UNITED STATES
NUCLEAR REGULATORY COMMISSION
WASHINGTON, D.C. 20555

OFFICE OF THE
LSS ADMINISTRATOR

October 10, 1991

MEMORANDUM FOR: Dade W. Moeller, Chairman, ACNW
B. Paul Cotter, Jr., Chairman, ASLBP
William C. Parler, General Counsel
Robert M. Bernero, Director, NMSS
Eric S. Beckjord, Director, RES

FROM: Lloyd J. Donnelly
LSS Administrator

SUBJECT: STRUCTURING THE LSS FOR EFFECTIVE ACCESS TO RAW DATA AND OTHER NON-TEXTUAL DOCUMENTARY MATERIAL

Enclosed is a recent report from the Center for Nuclear Waste Regulatory Analysis (CNWRA) which examines a very important LSS design consideration. The CNWRA's report provides recommendations on how the LSS might be structured to provide effective access to millions of pages of raw data and other non-textual documentary material being accumulated in data packages at the Yucca Mountain Project Office. These data are primarily the product of scientific investigations that contribute to technical reports and other finished products that DOE will use to support its HLW repository license application.

I am soliciting your office's review and comment on the issues and recommendations contained in the Center's report. Your input is important if the LSS is to effectively serve both the technical and legal communities. My office is also examining the report and we want to fully consider your comments/suggestions before making a proposal to the LSS Advisory Review Panel. I believe a joint meeting with representatives from all our offices and the CNWRA would be a productive way to obtain your views and examine different options for effectively searching and retrieving the data under consideration. You will be contacted in the near future to arrange for such a meeting which I would like to target for the last week of this month. Betsy Shelburne, x24027, of my staff is available to answer any questions your staff may have prior to the meeting.

Enclosure:
As stated

cc w/o enclosure:
The Chairman
Commissioner Rogers
Commissioner Curtiss
Commissioner Remick
J. Taylor, EDO
S. Chilk, SECY
G. Cranford, IRM

ALTERNATIVE WAYS OF MAKING PACKAGED DOCUMENTARY MATERIALS ACCESSIBLE WITHIN THE LICENSING SUPPORT SYSTEM

TECHNICAL REPORT

Prepared for

Nuclear Regulatory Commission
Contract NRC-02-88-005

Prepared by

Center for Nuclear Waste Regulatory Analyses
San Antonio, Texas

Center for Nuclear Waste Regulatory Analyses

P.O. DRAWER 28510 • 6220 CULEBRA ROAD • SAN ANTONIO, TEXAS, U.S.A. 78228-0510
(512) 522-5160 • FAX (512) 522-5155

September 27, 1991
Contract No. NRC-02-88-005
Account No. 20-3705-001

U. S. Nuclear Regulatory Commission
Attn: Ms. Betsy Shelburne
Office of Licensing Support System Administrator
Mail Stop EWW571
Washington, D. C. 20555

Subject: Interim Report on "Alternative Ways of Making Packaged Documentary Materials Accessible Within the Licensing Support System," Intermediate Milestone No. 20-3705-001-216-001.

Dear Ms. Shelburne:

Enclosed for your review and comments is the subject report. The report consists of two volumes; a Technical Report, and Exhibits. The body of the technical report was written as concisely as possible to permit timely reviews by the potentially large group of reviewers. Therefore, the treatment of subjects requiring more elaboration have been included as appendices. Of course, it is intended that anyone doing a serious review of the report and involved in the decision making will review and comment on both the main report and appendices. Additionally, the Exhibits volume contains actual documentary materials and is intended to give the reviewer the most realistic view possible of the existing situation. We hope you find the information provided in this report complete and very significant relative to this important issue of unitization of packaged documentary materials.

We look forward to your review and comments and the opportunity to discuss the Center's recommendations with you at your earliest opportunity.

Sincerely,

Rawley D. Johnson

Rawley D. Johnson, Director
Information Management Systems

RDJ/mag
f:AltWays.ltr
Enclosures
cc: J. Funches
S. Fortuna
B. Stiltenpole
S. Mearse
L. Donnelly
C. Cameron
J. Latz

CNWRA Directors/Element Managers
S. Young
R. Marshall
S. McPaddin
C. Acree
J. Cooper
S. Rowe

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

ALTERNATIVE WAYS OF MAKING PACKAGED DOCUMENTARY MATERIALS ACCESSIBLE WITHIN THE LICENSING SUPPORT SYSTEM

Prepared for

Nuclear Regulatory Commission
Contract No. 02-88-005

Prepared by

Rawley D. Johnson
Stephen R. Young
Charles L. Acree, Jr.
Joseph H. Cooper

September 1991

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

LIST OF FIGURES

LIST OF TABLES

EXECUTIVE SUMMARY

A massive amount of documentary material relating to the licensing of a high-level waste repository is expected to be unsuitable for entry into the Licensing Support System (LSS) in the form of searchable full text. Some of this "raw data" - the portion of it that consists mainly of graphic and handwritten materials, can at least be scanned for viewing as images on LSS screens. The rest, which consists mainly of magnetic media, will have to be examined by requestors through means other than electronic display.

Since requestors will be unable to avail themselves of LSS text-search techniques to find information that these materials contain, they will have to rely entirely on the adequacy of the bibliographic "headers" (indexes) that will be written to describe them.

The governing LSS Rule provides for the "packaging" of all types of documentary material, as a method of maintaining the collective integrity of the items pertaining to a particular investigative activity by storing them together. The Rule does not specify, however, how packaged material should be "named and identified," aside from the submittal of a bibliographic header based upon a package's table of contents.

This raises the question addressed in this report: Is a single header per package sufficient, or should supplementary headers be applied to identifiable "units" within a package, to maximize the likelihood that a requestor will be able to retrieve all needed information?

The Department of Energy's Yucca Mountain Project Office (YMPO) has been packaging large portions of its documentary material, which it calls "technical data," and has been "unitizing" and indexing its packages in a way that it believes will be compatible with LSS requirements.

Alternative approaches to the unitization of packages are best evaluated in terms of the merits of separately indexing the materials that are normally included within packages: finished investigative products (mostly technical reports), textual commentary on those products (reviews and authorizations), scannable graphic/handwritten material, and non-scannable "machine dependent" material (mostly magnetic media).

It is recommended that only the finished products and machine-dependent items should be assigned individual, supplementary headers - primarily on the basis that the commentary and the graphic/handwritten material that is included can be adequately described for retrieval purposes within a package's text-searchable table of contents.

An appendix provides a model that interrelates the various classes of anticipated documentary material, in support of the recommendations. Other appendices expand on important concepts; explain how "hypertext" indexing can provide the capability to "turn" rapidly to the initial page of a particular packaged item; discuss related issues such as the timing of package submission and "scannability"; tell how descriptive headers may be written; and envision a scenario for the entry of packaged material into the LSS.

An extensive set of exhibits includes several tables of contents that have been copied from existing YMPO packages, appropriate (attached) examples of graphic/handwritten material, and recent editions of YMPO catalogs and publications that illustrate the abundance and variety of technical data.

1. BACKGROUND

The Licensing Support System (LSS) Administrator (LSSA) has tasked the Center for Nuclear Waste Regulatory Analyses to develop recommended access protocols for those portions of documentary material related to the licensing of a high-level waste (HLW) repository that will be unsuitable for entry into the LSS in the form of searchable full text.

When the Center briefed the LSS Advisory Review Panel (LSSARP) on this task in October 1990, it pointed out that much of this material is currently being stored and indexed within "packages," a practice which required examination. Following receipt of the Center's preliminary report on its general task in February 1991, the LSSA asked the Center to write a special, interim report on alternative ways in which packaged material could be made accessible within the LSS. This report represents the Center's analysis of that subject

1.1. TEXT SEARCH

Searchable full text is important to the LSS because, in ordinary circumstances, the LSS will offer requestors two ways to find documentary materials that are potentially relevant to a query. One way will be to conduct a search through the documents' bibliographic headers, which can provide reference to them by title, author, and other attributes selected by the requestor. The other way will be to conduct a search, in effect, through every textual page of every document stored within the system, seeking to match specified keywords against the words that each document contains in order to find references to relevant materials. See Appendix A for more complete treatment of this subject.

The first method, bibliographic search, is traditional and has become commonplace for automated information storage and retrieval systems. Its success is entirely dependent upon the care that is extended to the preparation of bibliographic headers at the time that materials are cataloged into the system

The second method, full-text search, is becoming more widely used in computer systems as the costs of mass storage have declined and the capability to examine reams of material expeditiously has increased. The obvious prerequisite for this method is the availability within the system of searchable text, which a computer stores as strings of characters in standardized "ASCII" code. Some practical method of entering text into the system must, therefore, be employed. Simply keying it into the computer, when documents exist only on paper, is burdensome and would be cost prohibitive in the case of the LSS, which is expected to store millions of pages. Fortunately, there is an available technology, called optical character recognition (OCR), which is capable of reducing the input burden considerably.

The OCR technology that exists today cannot be expected to "read" everything that is presented to it in textual form. It can only recognize print fonts which it is designed to read, and, even then, has trouble identifying characters that are poorly imprinted, requiring clerical intervention to capture all of the assigned text correctly. While this technology has improved

during the past two decades and can increase its yield through parallel processing, it is expected to be handicapped in this essential way for many years to come.

Much of the text that is written these days, of course, within the HLW environment and elsewhere, is prepared on word processors or by computer, and is transferred via telecommunications links. Thus, it is available, at some point, in a digital form that could be input automatically and unerringly into a text file without any need for OCR equipment. The problem here is one of bringing this digital text together for LSS entry in a convenient, standard way that would be more satisfactory than printing it for OCR input.

Leaving these considerations aside, there is a huge amount of documentary material expected to be incorporated within the LSS framework that will clearly be unsuitable for text search - material such as handwritten notes, graphs, maps, photographs, sketches, numeric tables, and computer tapes. Exhibit I contains many examples.

By one estimate, approximately half of the existing and foreseen documentary material to be managed by the LSS will consist of items such as these, which will be inaccessible to requestors through text search and may therefore be found only through a search of bibliographic headers.

1.2. CHARACTERIZING AND LABELING NON-TEXT-SEARCHABLE MATERIAL

For the sake of convenience, "non-text-searchable" documentary material will be called simply that in this report. The Center has been unable to devise a descriptive, yet all-encompassing label that would be any less awkward.

The "LSS Rule" (54 FR 14925, dated April 14, 1989) distinguishes between two kinds of non-text-searchable material: 1) "graphic-oriented" material which can be captured in bit-mapped form by means of a digital scanning device and can be subsequently viewed by requestors as images on LSS screens, and 2) material which is not suitable for entry into the LSS in image form and will therefore have to be physically retrieved.

For the purposes of this report, it will be worthwhile to apply convenient labels to these two kinds of material that will be more descriptive than merely using the adjectives "scannable" and "non-scannable," or "imageable" and "non-imageable."

The scannable material cannot very well be labeled "graphic," because so many handwritten items are included - field notes, laboratory notes, daily logs, and filled-in forms. To call it "non-textual" would obscure the fact that it includes both handwritten text (viewed as mere scribbles by OCR equipment) and embedded/explanatory text. To call it "untyped" would ignore the thousands of partially-typed pages, typed-in forms, and tabular numeric-array presentations that are included. To call the items "non-documents" would be ambiguous because they definitely constitute documentary material.

This report will use the term "graphic/handwritten" for the scannable material, because it is predominantly graphic or handwritten.

The non-scannable material, which includes computer and film media, will be called "machine-dependent." Generally speaking, it requires the help of an electronic or optical/mechanical device to be examined.

Figure 1-1 displays these categories of documentary material as they relate to "hardcopy" textual material within the LSS and to packages. Appendix B provides a more detailed classification scheme, based upon a study of available materials, and provides a model of their inter-relationships.

It cannot be emphasized too strongly that the inability of LSS requestors to retrieve non-text-searchable materials (whether they are graphic/handwritten or machine-dependent) using text-search technique means that their successful retrieval will be entirely dependent upon the adequacy of the bibliographic headers making reference to them. That is why decisions which set the rules governing the creation of their headers are so vitally important.

1.3. PACKAGING

Given the vast amount of non-text-searchable material that must somehow be incorporated within the LSS framework, a concept called "packaging," also defined by the LSS Rule, offers significant potential for streamlined processing, storage, and retrieval by means of bibliographic headers. However, the packaging concept poses some important problems that need to be thoroughly debated and resolved.

The intent of a package of documentary material, quite simply, is to bring together, for purposes of convenient storage and expeditious retrieval, all of the items that pertain to a clearly identified investigative activity. The objective is to keep closely-related items together as an integral, mutually-dependent collection, on a consistent basis across packages, so that these items will not be inconveniently scattered or improperly disassociated from one another.

Because the purpose of a package is to assemble items on a given activity, the items themselves may include all kinds of documentary material, text-searchable and not. A "model" package, one that is ideally constructed if all parts are included, would feature a finished product (usually a technical report or a map), authorizations and reviews of that product, and all of the "backup" items (sometimes called "raw data") that were used as its basis - the numerous notes, logs, graphs, computer tapes, etc. that were generated during the investigation. All or most of the backup items are apt to be non-text-searchable, while a technical report and its commentary will invariably be suitable for text search.

Given the fact that the Department of Energy's (DOE's) Yucca Mountain Project Office (YMPO) has been packaging large portions of its documentary material during the past two years, the packaging concept has become a reality. From the Center's perspective, this practice

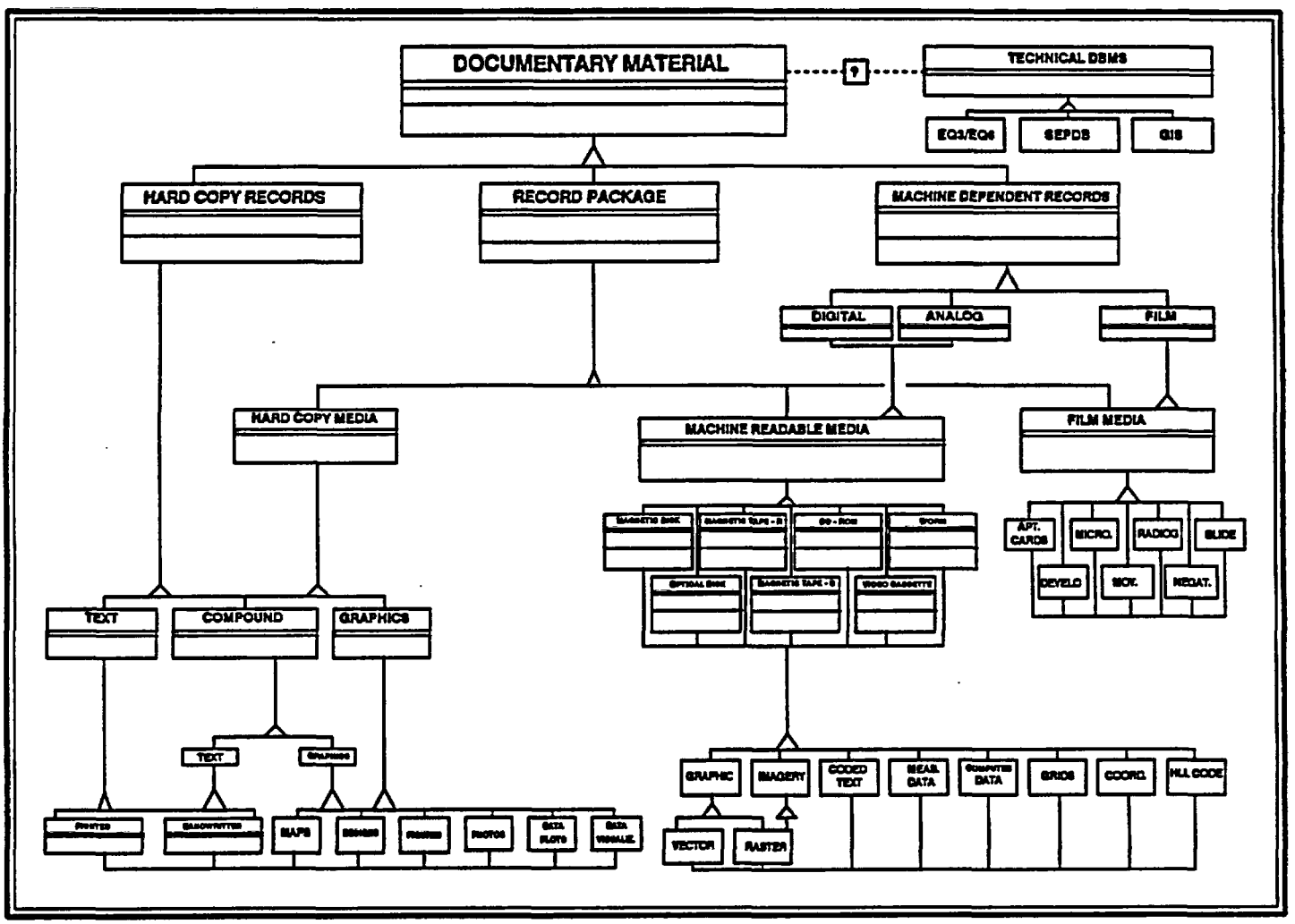


Figure 1-1. Categories of Documentary Material

of packaging has broadened its task to include the development of proposed access protocols for some text-searchable material - the portions of it that may be found within packages.

Exhibit I displays tables of contents for various kinds of packages that have been constructed by the YMPO and sample pages from those packages. Note that it is difficult to describe a "typical" package and that the "model" packages (which were conceived by the U.S. Geological Survey) often incorporate several hundred pages of non-text-searchable "raw data" (graphic/handwritten hardcopy) that comprise their bulk. Figure 1-2 shows how one of the packages would appear as a stack of paper.

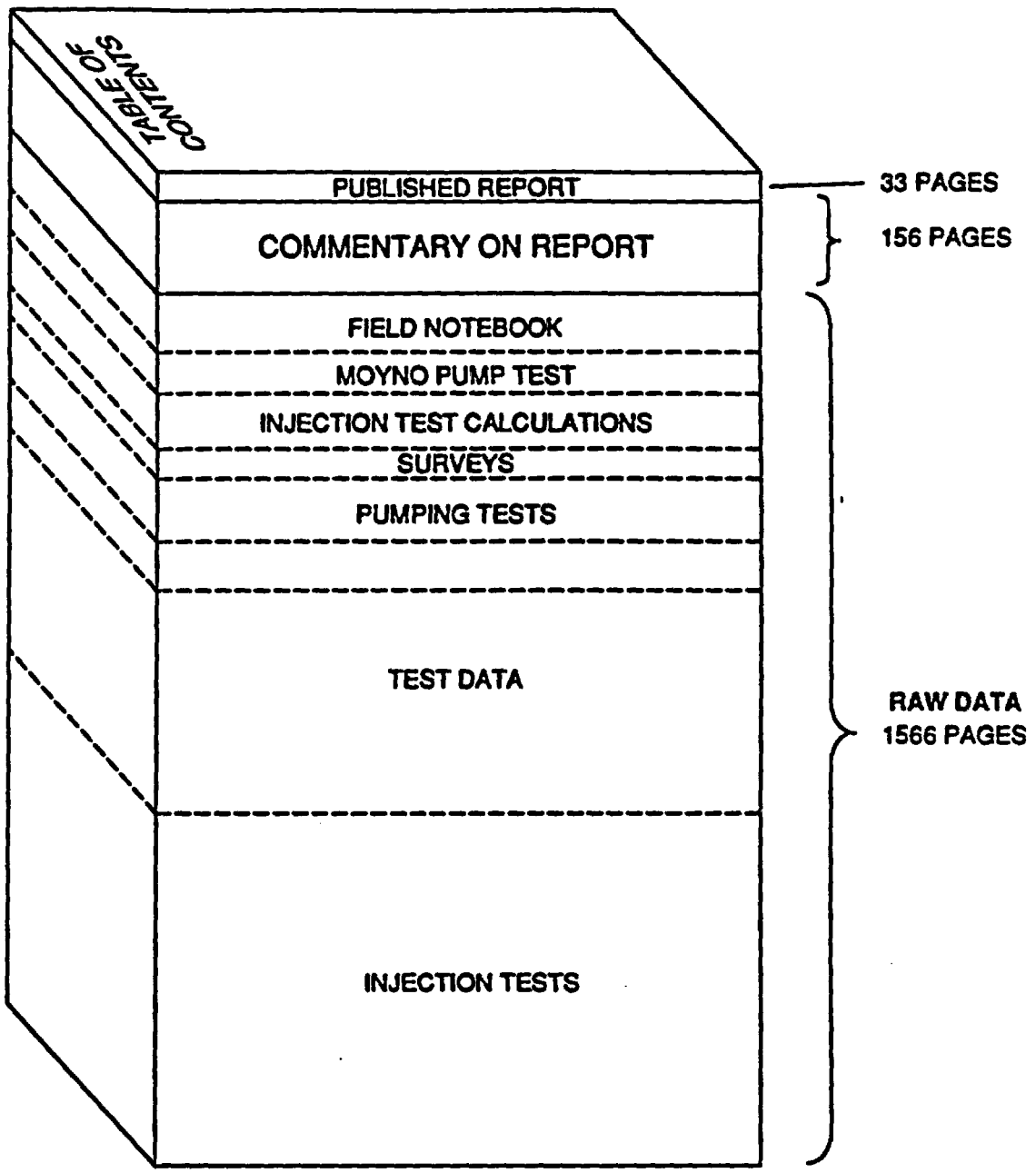


Figure 1-2. Package Seen as a Stack of Paper

2. THE ISSUE OF UNITIZATION

When closely-associated materials are combined within a package, is it sufficient to alert requestors to their collective existence by means of a single bibliographic header, making reference to the package as a whole, by subject and other attributes? Or is it essential to apply multiple, supplementary headers to the package, making separate bibliographic reference to each identifiable unit within the collection?

That is the primary issue which this paper is intended to address. To put it another way, to what degree should a package be "unitized" (subdivided for the application of bibliographic headers) in order to meet LSS requirements?

The issue calls for a fundamental managerial decision on a complex matter that requires attention to detail. It is, moreover, a records management issue that transcends all debate concerning the form that the system finally adopted to carry out the LSS function may take.

The LSS design concept has adopted a records management scheme built upon the principle of unitization, which dictates that all documentary material in the LSS will be subdivided and stored as units. The objective of unitization is to maximize the likelihood that a requestor will be able to retrieve all required information in the least number of units, while retrieving as little irrelevant information in each of them as possible. See Appendix C for further discussion of unitization.

In order to define alternative approaches that might be applied to packages within the LSS, it is first of all necessary to recall what the "LSS Rule" had to say about the matter and to examine existing packaging practices.

3. RELATED STIPULATIONS OF THE LSS RULE

The "LSS Rule" (54 FR 14925, dated April 14, 1989) states, in Sec. 2.1003(c)(1), that "Each potential party...shall submit...an image and a bibliographic header, in a time frame to be established by the access protocols...for all graphic oriented documentary material."

Continuing, in (c)(2), the Rule says that "Each potential party...in a time frame to be established...shall submit...only a bibliographic header for each item of documentary material that is not suitable for entry into the Licensing Support System in image or searchable full text."

