The Conservation Foundation

February 11, 1988

MEMORANDUM

TO: HLW Licensing Support System Advisory Committee Members

FROM: Howard Bellman, Tim Mealey, Matt Low and Kirk Balcom

SUBJ: NRC Actions and March 22-24, 1988 Meeting

Please find enclosed: a copy of the Federal Register notice issued by the NRC announcing its action to amend the charter of the HLW Licensing Support System Advisory Committee; a proposed agenda for the March 22-24, 1988 meeting; a copy of the technical staff report; and proposed consensus statements based on the committee's discussion of the issues at the December meeting.

The March 22-24, 1988, meeting will start at 9:30 a.m. on March 22nd and end at approximately 3:30 p.m. on March 24th. The meeting will be held at the Best Western Airport Plaza Hotel in Reno, Nevada. As its name suggests, the hotel is located near the Reno airport. Shuttle service is available on a call-in basis. When making reservations, call 800-648-3525 and ask for Christina Curtis. Be sure to mention that you are attending The Conservation Foundation/NRC meeting to assure receiving the reduced room rate (the rate we have been offered is $48.00 per night).

You will recall that we agreed to discuss all of the remaining issues at the March meeting in sufficient detail in order for the NRC to be able to prepare a draft rule in advance of the April meeting. With some issues the group is fairly close to achieving consensus and with others the group will be discussing an issue for the first time. We have structured the agenda to allow at least two hours discussion for each issue. Wherever possible we will strive for consensus but, unlike previous meetings, we will have to cut off discussion after a certain point in order to get through our agenda. Also, we have indicated how we think these agenda items correspond to the issues listed in the NRC position paper. Please review the proposed agenda and call Howard Bellman or Tim Mealey if you have any suggestions for improving it.

The enclosed technical staff report was prepared in response to questions that were raised at the December meeting. Even with the changes in committee membership, we felt that the requested

1250 Twenty-Fourth Street, NW Washington, DC 20037 USA 202/293-4800 Telex: 64505 PANDA
Affiliated with World Wildlife Fund

information would be useful to the membership of the newly constituted committee. If you have questions after reviewing the report, feel free to contact any one of the three persons listed in the introductory section. We have included a some time for questions you may have about this report in the agenda for the next meeting, but we are operating under the assumption that a presentation summarizing the report will not be necessary.

You will also recall that you agreed to send each other your suggestions for: 1) the types of documents that should be considered "discoverable;" 2) the subset of "prospectively generated" discoverable documents that should be excluded from automatic entry in searchable full text; and 3) the subset of "backlogged" discoverable documents that should be required to be entered in searchable full text.

Your suggestions should be sent to each other no later than March 4, 1988. Please be sure to send a copy of your these lists to Tim Mealey so that he can compile a master list in advance of the next meeting.

PROPOSED AGENDA

Meeting of the
HLW Licensing Support System Advisory Committee

Best Western Airport Plaza Hotel
1981 Terminal Way; Reno, Nevada
March 22-24, 1988

Time	Activity
Tuesday, 3/22/88	
9:30-9:45	Agenda Review
9:45-11:00	Organizational Issues: -- Adoption of Protocols -- Response to requests for participation -- Other(s)? Public Comment Period, if requested
11:00-11:10	BREAK
11:10-11:30	Questions and Answers Concerning the Technical Staff Report
11:30-12:30	Review proposed consensus statements on: What documents should go into the LSS and how?* (See issues 1, 2, 18 and 24 in the NRC Position Paper) Review suggestions for: 1) the categories of documents that should be considered "discoverable;" 2) the subset of "prospectively generated" discoverable documents that should be excluded from automatic entry in searchable full text; and 3) the subset of "backlogged" discoverable documents that should be required to be entered in searchable full text.
12:30-1:30	LUNCH
1:30-3:00	Continued discussion of lists and proposed consensus statement on what documents should go into the LSS and how.
3:00-3:30	BREAK
3:30-5:00	Review proposed consensus statement on how to handle privileged material* (See issues 3 and 4 in the NRC Position Paper)
5:00-5:15(?)	Public Comment Period, if requested

Wednesday, 3/23/88

8:30-10:30	Discuss how to handle drafts, handwritten notes and marginalia* (See issues 1, 2, 3, and 9 in the NRC Position Paper)
10:30-10:45	BREAK
10:45-12:45	Discuss the "mechanics of the LSS," including who is responsible for record entry* (See issues 6, 7, 10, 11, 12, 13, 15, 16, and 22 in the NRC Position Paper)
12:45-1:30	LUNCH
1:30-3:30	Discuss "access to the LSS," including who gets access and how* (See issues 20 and 21 in the NRC Position Paper)
3:30-3:45	BREAK
3:45-5:45	Discuss pre-hearing dispute resolution procedures, including the proposed use of a pre-hearing licensing board* (See issue 27 in the NRC Position Paper)
5:45-6:00	Public Comment, if requested

Thursday, March 24, 1988

8:30-10:30	Discuss LSS administrative oversight responsibilities* (See issues 14, 17 and 19 in the NRC Position Paper)
10:30-11:00	BREAK
11:00-12:30	Continued discussion of LSS Administration*
12:30-1:30	LUNCH
1:30-3:00	Wrap-Up - review and summarize any tentative agreements, discuss next steps
3:00-3:15(?)	Public Comment, if requested

* Discussion of these items will be limited to the times indicated, with the goal of achieving a tentative consensus whenever possible but, at a minimum, airing committee member's views on how these issues should be handled to allow the NRC to prepare a draft rule for review at the April meeting.

NUCLEAR REGULATORY COMMISSION

10 CFR Part 2

High-Level Waste Licensing Support System Advisory Committee (Negotiated Rulemaking); Change in Composition of Advisory Committee

AGENCY: Nuclear Regulatory Commission.

ACTION: Notice of the change in composition of the High-Level Waste Licensing Support System Advisory Committee and notice of next meeting.

SUMMARY: The Nuclear Regulatory Commission is changing the composition of the High-Level Waste Licensing Support System Advisory Committee. This advisory committee was established to develop recommendations for revision of the Commission's Rules of Practice (10 CFR Part 2) related to the adjudicatory proceeding for issuance of a license for the disposal of high-level waste (HLW) in a geologic repository. Specifically, the committee is attempting to negotiate a consensus on the procedures for the submission and management of records and documents for the HLW licensing proceeding. Recent amendments to the Nuclear Waste Policy Act have changed the site selection process for a HLW repository, narrowing the parties that may be affected by this rulemaking. Consequently, the Commission has revised the composition of the committee to reflect this change.

DATE: The next meeting of the revised HLW Licensing Support System Advisory Committee will be held on March 22, 23, and 24, 1988, beginning at 9:00 a.m., in Reno, Nevada. The location of this meeting in Reno, Nevada, will be announced at a later date. This meeting will be open to the public.

