Minutes of LSSARP Meeting

March 20-21, 1990

The second meeting of the Licensing Support System Advisory Review Panel (LSSARP or Panel) was held in open session in Bethesda, Maryland, on March 20, 1990, with a site visit and tour of the U.S. Patent and Trademark Office in Crystal City, Virginia, on March 21, 1990. Enclosure 1 is a copy of the meeting agenda. Enclosure 2 is a list of attendees.

ADMINISTRATIVE ISSUES

Mr. Hoyle began a discussion of administrative issues with the minutes of the December 19, 1989, meeting. Mr. Treby of the NRC's Office of the General Counsel asked that a change be made on page 2. In the third paragraph under the role of the LSSARP, the word "consensus" should be changed to "majority." Mr. Silberg, an attorney representing the utility group, said that consensus means the absence of an objection. Mr. Cameron of the LSSA staff asked for clarification of the definition for "consensus." Mr. Hoyle stated that "consensus means no dissent among us." The Panel would also provide advice on the basis of majority views, that is at least five of the seven members, with dissenting views attached. Mr. Hoyle noted that coalitions (several persons representing a group with only one vote) must agree among themselves and will have one vote. With this change, the minutes were approved and will be placed in the NRC's Public Document Room (PDR). Mr. Hoyle proposed that in the future he place a draft of the minutes in the PDR soon after each meeting and replace it with the final version when it is approved by the Panel. There was no objection to this proposal.

Since the LSSARP is a federal advisory committee, all meetings will be held in open session and minutes of each meeting will be placed in the PDR. Mr. Hoyle will send a letter report to the LSS Administrator after each meeting and will provide a copy to each Panel member.

Next Mr. Hoyle asked Mr. Lloyd Donnelly, the LSS Administrator (LSSA), about the report which must be sent to the Commission in June evaluating DOE's compliance with the LSS rule. Mr. Donnelly noted that this date was selected prior to the delay in the repository licensing schedule from 1995 to 2001. He is currently discussing the matter with DOE and will be notifying the Commission about future LSSA planning for compliance evaluation.

Mr. Hoyle suggested that transcripts be produced for each meeting. Some members felt that transcripts were unnecessary; others wanted them. After discussion, it was agreed that, on a trial basis, there will be a transcript for the next meeting.

STATUS OF LSS DEVELOPMENT

Next was a presentation on the status of LSS development by Dan Graser of DOE who is the project manager for the current LSS design effort and the contracting officer's technical representative (COTR) for the current LSS design contract with Science Applications International Corporation (SAIC). He will also be the COTR for any subsequent LSS procurements. His handout is included as Enclosure 3. The current design contract with SAIC, which is in the process of being modified to bring it to an orderly conclusion, will be used to develop the detailed design procurement specifications. The final contract deliverables, covering all functional areas of the system, are due in July, August, and September 1990 with a two- to three-month DOE review period before final acceptance. Because DOE has a very limited FY 1990/1991 budget, it intends to rely on technical support from the LSSA staff to assist in preparation of solicitation packages for the LSS procurement. In addition, DOE will attempt to augment its professional staff on LSS design by having DOD personnel detailed to DOE for an extended period of time.

DOE is endeavoring to obtain acquisition support assistance from the Federal Systems Integration & Management Center (FEDSIM), a GSA organization established to assist agencies in efficiently and effectively using their own information resources. FEDSIM has about 20 pre-approved contractors who are invited to bid on projects. This could save months in getting contracts in place. DOE may use FEDSIM to arrange a support contract for (1) development of specifications, (2) live testing and acceptance criteria, (3) support as tests are conducted, and (4) assistance in compiling results of tests. If DOE uses FEDSIM, and gets support from the LSSA staff, a request for proposal (RFP) could be issued in August 1991 with a contract for the LSS awarded in early-April 1992. Once the contract is in place, the vendor would begin working towards installation of the first node of the LSS in December 1992. A node has the full functional capability of the LSS for capture plus search and retrieval of both full text and images. By the next Panel meeting, DOE plans to have a final strategy and a schedule, and be able to present options that can be built in to create as stable a schedule as possible.

Ms. Barbara Cerny of DOE's Office of Civilian Radioactive Waste Management added that one of the ways to remain flexible is to go into a GSA program called "trail boss." Under this concept, a person is designated the trail boss for a particular procurement. The trail boss has authority for the entire procurement even though only a part of it has been specified at the time.

Under the previous procurement strategy which was discussed at the December 1989 LSSARP meeting, DOE would not have used just a single source or done only a single procurement for the entire LSS. After all the specifications were developed, there would have been additional procurements for the computer platform, the software, etc. Now, due to the delay in the repository schedule, DOE is going to combine the functional requirements into one solicitation and make a single award. The contract will require establishing one node, ironing out all the bugs, and making sure it operates properly before proceeding. Additional nodes of the system will be procured as required. Ms. Cerny reminded the Panel that the new schedule assumes availability of funding.

When DOE envisioned using approximately six capture stations, there was an assumption that once the capture stations had been used to eliminate the document backlog, all but about three would be decommissioned and surplused. At a cost of about $2 million per capture station, approximately $10 million would be surplused. DOE was willing to absorb that cost as a necessary expenditure to permit timely elimination of the backlog. Under the present plan, hardware architecture will be adaptable to whatever is being done, i.e., once the backlog has been processed, the equipment will be used for something else. Therefore, there will be an estimated cost saving of approximately $10 million. When asked about cataloging costs, etc., Mr. Graser responded that he estimates that at least 70% of the cost will be labor. SAIC has been asked to reexamine the estimates used and the status of the backlog of documents, look at the new schedule, and determine whether or not the same number of documents will be generated by the time the license application is submitted. Currently only 12% of the documents being generated within DOE HQ are relevant to the LSS. Until a decision is made on the permit issue between DOE and the State of Nevada, there is not likely to be a near term increase in document volume over the level anticipated under the previous reporting program schedule.

HEADERS

The next discussion was led by Ms. Betsy Shelburne of the LSSA staff. A copy of the slides used in her presentation are included as Enclosure 4. Her handouts, a letter from B. Cerny dated January 31, 1990, a letter from D. Graser dated September 21, 1989, and a letter from F. X. Cameron dated August 7, 1989, are included as Enclosures 5, 6, and 7, respectively. Ms. Shelburne talked about LSS header needs and the elements of information that could be picked up. She requested that a working group be established to develop recommendations for required header elements.

After a discussion, Mr. Hoyle proposed establishing a header working group with membership from Nevada, NRC, DOE, and National Congress of American Indians (NCAI). Industry and adjacent counties indicated they have funding constraints at this time and cannot participate. Mr. Kirk Balcom, representing Nevada, was appointed Chairman. Mr. Hoyle will assign an NRC participant. Ms. Cerny stated that she would send an SAIC employee as DOE's representative because DOE had no one available from its staff. The Panel elected to have Ms. Betsy Shelburne be a member of the working group. Mr. Donnelly offered to provide office space and clerical support. The recommendations of the working group should be provided to Mr. Hoyle by mid-May for review and approval by the full Panel before being forwarded to the LSSA. Meetings of the working group are not covered under the Federal Advisory Committee Act since the group is a fact-finding committee. Therefore, working group meetings need not be announced in the Federal Register.