Finally, in (c)(3), it says that "Whenever documentary material described in paragraphs (c)(l) or (c)(2) of this section has been collected or used in conjunction with other such information to analyze, critique, support or justify any particular technical or scientific conclusion, or relates to other documentary material as part of the same scope of technical work or investigation, then an appropriate bibliographic header shall be submitted for a table of contents describing that package of information, and documentary material contained within that package shall be named and identified."

Supplementary Information to Sec. 2.1003 says that "The access protocols should ensure that any collection or 'package' of documentary material...which relates to a study should be submitted reasonably contemporaneous with the completion of such a 'package,' including any quality assurance that may be required."

Sections (c)(l) and (2), specify clearly how individual items of non-text-searchable material are to be submitted and speak respectively to what this paper is calling "graphic/handwritten" and "machine-dependent" material. The wording of (c)(3) provides for the use of packages (characterizing them very well). It mandates that a bibliographic header will be applied to each package (specifically to its table of contents) and says that collected materials must be described.

It stops short, however, of specifying precisely how the materials must be "named and identified." In the table of contents alone? In the bibliographic header for the package as a whole? In separate headers for certain items or groups of items? In separate headers for each individual item - as the Rule would dictate if it existed as an independent entity having no connection with a package?

On the question of package unitization, then, the Rule may be interpreted in different ways. It was not, and probably should not have been, written in sufficient detail to specify a precise solution to this issue. The intent of this portion of the Rule, however, is clear: to provide LSS users the most convenient, reliable, expeditious, and consistent access possible to non-text-searchable material. Details of interpretation and amplification were left to subsequent resolution, which would be based upon further analysis of foreseen LSS holdings and the ways in which they might actually be packaged.

4. EXISTING PACKAGING PRACTICES

To date, the Department of Energy's Yucca Mountain Project Office (YMPO) is the only LSS participant that has adopted formal packaging procedures. It is, moreover, the producer of the vast bulk of documentary material. Consequently, it will be the organization that will be most affected by any procedural changes which may have to be imposed to satisfy LSS requirements in the area of packaging.

The YMPO will be particularly affected by any changes having to do with package unitization and the construction of LSS bibliographic headers. It intends to create LSS headers (once LSS capture operations begin) by converting, in as straightforward a fashion as possible, the headers it is currently creating to catalog its packages and other material within an internal computer-based index called the Records Information System (RIS). It will also operate an LSS capture station to scan its own materials and, when possible, interpret their text for the LSS.

The YMPO employs what it calls "data-record packages" to store most of its "technical data" - a term it uses to include acquired, interpreted and supplementary data related to its technical investigative activities. The YMPO also uses packages to collect materials pertaining to its audits, procurements, training, reviews, and drawings. However, with the exception of the drawing-record packages (currently few in number), only the data-record packages contain significant amounts of non-text-searchable documentary material and are therefore of prime concern here.

A YMPO data-record package amounts to a case file on a particular investigative activity - usually undertaken by a contracting organization. It is compiled systematically at a "Participant Data Archive" as materials arrive there intermittently from the responsible investigator. All relevant material to a particular activity is included in the package so that it will all be available in one place.

When a package is complete, an appropriate title and table of contents are created for it. Packages are then forwarded through Local Records Centers to the YMPO Central Records Facility for microfilming as records and indexing into the RIS. The RIS does not provide for full text search of documents or for images of them to be displayed. The YMPO records system, as it currently exists, is microform based, with a computer index to that film.

In the past, the YMPO has often assigned multiple accession numbers within each package, which resulted in the creation of a bibliographic header for every item thus designated. It did this so that selected materials could be retrieved individually as well as through retrieval of the package as a whole. During the past year, however, the trend has been to streamline the process by assigning fewer accession numbers within a package, thereby creating fewer headers leading to the material that a package contains.

Individual headers have always been applied to technical reports and other finished products within a package, in accordance with established YMPO procedures that remain in effect.

Project authorizations and reviews that are contained within a package (which are customarily the subject of correspondence), may or may not be assigned separate headers under evolving guidance. They have received headers in the past.

Another kind of material often contained within packages that has received individual headers in the past and will presumably continue to receive them, are what the YMPO calls "special process" (tapes, films), "one-of-a-kind" (non-duplicable or color-coded), and "oversize" (extra-large) items. None of these items are filmed for the microfilm record, although oversize items are filmed on aperture cards for best resolution. The items are removed from their packages and placed in secure physical storage. "Slip sheets" are inserted in their place within the package, acting as surrogates of the microfilm.

In summary, a YMPO data-record package is currently being assigned a principal header leading to the package as a whole, as represented by its table of contents. In addition, each non-microfilmable item associated with the package has been provided a separate RIS header. If the package contains a finished product (such as a technical report), a separate header is prepared for it also. Other headers have in the past been created for the individual items of commentary directly associated with the product.

The materials within a package that have seldom been assigned individual RIS headers by the YMPO (and will consequently lack convertible headers for the LSS) are the graphic/handwritten items, which often amount to several hundred pages within a given package. They are simply described by line-item on the table of contents.

The YMPO is continuing to refine its relatively new procedures for the management of technical data, which, given the tremendous complexity of its working environment, employing geographically dispersed personnel and multiple contractors, cannot help but benefit from continual efforts for improvement and greater consistency. In the process of refinement, the YMPO is planning a major revision of the RIS (software that is shared within its parent, DOE). In the fall of 1991 the YMPO plans to begin publishing an expanded Data Catalog, a quarterly published listing of acquired and developed data, describing its status and location. It will also upgrade its Automated Technical Data Tracking System (ATDT), which will henceforth produce the Data Catalog automatically and receive more timely update by means of a decentralized input process to be performed locally.

The YMPO is meanwhile struggling to incorporate backlogged material (estimated to amount to some four million pages) within its records system as expeditiously as it can, given tight resource constraints. The packaging that is occurring at the present time, involving dated, backlogged material, may differ somewhat from the packaging that will occur with respect to current investigations when they are renewed in earnest and when LSS document capture operations begin a few years from now.

See Figure 4-1 for a flowchart of the YMPO's technical data flow. For the sake of completeness, it includes two related facets of the YMPO operation: the Technical Data Base

(computer systems that provide for the manipulation and modeling of numeric data) and the Reference Information Base (a printed publication that summarizes fully interpreted data.). Future planning, including intended relationships to the LSS, is depicted in the figure by dotted lines.

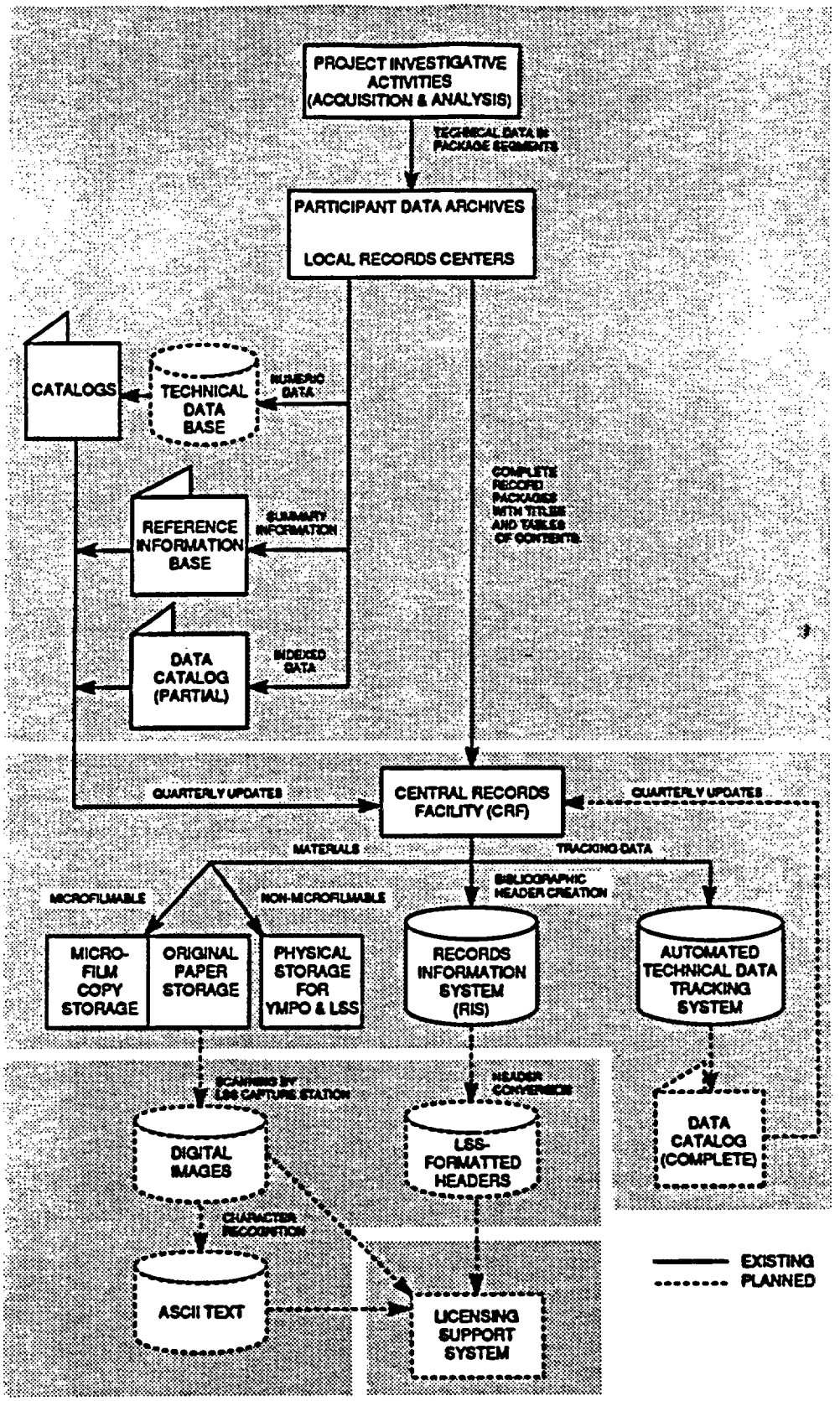


Figure 4-1. YMPO Technical Data Flow

5. ALTERNATIVE APPROACHES TO PACKAGE UNITIZATION

Given the fact that the LSS Rule is not specific regarding the ways in which packaged documentary material should be described for the LSS, and in light of existing packaging experience, what are the alternative approaches to package unitization for the LSS? Which among them would be most reasonable, in terms of practicality and cost?

What are the alternatives? A theoretical list of possibilities would be lengthy, ranging from indexing ("headering") the full package only, at one end of the spectrum, to indexing, additionally, every single item within it - every last lithologic log, weatherization data dump, field notebook entry, and so on.

In between these extremes are alternatives that would include the separate indexing of:

- Any technical reports (or other featured products) that packages may contain;
- Any commentary upon those reports that may be included;
- Any machine-dependent items; and/or
- Any graphic/handwritten items, which often constitute the bulk of a package - suggesting less-inclusive sub-options of:
    - Indexing these materials by groups of line-items, as listed on the package table of contents;
    - Indexing them by every line-item shown on the table of contents; or
    - Indexing them by a set of defined attributes that would, in effect, leave day-to-day decisions on this matter open to an indexer's best judgment.

It can readily be appreciated that it would be more confusing than helpful to debate the pros and cons of the full list of possible alternatives that would result from proposals to index additional items singly and in various combinations with one another.

Several of these permuted alternatives would be inherently illogical. For example, how could it be considered advantageous to index the report commentary and all of the graphic/handwritten material individually, but to refrain from indexing the reports themselves or the machine-dependent material?

That unsupportable example suggests that the best way to evaluate alternative unitization approaches would be to examine the respective merits of separately indexing the materials that

are normally included within packages: the reports, the commentary, the graphic/handwritten material, and the machine-dependent material.

6. BASIC ASSERTIONS

It would be customary at this point, before or during the examination of alternative courses of action, to list basic "assumptions" applicable to the options. In the context of the LSS, however, it is probably more appropriate to make what might be called some basic "assertions" that are essential to the discussion but are admittedly debatable. These are listed below.

- The packaging of LSS documentary material was adopted for sound reasons of efficiency and convenience. If packaging can be accomplished in accordance with the LSS Rule, and if inherent problems can be resolved, it would appear to be a worthwhile practice, advantageous to both document submitters and requestors, for the simple reason that all materials relating to a given activity will be conveniently kept and found in one place. (A computer system might actually store package parts separately for technical reasons but can collect them electronically upon command.)

- There is no sure, indisputable basis on which to predict that any particular portion of existing or anticipated documentary material will have little future value. Therefore, it would not be acceptable to relegate any portions of it to second-class treatment, either within packages or apart from them, on the basis of presumed lack of importance. A pertinent example of second-class treatment would be to exclude certain kinds of graphic/handwritten material from the image data base.

- All scannable documentary material must, without exception, be converted to LSS image form - available for convenient printing by requestors for the purposes of legal discovery or for further technical review. The voluminous graphic/handwritten material, therefore, cannot simply be stored somewhere, on paper or microform, for reproduction upon request. While this represents an accepted LSS design decision, based upon requirements of the LSS Rule, it is still contested by some, given the fact that scanning, storing and transmitting all of the scannable "raw data," which may amount to fifty percent of the total images in the LSS, will be expensive. The point here is that if any of the scannable material is scanned for images, it must all be scanned, to be consistent and to avoid any possible perceptions of certain data being made more readily available than other data. See Appendix D for further discussion.

- Every organization that will be submitting packages of information to the LSS can be expected to abide by the definition of a package that was established by the LSS Rule (cited above), which states that it is: "documentary material...collected or used in conjunction with other such information to analyze, critique, support or justify any particular technical or scientific conclusion, or relates to other documentary material as part of the same scope

of technical work or investigation." Thus, it may be expected that no arbitrary, expedient aggregation of material into inappropriate packages will occur.

- Since many different organizations will originate LSS bibliographic headers, and since all possible varieties of non-text-searchable material cannot be anticipated, indexing rules governing packages will achieve optimum results if they are kept as straightforward as possible, so that unusual occurrences can be accommodated. Complex or indefinite rules that would call upon diverse indexers to make value judgments regarding the unitization of packaged materials, exercising individual discretion, would be likely to result in unacceptable inconsistencies. Simple, uniform rules must therefore be set for personnel who cannot be expected to have the requisite expertise to appraise the substantive qualities of diverse packaged items.

- Requestors of documentary material through the LSS are likely to conduct their searches primarily on the basis of subject matter (e.g., "volcanism"), whether they are focusing on bibliographic headers or on full document text. It follows that packages will usually be retrieved through searches of the titles, keywords and abstracts that are contained in their headers and/or through the matching of subject terms that appear within the text of their tables of contents.

- When packages contain technical reports and commentary upon them, those materials will be made available to LSS requestors in text-searchable, as well as image, form.

- Given the critical importance to packages of their tables of contents - implicitly recognized by the LSS Rule, when it states that one will be written for every package:

    - Tables of contents must always be typewritten, so that their listings will become text-searchable within the LSS.

    - Tables of contents, as a group, must become text-searchable apart from the other, abundant document text that will be stored in the LSS, permitting searches to be directed exclusively toward packages when desired. This separation will not significantly affect the LSS design work that has been accomplished, which foresees that the LSS textual data-base will be partitioned, for the sake of efficiency and convenience, and that one of those partitions will provide for packaged material.

    - The basic format of tables of contents must be fundamentally consistent across organizational lines, thereby promoting their successful comprehension and use within the LSS.

- Tables of contents would benefit from a "hypertext" capability, which will permit a requestor to make selections from its line-items as if they were on an automated menu - "turning" quickly upon command to the initial page (image) of a desired line-item.

7. RESPECTIVE MERITS OF INDEXING CERTAIN PACKAGED MATERIAL

With these assertions in mind, the respective merits of indexing the four basic kinds of items that will be included in packages (reports, commentary, graphic/handwritten items, and machine-dependent items) can now be addressed.

In considering their differing characteristics, it is well to remember what they have in common. They not only represent the basic components of packages, but, with the exception of reports, one will find them almost exclusively within packages when querying the LSS. If they were to be assigned individual bibliographic headers, all of them would receive nearly identical header titles, because their foremost attribute is their intimate connection with the common investigative activity.

It is also important to bear in mind that a requestor will have two powerful tools that will facilitate the retrieval of packaged material: 1) a bibliographic header describing the package as a whole, and 2) a text-searchable table of contents which can be used as a selective menu. Given the power of these tools, the need for supplementary headers must be thoroughly justified.

The respective pros and cons of providing separate bibliographic headers to the four basic kinds of materials that are commonly contained within packages are addressed in Sections 7.1 through 7.4.

7.1. TECHNICAL REPORTS (OR OTHER FINISHED PRODUCTS FEATURED WITHIN A PACKAGE)

The essential characteristics of these products are that they will ordinarily be scannable and (for the most part) text-searchable, unless they are maps or other finished graphic products such as engineering designs. They will invariably be noted as specific terms in the package table of contents. While these products are usually associated with a package, they will not necessarily be contained in all packages. Ordinarily, there should not be more than one primary product to a package.

Pro (provide separate header)	Con (do not provide separate header)
These products are text-searchable and therefore must receive separate headers according to the LSS Rule, pure and simple.	As integral parts of packages, which consist mainly of non-text-searchable material, they do not require headers on absolute grounds. The LSS Rule is ambiguous on the point.

Pro (provide separate header)	Con (do not provide separate header)
A finished product cannot be adequately described on a package table of contents.	There is no need, because it will always be included in the package holding all of its associated correspondence and backup material. It is, and should remain, an integral part of its package.
Reports and other products need searchable specifics of their own, including authors, organizations, dates, abstracts, and descriptors, which require headers to provide this essential detail to a requestor.	Such detail will be provided by the package header. Also, full text search is available.
A requestor may want to retrieve all of the finished products on a given topic but not all of the cumbersome packages on that topic.	In that case, the requestor needs only to look at the products themselves, which will appear at the front of their packages.
Indexing the product in addition to the package will create little added cost, if any. Finished products will normally receive their own headers outside of packages and are likely to be attached to other indexed LSS material. Making them units initially will eliminate any need to index them again when they re-occur and are submitted to duplicate-check procedures at the LSS capture station.	Indexing two units, rather than one, costs twice as much in terms of work hours. (At the YMPO, it takes about three minutes for an indexer, typically paid $7.50 an hour, to process most records into the RIS - about forty cents each. Indexing reports takes the YMPO more time because lengthy abstracts and citations are included in the RIS header.)

7.2. COMMENTARY ON REPORTS OR OTHER FEATURED PRODUCTS (INCLUDING REVIEWS, AUTHORIZATIONS, AND RELATED CORRESPONDENCE)

The essential characteristics of these materials include the ability to be text-searched as well as scanned for images. When they exist, they are mentioned individually in the package table of contents. Obviously, they will only be included in a package when a product susceptible to authorization and review is itself there. At the YMPO, typically, there have been three reviews per packaged technical report.

Pro (provide separate header)	Con (do not provide separate header)
As text searchable items, they, too, should receive separate headers according to the LSS Rule.	Not necessarily. Again, the LSS is not explicit regarding packaged items.

Pro (provide separate header)	Con (do not provide separate header)
If the products to which they relate are separately indexed, they should be also, for the sake of consistency.	Consistency in such matters is not always a virtue. A definite need must be established.
The ability to find these commentaries within the LSS on the basis of the reviewers' names will be essential, particularly to intervenors, and that is best accomplished by the use of individual headers which will contain those reviewers' names.	The intervenor will find reviews adequately described for retrieval purposes on a package table of contents, which will provide reviewers' names and place them firmly within the context of the product they are reviewing, by subject. The table-of-contents data base may be conveniently text-searched for these names - combined, if one wishes, with product-author names.
A requestor might want to search commentaries on some other basis, which could be assisted by a header.	Full-text search of the commentaries themselves is available for that. Headers for them would add nothing worthwhile for a requestor, who would normally retrieve them simply as "reviews" of an investigation of interest.

7.3. GRAPHIC/HANDWRITTEN MATERIAL

The essential characteristics of these materials, sometimes grouped together in the "raw data" area of a package, are that they are scannable for images but are seldom text-searchable. When they exist as backup materials associated with a specific investigation, they will always be included in a package, together with other contributory items. Within their package, they are generally too numerous to be listed individually in the table of contents, often amounting to several hundred pages. Consequently, they are usually listed collectively as line-items (e.g., "pumping tests", "field notebook").

Pro (provide separate header)	Con (do not provide separate header)
If they are not individually indexed, these items will be essentially "buried" within packages.	They may be found by looking at package tables of contents and are generally of interest only as integral parts of their respective packages.

Pro (provide separate header)	Con (do not provide separate header)
It is worth the extra effort to index each of them individually, in case someone wants to search for them that way.	If they truly amount to fifty percent of the items to be incorporated into the LSS, indexing them individually would double the labor-intensive burden of creating LSS headers, a very costly enterprise.
Package tables of contents describe the items in too few words, without any detail.	The descriptions are nevertheless adequate. Further detail in their case is unnecessary.
Headers would provide more detail about each of them on an individual basis.	A header would be bound to feature the subject of the package as a whole in its title if that title is to have any meaning. Aside from their line-item description, what more could usefully be placed in the available fields of separate headers that would assist these items' retrieval?
To reduce cost, if it is impractical to index the items individually, they could be indexed as they appear as line-items on the package table of contents.	That would still be too costly, especially since there is no need to index them at all.
At the very least, line-items should be indexed selectively, when they qualify as stand-alone units. The pertinent questions to ask are whether an item, if it were to be retrieved outside the context of its package, would be independently understood, and whether it might be known individually to someone as an item of interest.	Even selective indexing is unnecessary, because the items will be adequately described in their text-searchable tables of contents. Separate headers could actually complicate a search by yielding misleading, multiple references to the same package, and work against the very purpose of unitization, which is to retrieve desired information in the smallest number of units.
A package contains a lot of other material which may not be of interest to a requestor.	The rest of the package (all on the same topic) is never irrelevant.
It will be hard to locate the items of interest within a package because a package can amount to several hundred pages (LSS images).	Hypertext indexing of the table of contents will enable a requestor to immediately "turn" from that listing to the first page (image) of a desired line-item.

Pro (provide separate header)	Con (do not provide separate header)
The requestor may still face a daunting task, if presented with a line-item containing a hundred pages or more.	It is doubtful that there will be many circumstances in which a requestor will know in advance which page of the line-item (of the field notebook line-item, for example) will be of interest.
Indexing the materials selectively by line-item would provide wider search options to locate various sets of images. A requestor might, for some reason, want to find these items in a query across packages.	While it is improbable that a requestor would want to find such things as field notebooks, pumping tests, lithologic logs, and so on, in general, across packages - unrelated to the specific investigative activities to which these functions are attached, they will nevertheless be findable through text searches of the package tables of contents, without any need for separate headers.
The items should be indexed for the sake of consistency, since graphic/handwritten material will not always be packaged and will have to be individually indexed when it is not, as required by the LSS Rule.	Consistency in this instance would be better served by relieving indexers of the need to apply value judgments to material that they will lack proper expertise to evaluate. If someone really wanted to pursue instances of "pumping tests" (in general, wherever they might have occurred outside of packages), the requestor could simply search for additional occurrences within other headers and text.