FOR FURTHER INFORMATION CONTACT:

NRC Staff

Francis X. Cameron, Office of the General Counsel, U.S. Nuclear Regulatory Commission, Washington, DC 20555, Telephone: 301-492-1623.

Kenneth L. Kalman, Office of Nuclear Material Safety and Safeguards, Washington, DC 20555, Telephone: 301-492-0428.

Facilitators

Howard S. Bellman
Timothy J. Mealey
Matthew A. Low
Kirk Balcom
Conservation Foundation, 1250 24th Street, Washington, DC 20037, 202-293-4800

SUPPLEMENTARY INFORMATION:

Background

On August 5, 1987 (52 FR 29024), the Commission announced the formation of an advisory committee ("negotiating committee") to develop recommendations for revision of the Commission's Rules of Practice in 10 CFR Part 2 related to the adjudicatory proceeding for the issuance of a license for a geologic repository for the disposal of high-level waste (HLW). Specifically, the committee is attempting to negotiate a consensus on the procedures for the submission and management of records and documents for the HLW licensing proceeding. These revisions relate to the development of an information management system (the "Licensing Support System" or "LSS") that would contain all of the data supporting the DOE license application, as well as all of the potentially relevant documents generated by the NRC and other parties to the licensing proceeding. Implementation of this system was intended to accomplish the following objectives—

- To facilitate discovery by providing comprehensive and easy access to potentially relevant licensing information;
- To establish the information base for the licensing proceeding, to the extent practicable, before the DOE license application is submitted and the three year statutory time period begins;
- To facilitate review of the relevant licensing information by all parties and eventually the boards through the provision, to the extent practicable, of full text search capability;
- To reduce the time associated with the physical submission of motions and other documents associated with the licensing proceeding by providing for the electronic transmission of these documents;

As stated in the August 5, 1987 Federal Register Notice, the Commission considered parties for membership on the negotiating committee on the basis of: (1) Whether they have a direct, immediate, and substantial stake in the rulemaking, (2) whether they may be adequately represented by another party on the committee, and (3) whether their participation is essential to a successful negotiation. The Commission also noted its concern that the negotiating committee be kept to a manageable size in order to maximize the potential for arriving at a consensus.

Based on the above criteria, the Commission invited the following groups to participate on the negotiating committee—

(1) State of Nevada
(2) State of Washington
(3) State of Texas
(4) Yakima Indian Nation
(5) Nez Perce Indian Tribe
(6) Confederated Tribes of the Umatilla Indian Reservation
(7) Department of Energy
(8) National Congress of American Indians
(9) Utah, Oregon, and Mississippi (jointly)
(10) Minnesota and Wisconsin (jointly)
(11) The Sierra Club, Environmental Defense Fund, and Friends of the Earth (jointly)
(12) Nuclear Waste Task Force, representing a coalition of local Texas nongovernmental groups
(13) Edison Electric Institute and the Utility Nuclear Waste Management Group (jointly)
(14) Nuclear Regulatory Commission
(15) U.S. Council for Energy Awareness
(16) National Conference of State Legislatures
(17) National Association of Regulatory Utility Commissioners
(18) State of Tennessee
(19) Penobscot Indian Nation

A coalition of local governments from the State of Nevada was added at a later date. The negotiating committee has met monthly since September 1987. Considerable progress has been made in defining and prioritizing the relevant rulemaking issues, and in developing a consensus on those issues.

The Nuclear Waste Policy Amendments Act of 1987

On December 22, 1987, the President signed into law a bill amending the Nuclear Waste Policy Act (the Nuclear Waste Policy Amendments Act of 1987, Pub. L. No. 100-203). The new law provides for a phase-out of site-specific

activities at all first round candidate sites other than the Yucca Mountain site in Nevada within ninety days of enactment. If the Yucca Mountain site is found to be unsuitable for a geologic repository, new legislative authority would be needed to begin characterization of any other site. In regard to the second geologic repository, no site-specific activity can be conducted unless specifically authorized by Congress. Such authorization cannot be considered until the Secretary of Energy reports to Congress. This report will not be submitted until January 1, 2007, at the earliest. The legislation nullifies the Department of Energy (DOE) proposal to locate a Monitored Retrievable Storage facility (MRS) on the Clinch River in Oak Ridge, Tennessee, as well as any of the alternative sites in the proposal. The Secretary of Energy is authorized to site, construct, and operate an MRS. However, the Secretary may not select a site for an MRS until after a number of conditions are met, including—

- After the report and recommendation of the independent MRS Review Commission is submitted to Congress on June 1, 1989;
- After the Secretary evaluates potentially suitable MRS sites;
- After the Secretary recommends to the President the approval of a site for the development of a geologic repository.

The primary effect of the legislation is to focus the Department of Energy's site characterization efforts on a single site in Nevada to determine its suitability as a site for a geologic repository. Efforts in regard to other first round sites for a geologic repository, and the search for a second geologic repository, have been terminated. The Commission's selection of the participants for the original negotiating committee was based upon the wide range of first and second round sites that were possible candidates for the location of a geologic repository under the existing statutory framework. With the change in the statutory framework, the Commission now believes it is appropriate to revise the composition of the negotiating committee to reflect the focus on characterizing the Nevada site. The members of the revised committee are—

- State of Nevada;
- a coalition of Nevada local governments;
- Sierra Club, Environmental Defense Fund, and Friends of the Earth (jointly), representing a coalition of nonprofit environmental groups;
- Edison Electric Institute and the Utility Nuclear Waste Management Group (jointly);
- Department of Energy;
- Nuclear Regulatory Commission.

Additional membership on the new committee will be governed by committee protocols. The Commission expresses its appreciation to all former participants for their service on the negotiating committee. Their participation has provided a solid foundation for the continuing work of the new committee. In addition, the Commission would welcome any former committee members to convey any remaining concerns on the rulemaking issues to the committee at future meetings.

Federal Advisory Committee Act

In accordance with the Federal Advisory Committee Act (FACA), 5 U.S.C. App. 1, the Commission has submitted an amended charter for the negotiating committee to the General Services Administration that reflects the change in committee composition. In accordance with the Commission's regulations in 10 CFR Part 7, advance notice of negotiating committee meetings will be provided in the Federal Register, the meetings of the full negotiating committee will be open to the public, members of the public will be able to submit written or oral statements to the committee, and detailed minutes of each meeting will be made available for public review and copying.

The next meeting of the negotiating committee is scheduled for March 22, 23, and 24, 1988. The previously scheduled meeting on February 11 and 12, 1988, has been cancelled.

Dated at Bethesda, Maryland, this 1st day of February 1988.

For the Nuclear Regulatory Commission.

Donnie H. Grimsley,

Director, Division of Rules and Records, Office of Administration and Resources Management.