LSS DESIGN

Mr. Balcom said that he is also interested in the current DOE efforts to design the LSS and feels that the Panel should have input into that now rather than in August or September when the SAIC design documents are received by DOE. Ms. Cerny said that there are many documents that, though they are final deliverables, are in no way fixed in terms of the final RFP (request for proposal) for the LSS. She suggested that these documents could be reviewed by the Panel and Panel comments could be incorporated in the RFP. Mr. Balcom felt that would be too late; Nevada, particularly, would like to be involved in the design process earlier, perhaps in monthly meetings. He stated that the design issue is as important as the header issue. Mr. Hoyle asked Ms. Cerny if a working group could be provided information to review as it becomes available. Ms. Cerny responded that DOE periodically conducts major design review meetings and recommended that the Panel either take part in the design meetings or thoroughly review the documents as they come out. She stated that she would welcome input to the design process.

Mr. Hoyle noted that a major concern is whether the Panel input will be injected into the design process in time to affect the design. He pointed out that section 2.1011(f) of the rule specifies that the Panel shall provide advice to DOE on the fundamental issues of the design and development of the computer system. When the design documents are delivered, a working group should review them and make a recommendation to the Panel for submission to the LSSA. Mr. Graser noted that there is about a four-month period for the FEDSIM contractor to become familiar with the design documents; the Panel could also review the

documents during that time. Ms. Cerny said that FEDSIM review period will occur after the documents are delivered to DOE, and agreed there would be no problem with the Panel's reviewing the documents then.

Mr. Donnelly reminded the Panel that DOE has the responsibility to design and develop the system, and must be given the opportunity to do that without undue interference by either his office or the LSSARP. He noted that when the Panel reviews the design documents, perhaps some of their current concerns will be alleviated.

OTHER FEDERAL AGENCIES' INFORMATION MANAGEMENT SYSTEMS

There were presentations by three Federal agency representatives who discussed their experiences with design, procurement and operation of large automated information management systems. First was Mr. Boyd Alexander of the U.S. Patent and Trademark Office (PTO). A copy of the slides used in his presentation are given in Enclosure 8. Mr. Alexander was asked about the costs of the system. The PTO charges a database user fee of $40 per hour for text search. They break even with that fee. The lifecycle costs for the total system approach $500 million. This includes future development, equipment, contractor costs, etc. PTO is now attempting to make a policy decision regarding whether information should be available to the public with no user fee.

Mr. Bill Holmes, the Director of the Archival Research and Evaluation Staff at the National Archives and Records Administration, discussed the automated data system at the National Archives and showed some of the ways they enhance images made from very poor quality originals. As examples he used documents from the Civil War period.

Mr. David Copenhafer of the Securities and Exchange Commission spoke about the EDGAR (electronic data gathering, analysis, and retrieval) system. The EDGAR system is a pilot project that SEC has found to be a useful testing ground for different approaches. The system contains about 64,000 records or about one million pages. Due to the nature of their business, SEC does not intend to put old data into EDGAR. When asked about costs, Mr. Copenhafer noted that members of the public can get anything in their Public Reference Room at no charge. Total system lifecycle costs are about $100 million.

Following the Federal agencies' presentations, Mr. Hoyle asked Mr. Donnelly if his office planned to do a "lessons learned" study summarizing the experience of other agencies with large information management systems. Mr. Donnelly stated that this would be done.

REVISION OF TOPICAL GUIDELINES

Mr. Treby discussed the revision of topical guidelines. The topical guidelines published in the rule were intended to be interim guidelines pending issuance of a Regulatory Guide. The LSS Internal Steering Committee (LSSISC) has established a task force to propose final guidelines. The LSSISC expects to send recommendations to the Commission in May and to provide them to the Panel by early July. The LSSISC task force's recommendations will be mailed to Panel members for a discussion at the fall meeting.

LSSA/DOE MEMORANDUM OF UNDERSTANDING

Mr. Cameron, the Deputy LSS Administrator, gave a brief update on the Memorandum of Understanding (MOU) between DOE and LSSA. The MOU will set forth mutual responsibilities for design, development, and operation of the LSS. The MOU will list the major procurement activities, the deliverables involved, and the schedule. The schedule will be included in a management plan attached to the MOU. The LSSA sent a draft MOU to DOE. DOE is incorporating its comments for response to the LSSA in April. Negotiations will begin in mid-April with hopes of completion by June or July. The MOU must be reviewed and approved by the Commission.

DOCUMENT LOADING PRIORITIES

Mr. Cameron also discussed the prioritized loading of documents into the LSS. Under the present schedule, the first node of the LSS will be ready for operation in late December 1992. Between 500,000 and 750,000 pages per year will be processed (captured) by that first node. With recommendations from Panel members on priority loading categories, the LSSA staff will compile a document loading priority schedule for circulation to the Panel and discussion at the fall meeting. The first node could be used for loading of priority documents, with the second, third, etc., nodes used to work off the backlog of lower priority documents. The Commission will review the priority loading schedule and a "costs and needs" analysis before any documents are loaded into the LSS beyond those that are needed to fully test and evaluate the first node. Mr. Cameron reminded Panel members that their priority recommendations on the Prioritized Document Production Schedule-LSS Participant Worksheet (which was distributed and discussed at the December 1989 Panel meeting) are due to the LSSA by September 1, 1990.

FUTURE MEETINGS

Mr. Hoyle proposed that a short Panel meeting be held in early June to review the header working group's recommendations. It was agreed that the next meeting will be June 7, 1990, in the Washington, D.C., area. He will attempt to set up videoconferencing equipment for the meeting.

A proposed planning agenda for future meetings was distributed. See Enclosure 9. The fall Panel meeting will include discussions of SAIC products, priority loading categories, access to technical data, revision of topical guidelines, and the compliance evaluation program. The Panel members agreed that the fall meeting will be a two-day session on October 10 and 11, 1990, in Reno, Nevada.

Mr. Hoyle reminded Ms. Cerny that the Panel must have the design documents as soon as DOE receives them. Ms. Cerny agreed that as soon as she receives the documents, she will send them to Mr. Hoyle. He in turn will send them to the Panel members with a request for written comments. Panel members' comments will be distributed and a discussion meeting arranged, if necessary.

The morning of March 21, 1990, the Panel met at Mr. Boyd Alexander's office at the U.S. Patent and Trademark Office for a demonstration and tour of their automated document system. Enclosure 10 is a copy of the handout used in the PTO demonstration.

Enclosures:
1. Agenda
2. Attendance List
3. D. Graser Handout
4. B. Shelburne Slides
5. B. Cerny letter dtd 1/31/90
6. D. Graser letter dtd 9/21/89
7. F.X. Cameron letter dtd 8/7/89
8. B. Alexander Slides
9. Planning Agenda
10. PTO - Automation

AGENDA

LSS ADVISORY REVIEW PANEL MEETING

MARCH 20 - 21, 1990

Tuesday, March 20, 1990

9:00 Agenda Overview and Panel Administrative Issues (John Hoyle, LSSARP Chairman)

10:15 Break

10:30 Status of LSS Development (Barbara Cerny - DOE)

10:50 Headers (Betsy Shelburne - LSSA)

12:00 Lunch Break

1:15 Information Items (LSSA/DOE Memorandum of Understanding (MOU); Revision of Topical Guidelines; Priority Loading Categories)

2:00 Automated Information Management Systems: Experiences from Other Federal Agencies:

- 2:00 Patent and Trademark Office (Boyd Alexander)
- 2:30 National Archives and Records Administration (Bill Holmes, Director, Archival Research and Evaluation Staff)
- 3:00 Break
- 3:15 Securities and Exchange Commission - EDGAR (David Copenhafer, Deputy Director, Office of EDGAR Management)

4:30 Schedule and Agenda Planning

5:00 Adjourn

Wednesday, March 21, 1990

9:00 Site Visit to U.S. Patent and Trademark Office, Room 916, Crystal Park 2, 2121 Crystal Drive, Arlington, Virginia (Convenient to Crystal City stop on either Blue Line or Yellow Line of Metro)