7.4. MACHINE-DEPENDENT ITEMS

The essential characteristics of these materials are that they will be neither text-searchable nor scannable for images. They will have to be retrieved through other means, upon request. They are more likely to be included within packages than not. When packaged, they will always be mentioned as individual line-items in the table of contents. There may be several of them within a package, or none at all.

Pro (provide separate header)	Con (do not provide separate header)
These items require separate bibliographic headers because they cannot be adequately described on a table of contents. They require descriptions of the types of media on which they are recorded and mention of the specific kinds of machines on which they are dependent. For the most part, they are magnetic media, which require rather detailed descriptions.	Granting that need, there is a better alternative than the use of headers. The placement within packages of scannable, text-searchable, surrogate information sheets would satisfactorily inform requestors of these items' specifications.
An information sheet would be an inconvenient way to provide adequate detail. Moreover, these materials need headers to be successfully found, because, as physical items (incapable of being scanned, filmed, or stacked with hardcopy), it is anticipated that they will be stored apart from the packages with which they are associated. Some field in the LSS bibliographic header, such as an added "storage location" field, must provide convenient notation of their actual location to assist timely accessibility.	The same surrogate information sheets within packages could inform requestors sufficiently of these items' storage locations, as well as specifications.
The items are not nearly so numerous within packages as graphic/handwritten items and would be worth the extra expense required to index them individually.	The materials are nevertheless fairly prevalent within packages. Indexing them individually would multiply total indexing expense by a significant (but unknown) factor, perhaps three or four.

8. RECOMMENDATIONS

The authors of this report believe that the most important decision criterion within the pros and cons listed above is whether or not the respective materials will be adequately described within package tables of contents and will therefore have no real need for bibliographic headers of their own.

On that basis, the authors are persuaded that finished products, such as technical reports, and machine-dependent items, such as computer tapes, will not be adequately described on tables of contents. For purposes of retrieval, they will require separate bibliographic headers, independent of the header that will be applied to each package as a whole.

There are other convincing reasons for assigning headers to finished products. As published items receiving wide distribution, they will repeatedly come to the attention of indexers outside of their associated packages, as items worthy of independent headers. This is likely to cause confusion and duplicative effort if they are not separately indexed in the first place. They should, moreover, be conveniently retrievable apart from their bulky packages. The additional costs to index them would be minimal.

There are additional convincing reasons also to index machine-dependent items. Formatted headers will provide a real convenience to requestors wishing to retrieve these items according to their specific machine dependencies and in consideration of their storage locations. Information sheets appearing as text-searchable images within packages would not provide that same convenience. The incremental cost of separate indexing would not be high.

Commentary and graphic/handwritten material, on the other hand, will be adequately described for retrieval purposes by package tables of contents and will therefore not require separate headers. Other arguments to index them are unconvincing.

A summary of the respective characteristics of these four kinds of packaged documentary material is presented in Table 8-1, which includes judgments concerning the adequacy of table-of-content description and the above indexing recommendations.

Table 8-1. CHARACTERISTICS OF PACKAGE COMPONENTS

Topic	Published Reports/ Products	Commentary on Products	Graphic/Hand written Material	Machine-Dependent Items
Available for Text Search	Yes	Yes	No	No
Available in Image Form	Yes	Yes	Yes	No
Incidence within Typical Package	1	Few	Numerous	Few
Individually Listed on Table of Contents	Yes	Yes	Grouped	Yes
Adequately Described by Table of Contents	No	Yes	Yes	No
Separate Bibliographic Headers Recommended	Yes	No	No	Yes

When LSS access protocols are finally adopted, the authors' recommendations, if accepted, may or may not equate with existing practices of the YMPO regarding its RIS system. This will depend on the outcome of YMPO reconsideration of its ongoing record-indexing procedures. Some changes in current YMPO procedures, however, are recommended so that LSS requirements for universal ease of accessing packaged materials will be satisfied.

- Package tables of contents should acquire a consistent text-searchable format. The same basic format should be used by all YMPO contractors, as well as by the (less prolific) non-DOE providers of documentary material who may submit packaged material to the LSS.

- Line-item entries on a table of contents should be adequately descriptive, in all cases, so that the listing may function successfully as an important retrieval mechanism.

- The word "package" should be deliberately included within the titles of bibliographic headers that are created for packages so that an LSS requestor will immediately be alerted to the fact that a package of material has been found, not simply a report having a nearly equivalent title. This can be done automatically during the document capture process.

The following recommendations were implicit in the section that enumerated "Basic Assertions".

- All scannable documentary material within packages must, without exception, be converted to LSS image form.

- All organizations must abide by the definition of packages provided by the LSS Rule.

- Technical reports and their commentary within packages must be rendered text-searchable.

- Tables of contents must be rendered text-searchable and be written in accordance with a consistent format. Their text should be searchable within a partitioned database. Furthermore, once a table of contents is retrieved, the capability of selecting line-items from it, for the convenient display of their beginning pages, should be available in the LSS by means of hypertext indexing. Hypertext will obviate the need for numbering the pages of a package (not currently done at the YMPO) in order to place beginning page numbers beside each listed line-item on the table of contents, which would otherwise be required to readily locate line-items within a large assembly of material. See Appendix E.

Resolution of the unitization issue is fundamental to the projected issuance of access protocols for non-text-searchable material by the LSSA. There are some closely related issues, however, which inevitably arise during the course of any discussion of this topic. The authors do not yet have firm recommendations in their regard (or solutions for other issues that have arisen in connection with the development of complete access protocols). However, a few suggestions are offered for consideration at this time, as presented in Appendix F (Other Packaging Issues) and Appendix G (Description of Technical Data Using Bibliographic Headers).

Appendix H is included to provide an appreciation of how the authors' recommendations would fit into the context of the submission and entry of documentary material into the LSS.

APPENDIX A

BIBLIOGRAPHIC AND TEXT SEARCH

A.1. INTRODUCTION

One of the principal objectives of the LSS is to provide electronic discovery for all participants. Accordingly the LSS rule provides for all relevant information to be entered into the system insofar as is technically feasible. When the LSS has been loaded, all participants will have immediate access to all information pertaining to a given subject or topic of interest which is contained in image and/or textual form in the system. This, of course, is a somewhat idealized and simplistic view of the LSS, but it is important to keep in mind that the primary purpose of the system is to provide rapid access to information relevant to the licensing of the repository for a wide range of participants.

A.2. THE CONCEPT OF THE INFORMATION HORIZON

For purposes of discussion, it may be helpful to introduce the concept of the information horizon. In every effort toward information retrieval there is a information horizon beyond which information cannot be effectively found or retrieved. A number of factors affect the characteristics of the information horizon. These factors include access protocols, indexing, retrieval efficiency, relational linkages, degree of specificity of the search parameters, etc. The point is that, just as there is a physical horizon which limits our direct knowledge and experience of the physical world, the information horizon of a system limits our ability to directly find and access information. Data may well be present beyond the information horizon and it may be properly formatted and maintained, but it is not useful because it cannot be effectively found and retrieved by the user.

A.3. EXPANDING THE INFORMATION HORIZON

The LSS will include extensive indexing and retrieval capabilities which will permit the user to find the desired information quickly and effectively. The indexing is intended to have two major thrusts:

- Bibliographic search
- Full text search

A.4. BIBLIOGRAPHIC SEARCH

Bibliographic headers will provide access to information by giving the LSS user the ability to scan fixed header fields and then select the desired documents for inspection and review. The process is very similar to searching a card catalog in a library which has been organized by title, author, subject, etc. But, of course, the electronic searching of bibliographic headers will be much more capable and much faster than searching through a card catalog.

Once the search of bibliographic headers has been performed, the user has really only retrieved a list of document accession numbers which, in turn, permit the retrieval of the text

or the image of the documents. The process of retrieving accession numbers in order to access the documents will not normally be apparent to the user. But it is important to understand that each document must have and will have a unique identifier which the system will use to access the information in that document, whether textual or image in format.

The limitation of bibliographic headers is that the content, accuracy, completeness and quality of the header determine the ultimate limits on the search capability. If the header is complete and if it accurately describes the document, then there is a high probability of the user being able to identify and retrieve the document. If, however, the header is incomplete, inaccurate or not adequately descriptive of the document, then the desired information cannot be easily identified and retrieved by the user. Thus, the quality and accuracy of bibliographic headers can have the effect of dramatically shifting the user's information horizon.

One of the most important criteria in evaluating a proposed system of bibliographic headers is the effect of the proposed headers on the user's information horizon. The absence of an adequate header or equivalent information which can be searched in order to identify the desired document has the effect of placing that document beyond the user's information horizon and makes it inaccessible for all practical purposes. A non-descriptive or incomplete bibliographic header similarly renders the document beyond the information horizon. It is rather like a tall ship beyond the visual horizon: you can see the tip of the mast and you know that something is out there, but you cannot really tell what it is.

In the case of a system like the LSS which will contain millions of pages, if the bibliographic headers are the primary search mechanism and are not reliable or sufficiently descriptive, then the user may be forced into a sequential inspection of individual documents. Such a sequential search is a practical impossibility which effectively places the desired documents beyond the information horizon and makes them "unavailable." Providing alternate search mechanisms such as full text searching can reduce the level of dependency upon headers, but the quality and completeness of the headers remains a significant concern, particularly for those documents which are not available for full text searching.

For this reason, the question of the quality of the bibliographic header is fully as important as the question of which documents or which components of a package will have a header. The issue of header quality is a major concern with any backlog of documents and with the procedures of packaging as well.

Because of his or her intimate familiarity with the information content of the document, the person best qualified to prepare the header is the person who originates the document. The same argument applies but even more stringently to the question of abstracting. The preparation of a meaningful abstract requires that the person preparing the abstract be familiar with the material in the document and also be able to meaningfully interpret its significance. Similarly, if a table of contents for a record package is to be truly informative, its preparation must be more than a purely mechanical or clerical process. Therefore, the further that the preparation of the bibliographic header (with abstract) or table of contents is removed from the creation of

the document in terms of space, time or personnel, the less likely it is that these important search tools will be complete and of high quality.

A.5. FULL TEXT SEARCH

Full-text searching will permit the LSS user to search for individual words, combinations of words, phrases, etc. which are embedded in the ASCII text of the documents. The effect of full-text searching is to dramatically expand the information horizon, but only for those documents which are entered in full-text (ASCII) format. This is because all documents which are full-text searchable are by definition within the information horizon defined by the content and extent of the text-searchable database.

But the content and extent of the database in terms of the information it actually contains is not the same as the user's effective information horizon, in terms of what information the user can actually find and retrieve. If the implementation of the full-text search is not fairly sophisticated, then the user is in the position of a very nearsighted person without glasses trying to see an object which may be close by or at an intermediate distance or even many miles away - the physical horizon of what could be seen is quite far away, but the effective horizon of what can be seen is very close. Just as the nearsighted person without glasses has difficulty in focusing on objects beyond his effective visual horizon, the sheer volume of "hits" returned by a query can easily obscure the desired information and render it beyond the effective information horizon of a system user.

Therefore, the user must necessarily take a balanced approach, using bibliographic headers for a "primary" search of the information base and using full text search to check for information contained in the documents which may not be fully indicated by bibliographic headers. Eliminating either bibliographic headers or full-text searching for a class of documents could be expected to significantly shift the user's information horizon and thereby compromise his ability to find and retrieve those documents.

As in the case of the bibliographic header, the result of a full-text search is a list of document accession numbers which in turn are used to retrieve the desired documents. Thus, there must be a process of assigning document accession numbers and associating those accession numbers with the retrieval indexes to permit rapid and effective access to the desired document. Such indexes when properly implemented have the ability to position the user's display to the selected document, page, line or even individual word. Therefore, the process of assigning accession numbers and associating them with the retrieval indexes is a matter of great importance for both bibliographic header indexes or contextual search indexes.

APPENDIX B

CLASSIFICATION OF DOCUMENTARY MATERIAL

B.1. CLASSIFICATION OF DOCUMENTARY MATERIAL

Although the Licensing Support System Rule (LSS Rule) does not explicitly address the issues of indexing and retrieval of technical data within the proposed bibliographic documents management framework, technical data comprises a large part of the existing records inventory. Technical data generated and developed by the investigative, analytical and interpretative activities (Figure B-1) of the Yucca Mountain Project (YMP) is distributed through all categories of documentary material considered by the LSS Rule. Technical data is not restricted to any specific type of record, nor is there any *a priori* physical dependence of technical data on record type. The type and occurrence of technical data throughout the inventory of project records is largely the result of conventional or convenient use of media, recording technology, and documentation or publishing standards that were fortuitously in use at the time the data was generated and/or used.

A considerable amount of technical data is likely to become available through computerized database management systems. The Yucca Mountain Project Office (YMPO) is currently using three technical database management systems. The Site and Engineering Properties Data Base (SEPDB), managed by Sandia National Laboratory, is for site-specific parametric data. The GEMBOCHS(EQ3/6) database holds more general geochemical and thermodynamic data and is administered by Lawrence Livermore National Laboratory. The ARC/INFO geographic information system (GIS) is for storage, display and mapping of spatially distributed data and information. EG&G has responsibility for managing the GIS. The LSS must accommodate this data, and decisions must be made regarding access routes and protocols.

Access to project technical data will be an important function of the LSS. The Rule anticipates use of the LSS as a technical review information system, and NRC staff have indicated a desire to use the LSS as an information management system for pre-licensing review and development of regulatory guidance. NRC staff have also indicated that LSS may supplant the NRC NUDOCS system currently in use.

Documentary material can be classified on the basis of the submission requirements specified in 10 CFR Part 2 Subpart J. The LSS Rule identifies a class of records that are printed on paper, and which can be converted to a raster (digital, or bit-mapped) image with a digital scanning device. Furthermore, the text contained in these records can be converted to ASCII code, directly from the raster image using optical character recognition (OCR) techniques. All of these records must also have a bibliographic header. This class of materials has previously been referred to as Category I documentary material by Acree and Young (1991) (Figure B-2).

Records printed primarily on paper (some may exist as blueprints, or on mylar film or vellum) that consist mostly of graphics or handwritten text (Category II documentary material) can be converted to a raster image by digital scanning. This class of records would not include material with a substantial amount of text, but would also require a header.

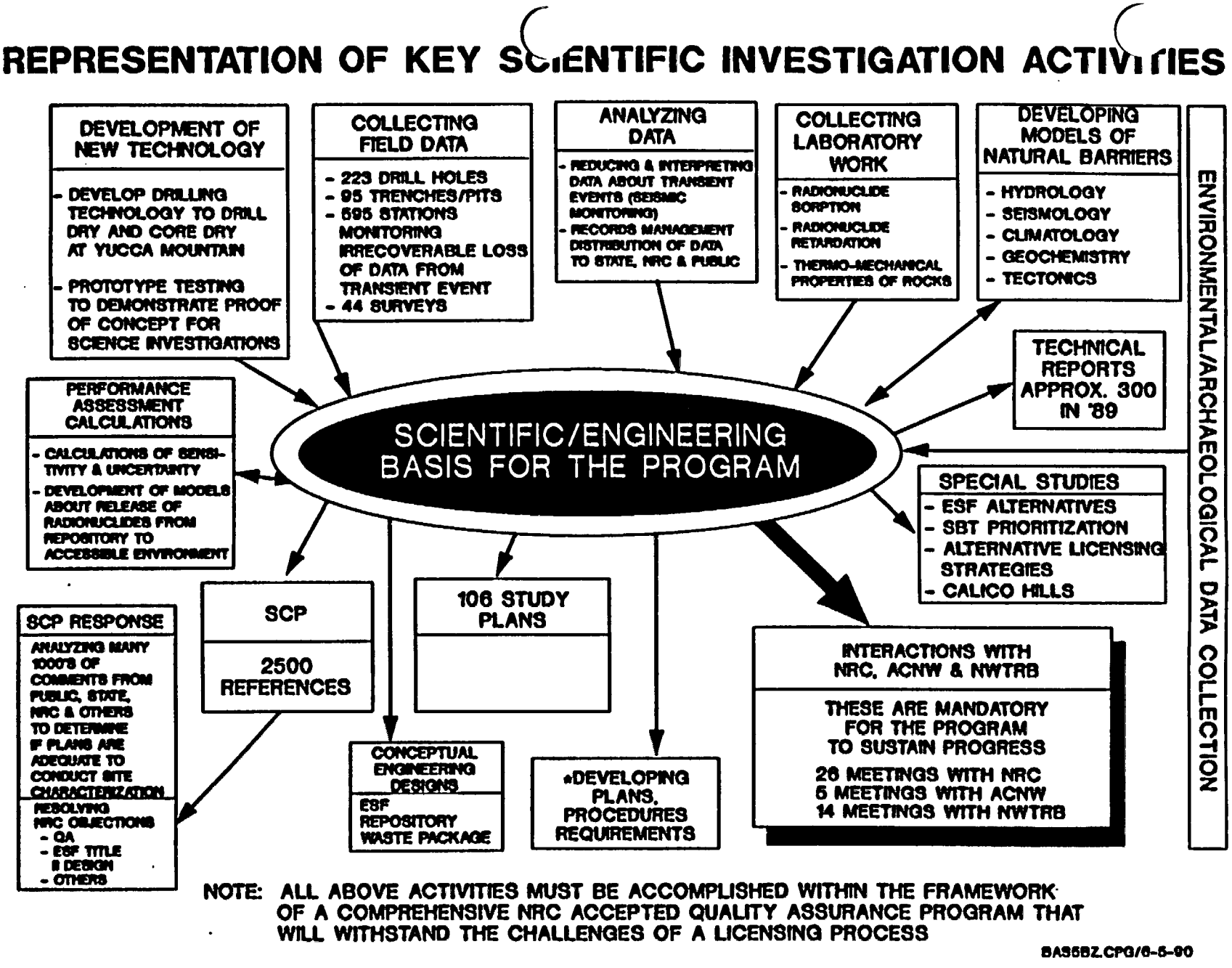


Figure B-1. Scientific Investigative Activities of the DOE High-Level Radioactive Waste Program (after DOE, 1990)

CATEGORIES OF DOCUMENTARY MATERIAL		LSS SUBMISSION REQUIREMENTS		
		ASCII	IMAGE	HEADER
I	DOCUMENTARY MATERIAL SEC 2.1003(a)(1)(2)(3) SEC 2.1003(b)(1)(2)	ASCII required	IMAGE required	HEADER required
II	GRAPHIC-ORIENTED DOCUMENTARY MATERIAL SEC 2.1003(c)(1)	NO	IMAGE required	HEADER required
III	DOCUMENTARY MATERIAL NOT SUITABLE FOR ENTRY SEC 2.1003(c)(2)	NO	NO	HEADER required

Figure B-2. Categorization of Documentary Material Based on Submission Requirements Described in 10 CFR Part 2 Subpart J. Categories are distinguished on the basis of the data form in which the material must be submitted for loading. Categories I - III are from Acree and Young (1991).

The class of records not suitable for direct entry into LSS (Category III documentary materials) comprises data recorded on machine-readable physical storage media, or records created by a photographic process. Instead of the record, a header is required to be submitted that is sufficiently descriptive to allow determination of the type and utility of the data. The Rule allows for the compilation of any combination of the above three categories into a "package of information" to be submitted with a header.

B.1.1. Categories of Documentary Material

The LSS Rule describes categories of documentary material with respect to the data form in which the material is to be submitted. Category I material is submitted as ASCII code and raster (bit-mapped) imagery; Category II as raster imagery only, and Category III is archived externally to the LSS. Each record must be referenced by a descriptive header. The requirement that a header be submitted for each record is the common thread of reference for all of the documentary material. That is, the header is the only indexing requirement that is common to all categories of material.

Currently, under the records management architecture of the Department of Energy/Office of Civilian Radioactive Waste Management (DOE/OCRWM) Yucca Mountain Project Office/Central Records Facility (YMPO/CRF), a considerable amount of documentary material is being compiled into Records Packages. The LSS Rule considers the use of a "package of information" (10 CFR Part 2 Subpart J sec. 2.1003(c)(3)), and states explicitly that "Whenever documentary material *described in paragraphs (c)(1) or (c)(2) of this section* has been collected or used in conjunction with other such information to analyze, critique, support or justify any particular technical or scientific conclusion, or relates to other documentary material as part of the same scope of technical work or investigation, then an appropriate bibliographic header shall be submitted for a table of contents describing that package of information, and documentary material contained within that package shall be named and identified." (Emphasis added by authors.)

B.1.2. Record Package

Each record "described in paragraphs (c)(1) or (c)(2)" of section 2.1003 is required to be submitted as ASCII code plus a raster image of the record, or as an image alone, respectively, and each will have a header. However, (c)(3) may require only that a descriptive header be submitted in reference to the package Table of Contents (TOC). It follows then that an issue may arise as to a requirement that documentary material that is compiled within a package be imaged, text-coded, and have a header. The rule is open to interpretation on this issue.

The most straightforward interpretation of Section 2.1003(c)(3) of the LSS Rule is that each item, or unit, of documentary material within a package is required to be submitted as a raster image, and to have substantial text content ASCII-coded. However, the 'record' is actually the package. As such, only the package needs to have a header. The individual

documents within the package would be listed on a comprehensive Table of Contents, but would not require an independent header.

The need, from the perspective of a potential user, to divide a package into more fundamental units for separate headering, is difficult to assess given the sampling of material currently available through the YMPO/CRF (Exhibits I and III). By far, most of the substantive data records packages currently at the CRF are from the U.S. Geological Survey (USGS). The USGS has produced, and will continue to produce a substantial volume of documentary material for the Yucca Mountain Project, but there are many other technical contractors and subcontractors (Figure B-3). Unitization of packages would benefit from uniformity among all of these contractors. Sufficient uniformity may not exist in backlog of documentary material that has accumulated to date, or that is accumulating currently.

Some commonality seems to exist between packages submitted by the USGS (Exhibit I B through I P) and those being held by the Data Records Management System (DRMS) of Sandia National Laboratory (Exhibit IV), in that both have a labeled 'Raw Data' section.

