Department of Energy
Washington, DC 20585

JAN 3 1 1990

Mr. John Hoyle
Chairman
LSS Advisory Review Panel
U.S. Nuclear Regulatory Commission
Washington, D.C. 20555

Re: Background Materials on LSS Headers

Dear Mr. Hoyle:

In response to the discussions held at the December, 1989 Advisory Review Panel meeting in Reno, we are forwarding materials related to LSS bibliographic header development. Four documents trace the header development process from late 1987 through May, 1988 and are enclosed for your information.

We also checked with Mr. Richard Pierce (of SAIC's LSS design project team), who participated in the technical working group during the negotiated rulemaking process, as to the status of the headers at the time when the rulemaking process was completed. His recollection is that the technical working group and the negotiating committee were able to develop a list of fields only for the submitters' headers but not the more comprehensive version required for the LSS environment.

He stated that the composition of the LSS header was an issue that was deferred, to be addressed at a later time by the Panel and the Office of the LSS Administrator. This seems to be consistent with the final rule, Sections 2.1011 (f)(1) where the Panel is to provide advice "...on the fundamental issues of the design and development...", and 2.1011 (f)(2)(i) where the Panel is to provide advice on "...format standards for the submission of documentary material...such as...bibliographic headers..." and with the broader mandate to the Office of the LSS Administrator and the Advisory Review Panel provided in Sections 2.1011 (d)(8) and 2.1011 (d)(14).

In reviewing our files, however, we noted that the documentation trail ends somewhat abruptly and that there is a "missing" piece of documentation -- that being some acceptance or affirmation by the negotiating committee of the final piece of documentation entitled "Draft Bibliographic Header Fields, Rev. 3, 5-17-88". I think it would be useful to all the potential participants if the Panel can definitize the list of fields for submitters' headers at the next meeting.

Please feel free to contact Dan Graser of my staff at 586-4589 if you require any further background information or assistance.

Sincerely,

Barbara A. Cerny
Director
Information Resources Management
Division
Office of Civilian Radioactive Waste
Management

Enclosures

1. "Draft Bibliographic Header Fields", Rev.3, 5-17-88
2. Draft "Minutes of the HLW Licensing Support System Advisory Committee Meeting", April 18-19, 1988, Washington, D.C.
3. "Information Retrieval Systems: A Tutorial" Prepared by Negotiated Rulemaking Technical Staff, February 3, 1988
4. Attendance List and Attachment 8 (Glossary of Terms), "Meeting of the HLW Licensing Support System Advisory Committee", November 19-20, 1987

cc:
L. Desell, RW-331

DRAFT BIBLIOGRAPHIC HEADER FIELDS
Rev. 3 5-17-88

The fields in the following list are considered by the Technical Staff to be either required to filled in by each participating organization submitting documents to the LSS, or in some cases are optional. They are expected to be a subset of the "full" header to be used in the LSS. Some fields are applicable to only certain types of documents, however. For this purpose a document is considered to be any document which can stand alone and could possibly be searched by a user, whether or not it is an attachment or enclosure to another document. A letter with three stand-alone attachments would require 4 bibliographic headers to be submitted - one for each of the letter and attachments. It will, of course, be necessary to develop detailed coding instructions on how to fill out the bibliographic header.

REQUIRED FIELDS:

Accession No.[2] (non-system) - This would be a unique alpha numeric consecutive number assigned by the submitting agency for two purposes:

1. To distinguish one agency's submitted documents from another's, thus allowing an agency to retrieve all of its documents.

2. To perform a control function, i.e., ensuring that every submitted document from an agency is received and entered into the LSS.

Submitter Center[1] - the office, site, division, etc. that is submitting the document to the LSS.

Document Type[1] - the format in which the information is presented, e.g. correspondence, report, regulation, etc.

Number of pages[2] - the length of the entire document represented as one number.

Title[2] - the title that appears on the document.

Description - in cases where there is no title or the title does not convey sufficient information, this is a brief description of the document, e.g., "letter concerning Negotiated Rule-Making Committee Meeting Agenda" or "Progress report for April 1988 - June 1988".

Author(s)[2] - the name of each individual authoring the article, report, etc.

Author Organization(s)[1] - the name of the organization, corporation, or agency producing the document or the corresponding organization, corporation or agency to which the author belongs.

Sponsoring Agency[1] - the agency(cies) who provided the funding for the work performed in the document.

DRAFT BIBLIOGRAPHIC HEADER FIELDS
(continued)

Recipient(s)[2] - the name(s) of those persons receiving the document either as the addressee(s), the distribution list, or the recipients of copies ("cc" or "bcc").