Enclosure 2

Attendance List

LSS Advisory Review Panel Meeting, March 20-21, 1990

Panel Members

Nuclear Regulatory Commission

John C. Hoyle, Panel Chairman
Stuart A. Treby
Phillip Altomare

Department of Energy

Barbara Cerny
Dan Graser

State of Nevada

Kirk Balcom

Local Government - Site

Steve Bradhurst

Local Government - Adjacent

Dennis Bechtel
Liza Vibert
Peter Cummings

National Coalition of American Indians

Loretta V. Metoxen

Nuclear Industry

Jay Silberg
Felix Killar

U.S. Patent and Trademark Office (Non-Voting Member)

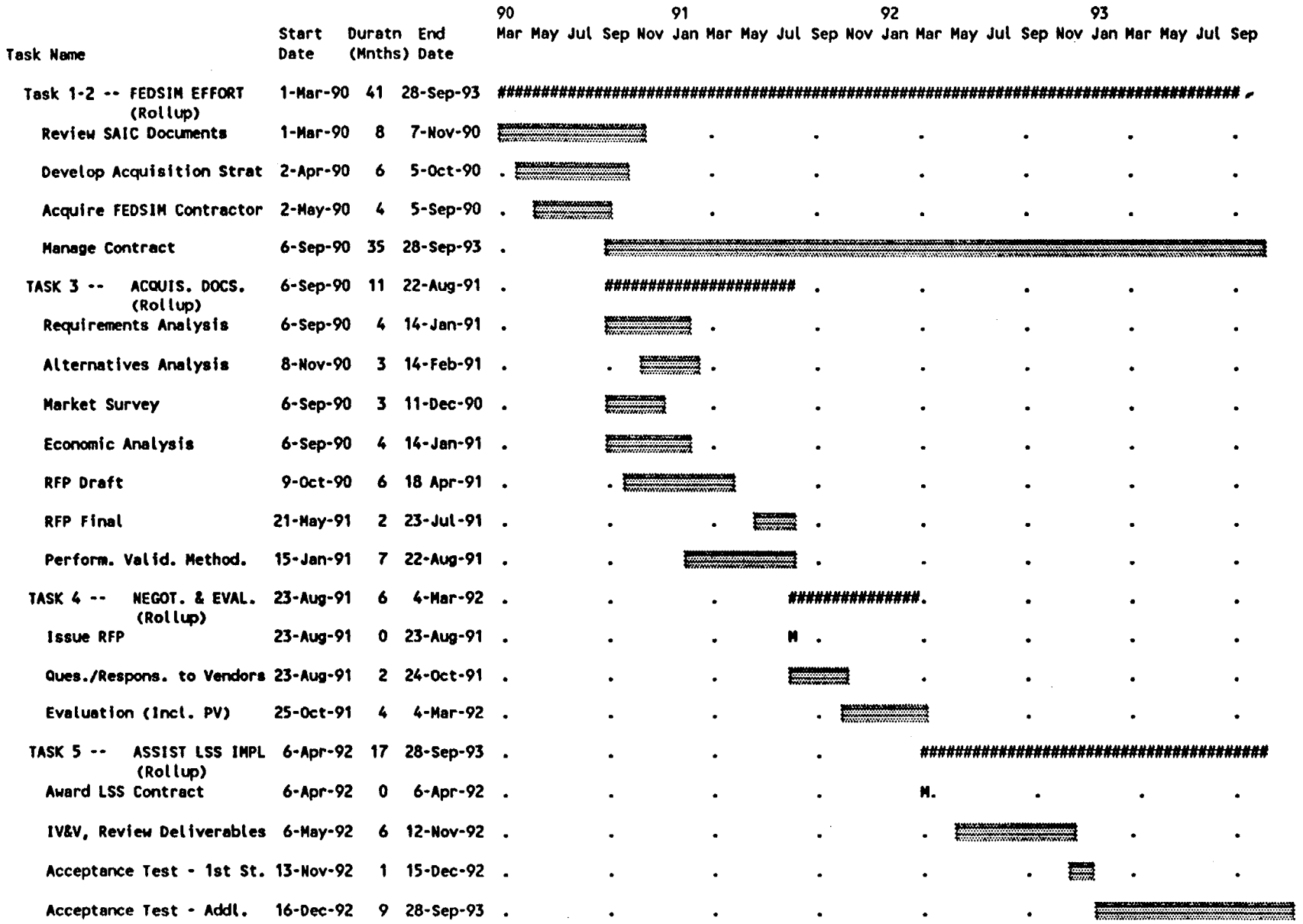
Boyd Alexander

Others

Lloyd Donnelly, NRC/LSSA
Chip Cameron,NRC/LSSA
Betsy Shelburne, NRC/LSSA
Avi Bender, NRC/LSSA
Lynn Scattolini, NRC/LSSA
Marilee Rood, NRC/LSSA
Rosetta Virgilio, NRC/GPA/SP
Jack Whetstine, NRC/ASLBP
John Frye, NRC/ASLBP
Kathryn Winsberg, NRC/OGC
Steve Scott, NRC/IRM
Susan Bilhorn, NRC/OCM/KR
Eileen Tana, NRC/NMSS
W. Richard Pierce, SAIC
Roger B. Bradford, SAIC
Stephen Spector, CNWRA
Robbie Cooke, Wang Labs, Inc
Jim Smith, Government Computer News

LSS DELIVERABLES
FROM THE REVISED SAIC CONTRACT

DESIGN EFFORT	
System Concept Feasibility Report	7/88
INFORMATION MANAGEMENT	
Archives Operations Procedures	3/89
LSS Thesaurus Maintenance Procedures	9/90
LSS Thesaurus (Draft)	3/90
Controlled Vocabulary	9/90
Controlled Vocabulary Maint. Procedures	9/90
LSS Prototype Cataloging Manual	1/89
PROTOTYPE	
Prototype Effort Analysis & Report	2/90
CAPTURE SYSTEM	
Capture System Design Document	3/89
Capture System (Stand alone) Specs	3/90
SEARCH SYSTEM	
Search System Design Documentation	9/90
Search System Text DBMS Functional Design	8/90
Search System H/W Configuration Designs	8/90
Search System Custom Applications S/W Design	8/90
Search System DBMS S/W Architecture Design	8/90
IMAGE SYSTEM	
Image System Design Documentation	8/90
Image System H/W Configuration Designs	7/90
Image System Custom Applications S/W Design	7/90
Image System DBMS S/W Architecture Design	7/90
WORKSTATIONS	
Workstation H/W Configuration Design	8/90
Workstation Applications S/W Architecture Design	8/90
COMMUNICATIONS	
Communications H/W Design	8/90
Communications Circuit Design	8/90