B.2. CLASSIFICATION OF RECORDS

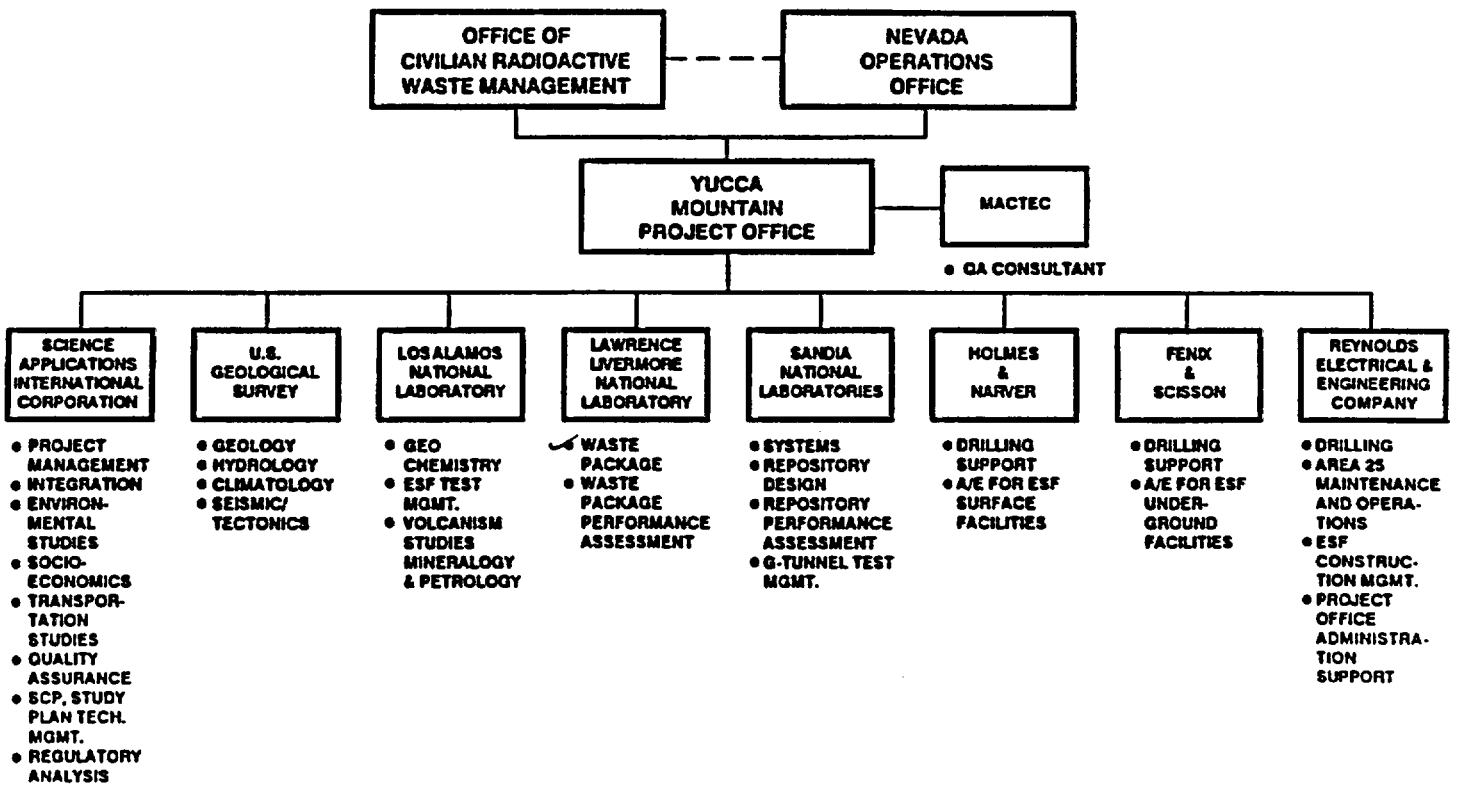
Each of the categories of documentary material considered by the LSS Rule are comprised of records that are produced in various formats and styles, archived on various physical media, and which contain a variety of information or data types. The authors have adapted a data modeling technique to the classification of records that comprise the documentary material produced by the YMP, and that is stored at the CRF. For the purpose of this classification, a record is considered to be the fundamental unit of the inventory of documentary material. The project records are then classified logically in a manner that is consistent with the categories of documentary material described in Subpart J.

For the most part, the CRF and each of the Local Records Centers (LRCs) manage documentary material on the basis of record type and technical subject. With respect to LSS submission requirements, record-type classification, and development of indexing procedures on the basis of type is of primary importance. Record type is less important from a retrieval point of view, where technical subject is likely to dominate the concerns of a potential user. Given these priorities, the data modeling approach taken by CNWRA at this time focuses on the current project record types and relationships between records and technical data content.

B.2.1. Data Modeling Approach

The authors have modified an object-oriented modeling and design methodology (Rumbaugh, et.al., 1991) for application to records classification. The approach is basically a data modeling technique that shows the properties of and inter-relationships between well-defined classes of records (Figure B-4). The choice of class and subclass names and descriptions is based closely on the documentary material currently managed by the YMPO/CRF.

ROLES OF PROJECT PARTICIPANTS



QARMS4.PM3/12 6 88

Figure B-3. List of Yucca Mountain Project Participants Through 1990 Showing the Main Technical Subject Areas of Responsibility (after DOE, 1988). These participants have contributed most of the current backlog of documentary material. However, additional participants (i.e., Raytheon) are joining the project.

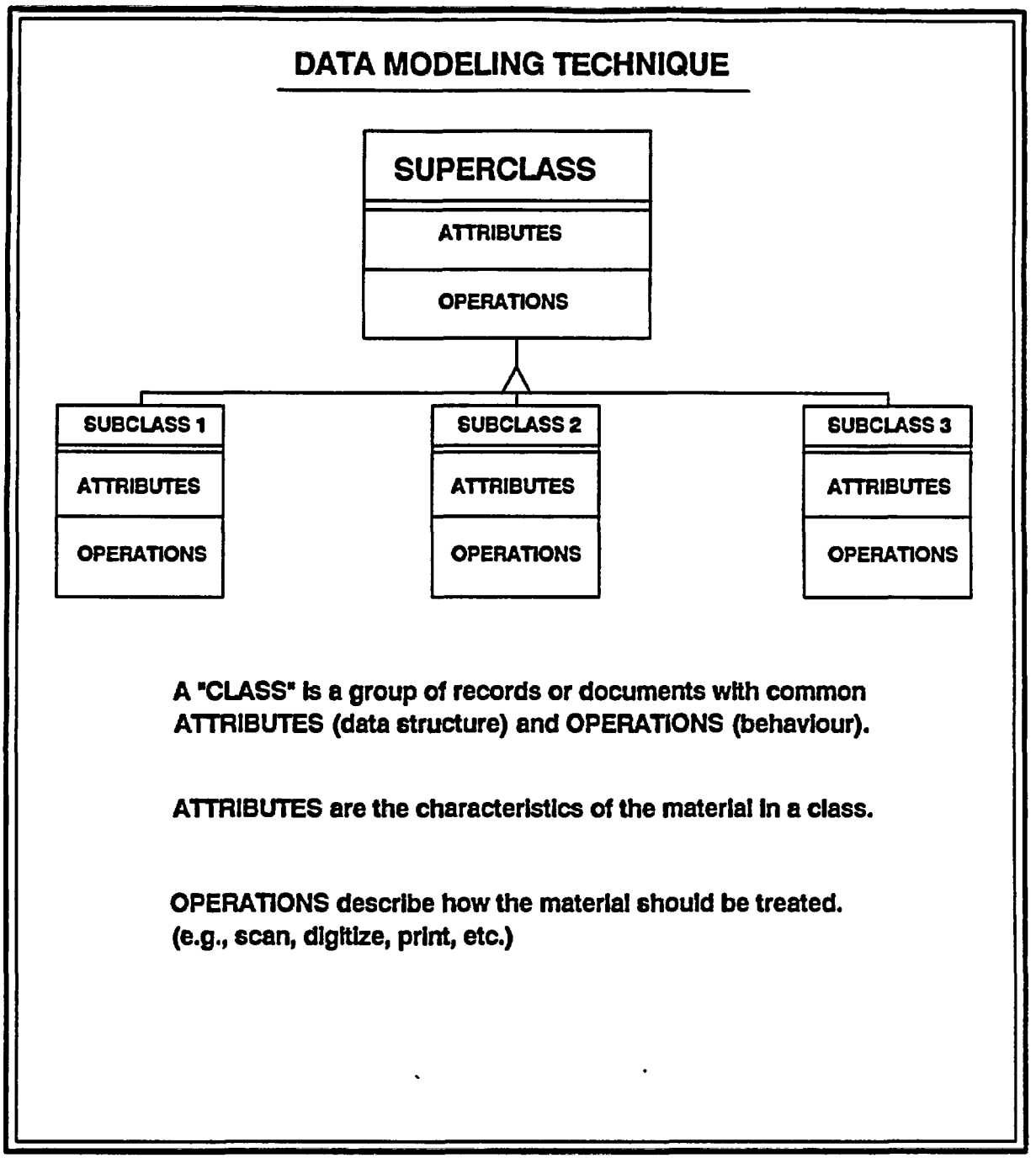


Figure B-4. Generalized Data Model Showing Relationships between Classes and Subclasses. ATTRIBUTES and OPERATIONS describe the properties of the classes and their behavior, respectively. Both attributes and operations are inherited fromthe SUPERCLASS (parent) to the SUBCLASS (child).

A *class*, in this modeling approach, is a group of records or documents with common *attributes* (data structure) and *operations* (behavior). All of the individual records in a class must share the set of defined attributes and operations. Attributes are, essentially, the characteristics of the material in the class. Operations define, or describe, how the material should be treated. More specifically, operations describe what the records in a class can do, or what can be done to them (e.g., scan, digitize, print, etc.). Accordingly, operations may be class dependent. That is, for example, a 'scan operation' may work a little bit differently for one class than it does for another. The 'print operation' is certainly different when applied to a computer file consisting of a high-level printer control language (e.g., HP pcl), than when applied to a photographic negative, although both of these records can be said to 'printable'.

Any class that is subdivided, or subclassified, is considered to be the superclass (or parent) of each of the subclasses (child). Attributes and operations are therefore inherited by the subclasses. Each of the records in a subclass must have the same sets of attributes and operations as the superclass, but have additional properties, not shared by the parent class, that are the basis of the subdivision. (Shlaer and Mellor, 1988; and Cardenas and McLeod, 1990.)

B.2.2. Characterization of Technical Records

Technical records (records with a substantial technical data content) occur in a great variety of forms depending on the type of data in the record, the recording media, and the format of the recording. To be certain, development of effective indexing instructions will require that records be grouped into classes. However, the classification structure of documentary material should be constructed differently, depending on perspective. From a submission, or indexing point of view, media and format are critical. From a technical users point of view, technical subject is of prime importance. These two differing perspectives must merge somewhere within the LSS.

Consideration of documentary material from the perspective of LSS content versus the current physical state of the material is a case in point (Figure B-5). This type of data model (Figure B-5) can be used to show a class of records called LSS DOCUMENTARY MATERIAL that represents a hypothetical case where the LSS has been loaded with some portion of the inventory of project documents. In a very broad sense, LSS DOCUMENTARY MATERIAL can be considered to be of two basic types, INTERNAL and EXTERNAL. INTERNAL material is held within the system in coded text (ASCII) or raster format, for retrieval, display or copy. The EXTERNAL material is the Category III material (Figure B-2) that is referenced by a header, but is not physically stored by the LSS.

The PROJECT DOCUMENTS class at the bottom of the model represents the documentary material as it actually exits at the YMPO/CRF and LRCs. This material can be subclassified on the basis of the type of media on which it is recorded. The line-and-circle connection between PROJECT DOCUMENTS and MEDIA TYPE is a one-to-many relationship. Any individual record in the superclass can be recorded on any one of a variety of recording media (e.g., paper, magnetic tape, film, etc.). This is a good illustration of why the project

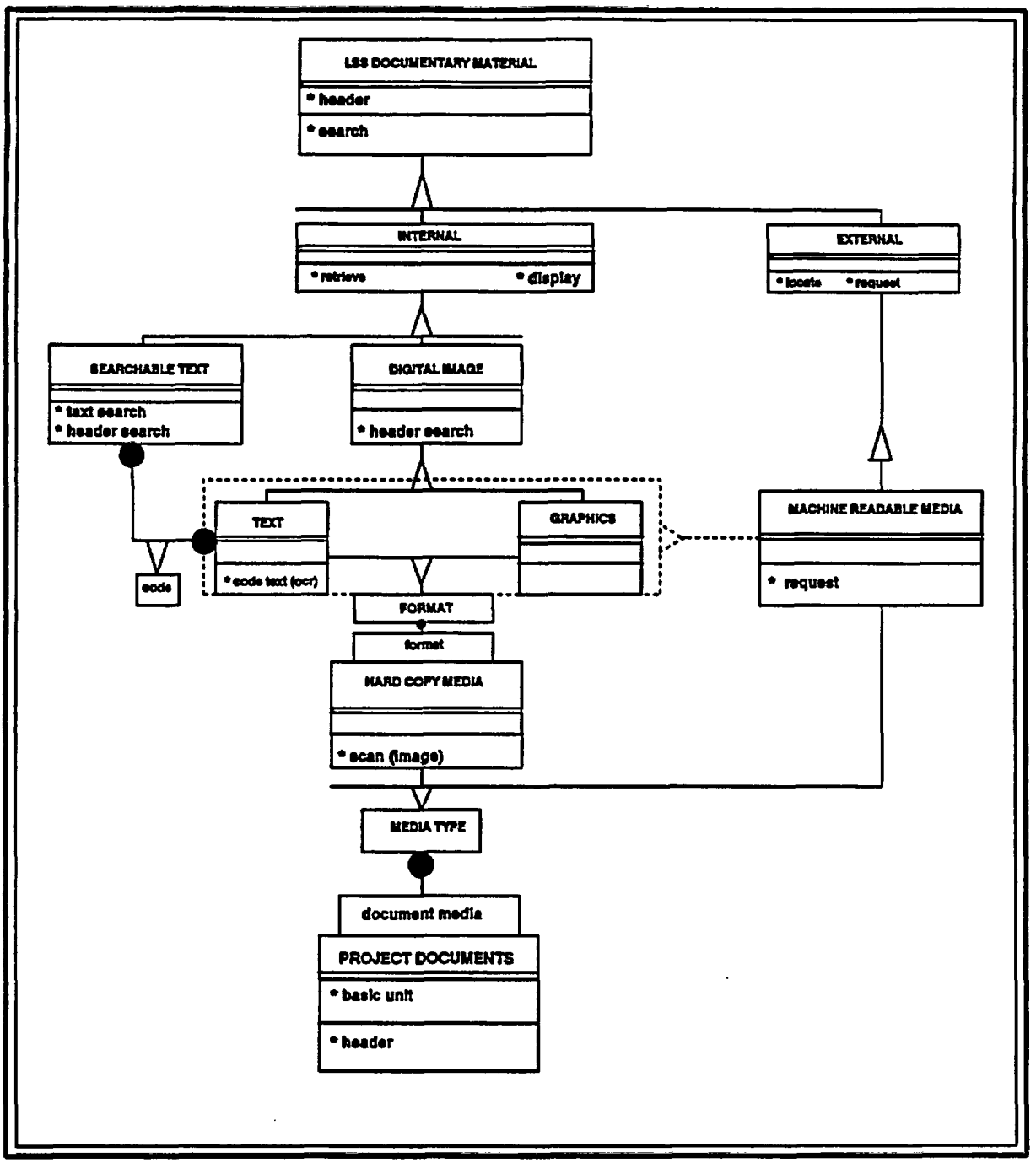


Figure B-5. Data Model Showing Classification of Anticipated LSS Documentary Material versus Project Documents. Subclassification of Project Documents is parallel to LSS Documentary Material at the FORMAT type level.

documentary material cannot be directly subdivided on the basis of media type.

MEDIA TYPE, however, does have a set of direct subclasses. HARD COPY MEDIA and MACHINE READABLE MEDIA both inherit the parent attribute of being physical storage media. HARD COPY MEDIA is subclassified on the basis of the basic format of the recording. It is at this level, for this simple classification structure, that classes of material within the LSS are concordant with classes of the same material outside of the system. The FORMAT classes of TEXT and GRAPHICS, and MACHINE READABLE MEDIA are correlative with the subclasses of SEARCHABLE TEXT, DIGITAL IMAGERY and MACHINE READABLE MEDIA referenced in the LSS. This type of convergence of system-related and real world classes of material needs to be demonstrated to show that the logical structure of the LSS is capable of adequately managing the real world material.

Attributes and operations are assigned to the class by writing them in the appropriate box below the class name. For LSS DOCUMENTARY MATERIAL, the key attribute is the header. All classes of LSS records have a header. This attribute is assumed to be inherited by the child classes of INTERNAL and EXTERNAL, and so it is not necessary to write it down again. In turn, all SEARCHABLE TEXT and DIGITAL IMAGERY will be referenced by a header. The *search* operation applies to all LSS DOCUMENTARY MATERIAL, but the *display* and *retrieve* operations apply only to the INTERNAL material, while *locate* and *request* apply to the class of EXTERNAL material. Note that the operations listed can be distinguished as being 'submission-oriented' under PROJECT DOCUMENTS and 'user- oriented' under LSS DOCUMENTARY MATERIAL. This is a reflection of the perspectives taken to structure the subclasses. The user-oriented class structure must converge with the submission-oriented class structure for optimal search and retrieval. It shows that the way in which the material is held in the LSS is compatible with the way the material actually exists.

B.2.3. Data Model of YMPO Documentary Material

Design of computer-assisted information management systems requires a thorough understanding of the actual structure and operational utility of the material that is to be represented logically. Efficient retrieval of technical data from a document management system based on bibliographic indexing presents some special design considerations. Technical data is distributed through various types of records. Logical design of the LSS must consider the class structure of the records inventory to be loaded into the system. However, realizing that in most cases, technical data content may ultimately be the primary retrieval criteria, the bibliographic header must be designed to accommodate all of the extant record types in addition to communicating the technical data/information content of the record. In actuality, a 'many-to-many' type of relationship usually exists between data and the records in which it is contained. At a fundamental level, many different types, or classes, of records can be used to archive, or represent, an individual class of data. Likewise, many different types of data may be stored on a single type of record.

The LSS Rule, by determining submission requirements in terms of the data form of the record, emphasizes record type as an organizational focus for design. It should also be noted that the YMPO information management procedures are heavily based on record type. Accordingly, the authors have developed a data model of the record class structure of the YMPO that shows how technical data is actually distributed through the record classes, and shows relationships between record classes based on technical data content (Figure B-6). The guiding premise here is that search and retrieval operations will be carried out by users thinking in terms of the real-world form and function of documentary material and data content, rather than in terms of how it is represented in the LSS. The purpose of the model is to show how actual records can be classified using attributes and operations specified by the LSS Rule. In particular, the set of operations used in the model are primarily those used to describe documentary material in the Rule (Figure B-2).

The basic unit of classification is the 'record'. A record is here considered to be the fundamental unit of documentary material to which specific attributes and operations can be applied. A class will always be characterized as a group of records that share a specific set of attributes, and to which a specific set of operations applies. This does not mean that every record in a class has exactly the same set of attributes and operations, but rather that every member of the class has a defined set of attributes and operations in *common*. Indeed, the class is defined by the set of attributes and operations that applies to it.

The initial superclass of records to be considered is basically the entire world of documentary material generated by the high-level waste project (Figure B-6). At this point, the primary attribute of this material is that it is required to be represented within the LSS. The main thing that needs to be done with each unit, or record, of this material at submission to the LSS is the headering procedure. The header is carried as an operation, rather than an attribute in this model because the focus is on the material as it actually exists, rather than how it is represented in the LSS. DOCUMENTARY MATERIAL consists of three basic classes of records (Figure B-6). Each of the child subclasses inherits the attributes and operations of the parent class. That is, every record in each of the three subclasses consists of material that must be submitted to the LSS with a header.

HARD COPY RECORDS are considered at this point to be records printed on paper, the definitive attribute. The distinction between, for example, the class of HARD COPY RECORDS and its parent class of DOCUMENTARY MATERIALS illustrates the utility of this modeling approach. HARD COPY RECORDS are the subclass of DOCUMENTARY MATERIAL to which the scan (or, 'imaging') operation applies. Addition of the scan operation essentially is the basis of creation of the subclass. By definition, every HARD COPY RECORD will be 'imaged', or submitted as a digital 'image'. This operation is not true for the superclass of DOCUMENTARY MATERIAL because this class contains records, in addition to the hard copy material, that cannot be scanned.

The RECORD PACKAGE class consists of documentary material that has been physically grouped together. Each item within the class is related to the others by the task under

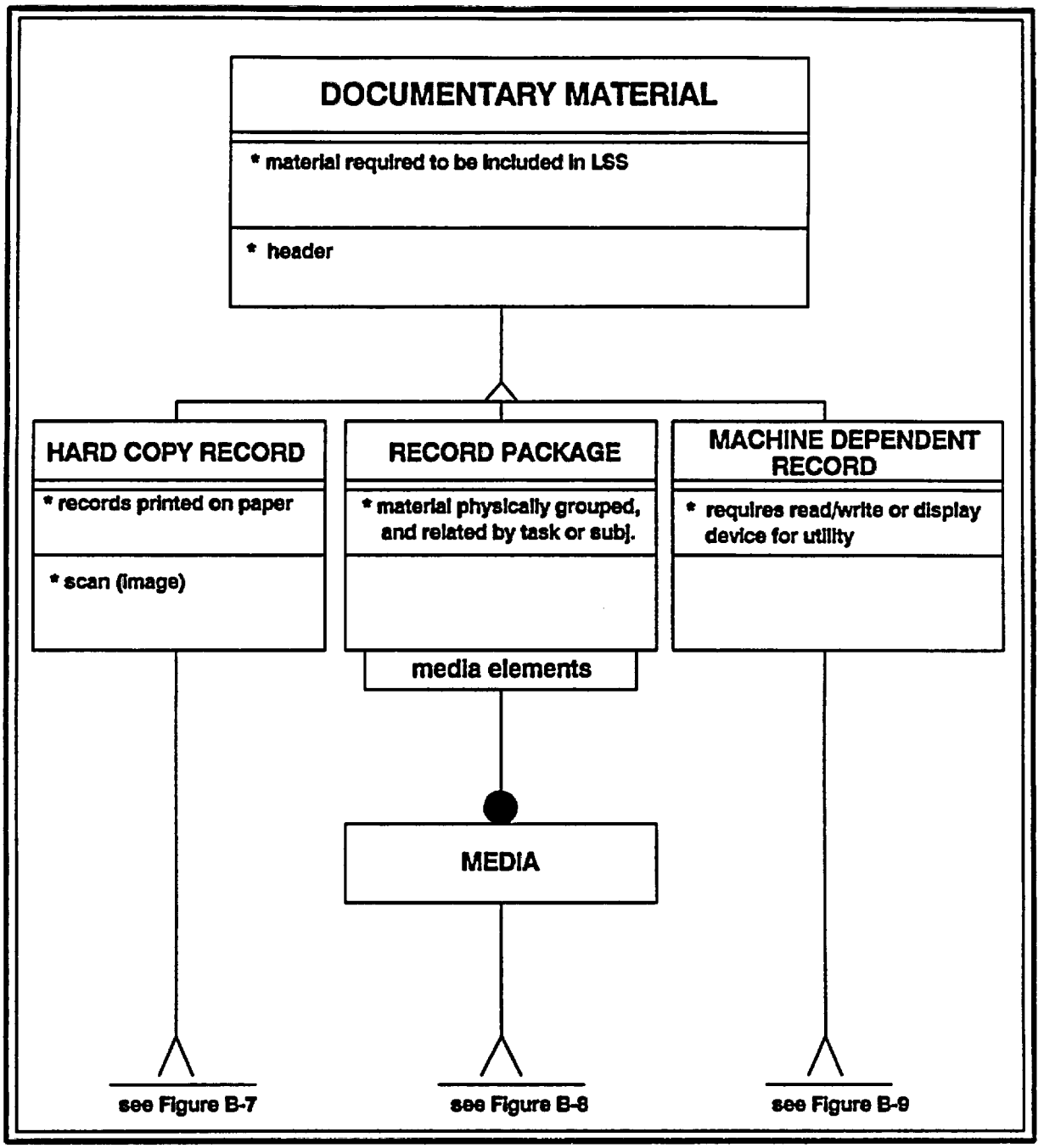


Figure B-6. Subclassification of the Initial Superclass of DOCUMENTARY MATERIAL.

which it was created, or by closely related subject. The RECORD PACKAGE class may contain all categories of documentary material. Operations are not assigned to the RECORD PACKAGE, as there is no operation common to all elements of the class. RECORD PACKAGE inherits the header operation from DOCUMENTARY MATERIAL, so it is not necessary to repeat the annotation. In this model, the package is the record. A package is not considered to consist of records, but rather to consist of units of documentary material distinguished on the basis of the media that it is recorded on. The 'media elements' tag on the bottom of the class box indicates that RECORD PACKAGE is not divided into different types of packages, but rather that the material within the package is subclassified on the basis of media type. The line-and-solid-circle connection between the media elements tag and the MEDIA box means that any individual unit, or item, within the package may occur on a variety of media (e.g., hard copy paper, magnetic tape, photographic positive, etc.). The line and branch below the MEDIA box indicates that media is divisible into subclasses. In sum, record packages are subdivided on the basis of the physical media on which the individual units of documentary material in the package is stored.