[FR Doc. 88-2429 Filed 2-4-88; 8:45 am]

BILLING CODE 7590-01-M

PROPOSED CONSENSUS STATEMENT #1

DISCOVERABLE RECORDS

Discoverable records means any record that is relevant or likely to lead to the discovery of information that is relevant to the licensing of a geologic repository for the disposal of high-level nuclear waste under regulations set forth in 10 C.F.R. Chapter 1, including but not limited to, (INSERT categories of records agreed upon by the committee).

PROPOSED CONSENSUS STATEMENT #2

CRITERIA FOR DETERMINING WHETHER
DISCOVERABLE DOCUMENTS WILL BE ENTERED INTO THE LSS
IN SEARCHABLE FULL TEXT VS. NON-SEARCHABLE FULL TEXT

1. Discoverable records will be separated into two categories:

   Prospective -- Records generated in electronic format. These records include any and all records generated after July 1, 1988 (the date on which it is presumed that all discoverable documents generated at DOE will be generated and captured in an electronic format), as well as all records generated prior to this date which, in fact, have been generated in electronic format.

   Backlog -- Records not generated in electronic format. These records include discoverable records generated prior to July 1, 1988, provided that such records have, in fact, not been generated in electronic format.

2. Electronic format, as the term is used herein, means documents generated and captured electronically in a format which meets basic compatibility requirements for subsequent entry into the LSS in searchable full text (i.e., machine readable).

3. All prospective records will be entered into the LSS in searchable fukl text except those which fall into specifically enumerated exclusionary categories (to be determined by the committee). Searchable full text may be enhanced by bibliographic headers and/or abstracts.

4. No backlog documents will be entered into the LSS in searchable full text except those falling in specifically enumerated inclusionary categories (to be determined by the committee). These records will entered into the LSS in searchable full text no later than (insert date to be determined by the committee).

5. Backlog documents which are not entered into the LSS in searchable full text will be entered in surrogate form, using either headers and/or abstract indexes. These indexes themselves will be searchable and the images of backlog records will be captured electronically for display, but not for searching, in full text.

6. The categories of backlog records to be entered into the LSS in searchable full text and prospective records to be excluded from entry into the LSS in searchable full text may be revised from time to time by potential parties to the licensing proceeding. (The committee will establish a process pursuant to which such revisions can be made

binding. This process could come within the ambit of a pre-licensing board, or some other means to be determined by the committee.)

7. Notwithstanding the above, potential parties to the licensing proceeding may request that another party's records be entered into the LSS in searchable full text if they determine that they or the other party may rely on such records during the licensing proceeding.

PROPOSED CONSENSUS STATEMENT #3

PRIVILEGED DOCUMENTS

A. ATTORNEY/CLIENT AND ATTORNEY WORK PRODUCT PRIVILEGES

1. The attrney/client and attorney work product privileges may be relied upon by all parties to the licensing proceeding to withhold records from the unprotected, searchable full text portion of the LSS.

2. The privileges will be applied in accordance with interpretations under the Federal rulles of discovery and evidence.

3. All discoverable records, for which these privileges are asserted, must be identified in the LSS with basic header information, to wit: author, recipient, date, title, brief description.

4. The basic headers which will be used to identify records for which this privilege is asserted must entered into the LSS no later than three (3) months (or some other date to be determined by the committee) after generation of the record by the party asserting the privilege.

B. DELIBERATIVE PROCESS PRIVILEGE

1. The deliberative process privilege will be available to the Federal agencies, Indian Tribes, states and local governments who are parties to the licensing proceeding.

2. To the extent that an administrative or judicial decision is rendered which places a limitation or preclusion on the applicability of this privilege to any record or category of records generated or possessed by any Federal agency, such a decision shall apply to records to be entered into the LSS.

3. All records for which a deliberative process privilege is claimed shall be identified in the same manner as A.3. above, reasonable contemporaneously with the creation of the record.

4. To the extent that the committee agrees on the use of a pre-licensing baord to hear and resolve discovery disputes, DOE agrees to be bound by decisions of that baord on the issue of whether the NWPA limits the availablity of the deliberative process privilege to DOE.

INFORMATION RETRIEVAL SYSTEMS
A TUTORIAL

Prepared By
Negotiated Rulemaking Technical Staff

FEBRUARY 3, 1988

CONTENTS

1.0 INTRODUCTION

This document has been prepared jointly by technical staff of the Conservation Foundation, the Nuclear Regulatory Commission, and Science Applications International Corporation (SAIC), the DOE LSS contractor. Opinions expressed in this document are those of the authors and are based on review of the literature and "hands-on" experience in designing and using on-line information and litigation support systems.

For further information or clarification, please contact:
Kirk Balcom (703) 476-1100
Avi Bender (301) 492-9914
Dick Pierce (703) 821-4350

1.1 PURPOSE

The purpose of this document is to provide the Negotiated Rulemaking Advisory Committee with a tutorial on basic information retrieval concepts and to establish a common framework and vocabulary for all future discussions. The document provides an explanation of search and retrieval methods, and a discussion of various storage, indexing and display techniques. This is followed by a description of common options for database creation and for the retrieval process. A glossary is included to define the most commonly used terms.

A very important system requirement, and the ultimate measure of success, is to provide accurate and timely access to all information within the LSS. There are other requirements as well and each imposes a different design specification. A major premise in developing this guide was to focus attention on a major technical driving factor, information search and retrieval concepts, and less on the hardware, cost and design aspects. These latter issues will be addressed at a later stage when more definitive requirements are established.

1.2 HOW TO USE THIS DOCUMENT

Section 2 of the report will guide you through the common ways to search and retrieve documents from an on-line database and will describe some of the advantages and disadvantages of each option. Section 3 describes how the information can be captured from hard copy or directly from word processing equipment in order to create the electronic database. Section 4 then takes you through the various options for cataloging and indexing. Storage options are described in Section 5 and document display and output options are described in Sections 5 and 6.

Using Section 2 as a menu, the reader can then turn to Section 8 to see the various options for creating a system to achieve the desired search and retrieval alternative. For example, if it is determined that only an abstract/bibliographic search will be required then all the options described under scenario E are possible. If enhanced full text search is the option then all the options under scenario F are possible. Closer scrutiny of scenarios A through F reveals redundancy of options in storage.

display, database creation indexing, display and workstations. Specific requirements such as "perform full text search and retrieve original highlighted ASCII text within 60 seconds and image within 24 hours" will begin to eliminate some of the options. Otherwise almost every conceivable scenario is possible but not necessarily practical. The actual approach for developing the LSS may involve some or all of scenarios A through F. Finally, while search and retrieval techniques are certainly important factors in determining system requirements, there are additional performance parameters which must be defined in order to specify a system. These are discussed briefly in Section 9.

2.0 SEARCH AND RETRIEVAL

Documents are searched and retrieved either manually through physical files, or electronically through computer searches of bibliographic headers, subject terms, abstracts, or full document text and are then available for review in electronic or hard copy readable form.