90
91
92
93
Mar May Jul Sep Nov Jan Mar May Jul Sep Nov Jan Mar May Jul Sep Nov Jan Mar May Jul Sep
Task Name
Start Date
Duratn (Mnths)
End Date
Task 1-2 -- FEDSIM EFFORT (Rollup) 1-Mar-90 41 28-Sep-93
Review SAIC Documents 1-Mar-90 8 7-Nov-90
Develop Acquisition Strat 2-Apr-90 6 5-Oct-90
Acquire FEDSIM Contractor 2-May-90 4 5-Sep-90
Manage Contract 6-Sep-90 35 28-Sep-93
TASK 3 -- ACQUIS. DOCS. (Rollup) 6-Sep-90 11 22-Aug-91
Requirements Analysis 6-Sep-90 4 14-Jan-91
Alternatives Analysis 8-Nov-90 3 14-Feb-91
Market Survey 6-Sep-90 3 11-Dec-90
Economic Analysis 6-Sep-90 4 14-Jan-91
RFP Draft 9-Oct-90 6 18 Apr-91
RFP Final 21-May-91 2 23-Jul-91
Perform. Valid. Method. 15-Jan-91 7 22-Aug-91
TASK 4 -- NEGOT. & EVAL. (Rollup) 23-Aug-91 6 4-Mar-92
Issue RFP 23-Aug-91 0 23-Aug-91
Ques./Respons. to Vendors 23-Aug-91 2 24-Oct-91
Evaluation (Incl. PV) 25-Oct-91 4 4-Mar-92
TASK 5 -- ASSIST LSS IMPL (Rollup) 6-Apr-92 17 28-Sep-93
Award LSS Contract 6-Apr-92 0 6-Apr-92
IV&V, Review Deliverables 6-May-92 6 12-Nov-92
Acceptance Test - 1st St. 13-Nov-92 1 15-Dec-92
Acceptance Test - Addl. 16-Dec-92 9 28-Sep-93

TASK AREAS TO UTILIZE NRC SUPPORT

PROCUREMENT/DEVELOPMENT

Plan:

- Develop SOW for FEDSIM Support Contractor.
- Obtain and Review SAIC's Design Deliverables.
- Support Contractor Turns Deliverables into "Turnkey Solicitation".

Requires:

- Participate in solicitation, evaluation, award of FEDSIM's support contractor (DOE/NRC must provide 2 staff for the FEDSIM solicitation process).
- Reviews of SAIC Deliverables done in conjunction with FEDSIM will ensure that they can be used for a "Turnkey" solicitation vehicle. (SAIC's deliverables and the "Turnkey" SOW drive LSS we will buy.)

INFORMATION MANAGEMENT

Plan:

- Establish screening criteria and plan for RIS current and backlog collection migrating to LSS.
- Revise and update OCRWM Indexing Manuals moving toward LSS submitter's profile.
- Survey, then develop plans, procedures, schedules for document & data backlog processing. (Includes information located with participants such as USGS.)
- RIS Keyword, Organization Codes, Thesaurus control with enhanced tools.

Requires:

- Relevancy criteria development.
- Prioritization criteria development.
- Header development for LSS submitter records within broader LSS Header Record.
- Thesaurus development, controlled vocabulary management, resolution and concordance of "document types".
- Operations procedure development consistent with QA requirements.

Enclosure 4

OFFICE OF THE
LSS ADMINISTRATOR

DEVELOPMENT OF LSSARP HEADER RECOMMENDATIONS

A BRIEFING BY THE OFFICE OF THE LSS ADMINISTRATOR IN SUPPORT OF THE LSSARP

LSS ADVISORY REVIEW PANEL MEETING
MARCH 20, 1990
BETHESDA, MARYLAND

OUTLINE OF THE PRESENTATION

- THE HEADER ISSUES
  - WHY HAVE A HEADER ?
  - WHAT CAPTURED ?
  - WHO CAPTURES ?
- NEXT STEPS

- BRIEFING HANDOUTS:
  - Cerny, DOE, to Hoyle, LSSARP, dated Jan. 31, 1990
  - Graser, DOE, to Cameron, LSSA, dated Sept. 21, 1989
  - Cameron, NRC, to Graser, DOE, dated Aug. 7, 1989

THE HEADER ISSUE:

WHAT INFORMATION ELEMENTS SHOULD BE CAPTURED BY WHOM ?

WHY START RESOLVING HEADER ISSUES NOW?

- PARTICIPANT PLANNING
  - → INPUT INTO SUBMITTER'S INTERNAL RECORDS MANAGEMENT SYSTEMS AND PROCEDURES
  - → BUDGETING (Staff & Dollars) AND CONTRACTING
  - → BEGIN PROCESSING DOCUMENTS FOR FIRST LSS NODE
- IDENTIFY AND RESOLVE ISSUES THAT IMPACT LSS DESIGN
- IDENTIFY AND RESOLVE ISSUES THAT IMPACT LSS OPERATIONS & MAINTENANCE
- THE MATRIX OF POSSIBLE WHO AND WHAT ALTERNATIVES HAVE SIGNIFICANT COST-BENEFIT RAMIFICATIONS
- THE PROCESS FOR REASONED CONSIDERATION OF THE ALTERNATIVES WILL TAKE TIME

- DECIDING ON HEADER NEEDS NOW MEANS DOING SO WITHOUT PERFECT INFORMATION, e.g.
  - DUPLICATE CHECKING ALGORITHM
  - SEARCH & RETRIEVAL MECHANISMS

- FINAL DESIGN NOT LIKELY TO IMPACT BIBLIOGRAPHIC HEADER, BUT COULD CHANGE ENHANCED HEADER REQUIREMENTS, e.g.
  - ABSTRACT -- HUMAN vs. SOFTWARE COMPOSED

DEVELOPMENTS TO DATE

- **PRELIMINARY LIST OF FIELDS FOR BIBLIOGRAPHIC HEADER (SUBMITTERS' HEADER) DEVELOPED BY TECHNICAL WORKGROUP OF THE ADVISORY COMMITTEE ON LSS RULE**
  - DATED MAY 1988
  - LISTED TWENTY FIVE REQUIRED FIELDS !! NOT SIMPLE !!

- **DOE/SAIC PROTOTYPE TEST CONDUCTED IN FALL, 1989**

- **DOE/SAIC PROTOTYPE TEST REPORT RELEASED IN FEBRUARY, 1990**
  - PROVIDES INPUT RELEVANT TO HEADER DESIGN AND CAPTURE PROCEDURES
    - → NOT ONLY THE WHAT?, BUT ISSUES RELATED TO HOW?

"IF A FULL-TEXT DATABASE, WHY DO WE NEED A HEADER ?"

- **IMPROVE USER'S SEARCH RESULTS -- RECALL and PRECISION**
  - PROVIDE ADDITIONAL ACCESS POINTS WHICH:
    - → MIGHT BE IN TEXT, BUT NOT IN CONSISTENT FORMAT (names, numbers, subject)
    - → MIGHT NOT BE IN FULL-TEXT (contract number, classification codes, project number)
  - IMPROVE RECALL GIVEN:
    - → VARIETY OF DOCUMENTS IN LSS COLLECTION
    - → UNSTRUCTURED TEXT
  - IMPROVE PRECISION - NARROW or EXPAND UNIVERSE PRIOR TO FULL-TEXT SEARCH
- **PROVIDE DESCRIPTIVE INFORMATION ABOUT DOCUMENTS**
  - FOR ON-LINE REVIEW OF SEARCH RESULTS
  - FOR PRINTED LISTINGS, BIBLIOGRAPHIES, ANNOUNCEMENTS OF NEW ENTRIES
- **IMPROVE SPEED OF CERTAIN QUERIES**

BREAKDOWN OF QUESTION: WHAT ELEMENTS ARE CAPTURED BY WHOM ?

WHAT?

MINIMUM BIBLIOGRAPHIC ELEMENTS -- Date, Author, Title, etc.

EXTENSIVE BIBLIOGRAPHIC ELEMENTS -- Contract and Report numbers, Project numbers, Witnesses, Sponsoring Organization, etc.

SUBJECT INDEXING

Descriptors -- Controlled Vocabulary (Thesaurus)

Identifiers -- Free Form Words and Phrases

ABSTRACTS

CLASSIFICATION NUMBERS OR CATEGORIES -- Based on Subject, Topical Guidelines, or DOE Mission Plan

WHO?

LSS PARTICIPANTS (SUBMITTERS)

OR

LSS ADMINISTRATOR'S CONTRACTOR

W H A T ?