The class of MACHINE DEPENDENT RECORDS consists of documentary material that requires some electronic or optical/mechanical device to read, write or display, the data or information contained on the record. This class of records consists solely of Category III documentary material, i.e., material that may not be directly represented in the LSS. It is possible to format, or reformat, most of this material so that is could be entered into the LSS, and made retrievable by query. The LSS Rule suggests that this was not the initial intent. However, this option will be considered in a later section on access to technical data.

B.2.4. Hard Copy Records

HARD COPY RECORDS are subdivided into TEXT RECORDS, COMPOUND RECORDS and GRAPHICS RECORDS (Figure B-7). The subclasses inherit the scan operation from the parent class, and the attribute that they are all printed on paper. Categories I and II (Figure B-2) materials are accommodated here because the net set of operations at this level of the model is, to write a header, scan the record, and code the appropriate text.

TEXT RECORDS are all 100 percent text. These records are subdivided into PRINTED TEXT RECORDS and HANDWRITTEN TEXT RECORDS. PRINTED inherits the scan and header operations and adds the code-text operation by optical character recognition of the raster image. HANDWRITTEN records inherit the scan operation but do not add a code-text operation.

COMPOUND (hard copy) RECORDS consist of units of documentary material that are composed of mixed text and graphics. COMPOUND inherits the scan and header operations and the printed paper attribute. These records are not subclassified but are considered to be composed of format elements. The line-and-solid-circle connection between the 'format elements' tag and the FORMAT box is another one-to-many relationship. Any individual document is composed of some combination of the TEXT format and GRAPHICS format

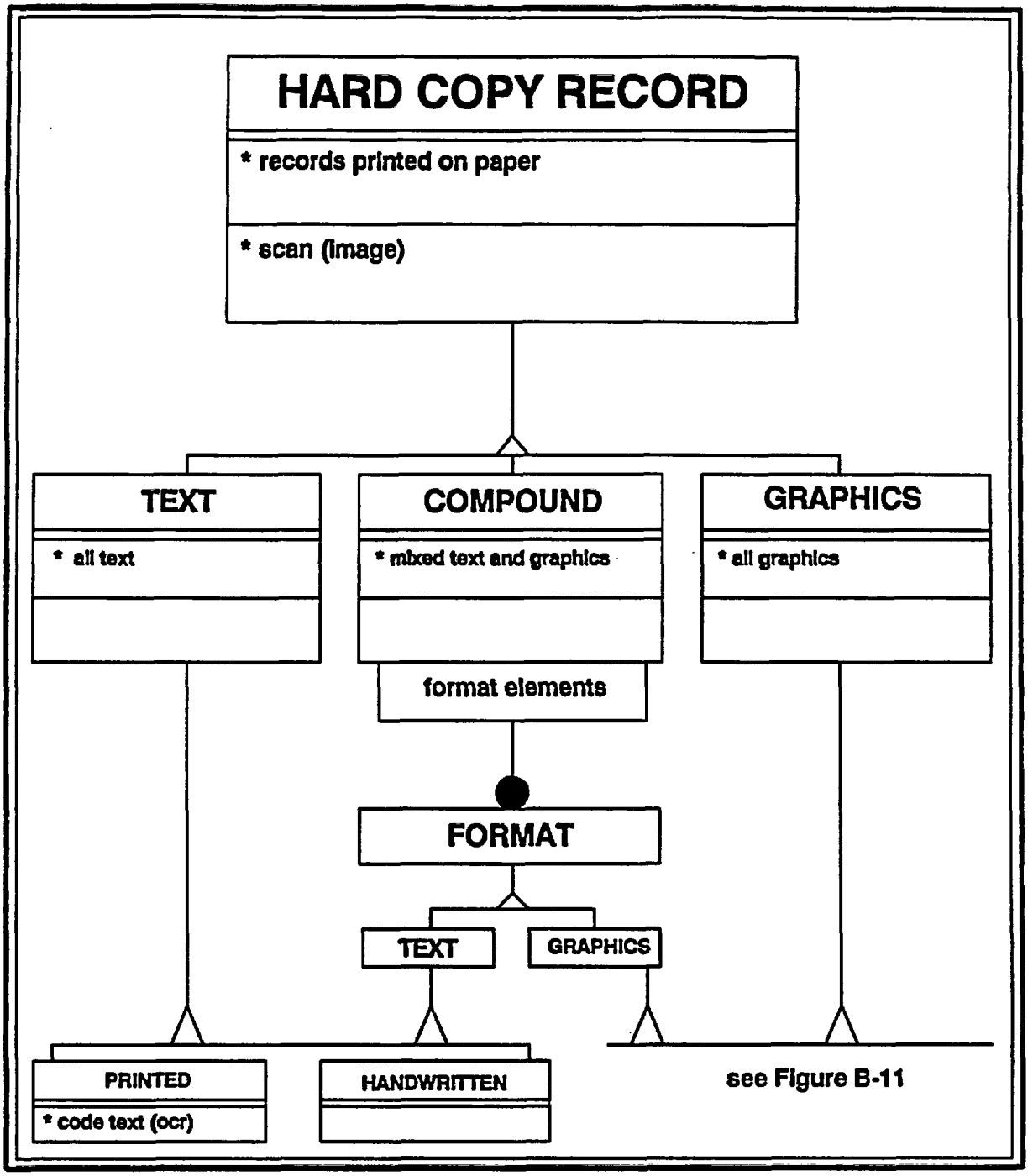


Figure B-7. Subclassification of HARD COPY Records. (OCR) indicates that the text coding operation in the PRINTED TEXT class is by optical character recognition of digitized text.

classes. The branch below TEXT format shows that either PRINTED or HANDWRITTEN text can occur in the compound record, and that if the text is printed, the code-text operation applies.

GRAPHICS RECORDS are all 100 percent graphics. This class inherits the scan and header operations. GRAPHICS and GRAPHICS-format COMPOUND RECORDS are further subclassified.

B.2.5. Record Packages

RECORD PACKAGE inherits the header operation and has the added attribute of a descriptive Table of Contents that lists all items of documentary material contained in the package. Units of documentary material contained in a RECORD PACKAGE may be recorded on any of three classes of MEDIA (Figure B-8). HARD COPY MEDIA inherits the scan operation but not the header operation. Since the package is considered to be the record, the package header suffices to reference the included material. The Table of Contents can then be used to determine the detailed contents of the package.

HARD COPY MEDIA within a package consists of a single subclass of PAPER media. Any individual document on paper media can then be discriminated on the basis of the class of record carried on the media. Subclassification of RECORD class here follows that of the HARD COPY RECORD class (Figure B-7).

MACHINE READABLE MEDIA requires the use of a read/write or display device. The header operation must be assigned at this level since the it is not directly inherited through the MEDIA class. A separate header is considered to be an essential reference for this material, even though it is in a package. The physical media require that these items be in controlled storage in anticipation of a potential user request.

Individual items of FILM MEDIA must also have headers assigned. FILM MEDIA records are generally considered to require use of a display/print device for most practical purposes.

B.2.6. Machine Dependent Records

MACHINE DEPENDENT RECORDS are subclassified by the type of data written on the record (Figure B-9). Both DIGITAL and ANALOG data can be recorded on machine readable media. Analog recordings are also commonly made onto photographic film plates or strips. This class of records inherits the header operation, and each subclass in turn inherits headering.

MACHINE READABLE MEDIA is divided into seven general subclasses (Figure B-10). Each subclass of recording media may hold any one of nine classes of DATA TYPE. This is the first occurrence of data type in the top-down hierarchy of the model. The model is not a flow chart, and therefore does not have any explicitly directional characteristics.

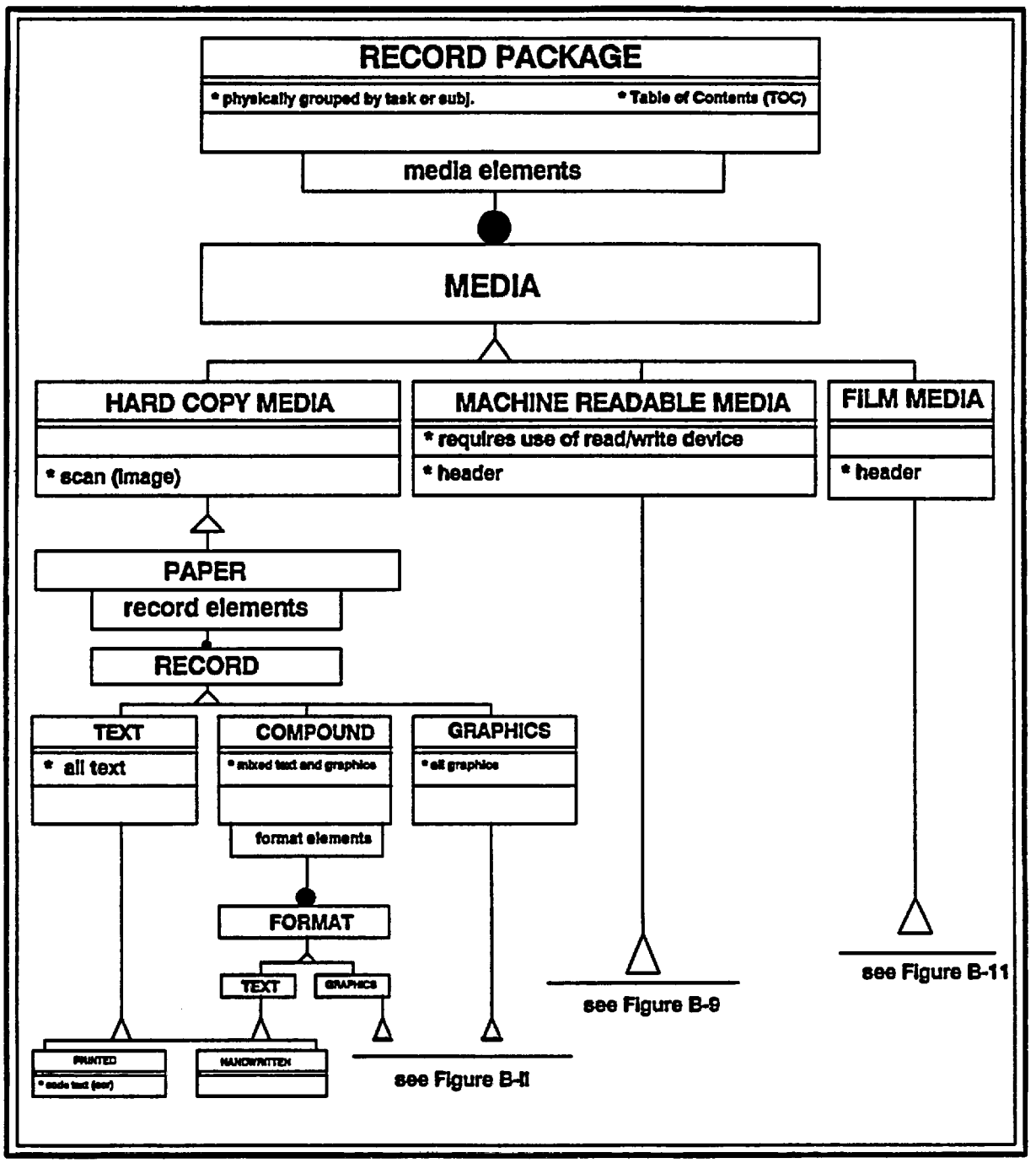


Figure B-8. Subclassification of RECORD PACKAGE

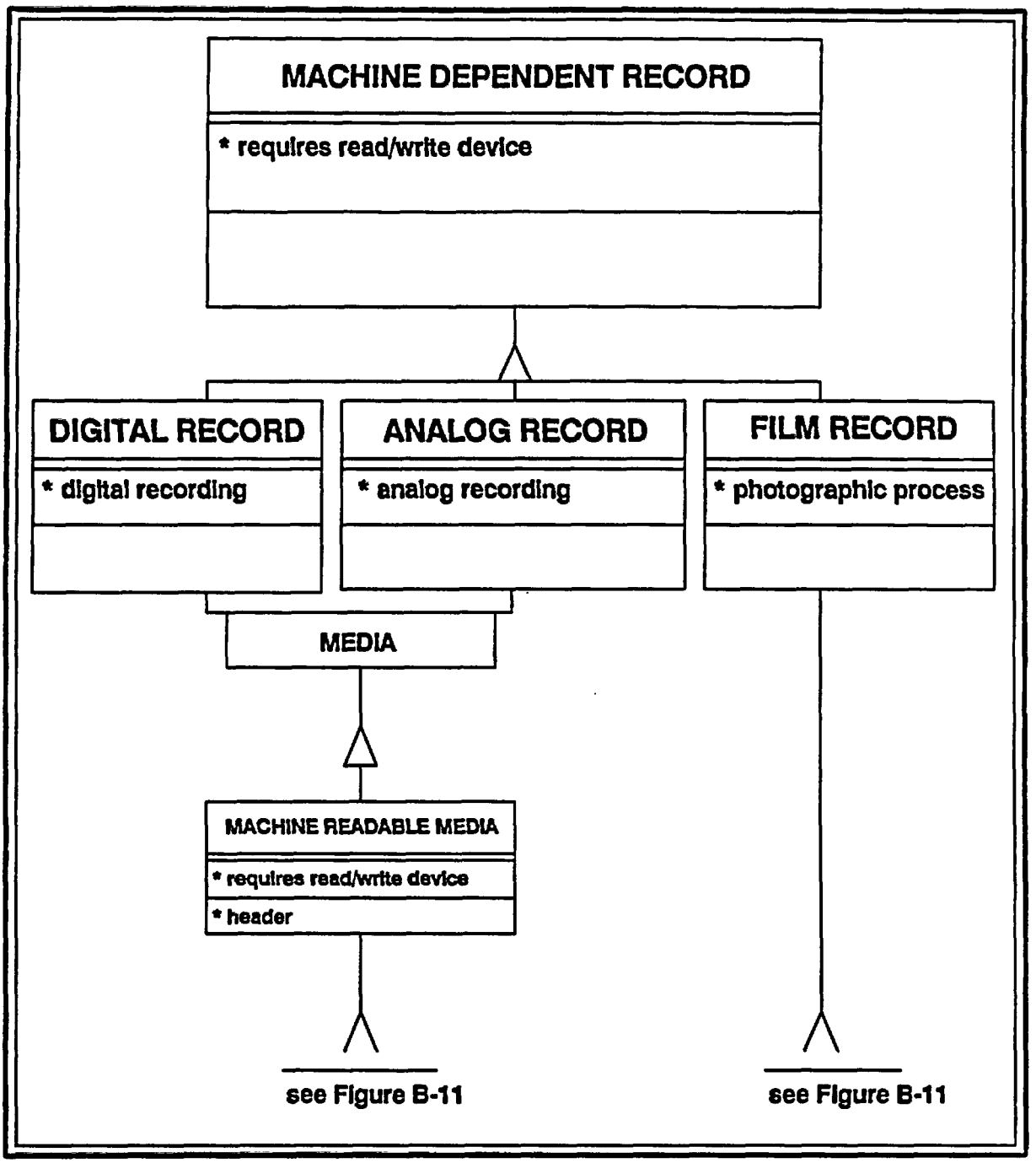


Figure B-9. Subclassification of MACHINE DEPENDENT RECORDS

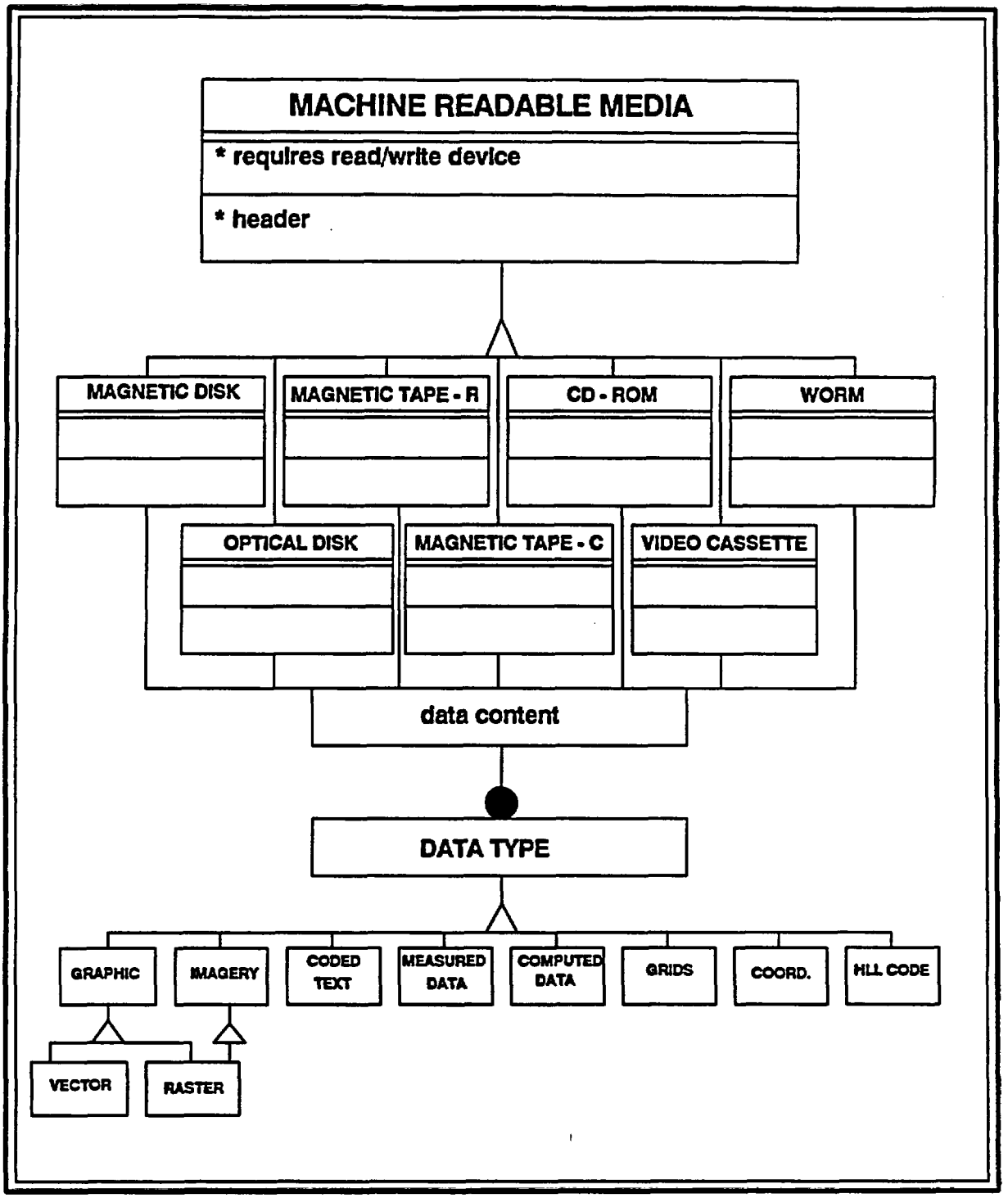


Figure B-10. Subclassification of MACHINE READABLE MEDIA

However, if a bottom-up perspective is taken at this level, it is clear that any one data type could also be recorded on any of the media. This type of many-to-many relationship largely precludes the inception of classes of technical data, based on media type, that have sufficiently unique sets of attributes and operations.

Most of the YMP documentary material can easily be classed into one of the DATA TYPES on the lower level of Figure B-10. However, this level of classification is sufficiently far removed from both submission requirements and retrieval criteria to be of only adjunct concern to either. Information about these data types is critical, however, to internal management of the LSS database(s).

FILM MEDIA have the attribute of being recorded by a photographic process and inherit the header operation (Figure B-11). FILM RECORDS have a one-to-many relationship with FILM MEDIA. An individual film record may be recorded on any one of at least seven different media.

B.2.7. Data Types

DATA TYPE has explicit many-to-many relationships with all MEDIA types and with the fundamental format elements of TEXT and GRAPHICS (Figure B-11). The line-and-solid-circle connections at the base of the model show that any one of the record FORMAT classes can be recorded on a variety of MEDIA classes using many different types of data, or can be recorded photographically on a variety of FILM MEDIA. This type of relationship precludes unique description of technical data using data type, format or media as unique descriptors. They are properly utilized as members of a set, or suite, of header fields, and are thus descriptive adjuncts to technical subject.

Printed text can be represented in many different ways without changing the technical or information content of the record. A text record may be digitally scanned and converted to a raster image. The bit patterns on the image can then be converted to a character code like ASCII. Alternatively the record could be photographed onto microfilm, and then reprinted back to paper. It all reads the same. Measured or computed data is a more complex case. The data may be represented as rows and columns of alphanumeric characters. The data may be plotted, or mapped to reveal useful graphic patterns. The same plot can be recorded as either vector (vertex coordinates) or raster (bit-mapped) data, or a hard copy of the plot can be photographed. Conversely, virtually any type of data can be transformed for direct representation at the binary level. Photographs, video imagery, maps, designs, data plots and various types of data visualizations can be represented as raster imagery, decomposed into constituent data types or photographed.