Recipient Organization(s)[1] - the corresponding organization, corporation, or agency to which the recipient belongs.

Journal Information[1] - if the document is an article from a journal, the name and other journal information that would distinguish the article.

Document Date[2] - the date contained on the document that is the date that the document was created or printed.

Errata Date[2] - if the document is an errata sheet, the date of these corrections.

Contract No.[2] - the contract number, if any, under which the work reported in the document was performed.

Document or Report No.(s)[2] - the number(s) assigned to the document by the producers and by the sponsoring agency(ies) if any

Edition - the version of a document, whether draft, revision, supplement, etc.

Meeting Date[2] - the date referenced in or included in the text of a document of a meeting that has taken or will take place.

Site of Activity[1] - the location, if pertinent, to which the work in the document pertains.

Document Reference[1] - The document whose content or production is influenced by the submitted document.

Image/ASCII Identifier[2] - Microform frame number or file identification of corresponding image and file identification of corresponding ASCII file.

Protected[1] - The type of privilege or protection (if any) being claimed for the document.

Document Condition[1] - terms such as pages missing, illegible portions, attachments missing, marginalia present, etc.

Parent Document Identification[2] - Accession number of the parent document if this is a stand alone attachment or enclosure, or the accession number(s) of the stand-alone attachments or enclosures if this is a parent document.

Abstract for Non-Documents - a full description of the item including such information as dates, purpose, physical description, location, etc. For raw data - a full description of the data including such items as how the data was collected, format, purpose, type, dates, etc.

DRAFT BIBLIOGRAPHIC HEADER FIELDS
(continued)

THE FOLLOWING FIELDS ARE OPTIONAL:

Descriptors[1] - terms assigned from the LSS Thesaurus that best represent the content of the document. (Use of this field requires adherence to additional LSS coding procedures.)

Identifiers - terms that are not contained in the Thesaurus that the submitter believes will assist a user in retrieving the document; these may be "buzz words" or words representing new concepts that have not yet appeared in the Thesaurus.

Comments - any information not contained in the listed fields that would be helpful to the LSS catalogers.

Abstract - a summary of the contents of the document.

Notes:

1 - governed by an authority list

2 - governed by format rules

May 3, 1988

---------- D R A F T ----------

MINUTES OF THE HLW LICENSING SUPPORT SYSTEM ADVISORY COMMITTEE MEETING

APRIL 18-19, 1988
Washington, D.C.

MEETING LOCATION AND ATTENDANCE

The sixth meeting of the HLW Licensing Support System Advisory Committee (hereafter referred to as the committee) was held on March 18, 1988 from 9:00 a.m. to 5:00 p.m. and April 19, 1988 from 9:00 a.m. to 3:30 p.m. The meeting was held in the offices of The Conservation Foundation in Washington, D.C.

A list of committee members and members of the public who attended this meeting is appended hereto as Attachment 1.

APPROVAL OF THE MINUTES

As its first item of business, the committee discussed the draft minutes from the committee's March 22-24, 1987 meeting. Several committee members indicated that they had not had time to review these draft minutes in sufficient detail. Others indicated that they would provide suggestions to the facilitator for changes that they felt were relatively minor and non-substantive in nature. Thus, no changes to the draft minutes of the March meeting were officially approved by the committee.

EXPLANATION OF CHANGES MADE TO THE NRC'S DRAFT RULE

NRC representatives explained that the draft text of a new Subpart J to 10 CFR Part 2 that was distributed to committee

With no other general questions or comments, the committee agreed to take a recess to provide committee members who had not yet seen the newly revised text an opportunity to review it in detail. The committee also agree that upon reconvening, they would discuss the draft rule section by section.

DISCUSSION OF THE DRAFT RULE

Section 2.1000 - Scope of Subpart

NRC representatives explained that the intent of this section was to incorporate by reference certain provisions of Subpart G, NRC's rules of general applicability, to the rule for the HLW licensing proceeding which will be published as Subpart J in Part 2.

NRC was asked why sections 2.740 and 2.741 were not listed in the provisions of Subpart G that would be incorporated by reference. NRC representatives responded that these sections were essentially lifted verbatim, with minor changes to accomodate the special circumstances of the HLW licensing proceeding and the proposed use of the LSS into sections 2.1018 and 2.1019 of this draft rule.