QUESTION:

GIVEN ALL THE BENEFITS, WHY NOT DEVELOP THE MOST EXTENSIVE HEADER ?

ANSWER:

DIFFERENT LEVELS IN CODING HAVE DIFFERENT BENEFITS AND VERY DIFFERENT COSTS

HEADER ELEMENTS: SUMMARY BENEFITS AND COST FACTORS

TYPES OF HEADER ELEMENTS	EXAMPLES	BENEFITS					COST FACTORS		
		USES	RECALL		PRECISION		LABOR *		TRAINING, QC, and MAINT-ENANCE
			CONTENT	OTHER	CONTENT	OTHER	VOLUME PER HOUR	HOUR RATE, SALARY	
"BIBLIOGRAPHIC"									
▸ DESCRIPTIVE	Date, Author Pages, Title Report Number Condition	Structured Access & Presentation	low	high	low	high	high	average	low
▸ TAGS GROUPING LIKE DOCUMENTS ▸not always in text	Project No., Event Date, Contract No.	Structured Access & Scoping	average	high	average	high	average	average	average
▸ LINKINGS	References, Pointers	Structured Access	average	high	average	high	average	average	low
SUBJECT KEYTERMS: ▸ CONTROLLED	Thesaurus	Structured Access &Scoping	average	low	average	low	low	high	high
▸ UNCONTROLLED	Free Form	'Structured' Access & Thesaurus Update	average	low	average	low	average	average	low
ABSTRACTING	-Annotative, -Indicative, -Informative	Access & Presentation	- low - aver. - high	low	- average - average - high	low	average lowest lowest	average high high	average high high
CLASSIFICATION	Top.Guidelines,	Access, Scoping, Presentation	average	average	high	low	high	average	average

*COST PER DOCUMENT = $\frac{\text{STAFF HOURLY PAY RATE, INCLUDING OVERHEAD}}{\text{DOCUMENTS PROCESSED PER HOUR}}$

Backup for File Room Space Projections for TWFN.

THE AEROSPACE CORPORATION

Suite 4000, 955 L'Enfant Plaza, S.W., Washington, D.C. 20024-2174 Telephone: (202) 488-6000

6812-04.86.rlj.15 '86 MAR 25 A11:56
24 March 1986

WM-Rec
Wtl Docket File
A4167
AFSD

WM Project 10,11,16
Docket No.
PDR ✓
LPDR ✓ CB, NL

Distribution:
ABender PAltomare
(Return to WM, 623-SS) SL

Mr. Avi Bender - WMPC
Policy & Program Control Branch
Division of Waste Management
Office of NMSS
Mail Stop 623-SS
U.S. Nuclear Regulatory Commission
Washington, D.C. 20555

Dear Mr. Bender:

TRANSMITTAL OF REVISION 2
REQUIREMENTS DEFINITION FOR A LICENSING INFORMATION
MANAGEMENT SYSTEM FOR NUCLEAR WASTES

Reference: Draft Report Requirements Definition for an Information Management System for Nuclear Waste, Aerospace Corporation, 31 January 1986 (6812-04.86.rlj.05)

12

Enclosed are ten draft copies of the subject report incorporating the definition and rationale for the requirement of full text storage and retrieval of LIMS records. There will be a final version of this report in the late Spring following the Pilot Project Demonstration Tests. The final draft will refine the requirements determined during the demonstration program. So far, these include: (1) an update on the projected number of future records with an estimate on how many would be in the NRC system and in the DOE system, (2) a new section on applicable standards, (3) a new section on the functional requirements of document capture, and (4) any other relevant requirements that can be defined between your staff and ours.

Comments on this latest draft would be appreciated.

Very truly yours,

R. L. Johnson
Systems Director
Eastern Technical Division

RLJ:gbf
Enclosures

cc: P. Altomare - WMPC
G.E. Aichinger - SD/PMR (letter only)

Aerospace Report No.
WPR-85(5812-04)-1

DRAFT

Revision 2

Requirements Definition for a Licensing Information Management System for Nuclear Waste

Subtask 1. Task Order 002 of FIN 4167
Programmatic System Studies and Analyses

March 1986

13

Prepared for

Policy and Program Control Branch
Division of Waste Management
U.S. NUCLEAR REGULATORY COMMISSION
Washington, D.C.

Prepared by

Government Support Division
THE AEROSPACE CORPORATION
Washington, D.C.

Contract No. FO4701-83-C-0084

W H O ?

QUESTION: WHO CAPTURES WHAT ELEMENTS OF THE HEADER?

ANSWER: MUST GO BACK TO UNDERSTANDING AS REFLECTED IN THE LSS RULE

- PARTICIPANTS WILL SUBMIT A MINIMUM SERIES OF DESCRIPTIVE FIELDS

  (PARAPHRASE OF DEFINITION OF "BIBLIOGRAPHIC" HEADER IN THE LSS RULE)

- LSS ADMINISTRATOR WILL ENHANCE TO A FULL HEADER

  WITH SUBJECT TERMS AND OTHER INFORMATION, AS NECESSARY

SO WHY WORRY ABOUT THE "WHO" ?

- LSS RULE DOES NOT CLEARLY DRAW THE LINE WHERE MINIMUM ENDS & ENHANCED BEGINS
  - OLD WORKING GROUP RECOMMENDED TWENTY FIVE REQUIRED FIELDS FOR SUBMITTER'S HEADER
    - →→ IS THIS MINIMUM ?

CRITERIA TO CONSIDER

- KNOWLEDGE: SOME ELEMENTS ONLY KNOWN BY SUBMITTER
- QUALITY: SOME ELEMENTS "BEST" KNOWN BY SUBMITTER

  CONSISTENCY AND QUALITY MIGHT BE BETTER IF DONE BY CENTRAL STAFF -- LSSA
- COSTS: SOME ELEMENTS ALREADY DONE BY SOME SUBMITTERS IN THEIR OWN RECORDS MANAGEMENT PROCESSES -- WHY DUPLICATE EFFORT?

  BURDEN ON SUBMITTERS TO DO MORE SOPHISTICATED CATALOGING

CRITERIA NEED TO BE APPLIED IN TWO AREAS:

- ALL BIBLIOGRAPHIC ELEMENTS

- CLASSIFICATION CODES

SUGGESTED NEXT STEPS

- **FORM LSSARP WORKING GROUP TO RECOMMEND SUBMITTER & ENHANCED HEADERS TO LSSARP**

  - **GOAL:** MAKE REASONED STUDY
    - BASED ON PREVIOUS WORK TO DATE
    - NOT TO REINVENT AND REDO PREVIOUS WORK

  - **TASKS:**
    - WORKING GROUP DEVELOPS WORK PLAN,
    - SUBMITS HEADER RECOMMENDATIONS TO LSSARP MEMBERS FOR WRITTEN COMMENTS, and
    - REVISES RECOMMENDATION BASED ON MEMBERS' COMMENTS

SUGGESTED DECISIONS FOR TODAY

- MEMBERSHIP OF LSSARP WORKING GROUP
  - PANEL MEMBER ORGANIZATION REPRESENTATIVES HAVING KNOWLEDGE OF HEADER DESIGN & USE
  - SAIC REPRESENTATION
- LSS ADMINISTRATOR'S ROLE
  - WILLING TO SERVE AS WORKING GROUP MEMBER
  - WILLING TO PROVIDE SPACE AND CLERICAL SUPPORT
  - WILLING TO PROVIDE LIMITED TECHNICAL ASSISTANCE THROUGH CONSULTANTS
- TENTATIVE SCHEDULE
  - WHEN IS A FINAL DECISION NEEDED?
    - AS SOON AS PRACTICAL, TENTATIVELY SCHEDULED FOR FALL '90 MEETING

Department of Energy
Washington, DC 20585

JAN 3 1 1990

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

Re: Background Materials on LSS Headers

Dear Mr. Hoyle:

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

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

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

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

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

Sincerely,

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

Enclosures

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

cc:
L. Desell, RW-331

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

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

REQUIRED FIELDS:

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

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

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

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

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

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

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

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

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

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

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

DRAFT BIBLIOGRAPHIC HEADER FIELDS
(continued)

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

DRAFT BIBLIOGRAPHIC HEADER FIELDS
(continued)

THE FOLLOWING FIELDS ARE OPTIONAL:

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

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

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

Abstract - a summary of the contents of the document.