A search strategy generally retrieves one or more "hits" (those documents which meet the terms of the search query). The success of the search strategy is measured by two factors--recall and precision. Recall is the number of documents retrieved in relation to the number of documents that exist on the query. Perfect or 100% recall is retrieving all of the documents that satisfy the query. Precision is the number of retrieved documents that actually pertain to the query in relation to the total number of documents retrieved. Perfect or 100% precision means that there are no "false drops" (irrelevant documents). Retrieval systems are usually rated by how well they perform on recall and precision. In general, as recall improves, precision decreases. As the database grows, the user tends to reduce the number of hits by more restrictive searches, i.e. adding conditions which reduce recall. The third factor to consider is whether the amount of information displayed for each "hit" is sufficient to ascertain whether the "hit" is useful. Good system design as well as experience in using on-line databases are important factors in improving document retrieval.

2.1 BIBLIOGRAPHIC HEADER

A bibliographic header is composed of the essential parts of the document, such as author, title, date, etc., along with descriptive features, such as type of document, number of pages, etc. A search can be conducted on any word or date in the header. This type of system provides excellent recall and precision for such queries as "give me a list of all documents written by author x" or "give me a list of all documents published in the year 19xx." The system does not lend itself to content based searches since a search term must appear in the header. Therefore recall and precision are poor for content based searches. In addition, while the display of information is sufficient for an author or date search, it gives little or no indication of the validity or usefulness of the document in a subject search. Generally a review of the document is needed to determine usefulness.

2.2 BIBLIOGRAPHIC HEADER WITH ABSTRACT

The addition of a searchable abstract to the header improves the recall and precision for subject searches, as well as the ability to determine the usefulness of each document. A searcher must take into account, however, all possible synonyms for the subject term in order to increase recall. A well-written abstract that includes those words most likely to be used for retrieving that document will also substantially increase recall. In some cases, an extensive abstract can actually eliminate the need for obtaining a hard copy of the document. As a whole, recall is poor to average and precision is about average for this system, while the display of information is greatly improved over a bibliographic header. This is a more costly system than the header-only system since the author or an abstractor is

needed to provide the abstract.

2.3 BIBLIOGRAPHIC HEADER WITH SUBJECT TERMS

This system adds subject terms to the header, also improving recall and precision for subject searches. However, the information displayed for each "hit" is a poor indication of the usefulness of the document as subject terms are frequently limited in number and therefore are only an indication of the subject matter of the document. A hard copy of the document is generally necessary to determine its usefulness in meeting the search criteria. Subject terms are also useful in eliminating ambiguities of words in the header. Overall, the system is about average for recall and precision and below average for display.

2.4 BIBLIOGRAPHIC HEADER WITH ABSTRACT AND SUBJECT TERMS

The addition of both an abstract and subject terms to the header allows for a greater degree of recall than the previous systems. A searcher can also improve precision by looking at keywords assigned to a useful document and limit a search by using the same keywords. Again, the abstract assists in determining whether the document is useful. Recall is rated average to good, precision is average, and display is above average.

2.5 FULL TEXT

Full text indexing allows the searcher to search on every word within the document. If such a search is performed in conjunction with a synonym file, the resulting recall of documents may be higher than any of the preceding methods but with a relatively lower than average level of precision. Without the benefit of a synonym file the researcher (unless very knowledgeable in the field) will run into problems of semantics. For example, searching on volcanic may not result in documents using the words earthquake, ground movement, slip fault, tectonic...

Full text search is a superior method for content based searches used to identify places, people, and terms with the documents. Searching for concepts, however, is not an easy matter since concepts generally do not appear as words in the text. Full text indexing without any enhancement can create an unwieldy document retrieval situation where instead of finding the needle in the haystack the user retrieves the needle and the haystack. Depending on the software package used, display is generally above average since one can see the highlighted words within context. Built in term weighting algorithms are also available to display documents according to an importance ranking factor based on the frequency of the hit word within the document.

Compared to abstracts and subject terms, full text requires the least amount of human intervention during the database indexing process.

2.6 ENHANCED FULL TEXT

The approach that maximizes the virtues of all the preceding indexing schemes is enhanced full text. By combining bibliographic header, which provides a <u>structure</u> for the information before it enters the database, with

the full text which provides for content based searches, and subject terms which provide concepts, the resulting recall and precision is superior. The user now has greater flexibility to use either full text search, bibliographic header, subject terms, or a combination of the three.

2.7 RETRIEVAL ENHANCEMENTS

Regardless of which system is chosen for a database, there are certain retrieval enhancements that should also be considered to improve searching. These include:

a) Boolean Logic - the use of connectors such as "and," "or," and "not."

b) Range Searching - the use of phrases such as "from ... to ..." or "between ... and ..." and other similar phrases for searching date or other ranges.

c) Field Searching - the capability of limiting the search to a specific field, such as author, date, title, etc.

d) Phrase Searching - the ability to use phrases such as "nuclear waste" or "nuclear power plant."

e) Proximity - searching for a word within x number of words of another word, e.g., the word "nuclear" within 3 words of "power."

f) Sorting - sorting the output chronologically, alphabetically by author, etc.

g) Limiting - limiting the output to certain years, a specific language, a geographical area.

h) KWIC or keyword in context format - displays the keyword surrounded by the 25 or so words before and after.

These are only some of the major enhancements to be considered.

3.0 DATA CAPTURE

Data capture is the process by which documents and information become a part of the LSS. The process can take several forms including placing documents into a file cabinet, entering the full text of a document into machine readable (ASCII) form, and capturing the image on a microfilm or in an electronic (bit-mapped) image file.

3.1 IMAGES

3.1.1 Electronic

Capturing an electronic image of a document from hard copy (paper) is a straight-forward process consisting of feeding documents in to a scanning device, checking the resultant image, and entering a file identification of the document. The image is a replica of the original, including margin notes, signatures, graphics, date stamps, etc. which can not be captured in ASCII form. Images are the only reasonable method of capturing graphic oriented documents.

Electronic images require relatively large amounts of storage, typically 50,000 to 100,000 bytes per 8 1/2 x 11 inch page, as compared to ASCII at 2500 to 3000 bytes per page. Thus the use of images requires high density storage devices such as optical disks.

Although images are electronic, the characters or words on the page cannot be recognized by the computer until the image is processed by optical character recognition.

3.1.2 Microform

Microform is used to describe all of the reduced size photographic capture processes such as microfilm and microfiche. This type of document capture has been used for several years and is fairly automated and inexpensive. Retrieval of the proper image must be assisted by a computerized index if the files are large, and viewing of the document is usually accomplished by a projection process. Recent developments have combined the storage capabilities of microfilm with the versatility of electronic images. In this configuration, a microfilm image is located automatically in a storage device, scanned electronically, and transmitted to a terminal for viewing. This process is slower than retrieving electronic images from optical disks.

