstation near (or on the same floor as) their office to perform background research on a technical subject with which they wish to become familiar. The user begins with a list of terms and subjects relevant to the topic and possibly a list of important authors, as well as some idea of the time frame for research of interest. The scientist or engineer begins by working primarily with catalog data. Specific classes of documents, e.g., memoranda, letters, progress reports, can be excluded from the search. Publication or issue date can be used to define a period of interest.

Using primarily subject and/or keywords, terms in the subject field, the user will perform an initial search, view only the number of documents selected (but not yet the catalog information itself), and sequentially refine the retrieval set until the number of selected documents is in the range of 50 to 100. During this process, the user may refer to the thesaurus to find legal keywords.

Initial queries might contain logical ORs, while the result set refinement would be more likely to contain ANDs and NOTs. From time to time the user would be likely to request intersection (ANDing) of the results of two independent sequential refinement searches. Next, histograms by subject and date may be requested.

When the result set is reduced to a set of 50 to 100 documents, the user may change strategies and begin to organize information by author name, and may request to see a histogram of the count of hits by author or may restrict the search to include or exclude selected authors. Only now will catalog information (probably sorted by author) be displayed, both as descriptive and subject catalog information will be requested, followed by abstract text for selected documents.

Having reviewed several documents, the scientist/engineer may change strategies and perform a series of full-text searches using proximity searches with 3 to 5 terms per query. These searches will be performed against a previously-selected subset of documents.

The scientists/engineer will review the text of selected documents at the workstation, typically looking first at the title page, possibly the table of contents, then skipping to the end of the document to review

conclusions, references, and appendices. The reader will page both forward and backward through the document and view equations, figures and diagrams as well as text. From 10 to 20 documents per session may be reviewed in such a manner.

The user may request up to 50 pages of printed material perhaps once during the session. Overnight delivery is acceptable. The user will want to save intermediate results from one session to another.

APPENDIX C
ABBREVIATIONS USED

ANSI	American National Standards Institute
ASCII	American Standard Code for Information Interchange
ASME	American Society of Mechanical Engineers
bps	bits per second
CD-ROM	Compact Disk Read Only Memory
CFR	Code of Federal Regulations
DOE	Department of Energy
dpi	dot per inch
FFRDC	Federally Funded Research and Development Center
LAN	Local Area Network
LSS	Licensing Support System
M&O	Management and Operations contractor
NNWSI	Nevada Nuclear Waste Storage Investigations
NRAC	Negotiated Rulemaking Advisory Committee, officially known as the HLW Licensing Support System Advisory Committee
NRC	Nuclear Regulatory Commission
NTIS	National Technical Information Service
NWPA	Nuclear Waste Policy Act of 1982
OCR	Optical Character Recognition
OCRWM	DOE Office of Civilian Radioactive Waste Management
OMB	Office of Management and Budget
OSI	Open System Interconnection
PC	Personal Computer

ABBREVIATIONS USED
(continued)

QA	Quality Assurance
QC	Quality Control
RFP	Request For Proposal
SCP	Site Characterization Plan
SIMS	Sample Inventory Management System