Notes:

1 - governed by an authority list

2 - governed by format rules

May 3, 1988

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

MINUTES OF THE HLW LICENSING SUPPORT SYSTEM ADVISORY COMMITTEE MEETING

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

MEETING LOCATION AND ATTENDANCE

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

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

APPROVAL OF THE MINUTES

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

EXPLANATION OF CHANGES MADE TO THE NRC'S DRAFT RULE

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

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

DISCUSSION OF THE DRAFT RULE

Section 2.1000 - Scope of Subpart

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

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

Section 2.1001 - Definitions

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

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

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

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

INFORMATION RETRIEVAL SYSTEMS
A TUTORIAL

Prepared By
Negotiated Rulemaking Technical Staff

FEBRUARY 3, 1988

CONTENTS

1.0 INTRODUCTION

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

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

1.1 PURPOSE

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

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

1.2 HOW TO USE THIS DOCUMENT

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

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

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

2.0 SEARCH AND RETRIEVAL

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

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

2.1 BIBLIOGRAPHIC HEADER

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

2.2 BIBLIOGRAPHIC HEADER WITH ABSTRACT

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

needed to provide the abstract.

2.3 BIBLIOGRAPHIC HEADER WITH SUBJECT TERMS

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

2.4 BIBLIOGRAPHIC HEADER WITH ABSTRACT AND SUBJECT TERMS

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

2.5 FULL TEXT

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

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

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

2.6 ENHANCED FULL TEXT

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

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

2.7 RETRIEVAL ENHANCEMENTS

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

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

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

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

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

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

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

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

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

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

3.0 DATA CAPTURE

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

3.1 IMAGES

3.1.1 Electronic

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

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

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

3.1.2 Microform

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

3.2 FULL-TEXT

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

3.2.1 Optical Character Recognition (OCR) Process

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

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

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

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

The advantages of the OCR process is that it is relatively automated and can be performed without much human intervention up to the point of review and correction. If correction is minimal or not required (i.e. high quality documents), costs can be as low as $.20 to $.40 per page. With many corrections (i.e. low quality documents), costs can be as much as $2.50 to $3.00 per page. If the total costs exceed $3.00 per page, it can be less expensive to key in the document directly.

Continuous improvements are being made in OCR technology which will increase speed of production and reduce the error rate. Presently OCR of an image made from scanning of a good quality paper copy can be reasonably performed, however OCR from an image produced by blow-back of a microfiche or microfilm is not considered feasible.

3.2.2 Rekeying

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

3.2.3 Word Processing

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

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

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

3.3 HARD COPY

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

4.0 CATALOGING AND INDEXING

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

4.1 HEADERS

4.1.1 Bibliographic Headers

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

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

4.1.2 Subject Terms

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

4.1.3 Abstract

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

4.2 FULL TEXT

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

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

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

5.0 STORAGE

5.1 HARD COPY

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

5.2 MICROFORM

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

5.3 ELECTRONIC

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

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

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

5.3.1 Optical Disk

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

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

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

5.3.2 Magnetic Tape

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

5.3.3 Magnetic Disk

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

6.0 DISPLAY

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

6.1 IMAGE

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

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

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

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

6.2 ASCII TEXT

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

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

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

6.3 HEADER

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

7.0 DOCUMENT OUTPUT

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

7.1 HARD COPY

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

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

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

7.2 MICROFORM

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

7.3 FACSIMILE

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

8.0 REPRESENTATIVE SCENARIOS

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

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

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

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

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

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

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

A. BIBLIOGRAPHIC HEADER

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

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

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

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

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

B. BIBLIOGRAPHIC HEADER WITH ABSTRACT

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

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

C. BIBLIOGRAPHIC HEADER WITH SUBJECT TERMS

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

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

D. BIBLIOGRAPHIC HEADER WITH ABSTRACT AND SUBJECT TERMS

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

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

E. FULL TEXT

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

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

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

Storage
Same options as for Scenario A.

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

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

F. ENHANCED FULL TEXT

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

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

9.0 ADDITIONAL SYSTEM PARAMETERS

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

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

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

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

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

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

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

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

8) Cost.

APPENDIX

GLOSSARY OF THE
HLW ADVISORY COMMITTEE

GLOSSARY

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

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

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

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

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

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

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

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

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

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

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

CODING See CATALOGING

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

ATTENDANCE LIST

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

COMMITTEE MEMBERS (Including Spokespersons and Alternates)

Priscilla Attean
Penebscot Nation

Dennis Bechtel
Clark County, Nevada

Steve Bradhurst
Nye County, Nevada

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

Barbara Cerny
DOE

Don Christy
Nuclear Waste Office
State of Mississippi

Bill Clausen
State of Minnesota

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

Kevin Gover
Special Counsel
Nez Perce Nuclear Waste Program

Ronald T. Halfmoon
Nuclear Waste Program
Nez Perce Tribe

Robert Halstead
Radioactive Waste Review Board
State of Wisconsin

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

ATTACHMENT 8

GLOSSARY OF TECHNICAL TERMS

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

Header

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

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

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

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

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

Searchable Header

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

Hard Copy Document

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

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

Department of Energy
Washington, DC 20585

SEP 2 1 1989

Mr. Francis X. Cameron
Office of the LSS Administrator
U.S. Nuclear Regulatory Commission
Washington, D.C. 20555

Re: Your Letter of August 7, 1989
Comments on Prototype System Cataloging Manual

Dear Chip:

I have reviewed the above noted letter and its enclosure in some detail, and have forwarded them to SAIC for their consideration. I will be happy to review the next version of NUDOCS header design, and I agree that we should cooperate on the coordination of header design efforts. However, I feel that a better defined effort than merely exchanging preliminary study documents, internal system design/redesign, etc., is needed. We need to move toward a definition of LSS header record content. A focused work group should begin work on the development of header record designs so that all potential parties have a more definitive statement of the formats they should be moving toward. This initiative should commence sooner rather than later.

Regarding the comments you have forwarded, I would like to address what seems to be a persistent tendency within NRC to assume that the processing protocols, headers, and other record fields utilized in the instrumented test bed processing may pre-determine the eventual LSS header design. Likewise, there seems to be a tendency to perceive our test bed environment as having more objectives than, in fact, it does. Your letter implies that the headers used for the instrumented test bed reflect, or will reflect, the failure in LSS "to ensure the completeness and the unique identification of this critical set of documents" by our treatment of the Document Type and Detailed Document Type fields for the instrumented test bed. Let me assure you that the instrumented test bed header treatments are not pre-determinative of the LSS header formats.

The information management questions being addressed by the instrumented test bed can be summarized by the following:

- How will the system be used?
- What aids or hindrances are evidenced in our overall concept designs?
- What are the effects of partitioning text?

- How will header fields be utilized in conjunction with text search capabilities?

- How will descriptors be used in full text search?

- How effective are printed aids such as a thesaurus and a retrieval manual?