3.2 FULL-TEXT

The full text of a document may be entered into the LSS to be available to browse or read as part of the document selection process, or more likely to be used for full-text search by software or hardware. The three processes which are used to enter the full text of a document into the system are optical character recognition, rekeying, and conversion from machine readable form from word processing.

3.2.1 Optical Character Recognition (OCR) Process

The OCR process converts an electronic (bit-mapped) image of a page into

ASCII text (a bit pattern for each character and punctuation). The quality of the text produced is highly dependent on the quality of the image which 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 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:

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

The advantages 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. If correction is minimal or not required (i.e. high quality documents), costs can be as low as $.20 to $.40 per page. With many corrections (i.e. low quality documents), costs can be as much as $2.50 to $3.00 per page. If the total costs exceed $3.00 per page, it can be less expensive to key in the document directly.

Continuous improvements are being made in OCR technology which will increase speed of production and reduce the error rate. Presently OCR of an image made from scanning of 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.

3.2.2 Rekeying

Keying a document into a computer is accomplished simply by typing the characters directly on the keyboard. This rather low-tech approach is also the most costly method. At typical local service center rates of $1.00 per 1000 characters, a readable page will cost $2.50 to $3.00 to enter in ASCII form. Rekeying is the only reliable method for poor quality documents such as those produced from microform or deteriorated paper.

3.2.3 Word Processing

Documents which have been prepared on a computer by word processing software, for example, are already in machine readable format. However due to the fact that most full-text programs require that files be entered in ASCII form and computer communications are not standardized, some conversion is required. Generally speaking, tools are available for this purpose.

The major problem with receiving data in machine readable format is the quality assurance. It is necessary that the machine readable version of the document be verified as a true representation of the hard copy. (In many cases last minute changes to a document are made on a typewriter.)

Costs for this process can be minimal if the document is produced on the same computer and the conversion process is automated. Given the variety of parties and contractors associated with the repository, it is not expected that costs will be negligible for this method, but they will certainly be less than rekeying and probably less than OCR with correction.

3.3 HARD COPY

Filing of information in hard copy is the simplest and most direct form, however it is probably the most unwieldy. Given the geographic distribution of retrieval, at least two, and probably more copies of the data would be required. As with microform capture, a computer aided index is a requirement for large databases. One of the major problems with hard copy storage is security. Documents are not always returned to the files or may be misfiled. Hard copy, provided the copy is faithful to the original, is easy to read, requiring no projection device or display terminal.

4.0 CATALOGING AND INDEXING

Cataloging and indexing are processes for preparing the LSS records for retrieval. The type of cataloging is directly related to the search and retrieval techniques to be employed.

4.1 HEADERS

4.1.1 Bibliographic Headers

Bibliographic cataloging is the simplest form of a description of a document. It results in a series of descriptive terms, usually objective in nature, which can be assigned by relatively unskilled clerical personnel. Examples are author, recipient, date, title, type of document, etc. The bibliographic header represents the minimum information which might be entered into an information system about a document. It is the opinion of the technical staff that all records in the LSS should have a bibliographic header, even if more complete indexing including full-text is used.

The bibliographic header is generally typed into a "fill in the blanks" form as a document is entered into the system. The information could conceivably be provided by the organization submitting the document as part of the submission process.

4.1.2 Subject Terms

Subject terms represent an addition to the header which provides information about the material in the document. They are particularly useful for technical reports and similar lengthy documents and less important for correspondence. There are differences of opinion over the best method to assign subject terms to a document, whether by an information management (librarian) specialist, the author, an independent subject expert, or some combination. The assignment of subject terms to a document, if it is to result in successful retrieval, should be made by a highly skilled individual together with such tools as an authority list and controlled vocabulary. Cost may therefore be a major factor in considering the utility of adding subject terms to the header. While the assignment is subjective and dependent upon the skill of the individual, subject terms can enhance retrieval by incorporating terms which are not used in the text itself but are the terms normally used by the searcher. Subject terms are typically entered into fixed fields of a structured database.

4.1.3 Abstract

Adding the abstract to a header can be less costly in cases where it has been provided as part of the document. If the abstract must be created for the header, costs and the requirement for skilled individuals become a consideration. Most database programs have text fields which are sufficiently large to hold the abstract. In effect the abstract is searched in "full-text". If a document contains an abstract and is entered in searchable full-text, the abstract will of course be included automatically as a search mechanism.

4.2 FULL TEXT

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

Even after the inverted file has been created, new documents can be added to the system and the index modified to accommodate the additional information. Eventually, however, a modified index becomes inefficient to use, and a reindexing of the entire file is required.

Full text indexing, although not labor intensive, requires major computer resources and time to process large files. There are several examples, however, of commercial and government full text retrieval applications that are large and complex and still deliver reasonable indexing and retrieval response times. The files will require segmentation, although this may be invisible to the user.

5.0 STORAGE

5.1 HARD COPY

Hard copy (paper) is one possible mechanism for the information required in the LSS. The major problems with this method are the difficulties of locating documents, missing documents and pages due to misfiling or borrowing, and the space required. For 10 million pages approximately 600-700 filing cabinets occupying 4000-5000 square feet would be required. Advantages of hard copy include the readability of the document and the fact that the document is a true representation of the original including signatures.

5.2 MICROFORM

Storage in microfilm or microfiche provides a more condensed medium and therefore reduces the storage volume. Automated machinery is available to assist in locating a specific frame, but once it is found, a projection device is required in order to read the page. Quality of microform varies widely in readability and depends to a great extent on the quality of the original document. Missing documents can also be a problem with microform, but missing pages are not typical assuming the whole document was originally captured.

5.3 ELECTRONIC

To understand the electronic storage requirements for various techniques of capture and retrieval, consider an example document consisting of 5 pages of text and one page of graphic information. Storage requirements for the various cataloging and indexing forms are as follows:

	Assumption	Bytes
Bibliographic header	1500 characters	1500
Index to bibliographic header	Not all terms indexed	1000
Subject terms	10 phrases at 30 char/phrase	300
Index for subject terms	All terms indexed	300
Abstract	One-half page	1500
Inverted file of abstract	Abstract full-text searchable	1500
ASCII text of document	3000 characters/page	15,000
Inverted file of text	Full-text searchable by software	15,000
Image of graphic page	300 dpi compressed @ 20:1	55,000
Image of text pages	300 dpi compressed @ 20:1	275,000
	TOTAL	366,100

From this example, one can judge the relative impact on storage requirements of various search, retrieval, and display options.

5.3.1 Optical Disk

Optical disks represent the least cost electronic medium of storage for large volumes of data. Current optical disk technology is "write-once-read-

many" (WORM), which means that the information cannot be erased or changed. Such a medium is ideal for archival documents. Erasable optical disks are now arriving on the market, but the technology and storage density is not as advanced as WORM. A 12" optical disk storing 6.4 gigabytes can contain 100,000 pages in image form, 1,000,000 pages in indexed full-text, or headers for about 1,000,000 documents.