B.3. ACCESS TO TECHNICAL DATA

Since the primary intent of the LSS is to shorten the time required for license review by making documentary material immediately available for discovery or technical review, every

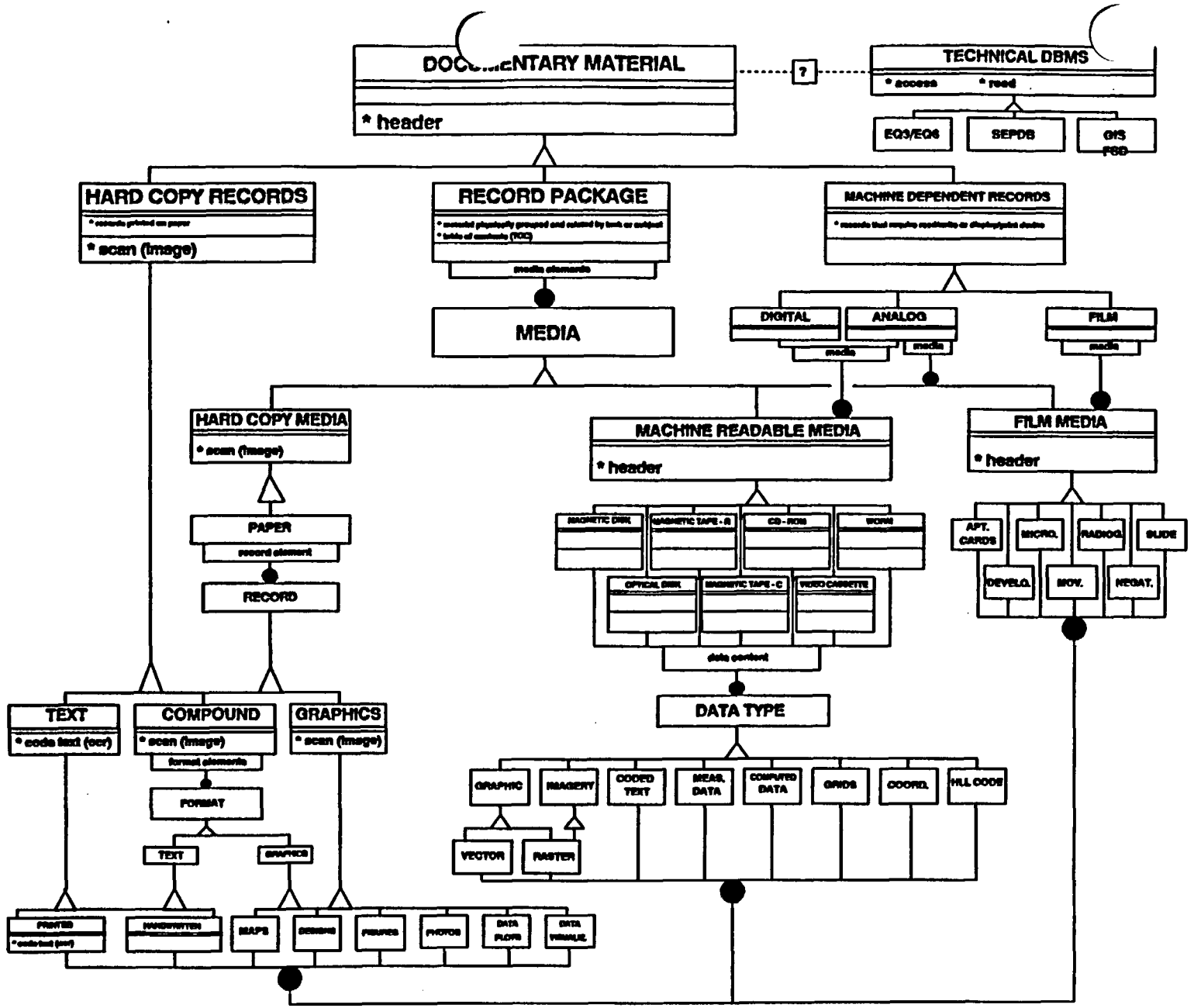


Figure B-11. Full Data Model of DOCUMENTARY MATERIAL

effort should be made to image as much as possible. The authors do not anticipate that major time delays will involve retrieval of the reference to a record, but rather the delays will occur when an image is not available for examination on the LSS workstation. Most of the time involved in the retrieval is in waiting for a copy of the record to examine. Accordingly, all hard copy material should be scanned. Only unusual machine-readable media should be excluded from direct entry into the LSS.

Effective and efficient access to technical data through the LSS will depend on adequate description of the data within the framework of the bibliographic record header. The set of header fields most suitable for accommodating this description are: i) TITLE/DESCRIPTION, ii) (TECHNICAL SUBJECT) DESCRIPTORS, and iii) MEDIA. Review of YMPO/CRF technical reports and data record packages by the authors, and use of the CRF Reference Information System (RIS) has shown that the most successful query procedure is centered on the use of and-string arguments in the TITLE field of the RIS header. There are some reasons for this which are relevant to classification and retrieval of technical data in the LSS.

Of primary importance, is the style and technical information content of the titles of technical records. Much of the technical material currently in the RIS are geoscience reports. Scientists in general, and perhaps geoscientists in particular, tend to be very descriptive in the formulation of titles. A title such as "Preliminary Geologic Map of Yucca Mountain, Nye County, Nevada, with Geologic Sections", seems to contain all the information most potential users would need in order to determine whether to examine it further. It is easily found by using a title query of GEOLOGIC&MAP&YUCCA&MOUNTAIN. This approach does not depend on exhaustive classification of the report, or the map, or the geologic cross sections contained in the report. It is not even particularly necessary or important to place a lot of information on this report in the header. If a user is interested in geologic maps and sections of Yucca Mountain, that user is probably going to take a cursory look at most of the geologic maps retrieved using a certain query. Therefore, it is of utmost importance to the user in this case to be able to easily view an image of the map.

Rapid and straightforward access to the image database is paramount. Design decisions should take this into account. Accommodation of the imagery is of prime importance. Therefore, submission protocols should ensure that every reasonable effort is made to transform graphic records to raster imagery.

An equivalent search for this report using key-word descriptors is not as simple. The document descriptors used by the CRF tend to be overly generalized, and in some cases a little obscure. Descriptors used for the report mentioned above include: GEOLOGY, MAPS, YUCCA MOUNTAIN, AND NYE COUNTY. Consequently a query for GEOLOGIC MAPS, which is the way it would normally be phrased, can be problematic. Document descriptors for the data record package titled; "Geohydrologic & Drill Hole Data for Test Well USW H-3 Yucca Mountain, Nye County, Nevada" are listed as BOREHOLES, DATA, and USW H-3. A title query, composed of the descriptors, such as: BOREHOLES&DATA&USWH-3 would not find

the record. Alternatively, the record is difficult to hit using 'GEOHYDROLOGIC, DRILL HOLE AND USW H-3' as descriptors.

The particular form, format and storage media of the record is not especially important at this point. It is important that a user be able to, i) find the record, or discover that it exists; ii) quickly look at the record, text-search/read the record/abstract; iii) quickly glance through (view images of) the figures; iv) decide if this record contains the type of technical data that is needed.

More detailed descriptions will be necessary for records on machine readable and film media that are not immediately viewable as an image file. Much of this material is contained within the inventory of data record packages.

B.3.1. USGS Data Record Packages

Most of the data record packages (DRPs) on file at the CRF that have a substantial amount of technical data content, have been submitted by the USGS. Record packages submitted to the YMPO/CRF by the USGS Local Records Center (LRC) are fairly consistent in format and style. The packages, however, may encompass material that would be considered as Categories I, II or III, if that material was not within the package. The YMPO currently treats DRPs as a discrete record, and so creates a single RIS header to reference it.

USGS data record packages usually contain supporting documentation for a published technical report. A typical DRP (Exhibit I B) will have the technical report (section A), a 'published report package' (section B) consisting of in-house reviews and correspondence, and a set of supporting 'raw data documents' (section C). The bulk of the technical data is in the technical report (usually an Open File Report or a Water Resources Investigation Report), and in the raw data section. Documentary material that comprises the package can be classed as HARD COPY MEDIA, MACHINE READABLE MEDIA or FILM MEDIA (Figure B-11). The published report is primarily a hard copy printed paper record, and as such is scanned and text-coded as appropriate. Likewise the published report package section. The raw data documents are usually a mix of record types.

For example, the raw data section of OFR 84-149 (Exhibit I A) contains a variety of field notes, test results, survey measurements, instrument calibrations, and observations. Most of this material is on 8.5 x 11" paper, and is composed of graphics or text not suitable for coding. This material is adequately referenced by the package header and Table of Contents, and only requires scanning for entry as an image file.

Much of the technical data currently under the control of the YMPO/CRF is directly associated with technical reports for which data record packages have been compiled. A variety of documentary material compiled in support of an investigative, analytic or interpretative study is included in these data record packages.

Exhibits I A and I B are representative examples of the diversity of material that can occur in a package. The technical report which heads the package (section A, in the package Table of Contents) can, of course, include a considerable amount of basic and interpreted data. The raw data section (section C), however, contains the bulk of the predominately hard copy graphic material that is not appropriate for text coding, and the material that is on machine readable media which may not be suitable for entry into the LSS.

YMPO/CRF procedures currently direct that hard copy records which cannot be easily microfilmed, be physically separated from the package and either placed on aperture cards (Exhibit II) and archived separately or, if not appropriate for an aperture card, just archived separately. Oversized or non-routine graphic oriented records such as large maps, borehole geophysical logs, photographs, and graphics printed or drafted onto vellum or mylar are included. A 'slip sheet' is placed in the package in place of the removed record.

The slip sheets currently included in the sampling, made by CNWRA, of data record packages are generally not sufficiently descriptive of the records they replace. LSS participants intending to use the LSS for technical review will not find these slip sheets of much use. Line item #3 in the Raw Data Documents section of USGS OFR 84-149 (Exhibit I B) references the slip sheet used to replace geophysical logs for the USW H-3 borehole. The slip sheet states that it is the responsibility of Fenix and Scisson (Birdwell Geophone Survey) to submit the logs. No other reference is given. All geophysical logs that are hard copy prints on paper or film should be scanned and held as an image file in the LSS. This includes field prints (printed by the logging unit real-time as the instrument is run out of the borehole) and final, or processed, prints (includes processing done after the log run). Geophysical log data currently held on tape should at least be printed to hard copy for scanning. The authors are currently considering the relative merits of reading the log data tapes directly into the LSS.

Line item 1 of the Raw Data Documents section of USGS OFR 83-669 (Exhibit I C), and of USGS OFR 81-1086 (Exhibit I D) is 7 slip sheets, and six slip sheets, respectively, for Develocorder Films containing earthquake seismicity data. This type of record is difficult to transform to a digital record, and so will be archived separately. The Develocorder film record should be represented in the LSS by a header. This is a typical example of the type of documentary material that, when stored externally from the LSS must have a header, even though it is listed in the Table of Contents of the package. Slip sheets alone are not adequate for effective retrieval of the data once it is identified.

The slip sheets for the develocorder data, for instance, contain an explanation of how and where the films are stored (see slip sheet entitled: Explanation of Seismic Film Record Storage) that includes the address of a warehouse in Jefferson County (Denver), Colorado and the lock combination to the door (Exhibit I C). Certainly, this material needs much more comprehensive, and up to date, description as to the data content and applicable access protocols. Users of the LSS must have clear direction for retrieving this data in a timely fashion once the record has been identified.

Table of Contents line item C - 5 of the same package (USGS OFR 83-669: Southern Great Basin Seismological Data Report for 1981 and Preliminary Data Analysis (Exhibit I C) is for 228 computer file header printouts describing archive tapes containing earthquake seismicity data. Contacts and tape numbers are handwritten on the header printout page, dated June 29, 1987. The authors recommend that photocopies of the tape cartridge face be inserted into the package either in place of, or in addition to the file header printouts. A recommended LSS header is described in a later section of this report (HEADERS FOR MACHINE READABLE MEDIA).

Drawing Packages (Exhibit III) are treated similarly to Data Record Packages in that oversized (greater that 8.5 x 11" page size) hard copy records are physically separated from the package and photographed onto aperture cards. Slip sheets included in the package in place of these drawings are of no use to a potential user and should be replaced with the original drawing document.

The separation of 'oversize' material from a package, by the CRF may be of some concern to the LSSA because opportunity exists here for omission of the separated material. Enclosures like maps, charts and photographic plates should be recovered and included with the parent record package or record. Separation of packaged material apparently requires dependence on the RIS to find the separated material. This may not be a direct concern of the LSSA, as the participant is required to submit complete records. However, the authors believe that special attention may need to be given to verification that all separated material has been recovered.

The Data Record Package for USGS OFR 88-560: Location Refinement of Earthquakes in the Southwestern Great Basin, 1931-1974, and Seismotectonic Characteristics of some of the Important Events (Exhibit I E), is a case in point. Section D, line item 1 is the slip sheet for Plate 1 of the technical report (USGS OFR 88-560) for which the package was compiled. The "No.-of-pages" for this line item is 0 (zero), and there is no slip sheet in the package, nor aperture card number on the RIS header for the package. Checking the references on the package RIS header leads to reference number NNA.89043.0054, which is the separate open file report (88-560). Now checking the RIS header for the separate report yields the aperture card number; 90000018894041, which is Plate 1. However, the aperture card number for Plate 1 is also not on the slip sheet for the separate technical report. In other words, Plate 1 is only listed on the RIS header. It is nowhere on the slip sheets of either the package or the report, nor is it on the package RIS header. Submission of the Data Record Package with its included technical report may allow for omission of Plate 1, unless the indexer is diligent in running down the aperture card reference. If an indexer did not read the Table of Contents to learn that Plate 1 needed to be included, it may end up missing in the LSS image file.

Line item D of Data Record Package USGS OFR 82-466: Electrical Studies at the Proposed Wahmonie and Calico Hills, Nuclear Waste Site, Nevada Test Site, Nye County, Nevada (Exhibit I F) lists 21 maps sent under separate cover. The package slip sheet does not describe the material or give the location or information needed for access. The RIS header for

the package lists the aperture card numbers 9000003530 - 9000003550 (twenty cards, not twenty one). A situation can be envisioned in which an indexer may not read the Table of Contents of the package carefully, and would not check the references on the RIS header. The package would be scanned into the LSS without the 21 maps, which are likely to be very important to the content of the report. If the maps were kept with the package, it would not be necessary to go through these extra steps in order to make sure that the maps are submitted to the LSS. An indexer should read the Table of Contents, and/or check the RIS header, to verify that enclosures and plates have not be omitted?

The authors are currently considering a recommendation to the LSSA that all technical reports available as published material from the originating agency (such as the USGS, Sandia National Laboratory SAND Reports, etc.) be acquired from that agency, or national laboratory, for entry into the LSS. Comparison of the acquired inventory with the YMPO/CRF RIS would identify any discrepancies that could then be filled from either the agency or the CRF. The CRF is still required to submit all project documents, but the published reports would not need to be scanned again.

Line items C-12 and C-13 on the Table of Contents for Data Record Package USGS WRIR 84-4272: Geohydrology of Test Well USW H-3, Yucca Mountain, Nye County, Nevada (Exhibit I G) are slip sheets for various documents including geophysical logs, drilling reports and pump tests. Reference to the slip sheets will show that there is no information on these sheets of use to potential users. This material should be recovered and submitted with the package. It does not do a technical reviewer or regulatory analyst any good simply to know that these things exist. The user needs to be able to view an image of the document to assess its utility (see Appendix G).

Exhibits I H through I P are additional representative samples of USGS Data Record Packages which give examples of removed figures, maps, plates, and other oversize material that were determined to be not suitable for microfilming at the time of submission to the CRF. Review of the Tables of Contents of the packages, with reference to the slip sheets attached shows that aperture card reference numbers are commonly omitted from both the slip sheets and the RIS header for the package. In several instances, access to the RIS header for the separate technical report was necessary to find the aperture card reference numbers. There seems to be substantial opportunity here for inadvertent omission of material from the package when submitted to the LSS. In addition, there are instances of a substantial number of figures removed from the package and represented by slip sheets which give a physical address for the missing material.

Exhibits I Q and I R, are a sampling of other types of Data Record Packages. These packages have not been reviewed in detail, however the same potential problem exits here, in that slip sheets are not sufficiently informative.

Documentary material that is part of a package, and held as digital imagery within the LSS is accessible through the package header and the package table of contents.

Accordingly, the authors intend to recommend that hard copy records that are part of packages not be separately headered. Further, the authors intend to recommend that all of the hard copy graphic material that is currently held separately from the data record package be physically included in the package as submitted to the LSS. This material should be scanned into the LSS, and text-coded where appropriate.

B.3.2. Sandia National Laboratories (SNL) Data Record Packages

The Yucca Mountain Project Data Records Management System (DRMS) produces a Sandia Data Set (SDS) that is roughly equivalent to the Data Record Package produced by the USGS for the YMPO/CRF. The DRMS is organized by File Guide (Exhibit IV), and divided into Laboratory Experiments (type L) and Field Experiments (type F). DRMS further appends a technical subject designator to the SDS. File guide numbers with technical subject designator are listed with the technical subject name in Exhibit IV A. Specific records are discrete reports with pertinent data attached, and are designated by the start date of the project that generated the report. For example: for Sandia Data Set ID# 51/L01A - 7/24/78; 51/L01A = SDS technical subject designator for studies of Thermal Conductivity; 7/24/78 = record (report) designator, and is the start date of the project.

The DRMS Data Catalog lists each record in the SDS inventory by file guide-record designator (Exhibit IV B), and gives a brief description of the type of data acquired. The DRMS Data Catalog and the associated Tables of Contents for each record listed are in a computer database. The authors intend to recommend that both the Catalog and the Tables of Contents listing be included in the LSS, in addition to each of the Data Sets.

An inventory of machine-dependent records at the Sandia DRMS/LRC (Exhibit V) lists the type of record by media and associated Data Set.

B.3.3. Reference Information Base

The Reference Information Base (RIB) is produced and maintained by Sandia National Laboratory as a synthesis of YMPO data acquisition and analysis activities (Exhibit VI). Assumptions made in analysis and modeling, limitations on use of the data, and recommendations on how to interpret the data are explained in a narrative section on each subject.

B.3.4. Technical Database Management Systems

B.3.4.1 Site and Engineering Properties Data Base (SEPDB)

The SEPDB (Exhibit VII) is developed and maintained by Sandia National Laboratory as a central source of numeric parameter data generated by the YMP. As the name implies, these data are site specific. They are generated by sampling and investigative studies carried out at the Yucca Mountain site and in the nearby area. The authors intend to

recommend that SEPDB data catalogs be submitted to the LSS as they are published by the administrator (currently SNL), and that DATA CATALOG be added to the field code list for the DOCUMENT TYPE field.

The authors intend to recommend that the LSSA consider the implications of direct on-line access to the SEPDB through the LSS. Read-only, on-line access to this database would greatly enhance the value of the LSS as a technical review tool, both in the prelicensing and license application review periods. NRC staff would benefit from the capability to down-load data from the SEPDB for display, analysis and verification.

B.3.4.2. Geographic Information System (ARC/INFO GIS)

The ARC/INFO GIS is operated by EG&G for the YMPO. This system is likely to become the largest and most comprehensive of the technical database elements. The purpose of this system is to create maps and other types of displays of spatially distributed data. The 'records' created by this system are all digital. Of course any of these records can be printed, or plotted to hard copy for scanning. However, the authors intend to recommend that an appropriate interface be created for the LSS to directly read and store the GIS database (possibly in INGRES format). In addition periodic (quarterly?) data catalogs should be requested from the YMPO by the LSSA for submission to the LSS.

B.3.4.3. Geochemistry and Thermodynamics GEMBOCHS (EQ3/6)

The EQ3/EQ6 database contains a more general set of data on the geochemical and thermodynamic properties of various minerals, aqueous chemical species, gases, and an extensive radionuclide database. Periodic data catalogs should be submitted to the LSS by the administrator of the database (currently LLNL). Exhibit VIII is a representative print-out of a small portion of the EQ3/EQ6 database.

The authors intend to recommend that the LSSA consider the implications of direct, on-line access of the EQ3/EQ6 database. Direct access to this data would enhance the utility of the LSS as a technical review tool.

B.4. REFERENCES

Cardenas, A. F., and D. McLeod. 1990. *Research Foundations in Object-Oriented and Semantic Database Systems*. Prentice-Hall. Inglewood Cliffs, NJ.

Rumbaugh, L., Blaha, M., Premerlani, W., Eddy, F., and W. Lorensen. 1991. *Object-Oriented Modeling and Design*. Prentice-Hall. Inglewood Cliffs, NJ.

Shlaer, S., and S. J. Mellor. 1988. *Object-Oriented Systems Analysis*. Prentice-Hall. Inglewood Cliffs, NJ.

APPENDIX C

UNITIZATION

C.1. INTRODUCTION

The LSS is intended to provide access to information relevant to the licensing of the repository for all participants. This means that the LSS must support the needs of individual users by providing effective search facilities and electronic access to information. Thus, it is important for the LSS to store documentary materials in a manner which permits users to effectively and efficiently identify and retrieve them. It is also important for the increments of documentary materials in the LSS to be relatively small and specific so that they may be effectively retrieved and presented.

C.2. THE NEED FOR IDENTIFYING AND RETRIEVING UNITS OF INFORMATION

If a search request results in the selection of several thousand pages of documentary materials which are then displayed consecutively on the screen, the user will find it difficult to absorb and use all of the information. Thus, it is important for the user to structure the query so that relatively small increments of information are selected for display. It is consequently important for the system to store information in such a way that it is accessible and retrievable in relatively small increments. Sophisticated search algorithms involving a combination of header and full text searches can be used to focus the user's query and select a very specific set of documents. But if the information in the database is not sufficiently unitized, the search results may be unmanageable from the user's perspective regardless of the sophistication of the search algorithms.

This requirement speaks to the need for considerable unitization of the information in the LSS. If, for example, a user's query resulted in the selection of a non-unitized package of information which contained several thousand pages of images, the user would have no alternative but to display all of the pages sequentially. If the desired information were actually at the end of these several thousand pages, then the user would have to bypass the irrelevant pages one by one until the desired page appeared on the screen. Clearly, this situation would become very burdensome for the user, and it would also be very inefficient and expensive from the perspective of computer and communications resources. Therefore, it is important to unitize the information in the LSS to an extent that will permit the user to select the smallest possible increment of information which will satisfy his or her information needs.

APPENDIX D

ELECTRONIC DISSEMINATION OF INFORMATION

D.1. INTRODUCTION

One of the principal objectives of the LSS is to provide rapid, convenient access to information relevant to the licensing of the repository for all participants. This means that the LSS must support the needs of individuals by providing effective search facilities and also by providing an electronic method of information dissemination.

D.2. THE NEED FOR ELECTRONIC DISSEMINATION OF INFORMATION

The LSS is intended to provide for electronic dissemination of information to individual participants and users. When a user has searched through the LSS and "discovered" the existence of some desired information, the system is intended to include facilities to make that information available to the user electronically. This requirement for electronic dissemination of information is the driving concept behind the maintenance of bit-mapped images of the documents. In effect, the LSS is intended to provide a combination of an electronic reproduction machine and an electronic mail service. The user should not only be able to search the LSS to find out about the existence of a document but he should also be able to obtain, view, and use an electronic copy of the document.

The implementation of this facility for electronic dissemination of information to the LSS participants should be very responsive and should also be relatively transparent to the user. Once the information has been identified and selected, the user should gain access to the image of the document or to the ASCII text simply by pressing a "view" or a "text" key. The user should also be able to use the electronic copy of the document to make a physical copy of the currently displayed page by simply pressing a "print" key, and should be able to obtain a copy of the entire document by pressing a "print document" key.

It is this requirement for the rapid dissemination of information which renders the maintenance of large amounts of non-electronic document media unacceptable in the LSS. A paper-based or microform-based LSS would not satisfy the needs of the users. Therefore, it is essential that, insofar as possible , all documentary material must be captured in an electronic form. This means that all documents must be scanned and, where possible, their images must be processed to produce searchable text.

APPENDIX E

THE APPLICATION OF HYPERTEXT TECHNIQUES TO PACKAGES

E.1. INTRODUCTION

A relatively new technique called "hypertext" has been developed which permits computer users to move quickly and efficiently between displays of related records and text. This technique permits rather complex relationships between related records to be presented clearly and efficiently, and permits the user to navigate freely between units of information without having to deal with conventional menus and selection lists. Within the LSS, the storage and retrieval of information within packages is a very promising candidate for the application of "hypertext" techniques.