The instrumented test bed, by intent, does not have the validation or testing of the specific level of document type treatment as an objective, although such by-products will be duly considered.

I would also like to make a general observation which is meant to be a constructive one. Your letter notes that NRC's upgrade of NUDOCS makes the continuing dialogue on the issues related to header design particularly important, and that you would like to ensure consistency between the LSS and NUDOCS headers. I read this, in conjunction with the detailed comparisons with 'the way things are done in NUDOCS' that are found throughout the 11 pages of comments you have provided, and am left with the impression that NRC perceives the LSS to be simply a restatement of NUDOCS in a new hardware and software environment. For example, take the question from the comments: Should the LSS detailed document types "be mapped to NRC document type codes"? Why not ask, rather, "To what degree will one be the subset of the other after we have met our design objectives?" It should not become a question of whose system drives the header design: records from all the participants need to be entered; the DOE collection will be preponderant in volume; NRC will be the critical user during the hearings; and, nothing in the LSS implementation should prevent the use of LSS as a records system by a given party.

The LSS is to serve multiple purposes which include its use as a surrogate for discovery, a tool to support motions practice, and, the Commission's docket and official record for the licensing proceeding. We are now attempting to design an LSS which meets all of these objectives. Perhaps NUDOCS already meets most of these design requirements, but, it is my observation that NRC's existing methodologies, document type codes, detail document type, and other treatments are, in fact, constrained by NRC's existing hardware configurations and software capabilities -- as would be the LSS if it simply mirrored NUDOCS (or ARS, for that matter). The point is that these are already dated technologies to some extent, whereas we have a unique opportunity to let our required functionality drive the hardware and software we procure (rather than having to build an application using whatever computers and software happen to be available). Decisions about header design should be made in light of the LSS' unique objectives and what the LSS will allow us to do with the 'tabula rosa' of new technology. During design stages it is important to remember that the LSS does not have to inherit the baggage of DOE, NRC, and other parties' existing systems' limitations, be they hardware, software, or limitations inherent in a system designed for other purposes. At the same time, we recognize that products already developed, as represented by existing systems, can and must be used in building the LSS data base.

I am suggesting that it is more important to determine what fields, field contents, and field formats are necessary to support the organization, search, and retrieval of a record in the LSS header and text environment. We need to do this with the intent of fully utilizing and maximizing the retrieval software's capabilities, as much as they may be anticipated. If we provide this sort of definition to the potential parties, each can begin the process of moving toward the acceptable LSS header record format with minimal rework being necessary at a later time.

A review of existing systems, such as your NUDOCS redesign effort, is useful in that it may provide a checklist of items that need to be addressed and is a source for lessons learned. On the other hand, close scrutiny of cataloging procedures used for our instrumented test bed is premature since the LSS header record formats are not as yet defined. The prototype cataloging procedures are not even a worthwhile point of departure for such a definition because the test bed environment does not attempt to define the anticipated LSS hardware or software environment -- it only emulates anticipated functionality in its study of the attributes which affect that environment.

I hope that these observations will be helpful in our mutual efforts to maintain the perspective of what our LSS design efforts should be based upon. We look forward to participating in the initiative where developing the LSS headers needed to meet LSS functionality is the primary design objective.

Sincerely,

Daniel J. Graser
Program Analyst
Information Resources
Management Division
Office of Civilian Radioactive
Waste Management

cc:
B. Cerny, RW-14

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

Enclosure 7

August 7, 1989

Mr. Daniel J. Graser
Information Resources Management
Office of Civilian Radioactive Waste Management
U.S. Department of Energy
Forrestal Building
1000 Independence Avenue
Washington, D.C. 20006

Dear Mr. Graser:

As part of the NRC efforts to review the design of the LSS, I am enclosing NRC comments on the SAIC reports "LSS Prototype Header Design," and "LSS Prototype Cataloging Manual." Although these reports focus on the LSS Prototype, our comments will need to be considered in establishing the header design and cataloging manual for the final LSS. Most importantly, the manner in which the NRC adjudicatory record has been incorporated into the Document Type and Detailed Document Type fields fails to ensure the completeness and the unique identification of this critical set of documents. In this regard, we would be interested in discussing the resolution of our comments and questions at your convenience.

A continuing dialogue on the issues related to header design is particularly important in light of the NRC upgrade of its document control system (NUDOCS). Part of the upgrade process is a re-evaluation of the headers, indexing manuals, and authority files for NUDOCS. We would like to ensure consistency between the LSS and NUDOCS headers and would encourage coordination of these header design efforts. In this regard, we would invite you and your contractor to evaluate the next version of the NUDOCS header design which will be ready for review in October 1989.

If I can provide any further information on our comments, please feel free to contact me.

Sincerely,

Chip Cameron

Francis X. Cameron
Chairman
LSS Internal Steering Committee

Enclosure: As stated

7/26/89

COMMENTS ON SAIC REPORTS ENTITLED

"LICENSING SUPPORT SYSTEM PROTOTYPE HEADER DESIGN"
March 7, 1989 version

and

"LICENSING SUPPORT SYSTEM PROTOTYPE CATALOGING MANUAL"
March 14, 1989 version

I. GENERAL COMMENTS AND QUESTIONS:

1. The Cataloging Manual (CM) states that 120,000 pages of documents will be captured. We understand that this represents about 2,600 documents including the SCP, its references, some of the "administrative record" and some handwritten notes. We are concerned that these documents are not a representative sample of the document types that will populate the system later on. At a minimum, this will affect the validation of the Document Type authority files. Also and more important, it will limit the ability of the various classes of searchers to fully evaluate the prototype in the "test phase". In what areas do you expect the header might change as the true makeup of the database evolves?

2. What is the source or basis for some of the specific format requirements in the Cataloging Manual? Is it patterned after any existing system, such as the DOE's ARS? NRC has provided SAIC with the NRC's NUDOCS header record layout, indexing manuals, and authority files. What, if any, are the reasons why some of the NRC conventions (such as Document Type structure and Affiliation codes) were not adopted?

3. How are numeric and alpha-numeric fields structured so as to allow for sorting and listing? Will indexers have to "zero-fill" or will the software justify appropriately?

4. What procedures are envisioned for the modification and update of the authority files based on submitter's suggestions and needs?

5. **There needs to be much more discussion internally within NRC and between the parties about the following issues:**

   A. **One issue that is not addressed to any degree in the header design document is the extent to which the submitter's authors or authoring offices as opposed to the submitter's catalogers will complete portions of**

the header, specifically the title and/or abstract. From the experience with NRC's NUDOCS, it may be better for the submitting office to at least "propose" as much of the subjective information as possible in order to limit the number of errors or misrepresentations committed by the catalogers unfamiliar with the context or the subject matter. From the description of the bibliographic fields, it appears that most can be completed by the submitting office with party's catalogers performing review functions for quality control and for format or classification consistency. There are cost/benefit issues to debate.

B. Abstracting
   * the need and purpose of an Abstract
     - for all documents or
     - for just selected documents by type
       - if so, which types of documents.
   * who, personally and organizationally, will prepare the abstract, depending on document type.
   * When will the abstract be prepared.
   * Any differing considerations on above issues between "backfit" phase versus the "real-time" phase when the timeliness of entry requirement will compete with the requirement for the quality of indexing for long-term retrieval.

C. More work must be done on the Document Type classification scheme. See Fields 5 and 6 for more information.

6. Section 3.1 of CM. Windows and pull-down menus are high-tech. They will certainly help new indexers and eliminate inconsistent entries. However, experienced indexers may want faster entry. Will it be possible for such authorized indexers to bypass windows and enter directly. Entries in specific fields could then be automatically checked against authority files at "end" before record is 'closed out'.