Optical disks can be searched randomly for files, thus resulting in faster response than serial devices such as microfilm.

5.3.2 Magnetic Tape

Magnetic tape is a relatively low cost storage medium, however it requires manual intervention (to mount the right tape on the tape reader) and retrieval is relatively slow. Magnetic tape is therefore not often used for information which must be accessed frequently, but is well suited for backup storage which is only accessed in the event of failure of the primary storage media.

5.3.3 Magnetic Disk

Magnetic disks are probably the highest cost storage media for large (gigabyte) storage requirements. Its advantage is primarily the speed of retrieval.

6.0 DISPLAY

All retrieval techniques will result in a list of "hits", i.e. documents which meet the query. Since no query technique is 100% efficient, additional review is probably required to make the final determination if the hits are indeed documents of interest to the user. This may be done on the screen by reviewing additional information on each document which may be stored in the system. Such information could be the image of each page, the ASCII text, the header, or a report such as a list of all documents by a specific author.

6.1 IMAGE

The electronic image of the page, displayed on a high-resolution terminal, provides a true representation of the original document in a form which can be read or skimmed. All markings on the page, including marginalia, signatures, and date stamps will be reproduced in the image as well as figures and graphics which cannot be stored electronically in any other form.

Images must be viewed on a high-resolution (100 dots per inch minimum) screen to be readable. The interface device between the screen and the computer will include a compression/decompression board which permits the storage of the image to be in a compressed form, approximately 1/10 to 1/30 of the original scanned image. This hardware is of course more expensive than standard monochrome monitors and interface devices.

Due to the fact that images, even in the compressed form, require some 50,000 to 100,000 bytes per page, remote transmission of images is not very practical. One page transmitted over a 2400 baud modem would take about 4 minutes.

Images can also be provided in microform and projected locally on a microfilm or microfiche reader.

6.2 ASCII TEXT

The text of the document may be available in machine readable form or it may have been created by the OCR process for the purpose of indexing the text for full-text search. If this ASCII form of the text is stored in the system, it can be viewed on demand in order to help determine if the document is indeed of interest. Note that even if the document is available for full-text search, it is the index of the text that is used by the software and the ASCII text is not necessarily maintained.

ASCII code is relatively compact storage compared to images, incorporating compression techniques to provide even more efficiency. Thus remote transmission of text is reasonable to accomplish. If the text can be transmitted to a personal computer, it can be stored, printed, and extracted for inclusion as quotes in other documents.

The text of a document contains only the alphanumeric characters and punctuation which were contained in the original document. It will not include signatures, hand-written notes, figures, or graphics.

6.3 HEADER

Output of the entire header of a document, including subject terms and abstract if they have been included, may be sufficient to determine if the document is of interest. This information will require the least amount of storage and transmission time of the possible screen outputs, and like ASCII text, will contain only alphanumeric characters.

7.0 DOCUMENT OUTPUT

Once it has been determined that a document is of interest and a more permanent record of the document is desired for detailed reading, it can be obtained in hard copy or microform.

7.1 HARD COPY

A copy of the document can be obtained in several ways:

- If the stored copy is in paper form, a photo copy can be made.
- If the stored copy is in electronic image form, a copy can be printed on a laser printer.
- If the stored copy is in microform, a "blowback" of the frame can be printed.

Any of these copies could be obtained at the LSS site, the user site, or sent by express or regular mail.

7.2 MICROFORM

A microfiche or microfilm copy of the document can be made from any of the stored forms noted above, and similarly transmitted to the user. Although storage space requirements of the user are reduced when the documents are in microform, a reader or reader/printer will be required.

7.3 FACSIMILE

Particularly when time is critical, copies of the selected documents can be transmitted to the user by facsimile devices. Cost of this alternative will be the highest, requiring not only transmission costs but also the requirement for a receiving device.

8.0 REPRESENTATIVE SCENARIOS

In this section we have attempted to define certain scenarios based on the search and retrieval techniques presented in section 2. The alternatives listed in section 2 through 7 can be combined in many forms to represent a system. These scenarios define the choices which must be made for each search and retrieval option, still leaving open the various remaining options. A possible set of scenarios are as follows:

A. A system which provides for search and retrieval on information contained in bibliographic headers only. The document could be stored on microform, electronic images, or hard copy.

B. In addition to the capabilities described in A., an abstract is added to the header which can be searched in full text.

C. In addition to the capabilities described in A., subject terms are added which can be searched.

D. A combination of B. and C. which permits searches on all header information including bibliographic, subject terms, and abstract.

E. A system which provides for full-text search of documents along with an abbreviated header. The document could be stored on microform, electronic image, or hard copy.

F. A combination of the system described in E with the capability to search headers with subject terms (C).

A. BIBLIOGRAPHIC HEADER

Document Database Creation
Options include:
- Scan pages to capture bit-mapped image
- Film pages for microfilm or microfiche
- Maintain hard copy

Cataloging/Indexing
Bibliographic header comprised of objective fields such as author, title, date, document type, accession number, etc.

Storage
Options include:
- Magnetic disk
- Magnetic tape
- Optical disk
- Microform
- Hardcopy

Display
Standard alphanumeric monitor for header information and interaction with the data base.
Optional high resolution monitor for electronic images and/or microform reader.

Document Output
Options include:
- Microform or hardcopy by mail or express
- Microform available at local workstation and printed locally
- Electronic image available at local workstation and printed locally
- Copy via facsimile device

B. BIBLIOGRAPHIC HEADER WITH ABSTRACT

All categories and options remain the same as Scenario A. except for:

Cataloging/Indexing
Bibliographic header comprised of objective fields plus the preparation of an abstract of the document.

C. BIBLIOGRAPHIC HEADER WITH SUBJECT TERMS

All categories and options remain the same as Scenario A. except for:

Cataloging/Indexing
Bibliographic header comprised of objective fields plus the selection of subject terms.

D. BIBLIOGRAPHIC HEADER WITH ABSTRACT AND SUBJECT TERMS

All categories and options remain the same as for Scenario A. except for:

Cataloging/Indexing
Bibliographic header comprised of objective fields plus the preparation of an abstract and the selection of subject terms.

E. FULL TEXT

Document Database Creation
Preparation of machine readable (ASCII) text of the document by conversion of hard copy using optical character recognition process or rekeying and conversion of documents available in word processing files.

Image of the document may optionally be prepared by:
Scanning pages to capture bit-mapped image,
Film pages for microfilm or microfiche,
or maintaining hard copy.

Cataloging/Indexing
Preparation of a bibliographic header which may be less detailed than in Scenarios A through D.
Indexing of the full text if software full text retrieval is employed.