E.2. WHAT IS HYPERTEXT?

The term "hypertext" is used to describe an information storage and retrieval technology which permits logical relationships between different text records to be embedded within the textual data itself. In a hypertext application the logical relationships, called hypertext links, are constructed such that certain words and phrases in the textual data are internally associated with other records and processes. Typically, the textual data is presented to the user on a display screen and certain words or phrases within the text are highlighted. These highlighted words or phrases are the hypertext links. When the user "points to" and selects one of the highlighted words or phrases using a mouse or other pointing device, the system responds by activating the hypertext link and executing the procedure associated with the selected word or phrase so that the related block of textual data is retrieved and displayed.

Hypertext applications can be very straightforward or they can be quite complex depending upon the structure of the data relationships and the user requirements which are being addressed. In the simplest implementations, the hypertext link occurs as a single level index. A primary text document is linked directly to other related documents, but there are no further links between the related documents and other records. In more complex hypertext applications, the links between documents are extended to a more or less unlimited depth. In these complex hypertext applications, the user may move quite far down the hypertext link, traversing from record to record across multiple links.

E.3. APPROACHES TO THE UNITIZATION OF INFORMATION WITHIN PACKAGES

Ideally, a unitized package should be accessible at the lowest level of its component documents, and those documents in turn should be accessible, if possible, by individual pages. A user should be able to perform a search which would lead to a package of documents, which in turn would display the table of contents of the package. After examining the table of contents, the user should then be able to use it to select any component document within the package for display.

The problem, however, is how to adequately unitize the contents of packages for the LSS users and how to permit effective searching and retrieval of them. A single bibliographic header

for each package, standing alone, will not suffice because the LSS user will need to know more than the fact that a package of information exists. The user will also need to know fairly specifically what is in the package. This requirement coupled with the need for electronic searching of package contents suggests a need for greater unitization of package contents.

There is justifiable reluctance to approach the full unitization of the packages by decomposing them and preparing bibliographic headers for all of the individual component documents, primarily because of the cost of preparing the headers and the impact that such extreme unitization of the packages would have upon the loading schedule for the LSS.

Conceptually, "unitization" of packages includes two separate but related processes:

- The logical decomposition of a package of documentary materials into its component parts such that individual component materials may be identified and retrieved;

- The preparation of suitable headers such that individual component materials may be found by searching these headers.

The reluctance to "unitize" packages does not arise from the difficulty of logically decomposing the packages into their component documents and materials. That process could be accomplished as a byproduct of processing the package through the capture station. The significant cost and schedule impact of the "unitization" of packages is associated with the preparation of suitable headers for the component materials. What is needed, then, is an approach which will permit substantial unitization of the contents of packages without a substantial impact on the cost or schedule of implementation and loading of the LSS. In other words, an alternative needs to be found which will avoid the preparation of full headers for all of the component materials within a package while still permitting efficient and effective searching for those materials.

E.4. THE USE OF THE TABLE OF CONTENTS AS AN ALTERNATIVE TO HEADERS IN UNITIZED PACKAGES

Each package of documentary materials includes a table of contents which contains an entry for each "line-item" unit of information within the package. The table of contents will be entirely textual in format and can, therefore, be captured both as a bit-mapped image and as ASCII text. Each entry in the table of contents should contain an adequate description of the line-item to which it refers. At a minimum, this table of contents entry should include a short description of the item, focusing upon its general information content. In concept, the descriptive table of contents entry could include, in a free-form format, much of the information which would otherwise be found in a bibliographic header. It certainly should contain the information from "header" fields which would be most pertinent to a search for the item.

When such descriptive table of contents entries are captured and converted to ASCII text, they could then be searched individually or as a group by the full-text search capabilities of the LSS. In this way, the table of contents entries could act as effective substitutes for the bibliographic headers which otherwise would be required to unitize the packages. If the table of contents were used in this way, then the user could locate a package and its table of contents entry by simply performing a full-text search on the table of contents partition of the LSS.

E.5. THE TABLE OF CONTENTS AS AN EXAMPLE OF HYPERTEXT

A fairly straightforward application of hypertext may be illustrated by the example of the table of contents of a book. In a typical table of contents, one finds a list of chapters or topics which are arranged sequentially to represent the organization of the book. Associated with each entry in the table of contents, there is a page number which tells where in the book the beginning of the text for that entry may be found. So when one wants to find some information in a book, the typical approach is to open it to the table of contents, find the desired entry, and then turn to the appropriate page. In other words, a logical relationship is established between each entry in the table of contents and the corresponding text for a particular chapter or topic. The page number is the vehicle used for traversing that logical relationship.

In a hypertext implementation of a table of contents, the same logical relationships would exist, but the linkages between the table of contents and the chapters would be embedded in the table of contents itself. A hypertext table of contents would not need to show the page numbers, but rather would have them embedded in the individual entries. By "pointing to" or selecting the desired entry in the table of contents, the user would be able to select and display the desired text itself. In effect, the hypertext application would automatically "turn the pages" so that the first page of text for the selected table of contents line-item entry would appear immediately.

E.6. THE USE OF A HYPERTEXT TABLE OF CONTENTS IN LSS PACKAGES

This type of hypertext implementation of the table of contents could have very direct application to the problems associated with packages of documents in the LSS. All packages in the LSS will contain a table of contents which fully describes what is in the package. In its simplest form, a hypertext approach to the storage and retrieval of packages would automate this table of contents. The package as a whole would have a bibliographic header, and when the package was selected a hypertext table of contents would appear on the screen. The user could then "point to" and select any desired entry in the table of contents, and the system would automatically retrieve that component item from the package. This approach realizes many of the benefits of full unitization without incurring the cost and schedule impacts of preparing full bibliographic headers for each component document within the packages. In effect, hypertext links would be used as automated substitutes for the bibliographic headers of the component documents within a package.

E.7. ESTABLISHING THE HYPERTEXT LINKS BETWEEN THE TABLE OF CONTENTS AND THE PACKAGE CONTENTS

Unlike bibliographic headers, the hypertext links between the table of contents and the component materials within a package could be established as a byproduct of the document loading process. For example, because the table of contents is the first document captured for a package, it can be displayed on the capture station screen immediately and used to "guide" the capture process for the remainder of the package. When displayed on the capture station screen, the table of contents would tell the operator what was in the package and thereby suggest a normal sequence for document loading. As each component document is loaded, the operator would point to and select the corresponding entry from the table of contents displayed on the screen using a pointing device such as a mouse. This simple process would permit the capture station software to associate the component documents within a package with the package's table of contents and to automatically build the required hypertext links.

When the table of contents is scanned and converted to bit-mapped form and ASCII text, an accession number would be assigned to it and associated with the accession number of the package header. Similarly, as each component item within the package is scanned, it would receive its own accession number which would also be associated with the table of contents and the package when the operator pointed to the corresponding entry in the table of contents on the display screen of the capture station. In this way, the hypertext links for the table of contents could be established as a byproduct of the capture process itself.

E.8. AN EXAMPLE OF UNITIZATION USING THE TABLE OF CONTENTS AS A HYPERTEXT LINK

Figure E-1 represents a hypothetical package which contains a final report, three reviews, a field notebook, slip sheets for four magnetic tapes not included the package, three data printouts, a pumping test, surveys, injection tests, a map and a slip sheet for a document which cannot be scanned and which is not included in the package. Using this hypothetical package as an example, the capture process would produce a header for the package (for the table of contents). The table of contents would be automatically linked to the accession numbers of the individual component documents during the capture process in such a way that the individual images of the component documents could be selected directly from the display of the table of contents.

Using either a bibliographic header or textual keyword search, or possibly even a combination of the two search methods, the user would select the package and the table of contents would be displayed. The user might then view the table of contents and decide to look at the images of the field notebook. Once the user has selected the field notebook entry from the table of contents, the first page of the notebook will be displayed. The user then should be able to page sequentially (forward and backward) through the images of the notebook pages or, alternatively, select any desired page by pressing a "go to" key and specifying the number of the desired page.

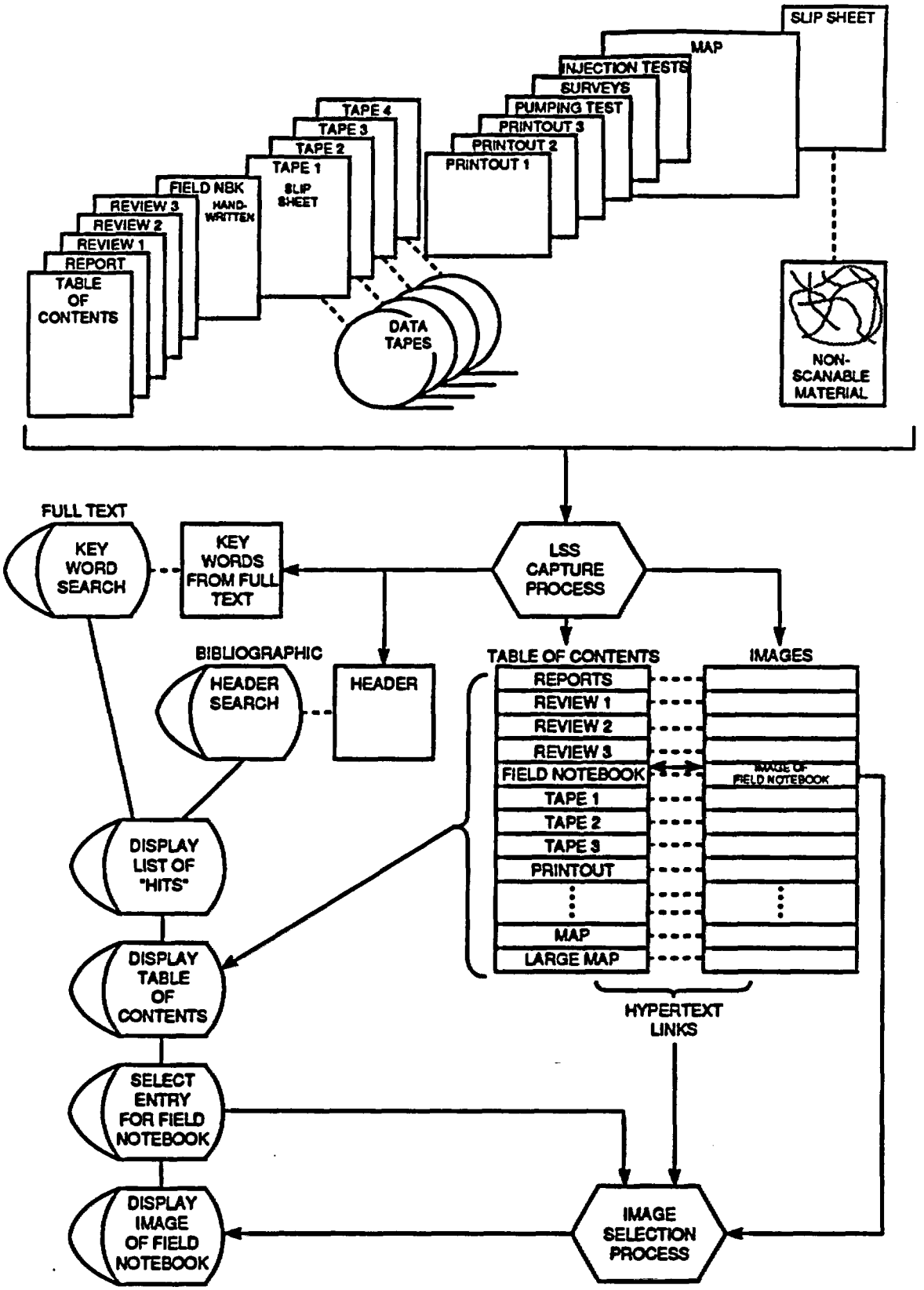


Figure E-1. Using the Table of Contents as a Hypertext Link

In this example, the magnetic tapes and the materials which cannot be scanned do not physically reside in the package, and therefore have no images of their own. They are represented by slip sheets and are physically stored in a special archive. Under the proposed scheme for preparing headers, these materials would each have its own header. Therefore, the headers for these materials could actually be accessed in two ways:

- A direct header search could be performed;
- A header search for the package as a whole could be performed and the headers for the materials could be selected from the table of contents using a hypertext link.

This example illustrates the effect of unitizing the information in the database at the level of the component document within a package and internally indexing it to the level of the page frame within the component document. This level of indexing would permit the system to respond to a user's request by selecting and transmitting only the single page to be displayed at any given instant in time. The result would be much faster response from the system and significantly reduced communication and processing costs.

E.9. ALTERNATIVE HYPERTEXT LINKS

It is important to note at this point that, just as a book may have both a table of contents and an index, it is also possible and practical to establish multiple hypertext links to a single document. Just as one might find a particular page of a book either through the table of contents or through the index, so could a hypertext implementation retrieve information either from a hypertext table of contents or from a hypertext index. The index of a book, or for that matter a hypertext index, differs from the table of contents only in that it is organized by subject rather than sequentially by chapter or item.

The document capture process for the LSS could actually create multiple hypertext links to each document. As we have seen, the table of contents could be processed in such a way as to automatically produce suitable hypertext links between the table of contents and the component documents within a package. Additional hypertext indexes oriented toward individual subjects or words could be created during the capture process as well. The documents in a package will be captured first as bit-mapped images and then those bit-mapped images will be processed digitally to perform the OCR conversion to ASCII text. There are some significant advantages to this approach. The accession number of the package and each of its component documents will have been assigned at the time of capture. The hypertext links between the table of contents and the package's bibliographic header and the bit-mapped images could also be established as part of the capture process. Then the bit-mapped images will be processed digitally to obtain the ASCII text and a hypertext link could be established automatically between the table of contents and that ASCII text as well. When the ASCII text is subsequently processed to create a contextual search index, the hypertext links could be extended to link individual words to the ASCII text, the bit-mapped image and the table of contents. In this way, the

hypertext links for a subject index could also be established as a byproduct of the capture process itself.

E.10. ALTERNATE ACCESS PATHS TO A DOCUMENT

When utilizing the full-text search capabilities, there are several opportunities for improving the functionality of the system from a user perspective which are not fully available when searching by bibliographic headers. These opportunities for increased functionality, however, are entirely dependent upon appropriate design and implementation of the index structure associated with the full-text search facility.

Figure E-2 illustrates a possible index and access structure which supports both the full-text and bibliographic headers. This illustration is based upon an index structure which contains the package identifier, document identifier, image page frame and text location for all entries. This index structure could be implemented in a number of ways, but the effect would be a unique entry for each page frame of each document and a pointer with an appropriate displacement for each keyword. Clearly, this would result in a very large index structure for the LSS, but it would permit access to information at the level of the package, document, page frame, or even key word.

E.10.1 Bibliographic Header Search

Using the example illustrated in Figure E-2, a bibliographic header search would involve the following sequence of operations:

A The user would enter a search argument using one or more header fields.

B The header search process would be invoked.

C The index structure for the header would be traversed, selecting entries which matched the search argument.

D As a result of traversing the index structure the index search facility would return the package identifier and/or the document identifier for every "hit".

E The system uses the package identifier and/or the document identifier to retrieve the document header and/or table of contents.

F If the table of contents is available, a hypertext link may be selected by the user to traverse to the desired document image or text.

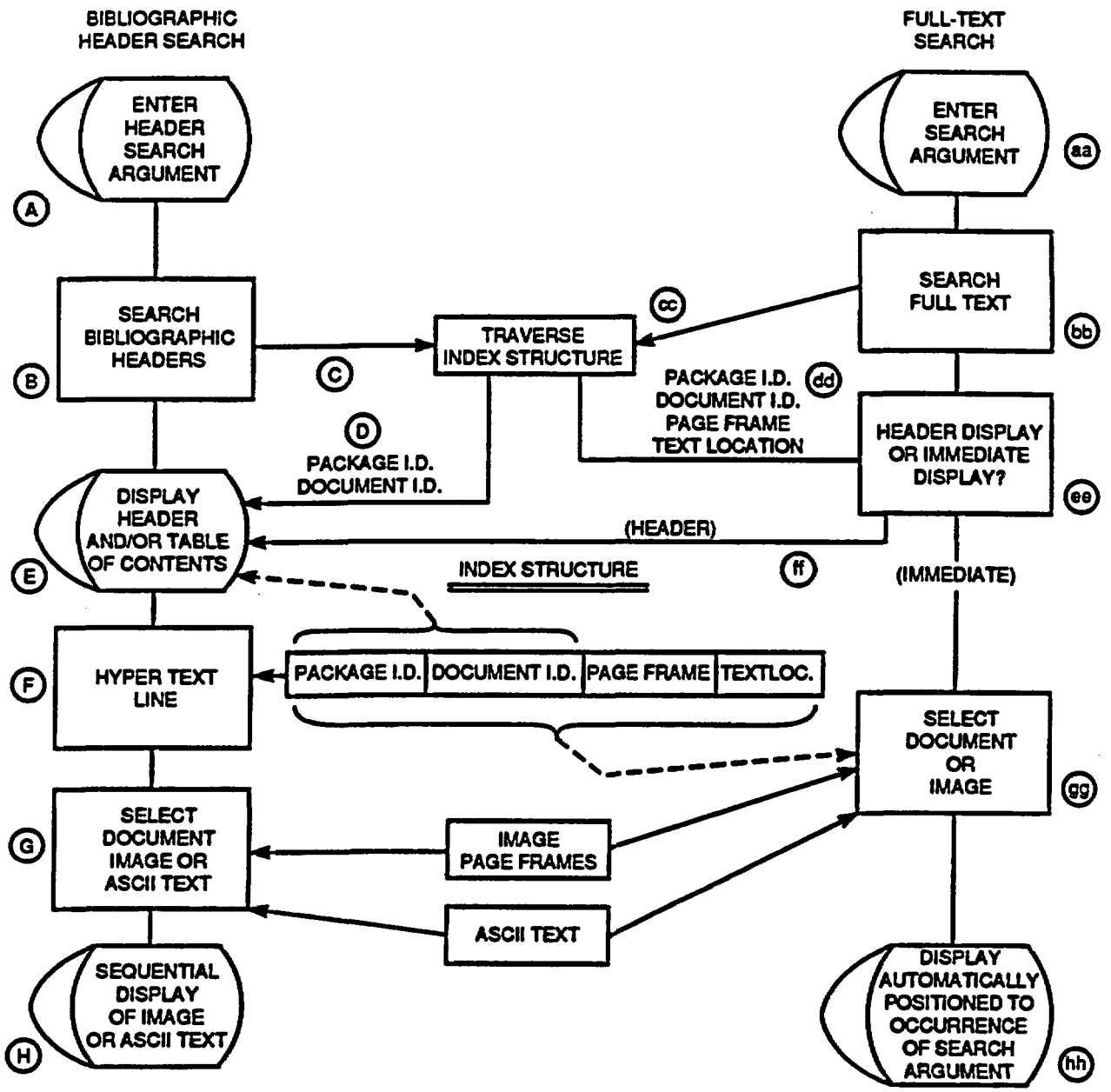


Figure E-2. Alternate Access Paths to a Document

G The system uses the resulting beginning document identifier and page frame or text location to retrieve the information to be displayed.

H The display is positioned to the first page frame of the image or the first page of the ASCII text. Then the user may sequentially page forward and backward through the display.

E.10.2 Full-Text Search

Using the example illustrated in Figure E-2, a full text search would involve the following sequence of operations:

aa The user would enter a search argument using one or more key words or phrases.

bb The full-text search process would be invoked.

cc The index structure for the full text search would be traversed, selecting entries which matched the search argument.

dd As a result of traversing the index structure, the index search facility would return the package identifier (if present), the document identifier, the page frame and the text location for every "hit".

ee At this point, the user may elect to use the information from the full-text index to perform either a bibliographic header display or an immediate display of the document.

ff If a bibliographic header display is selected, then the system passes only the package identifier and the document identifier to the bibliographic display procedure which continues with step E as indicated above.

gg In an immediate display of the document is selected, the system uses the document identifier, page frame and text location to retrieve the document image and/or text.

hh The display is positioned to the first occurrence of the search argument in the page frame or ASCII text as appropriate. Then the user may sequentially page forward and backward through the display.

E.11. THE APPLICATION OF HYPERTEXT LINKS TO DOCUMENTARY MATERIALS OUTSIDE OF PACKAGES

Although it is beyond the scope of this paper, the further applications of hypertext to non-packaged LSS materials deserves consideration. The current design documents for the LSS suggest an implementation which relies upon conventional menus and selection lists to identify and retrieve documents. But, just as the application of hypertext links would greatly improve the functionality and usability of table-of-contents entries for packages, hypertext links could profitably be associated with the bibliographic headers for all documentary materials in the LSS.

APPENDIX F

OTHER PACKAGING ISSUES

F.1. THE RETROACTIVITY ISSUE

The YMPO has already indexed into its records system several hundred packages that contain large volumes of non-text-searchable documentary material. This raises a question: Would the procedural changes recommended above have to be applied to those processed packages for the benefit of the LSS?

The authors would argue in the affirmative. Those packages' tables of contents should be rewritten. Otherwise, their importance would, in essence, be devalued. The necessary work could be accomplished centrally by the YMPO's Central Records Facility sometime before the material is loaded into the LSS.

F.2. THE PACKAGE-DELAY ISSUE

Packages often take several months to complete, depending upon the duration of the investigation they document. How, then, will the gradually-packaged materials be made available to requestors through the LSS before an investigation is finally accomplished?

The YMPO has attempted to rectify this problem by requiring quarterly submissions of packages in "segments," but has found that procedure to be troublesome, requiring so many exceptions (specially-approved submission schedules) to be granted that they have practically become routine.

Firm submission deadlines will be essential to achieve compliance with the LSS Rule, particularly during the three years of repository licensing review.

The authors are convinced that if an investigation by any submitting organization is expected to take several months to complete, it will be possible to divide it into logical phases that will permit timely submission of packaged material to the LSS. The segments representing the phases would be inter-related as they are submitted, and the final segment would contain the finished product, if any, and the commentary upon it.

An additional way of attacking the problem with respect to widespread YMPO investigations would be to make it clear to LSS requestors that, for the most recent material, they should place reliance upon the YMPO's quarterly Data Catalog, which will be input every three months into the LSS in text-searchable form.

F.3. THE TEXT-CONVERSION ISSUE

As recommended above, when packages contain technical reports and their commentary, those materials should always be made available to the LSS in text-searchable form, just as they would be if they were disassociated from a package. Otherwise, one method of securing their retrieval would be lost to requestors. The other materials commonly collected and physically

retained within packages are generally non-text-searchable because they are mostly graphic or handwritten. However, some of these other materials may be text-searchable in part.

Addressing itself to "graphic-oriented" documents, the LSS Rule says, in Section 2.1003(c)(1) that "Text embedded within these documents need not be separately entered in searchable full text."

There are two fundamental approaches which might be taken:

- The first would be to attempt to read the text of all packaged materials that are scanned by a capture station. That course of action, if performed conscientiously, would be time consuming because many interruptions to the process would surely occur, requiring manual intervention to resolve interpretive ambiguities.