Section 2.1001 - Definitions

Bibliographic Header The representative of the environmental coalition stated that the definition used for this term might be a problem because of the limitations that are placed on public access to the LSS under Section 2.1007. The facilitator briefly reported on the activities of the technical

work group which, he explained, is likely to recommend that the parties be required to complete a simple "bibliographic header," which would include information on such item as the date, author, recipient and subject of the document, and that the LSS Administrator would be required to prepare a more complete header for the document which would include more information than that supplied by the party. This additional information might include such items as keywords and an abstract of the document. NRC representatives explained that their intent was to leave this issue open for now and resolve it at some later date through the LSS Administrator and the use of the proposed advisory review board which will make recommendations to the LSS Administrator. No specific changes to this definition were suggested.

Document NRC representatives were asked what the phrase "associated with the business of" was meant to imply. They replied that they intented that this phrase would make it clear that contractor documents as well as agency documents were meant to be included in the LSS. The committee agreed to stike the part of this definition that was added by the NRC negotiating team from the definition used in the original text, such that the definition would read: "Document means any written, printed, recorded, magnetic, graphic matter or other documentary material, regardless of form or characteristic."

EEI representatives stated that the term documentary material was not defined in this definition section but it was defined in the text of the rule under Section 2.1003. The committee ageed that the sentence which defined this term in

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 B 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.

5.3 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.

WM Record File
D1056

WM Project
Docket No.
PDR
LPDR

Distribution:
Kalman
Altomare
JOB

(Return to WM, 623-SS)

ATTENDANCE LIST

Meeting of the
HLW Licensing Support System Advisory Committee
November 19-20, 1987

COMMITTEE MEMBERS (Including Spokespersons and Alternates)

Priscilla Attean
Penebscot Nation

Dennis Bechtel
Clark County, Nevada

Steve Bradhurst
Nye County, Nevada

Francis X. Cameron
Office of the General Counsel
U.S. Nuclear Regulatory Commission

Barbara Cerny
DOE

Don Christy
Nuclear Waste Office
State of Mississippi

Bill Clausen
State of Minnesota

Stan Echols
Office of the General Counsel
U.S. Department of Energy

Kevin Gover
Special Counsel
Nez Perce Nuclear Waste Program

Ronald T. Halfmoon
Nuclear Waste Program
Nez Perce Tribe

Robert Halstead
Radioactive Waste Review Board
State of Wisconsin

Alice Hector
Attorney for the Texas Nuclear
Waste Task Force
Hector and Associates

ATTACHMENT 8

GLOSSARY OF TECHNICAL TERMS

The following represents an initial consensus on the definition of technical terms following the November meeting in Denver. It is not complete and will be enlarged as the participants request clarification. In come instances, the terms are somewhat specific to the HLW terminology already developed, rather than the most representative or precise definition in current "discovery" or "litigation support" glossaries.

Header

Technique of coding a document, process or materials by describing its parts, usually know as "fields":

Bibliographic Header (simple coding)
- Document Number
- Date
- Author(s)
- Addressee(s)
- Copies Sent To
- Title
- Description (if title not clear)
- Document Type

Enhanced Header (usually includes some subjective analysis of the content of a document)
- Abstract
- Thesaurus, taxonomy
- Subject Terms

Additions Case-specific Fields, e.g.,
- Docket File Code
- Contract Number
- Report Number
- Concurrence List

Headers are also know as surrogates, DCF's, "coding forms", or bibliographic citations. The term "identified in the LSS" has been used in the NRC Position Paper to signify the use of a header.

Searchable Header

The information in the header after it has been indexed by a computer program and made available for searching on a computerized retrieval system

Hard Copy Document

The paper document or copy of it ("hard copy")

Image	The microfilm, microfiche or optical disk ("bit-mapped") version of the hard copy document
Full Text	The version of the document as it reside in a computer system for display ("linear file" in retrieval terms)
Searchable Full Text	All the words (except "stop" words) in the document after it has been indexed by a "full text" computer program and made available for searching on a computerized "full text" retrieval system ("inverted file" in retrieval terms)
Enhanced Full Text	Full text plus header or some additional way of describing a document
Keywords	Words in the searchable full text document; to avoid confusion, not used here to refer to a field in a header
Subject Terms	Words, terms and phrases created especially for a specific case or fact situation; usually included in an "enhanced header"
Fields	Parts which make up headers, e.g., author, title, date, abstract
OCR	Optical Character Reader; a device which converts hard copy text into computer-readable words
Optical Disk	A media (plastic disk) for storing large quantities of electronic data in the form of images, text or searchable words and phrases
CD-ROM	A form of optical disk commonly used for storage of electronic data
E-Mail	"Electronic Mail"; creation, storage and transmission of word processing documents from computer to computer
Record	e.g., hard copy document, geologic core sample, photograph, image, magnetic tape or disk