Storage
Same options as for Scenario A.

Display
Standard alphanumeric monitor for header and text information and interaction with the data base.
Optical high resolution monitor for electronic images and/or microform reader.

Document Output
Options include:
- Microform or hardcopy by mail or express
- Microform available at local workstation and printed locally
- Printing of ASCII text on local printer
- Downloading of ASCII text to local workstation
- Electronic image available at local workstation and printed locally
- Copy via facsimile device

F. ENHANCED FULL TEXT

All categories and options remain the same as Scenario E. except for:

Cataloging/Indexing
Preparation of a bibliographic header plus the selection of subject terms.
Indexing of the text if software full text retrieval is employed.

9.0 ADDITIONAL SYSTEM PARAMETERS

The preceding sections have focused on the search and retrieval aspects of the LSS system, including the impact of certain aspects on system design. There are several additional parameters which have significant effect on the system, and since they are related to aspects of search and retrieval or display, we will mention them here. Decisions on these aspects must be made as well before the system requirements can be complete and design specifications can be formulated. These parameters include:

1) Data volume - total number of documents and pages.

2) Response time - time to respond to a request such as a query or a request to print.

3) Geographic distribution - locations of end users and data input.

4) Number of users - especially the number who may use the system simultaneously.

5) Type of users - which will affect types of queries and the user interface.

6) Centralized versus distributed - location(s) of the data base.

7) Technology - constantly providing new capabilities and lowering the cost of existing capabilities.

8) Cost.

APPENDIX

GLOSSARY OF THE
HLW ADVISORY COMMITTEE

GLOSSARY

ABSTRACT

Summary of the main points in a document, usually organized around the theory of the case or subject matter at issue; also called digest; most common use in discovery systems is to summarize portions of transcripts.

ASCII

ASCII is the acronym for American Standard Code for Information Interchange. This is the system by which letters, punctuation characters, spaces, some special symbols and control codes are encoded into numeric values for interpretation and storage by a computer.

ASCII FILE

An ASCII FILE is a TEXT FILE containing the ASCII codes which represent characters and symbols (as opposed to an IMAGE FILE which contains the data to actually draw these characters). See also BIT-MAPS.

BIT

BIT stands for BInary digiT. It represents the smallest unit of information in a digital computer. It can have a value of either 1 or 0, and can be represented by a switch (which is either on or off).

BIT-MAP

Rather than storing the information on a page of text as a series of ASCII codes which represent the characters on that page, an IMAGE of that page may be created and stored in a computer. This IMAGE consists of a large number of BITS (ranging from x to y per page of typed text), where the zeros and ones stored by the BITS represent the white and black portions of the page at high RESOLUTION. Such an image is called a BIT-MAP. When displayed, a BIT-MAP can be interpreted only by a human user who "reads" the image; it is not meaningful to computer programs. A FILE containing a BIT-MAP may be copied, moved, displayed or printed by a computer system.

BOOLEAN LOGIC

Boolean logic (or Boolean algebra) is a system of logical functions and operators which permit computations and operations on binary (true/false) operations. This system was developed by and named after George Boole, an English mathematician (1815-1864).

BYTE

A BYTE is the basic unit of data storage. A BYTE is made up of a certain number of BITS. This number depends on the architecture of the computer, but is always divisible by two (with no remainder). The full ASCII code requires at least 8 BITS per BYTE, which is the minimum number found in conventional computers.

CATALOGING

CATALOGING is the process of describing a document being entered into a collection (e.g. a library or DATA BASE management system). The object of CATALOGING is to extract (or assign) the information necessary to access (find) the document without having to examine

sequentially each document in the collection. CATALOGING information may be used in INDICES of the collection. (See HEADER)

CD-ROM (or Compact Disk - Read Only Memory)
Some OPTICAL DISK systems use disks which have had data written to the disk by special reproduction equipment, and can only been read by the computer system onto which they are installed. When such disks (or disk systems) are Compact Disk format, they are called CD-ROMs.

CD-WORM (or Compact Disk - Write Once, Read Many-times)
Some OPTICAL DISK systems can write to disks as well as read them. Unlike magnetic disk storage devices, these systems can not erase and re-write information. When such disks (or disk systems) are Compact Disk format, they are called CD-WORMs. To modify a FILE stored on such a system, the entire file (including the correction) must be re-written. The new and old versions are distinguished by VERSION NUMBERS.

CODING See CATALOGING

CONTROLLED VOCABULARY
List of terms or phrases which are maintained for continuity of spelling and usage, such as authors, addresses, organizational abbreviations, document types, subject terms. (Also known as authority list)

CHARACTER RECOGNITION ENGINE
A device designed to convert a BIT MAP IMAGE of a document into an ASCII file is called a CHARACTER RECOGNITION ENGINE. Simple versions are designed to recognize specific character sets (font recognition devices) while more complex versions are programmed to recognize specific characters by their unique topology.

DATA BASE
An organized body of information on a pre-determined topic is a DATA BASE. Related DATA BASES can be logically or physically combined to constitute a larger and more detailed DATA BASE on a broader subject. A DATA BASE can be envisioned as a set of file cabinets, containing completed forms of a given kind. Each completed form is called a RECORD, each question on the form is a FIELD, and each completed question is the contents of that FIELD.

DOCUMENT FILES
A DOCUMENT FILE (or simply a "document", when this usage would not confuse the FILE with the physical document it represents) is the basic type of data stored in a computerized archive system such as the LSS. A DOCUMENT FILE is a TEXT FILE which contains the contents of a physical document; it and may also contain a HEADER.

E-MAIL
"Electronic Mail"; creation, storage and transmission of word processing documents from computer to computer.

FIELD

A RECORD may be subdivided into FIELDS, just as a form can consist of a number of blanks into which information can be entered. The data to be entered in a FIELD is determined by the FIELD'S definition. A completed set of FIELDS is called a RECORD. Examples include author, date, title, abstract.

FILE

A FILE is a unit of data storage. A FILE is identified by a FILENAME, and contains a collection of related data. These data need not be further organized (i.e., they may simply be a STRING of BYTES) or they may be subdivided further into named FIELDS.

FILENAME

Each FILE stored on a computer system can be identified by a FILENAME. Such a name is either unique to a FILE, or files with the same name can be distinguished by their location within the computer's FILE STRUCTURE, or by the VERSION NUMBER of the FILE.

FULL TEXT

The version of the document as it resides on a computer system for display ("linear file" in retrieval terms).

FULL TEXT SEARCHING

FULL TEXT SEARCHING is a computerized text processing technique which locates the occurrence of specific words or groups of words within a TEXT FILE. Logical relationships can be specified by Boolean logic expressions when stating the search condition (e.g. "Find places in the text where 'hot' and 'cold' occur within the same physical paragraph") and proximity expressions. Software FULL TEXT SEARCHING techniques require INVERTED FILES while hardware techniques stream the entire portion of the DATA BASE being examined through a hardware comparator, and do not require such files.