- The second approach would be to interpret the LSS Rule in a broad sense - to hold that any readable text that may emerge within these bundles of graphic/handwritten "raw" or "backup" data should be considered embedded text that need not be separately entered into the LSS.

Granting that these packaged materials will be sufficiently identified for retrieval purposes by the table of contents, the authors are inclined to recommend the former approach, on the condition that no manual intervention is applied. The materials should simply be text-read as successfully as possible through the automated process.

F.4. THE "SCANNABILITY" OR "IMAGEABILITY" ISSUE

The LSS Rule requires the submission of images of all "graphic-oriented" documentary material. It may be considered that some of it will be too large or color-dependent for useful scanning, given the capabilities of the conventionally-sized, black-and-white, image-display equipment that is sometimes contemplated for LSS use.

LSS design documents have suggested that extra-large items (which are expected to include maps for the most part but may include other items, such as engineering designs, will need to be divided into 8 1/2 by 11 inch parts for "tiled" scanning, so that details on them will not be too minute to discern. An item with dimensions greater than 17 by 22 inches would, in this scheme of things, be considered too large for scanning because the maximum number of inter-dependent computer images that can be conceptually related was considered by LSS designers to be only four.

Currently, the YMPO is filming its "oversize" graphics on 35mm aperture cards in order to achieve the best possible resolution. It is placing the originals in storage, along with color-coded maps and overlays that are not deemed suitable for black-and-white microfilming. In addition to noting the existence of these kinds of materials in tables of contents, since most

of them belong to packages, the YMPO is creating a separate RIS bibliographic header for each of them, just as it is creating separate RIS headers for its machine-dependent items.

The variable here is the emergence of vastly more capable graphic scanning and display technology that may be adopted by the LSS by the time the system is implemented. Even today, extra-large graphic images are customarily handled not by tiling, but by focusing a "zoom lens, in effect, on interesting parts of the graphic. Color graphics, of course, are commonplace.

Given the rapidity with which imaging technology is being developed for everyday use, the authors are not inclined to concede that it will be ignored by the LSS and therefore require the system to use only size-limited, black-and- white, tiled images. That is why the authors have defined "non-scannable" material as machine-dependent, excluding from the definition other material which the YMPO refrains from microfilming.

Actually, there is practically nothing that cannot be scanned to create computer images. The various types of film media (radiographs, mylar, helicorder, etc.) can all be printed and scanned. Video tape can be viewed on computer screens. Movie film can be converted to videotape. Coded numeric data, useful only in the context of a computer program (making it machine-dependent) could be printed and scanned. Scannable photographs can be taken of physical objects. And so on. It is a question of the usefulness of the appearance of such images on LSS screens and the cost-effectiveness of placing them there.

The authors are not prepared at this time to suggest the best ways to incorporate the above-mentioned material into the LSS. For the moment, it is sufficient to emphasize the definitional problems that are inherent in the term "scannability."

Anything not scanned will, of course, have to be stored for retrieval upon request. As recommended, all such items should be assigned bibliographic headers, packaged or not, in order to permit their physical retrieval.

F.5. THE DUPLICATE-SCANNING ISSUE

Is it possible that the items that are scanned within a package might be scanned again in other contexts, apart from their packages or within other packages that relate to subsequent investigations drawing upon the same items? The answer is yes, it is possible.

With regard to finished products and machine-dependent items, to which multiple references are most likely to be made, this will not present a problem if this report's recommendations are followed. Duplication of these materials will be avoided by means of automated duplicate-checking procedures that will occur when packages and their headers are submitted to the LSS capture process. But what about the commentary and graphic/handwritten materials, which cannot be automatically checked for possible duplication since they will not receive headers of their own, permitting the duplicate-check mechanism to operate?

The authors would argue that report commentary does not present a problem because any multiple existence is likely to be negligible from a cost standpoint, given the relatively minor number of pages involved, and should cause no harm from the point of view of retrieval.

The inclusion of the same graphic/handwritten material within more than one package (seismic data, for example, that is found to be useful in more than one investigation) is also apt to be infrequent and inconsequential to retrieval. If a requestor desires to inspect graphic/ handwritten material, that requestor will presumably want to observe it within its respective investigative context.

F.6. THE REPETITIVE SUBJECT-SEARCH ISSUE

It has been asserted that LSS requestors are likely to conduct their searches primarily on the basis of subject matter (e.g., "volcanism"), whether they are focusing on bibliographic headers or on full document text. If the LSS is not carefully designed, it is possible that a requestor could be inconvenienced by having to look in several different places for the same subject term.

If a requestor is seriously interested in the subject of volcanism, for example, he or she must resolve to look in both headers and text for it - actually in three places, given the recommendation that the free-text of tables of contents be separately stored. Within headers, the requestor might have to search the title, descriptor, and abstract fields separately, one at a time, for occurrences of the desired term, since it might appear in any one of them within a particular header, but not necessarily in all of them.

Obviously, the LSS should be designed to eliminate duplicate references routinely from the search "hit list." The authors would urge that the LSS also be designed to:

- Allow a single search against multiple partitions of the textual data base when desired;
- Search multiple fields at once, when desired; and
- Search headers and full text at once, when desired.

APPENDIX G

DESCRIPTION OF TECHNICAL DATA USING BIBLIOGRAPHIC HEADERS

Successful, long-term use of headers to describe the technical data content of a complex, evolving records archive will require a generic and extensible approach. The approach should be as generic as possible to allow accommodation of various record types, and as extensible as possible to allow modification of the header for unanticipated record types. Technical subject will be the most important indexing parameter to most users, so a comprehensive list of subjects will be very helpful. The current proposed LSS Thesaurus is a reasonably good start, but it needs to be more comprehensive.

G.1. HEADER FIELDS REQUIRED FOR ACCESS TO TECHNICAL DATA

The header fields considered necessary for access to technical data referenced in the LSS are as follows:

- TITLE/DESCRIPTION: In general, the titles of technical reports are quite descriptive with respect to technical subject, and will be the primary field of interest.

- DESCRIPTORS: Technical subject descriptors can be very useful if care is taken to carefully extract key words from the abstract and body of technical reports and data record packages. Ideally, key words should be assigned by the author. Absent this, key words should be directly extracted from the record, rather than subjectively assigned based on an after-the-fact perception of the subject of the record.

- AUTHOR: Authors are commonly associated with certain areas of technical expertise. Searches for technical data in specific subject areas can be anticipated.

- COMMENTS: This field should be used to give details not critical to retrieval, but of potential interest to users.

- MEDIA: A full description of the physical recording media will assist the user in determining how the data should be used and what type of device may be required to read the data.

Following is a preliminary, but representative, list of field codes for MEDIA. This list illustrates the detail required for the MEDIA field.

FIELD CODES: Records produced with computers:

magnetic disk:	5.25" high-density
	3.5" low-density

magnetic tape (reel):	7-track (800bpi, 1600bpi) 8-track (800bpi, 1600bpi) 9-track (1600bpi, 6250bpi)
magnetic tape (cartridge):	.25" 8mm (capacity in Gb)
CD-ROM (compact disk-read only memory):	ISO-9600 High Sierra
WORM (write-once-read-many):	
optical disk (rewritable):	ISO (International Std. Org.) Sony Format

The following information should be placed in the MEDIA field of the header for all records created using a computer:

1.TYPE AND MODEL OF COMPUTER:	CRAY CDC-CYBER DEC VAX DEC20 DEC PDP IBM:360, and, etc... SUN: 386i 3/* 4/* SparcStation-* Other Workstations etc.
2. OPERATING SYSTEM:	UNIX:System V Release.* Berkeley Standard Dist. 4.* ULTRIX (DEC) AIX (IBM) AUX (APPLE) XENIX/SCO (Santa Cruz Op.) DEC VMS LTSS (Livermore Timesharing System) CTSS (Cray Timesharing System) CDC-NOS & NOS-B & NOS-BE IBM: CICS, MVS, VM370, ... etc.

3. FILE FORMAT:

UNIX TAR
UNIX BRU
UNIX CPIO
UNIX UUENCODE/DECODE
UNIX LIMPLE-ZEV COMPRESSED
VMS BACKUP
DOS
APPLE/MACINTOSH
TEXT:ASCII
EBCIDIC
GRAPHICS:DXF
TIFF
RASTER, etc.

4. APPLICATION:

WORD PROCESSOR:

WordPerfect V*.*
DisplayWrite
ProfessionalWrite
and others

GRAPHICS/DISPLAY/DESIGN:

AutoCad
Lotus Freelance
Designer
and others

MAPPING:

ISM
RADIAN CPS
ZYCOR
ARC/INFO
and others

Other machine readable records:

video cassette tape:VHS
beta
microfilm
microfiche
aperture cards
35mm photographic slide
photographic negative:[dimension (mm)]
movie film strip:[dimension (mm)]

develocorder
radiograph

- DOCUMENT TYPE: The DOCUMENT TYPE field should indicate the general purpose or area of technical/administrative work that the record is associated with. Information on the physical form should usually go in the MEDIA field.

Following is a preliminary list of DOCUMENT TYPE field codes.

- FIELD CODES:
  - correspondence
  - reports
    - technical
    - status
    - presentations
  - procurement
  - governing documents
  - data catalog
  - data record package
  - legal
  - computer software
  - software documentation
  - map
  - design drawing
  - seismic reflection record section
  - borehole geophysical log
  - survey
  - instrument readout
  - field/laboratory notebook
  - photograph
  - instrument calibration
  - instrument recording
  - data listing / table

G.2. HEADERS FOR MACHINE READABLE RECORDS

Documentary material generated by large, complex scientific and engineering projects can appear to be highly variable, even chaotic, with respect to form, content and type. Indeed, there are many variables needed to *completely* specify any individual document. However, the variables (fields) required to sufficiently specify technical data on machine readable media are rather limited.

Most machine readable records are adequately indexed by TITLE/DESCRIPTION and MEDIA. If the record does not have a title assigned by the author or originator, the subject and type of data on the record should be described in the TITLE/DESCRIPTION field. For example, a magnetic tape with earthquake seismic records recorded from the Southern Great Basin Seismic Network geophone network could be described as: earthquake seismic data recorded by SGBSN, 56 stations (channels), events greater than M = 4.5, period 3/14/84 - 3/15/84. The source of this type of descriptive information is problematic. Sometimes it is annotated on the tape-header that is affixed to the reel, or it may be on the YMPO Technical Data Information Form. If it is not, the indexing technician must contact the originator for an adequate description. Usually, the data content of computer generated records can only be adequately described by the originator.

It is important to prominently note in the TITLE/DESCRIPTION field that the record is a *MACHINE READABLE RECORD* or is on *MACHINE READABLE MEDIA*, and to note what type of record it is. In some instances, this is done simply by noting what the record is, e.g., a 35mm slide of drilling rig at borehole site UE 25 A-1, Oct. 3 1985, or an 8 x 10" color photograph of core from interval 1525m-1526m in borehole UE 25 A-1, date, photographer, etc. In other cases the approach is to annotate: *MACHINE READABLE MEDIA*: 9-track tape with geophysical logs (density, gamma ray, caliper) from borehole UE 25 A-1, and then put all the media and machine detail in the MEDIA field.

Data and file format, and machine-specific information should be placed in the bibliographic header for this material. Potential users will need to determine the type of hardware and software systems required to read, copy or display the data recorded on the various machine readable media. This information should be placed in the MEDIA field of the header.

If a computer software application that uses a non-ASCII data format to write, or backup, data on magnetic storage media was used to record the data, then the name and vendor/developer of the system needs to be noted. A notation should be made in the COMMENTS field if the application is needed to read the data.

Examples:

The following examples indicate the type of information and level of detail for the key header fields.

Example 1. Machine Readable Records: Film

USGS Open File Report #83-669 - Southern Great Basin Seismological Data Report For 1981 and Preliminary Data Analysis. This report contains earthquake data for 1981 and associated analysis of the data (Exhibit I C). The Table of Contents of the data record package for this report lists two types of machine readable record. Line item number 1 under Raw Data Package (section C) is a set of 7 slip sheets for develocorder film.

Proposed Header:

TITLE/DESCRIPTION: *MACHINE READABLE MEDIA.* Develocorder Film Records. Earthquake seismic data from the Southern Great Basin Seismic Network. Box Numbers 35-43, 8/22/81 - 9/26/82. Part of USGS OFR 83-669: Southern Great Basin Seismological Data Report for 1981 and Preliminary Data Analysis.

DESCRIPTORS: EARTHQUAKES, SEISMOLOGY, DATA, DEVELOCORDER, FILM, YUCCA MOUNTAIN, NEVADA.

AUTHOR: Rogers, A.M., Harmsen, S.C, Carr, W.J., and Spence, W.

MEDIA: DEVELOCORDER FILM. 16mm film strips require a 'Develocorder Film Reader' for viewing.

DOCUMENT TYPE: instrument recording.

COMMENTS: This record cannot be duplicated in microfilm format. Each record is a non-sprocket-punched, 16mm film strip that covers a 24 hour time period. There are 57 films in each of the boxes.

Example 2. Machine Readable Record: Magnetic Tape

Line item number 5 is a set of 228 slip sheets for "Contents of SGB (Southern Great Basin) Local Earthquake Archive Tapes (computer prints).

Proposed Header:

TITLE/DESCRIPTION: *MACHINE READABLE RECORD*. TAR archive tapes with local Southern Great Basin seismological data. TAR tapes L001 through L021. Part of USGS OFR 83-669: Southern Great Basin Seismological Data Report For 1981 and Preliminary Data Analysis.

DESCRIPTORS: EARTHQUAKES, SEISMOLOGICAL, DATA, SOUTHERN GREAT BASIN, YUCCA MOUNTAIN, NEVADA.

AUTHOR: Rogers, A.M., Harmsen, S.C., Carr, W.J., Spence, W.

MEDIA: MAGNETIC TAPE REEL (MGR). 9-track tape @ 6250bpi. Recorded with DEC PDP 11-34. Analyzed with DEC PDP 11-70. Tapes written with DEC VAX/VMS Version V4.2. Format is TAR.

DOCUMENT TYPE: instrument recording

COMMENTS: File is 110 blocks sequential. The longest record is 25 bytes. Records are variable length with implied (CR) carriage control.

Example 3. Machine Readable Record: Video Tape

Line item #13 in the Raw Data Package section (section C) of USGS OFR 84-15: Televiewer Log and Stress Measurements in Core Hole USW G-1, Nevada Test Site, December 13-22, 1981 (Exhibit I H) lists 31 video tapes containing borehole televiewer log data.

Proposed Header:

TITLE/DESCRIPTION: *MACHINE READABLE RECORD.* Memorex video tapes of televiewer logs for borehole USW G-1. Part of USGS OFR 84-15: Televiewer Log and Stress Measurements in Core Hole USW G-1, Nevada Test Site, December 13-22, 1981.

DESCRIPTORS: TELEVIEWER, LOGS, STRESS, BOREHOLE USW G-1, YUCCA MOUNTAIN, VIDEO TAPES.

AUTHOR: Healy, J.H., Hickman, S.H., Zoback, M.D., Ellis, W.L.

MEDIA: VIDEO TAPE, Memorex.

DOCUMENT TYPE: instrument recording.

COMMENTS: 20 minutes per tape. 31 tapes.

Example 4. Geologic Map

In some cases it is useful to have descriptive information in addition to that available in the title of the record. This adjunct information is not considered to be critical for search and retrieval of the record, but rather it is meant to enable users to prioritize their viewing of the image database. In the case of a geologic map, for instance, the scale, latitude-longitude and grid coordinates may be useful.

Proposed Header:

TITLE/DESCRIPTION: Geologic Map of the Topopah Spring SW Quadrangle, Nye County, Nevada. USGS Geologic Quadrangle Map, GQ-439 (1965), SCALE = 1:24,000 (7.5'quadrangle), LAT. = 36deg45min - 36deg52min30sec, LONG. = 116deg22min30sec - 116deg30min.

DESCRIPTORS: GEOLOGIC, MAP, TOPOPAH SPRING QUADRANGLE, YUCCA MOUNTAIN, NYE COUNTY, NEVADA.

AUTHOR: Lipman, Peter W., and McKay, Edward J.

MEDIA: paper

Example 5. Data Record Package

USGS OFR 84-149: Geohydrologic & Drill Hole Data for Test Well USW H-3 Yucca Mountain, Nye County, Nevada by W. Thordarson, F. Rush, R. Spengler, and S. Waddell.

Proposed Header:

TITLE/DESCRIPTION: USGS Open File Report 84-149: Geohydrologic & Drill Hole Data for Test Well USW H-3 Yucca Mountain, Nye County, NV

DESCRIPTORS: GEOHYDROLOGIC, DRILL HOLE DATA, USW H-3, YUCCA MOUNTAIN, HYDROLOGY, HYDRAULIC

AUTHOR: Thordarson, W., Rush, R., Spengler, R., and Wadell, S.

DOCUMENT TYPE: Data record package

APPENDIX H

SCENARIO FOR THE ENTRY OF PACKAGED MATERIALS INTO THE LSS

H.1. SUBMISSION PROCEDURES

All documentary material of the DOE, NRC, and other parties to the HLW licensing proceeding that may be relevant to the review process must be submitted for entry into the LSS so that it will be available to all interested parties on a timely basis. The LSS Rule tasks the LSS Administrator (LSSA) with the development of access protocols for the inclusion of material that is not suitable for entry into the LSS in searchable full text.

It is planned that two sets of protocols will be defined: submission protocols and retrieval protocols. The former will explain how non-text-readable material must be prepared for entry into the LSS by means of its capture system. The latter will explain how the material may be retrieved with the assistance of the LSS search-and-image system.

Neither of these sets of protocols have yet been written. However, the following scenario is intended to illustrate, briefly, how packaged materials would be entered into the LSS in accordance with the recommendations made above - in context with other materials. These procedures would form the basis for the submission protocols, binding upon all parties. Internal management plans and procedures of submitting organizations would have to be revised as necessary to accommodate them.

H.2. THE INVESTIGATOR

According to this scenario, an investigator working on behalf of an organization that is producing documentary material would take the following steps.

- Determine whether an item to be submitted to the organization's records center (and ultimately to the LSS) should be incorporated within a record package pertaining to a clearly identified activity. If so, it would be submitted with other collected materials according to an established schedule (yet to be determined). If not, it would be submitted as an independent item. Submissions would be made in accordance with the internal management procedures established by the responsible organization, including transmittal forms, duplicate copies, identifying numbers, etc.

- Provide a distinctive title that concisely describes each independent item or package (when completed) in order to facilitate future identification and timely retrieval. If a finished report is included within a package as the final product of an investigative activity, its title would be incorporated within the package title.

- Provide a descriptive abstract for each item or completed package which would adequately summarize what will be found when the item is retrieved.

- Submit a standard descriptive form with documentary material, packaged or not, which is likely to be unsuitable for conversion to images through digital scanning at the LSS capture station. Such material will include items which cannot usefully be scanned (to be determined) as well as magnetic, film, and other media which is machine-dependent. In the case of packages, the descriptive form would be included within the submitted materials and listed in the table of contents.

- Submit copies of duplicable items, such as magnetic tapes, if the investigator would like to retain them. These could be held until they are no longer needed and submitted at that time, provided that the descriptive forms forwarded in their stead state their temporary storage location.

H.3. THE RECORDS CENTER

Upon receipt of documentary material, the records center of the responsible organization would take the following steps.

- Process both packaged and unpackaged non-text- readable documentary material for LSS submission together with ordinary textual material, as materials of all types are likely to be intermixed when they arrive. Packages will remain intact, except for non-scannable items, which will be physically separated (but not disassociated) from their packages for the purposes of separate header creation and permanent storage.

- Review the materials for legibility, quality completeness, and authentication.

- Verify that the descriptive forms for non-scannable items adequately describe the materials received, that they contain required specifications, and that they satisfactorily identify the storage locations of non-scannable items which have been temporarily retained by investigators.

- Resolve all submission discrepancies with the responsible investigator, using internal procedures, within a specified timeframe, before proceeding.

- Compile packages from package segments. When all portions of a package have arrived and a package is therefore complete, the records center would type a table of contents, using a standard LSS format - listing all of the items, or groups of items, that it contains, by beginning page number, describing them sufficiently as required.

- Verify the investigator's designation of certain items as non-scannable and separate them from their packages (if packaged) for processing. It would complete a descriptive form for any item which the investigator deemed suitable

for LSS scanning but which it does not consider suitable. Conversely, if an investigator was incorrect in assuming that an item was unsuitable for scanning, it would destroy the form and keep the item itself physically within its package.

- Create an LSS bibliographic header for each cataloging unit -- one for each independent item, each record package, each published product associated with a package, and each non-scannable item (packaged or not), based upon its descriptive form. It would complete the header fields for each unit, as applicable, following the guidelines in the approved LSS cataloging manual. The information in the descriptive forms that were submitted with the non-scannable items would be incorporated into the header fields - providing both specifications and storage locations for the items.

- Transmit all scannable items and packages to the designated LSS capture station, according to schedule. It would also transmit the bibliographic headers for these materials and for the non-scannable materials that will be stored for expeditious retrieval by LSS requestors.

H.4. THE CAPTURE STATION

Upon receipt of the scannable items and the bibliographic headers from the records centers, each LSS capture station (planned to be operated in the Las Vegas, Nevada, area by the DOE, for the entry of its own materials,and by the LSSA, for the entry of materials submitted by other organizations) would do the following.

- Process each submitted unit in turn, without any need for separating packages from unpackaged items, non-text-readable material from textual material, or headers for submitted scannable material from headers for stored non-scannable material.

- Ensure that the material in each unit is legible and complete, containing all purported attachments, and that record packages conform strictly with their tables of contents. It would also verify that the quality of a unit submitted as scannable will indeed enable it to be captured successfully in electronic image form.

- Resolve any discrepancies with the responsible records center, within a specified timeframe, before proceeding with the processing of a unit.

- Prepare units for electronic scanning, dividing any submitted unit into smaller units, as essential, for the purpose of assigning a separate bibliographic header to each if it is perceived that such action will help LSS users retrieve desired information more conveniently and consistently, and if the separated units are not

integral parts of their original unit and can be understood out of context of that original unit.

- Use the capture station terminal to assign an LSS accession number to each unit, retrieve the header corresponding to the submitter accession number, verify the information it contains, ensure that all required fields have been properly completed, check for potential duplication of units, and, with the help of the LSS Search and Image System, follow rejection procedures to ensure that units already processed into the LSS will not be entered again unacceptably.

- Scan the units and convert the bit-mapped electronic images to ASCII text. This will be done using multiple OCR devices to assist accuracy. Non-text-readable material, by definition, will not be converted by the interpretive process. However, it will inevitably be mixed with readable material, within packages and without. The capture station operator would seek to render as much of a unit text-searchable as possible (given guidelines yet to be established). The text of package tables of contents must not only always be read successfully; it must be converted to a hypertext index.

- Finish its bibliographic header, including the fields designated for completion by the LSSA. Headers for the non-scannable units will also be completed.

- Perform a final quality-control check, making modifications as necessary to ensure that all units have been properly entered into the LSS.

- Forward (in the case of the DOE capture station) the resulting bibliographic headers, images and ASCII text to the LSSA capture station location in Las Vegas, which will transmit all of the accumulated headers, images and text to the LSS Search and Image System.