HARD COPY

A HARD COPY is a paper copy of a document. It can be the paper original, a photocopy or a telefax copy, for example.

HEADER

A TEXT FILE in a computerized archive system such as the LSS generally contains the contents of a physical document, stored as ASCII codes of the text within that document. In addition to this text, CATALOGING information can be appended to the beginning (or "head") of the document. Such a HEADER may contain a variety of information in FIELDS, which may be accessed directly by DATA BASE management software (for INDEXED SEARCHING) or may be accessed by FULL TEXT SEARCH software (either independently or along with the body of the text from the document). Headers are also known as surrogates, document coding forms, DCF's, bibliographic citations and "identified" in the NRC consensus document on the rulemaking issues.

IMAGE

An IMAGE of a page visually presents the information on that page. This image is meaningful only to a human user, and can not be

interpreted by computer programs. Examples of document images are photocopies, telefax copies, microfiche and BIT-MAP IMAGE FILES.

IMAGE COMPRESSION
The number of BITS in an uncompressed IMAGE FILE of a page of text is equal to the area of the page times the RESOLUTION of the IMAGE (plus a few additional BITS required by all FILES). The amount of memory required to store this IMAGE can be reduced by IMAGE COMPRESSION techniques.

IMAGE FILE
An IMAGE FILE is a computer FILE containing a BIT-MAP of a document IMAGE. The number of BITS in an uncompressed IMAGE FILE of a page of text is equal to the area of the page times the RESOLUTION of the IMAGE (plus a few additional BITS required by all FILES).

INDEX (plural INDICES)
There are a variety of logical ways to physically arrange a collection of documents (e.g. alphabetically by author or by title, chronologically by date produced or entered into the collection). Each of these ways is designed to help access (find) a document based on a specific strategy for finding it. Unfortunately, a collection cannot be organized simultaneously in each of these ways. In order to make each strategy possible, surrogate collections can be created which contain the key information (sorted appropriately) and the location of the document. In libraries, these surrogate collections are the author catalog and subject catalog. Such DATA BASE surrogates constitute INDICES of the collection.

INDEXED SEARCH
INDEXED SEARCHING, the conventional method used by DATA BASE management software to access data, searches INDICES constructed to support the specific type of queries. This is distinguished from FULL TEXT SEARCHING, which searches the TEXT FILE (or corresponding INVERTED FILE, in the case of FULL TEXT SEARCH software) that has not been otherwise organized for retrieval.

INVERTED FILE
Software FULL TEXT SEARCH techniques do not directly search a TEXT FILE at the time the search request is made (as do word processing programs when searching for a STRING). Rather, the TEXT FILE is pre-processed to create a file containing the words in the TEXT FILE and pointers to their locations. The INVERTED FILE can be searched much faster than the original FILE since it has been pre-sorted.

KEYWORD
Accessing documents in a collection can be facilitated by assigning KEYWORDS to the document (or a RECORD representing it in a DATA BASE) during CATALOGING. KEYWORDS are words that describe the document's contents and are best assigned from a CONTROLLED VOCABULARY, preferably with the aid of a THESAURUS.

KEYWORD IN CONTEXT (KWIC)
Words in the FULL TEXT document, including words located before and after the keyword.

KEYWORDING
A part of CATALOGING, KEYWORDING is the processes of assigning KEYWORDS. KEYWORDS are generally assigned from a CONTROLLED VOCABULARY, and are most useful when based upon a THESAURUS.

OCR (or Optical Character Recognition)
A device or process which converts HARD COPY text into an ASCII file by using a CHARACTER RECOGNITION ENGINE.

OPTICAL DISK
An OPTICAL DISK is a computer data storage system, such a CD-ROM or CD-WORM disk drive, which records BITS as the presence or absence of minute pits on a glass disk. The system is "optical" since laser light is used to write and read this data from the disk.

PIXEL
An IMAGE can be represented by a large number of small spots (usually in rows and columns). These spots, which can be either black or white, are called PIXELS (from "picture elements").

PROTOTYPE
In compiling the information necessary to design and build a large DATA BASE management system, a system PROTOTYPE can be used to estimate quantitative performance information about components of a larger system to be built, and can be used to quantify and evaluate the behavior and response of users to software while it is being developed. Such a PROTOTYPE consists of hardware test environment in which specific components can be interfaced and evaluated, a software environment which can run a simulation (or simplified version) of software to be used in the complete system, and a test DATA BASE (representative of, but significantly smaller than the final DATA BASE) which can be used to test user behavior, software and hardware performance and DATA BASE organization.

RECORD
A RECORD is a group of one or more related FIELDS, containing data. A DATA BASE generally consists of group of RECORDS, each containing a group of related data in the subject of the DATA BASE. These can be considered individual completed forms in a file cabinet which represents the DATA BASE.

RESOLUTION
The RESOLUTION of a BIT MAP IMAGE is the number of PIXELS per unit area. If no IMAGE COMPRESSION has occurred, the number of BITS needed to store an IMAGE FILE is equal to the number of PIXELS in the IMAGE.

SCANNER
A SCANNER is a device which converts HARD COPY text into a BIT-MAP IMAGE.

STRING
A character STRING is a series of characters represented by their ASCII codes.

SUBJECT TERMS
Words or phrases assigned to a document during subjective CATALOGING, to represent the overall concept presented by a document. SUBJECT TERMS are usually selected from a hierarchical CONTROLLED VOCABULARY list, such as the DOE Keyword Dictionary, and are assigned at the closest level of detail.

SYNONYM FILE
One aspect of a THESAURUS is to identify words (or phrases) which have the same meaning (synonyms), and to select one which is used to represent and replace the others during KEYWORDING. A FILE containing such groups of related words is a SYNONYM FILE. Such a FILE can be used with some sophisticated FULL TEXT SEARCH software, so that each synonym is found in a search if any of a group of synonyms from the FILE are sought.

TEXT FILE
A TEXT FILE has its characters stored as ASCII codes, as opposed to IMAGE FILES where the shape of the character is stored in BIT-MAP form. TEXT FILES in the LSS generally contain the the text of documents in the system, and are therefore often referred to as DOCUMENT FILES (or simply, "documents", when this would not confuse them with physical documents).

THESAURUS
A THESAURUS is a CONTROLLED VOCABULARY with embedded instructions and relationships which assist in assigning KEYWORDS or SUBJECT TERMS consistently and logically during CATALOGING. THESAURI can be used for developing a search strategy at a precise level of detail and may contain broader, narrower, and related terms (synonyms). Also called taxonomy and classification scheme.

VERSION NUMBER
When FILES are modified in many computer systems, previous versions of the FILE are retained under the same FILENAME. To distinguish between versions, VERSION NUMBERS are assigned.