7. In Section 3.1.3 of CM, the Query function as discussed seems cumbersome and unsophisticated. Can it handle multi-parameter searches? If so, how will it handle embedded Boolean statements within statements? How does it differentiate between: ([A and B] or C) versus (A and [B or C]) ?? Is it thought that the prototype cataloguers did not need such sophisticated search capabilities? In the full system, both cataloguers and searchers will need such a capability.

8. Section 3.2.1 of CM implies that documents come to the station with LSS Accession numbers already assigned. What are the pre-indexing procedures and rules? Who defines and determines the "cataloging units"? When and how are accession numbers assigned? Who and how are duplicates searched?

9. In Section 3.2.3 of CM, "Deleting a Record", it is stated that the phrase "Delete Number?" will appear before an entire record is deleted. This could be confusing to the cataloger in that he or she may assume that only the LSS accession number will be deleted rather that the entire record. Also, is it possible to archive these "deletions" at least temporarily instead of erasing them so that they can be recovered if needed?

10. In Section 3.2.4 of CM, "Using Query," on pages 17 and 19 of CM, the method of performing a search is described. It is assumed that the described method is only for the use of catalogers or other individuals who have extensive experience with the LSS. The search software for most LSS users must be much more helpful.

11. Section 3.3.2 of CM -- How are "batches" defined? What if more than one batch is done in a day? or if one batch spans more than one day? The command "After what date (YYMMDD) does not seem to allow for this.

12. Section 4 of CM - Quality Control -- there definitely should be more than one level of QC. Also, the initials of the QC persons should also be carried on the data record. Each submitting party will have their own Quality Control procedures. However, QC should be given a lot of attention and the responsibility should be a major line function, not relegated to a committee.

II. COMMENTS ON SAIC PROPOSED FIELDS AND ASSOCIATED CATALOGING RULES:

Field #1. LSS Accession Number:

NRC places their Accession Number in the lower left corner. It would be of interest to know if there was a reason for your decision to place the LSS number in the upper right corner. Many NRC documents have notations in the upper right corner. One alternative placement would be the lower right corner; although some organizations place page revision numbers there. Another alternative would be to place the number vertically in the middle of the left margin.

We are confused as to how the "Package" header will differ from the header of the "parent" document. Or will the parent records just carry two Accession Numbers? In the NRC systems, the Accession Number of the Parent or Mother is carried on the data record of all the "children" and the Parent document carries a flag to denote the existence of "children". How will your method effect the hit counts, the sorting, and printouts? More explanation and some examples are needed here.

Field #2. Title/Subject:

The "subject line" on correspondence is usually a very broad characterization with little thought given toward long term retrieval or distinguishing it from other documents. It is acknowledged that brief abstracts (NRC NuDocs has 4 lines) prepared by catalogers are time consuming and not always the best. It is also acknowledged that with the full-text of documents available for on-line searchers, this short abstract may not be critical for search and retrieval purposes. However, for the purpose of listings, bibliographies, announcements, court certifications, and for scanning large "hitlists" to determine the relevant documents for further review, something more than the "subject line" will be required. Remember, not all end-users of the LSS will be on-line. Also, most letters do not have a "subject line" like Memoranda. Maybe this is the purpose of the LSS "Abstract Field #22. If so, it is not clearly stated.

On page 5 of SAIC LSS Prototype Header Design Report (HD), the last two lines of the discussion of Field 2, Title, state that the title of an encompassing work will be in the "Bibliographic Citation" field. It is not clear to which of the fields this statement is referring.

Cataloging Rules:

- What is the proposed length of this field?
- Why are the format rules so very specific? Given the number of varied catalogers from different parties overtime, will it not be a real burden on the indexers to follow these strict professional-type cataloging rules and on LSS staff to assure compliance and consistency? For what end? If it is to do sorting (and filing) by title alphabetically, then couldn't some software routine be written to ignore preceding articles?

- p. 27 of CM, 2nd paragraph -- there should be more explicit rules about what to cover in title descriptions. Phrases like "meaningful" and "reflecting the content" are too vague. The NRC system has more specific rules on what aspects of the document content should be covered varying by document type. While they may not be perfect, at least they should be reviewed.

- Shouldn't the same convention as with Abstract Field be used to denote actual wording of the Title versus indexer-composed description.

Field #22 Abstract Field is discussed here due to its interrelationship with the Title/Subject Field.

It is unclear in the SAIC Prototype reports which documents will be abstracted.

The CM states that a "brief description on the content" will entered. In the final LSS design, much more must be decided and said about the Type of Abstract required or accepted. Also, what is the proposed length of this "brief" description?

Field #3, LSS Pointer

p. 30 of CM, 2nd paragraph of "instructions" -- how will be cataloger know that a 'revision' is already in the system? Can't or shouldn't that be caught in the pre-indexing review (duplicate check)?

Using this field and maybe others, how will marked up copies of the same document be handled, i.e. reviewers handwritten comments and editing on a report or "pen & ink" changes ? How will such a document be indexed? Also, how will a cover letter forwarding various/selected replacement pages to a previously-submitted document be handled? How will drafts, revisions, errata, etc. be linked together? The listed codes in the "controlled vocabularies" do not seem to cover such a case.

This information must be captured somehow and the search software must utilize it to notify searchers passively that previous or later versions and/or errata exist and are on the system.

Field #5. Document Type

This should be a repeating field!

Is this and the "detailed document type" scheme already in use at DOE? As you know, NRC has their own scheme based on their own terminology. Somehow a mutual interagency list should be devised.

!!!> There is a major concern regarding the instruction for completion of this field on page 34 of the CM. There, it is stated that the cataloger will select the first [and no other] document type that matches the form of the document from the provided list. The eighth document type in that list is Legal Materials which (as discussed in the description of Field 6, Detailed Document Type, on pages 36 to 40) includes those documents associated with the NRC adjudicatory record. Under the current instructions for the catalogers, a document that is part of the adjudicatory record may not be identified as such if the cataloger finds a document type in the list that matches the document prior to reaching the Legal Material document type. This is a serious problem and must be corrected as soon as possible.

THIS FIELD CAN NOT BE USED AS THE DELIMITER FOR THE HLW ADJUDICATORY FILES!! See Section III for proposed new field.

A "legal" document in some other proceeding may be submitted to LSS. It should get a "legal" document type BUT may or may not be part of the HLW adjudicatory record.

Any non-legal document type, i.e.. drawing, journal article, letter, etc, at first may be entered as they are. Then later that document becomes an exhibit. It then must also carry the Legal document type while keeping its original document type code.

Field #6. Detailed Document Type

This also should be a repeating field. Some of the elements here are not mutually exclusive within one document.

These codes should be mapped to NRC document type codes. If for no other reason than to test clarity of both systems. Specifically, more work must be done on the legal document types.

Field #7. Document Date

How will transcripts and minutes of meetings spanning multiple days be coded?

Field #8. Document/Report Number

Is this field only for numbers of the specific document being cataloged, i.e. (1) contract number for actual contract and amendments, not reports done under that contract; (2) the USGS or NUREG report and revisions, not other documents, memos, letters about the USGS or NUREG report? The description appears this way. Assuming this is true, how will documents commenting on or 'about' such reports be coded?

What is the purpose to preceding alpha codes listed on p.45 of the CM? This appears redundant to the Document Type Code. Does this not put a burden on the searcher to know what kind of number he/she has been given to search? If final retrieval software has a 'wildcard' character, then this problem could be eliminated, but I think the classification is not justified.

Rule 7 on page 47 of the CM states that common abbreviations should be used where possible. While this suggestion is acceptable in theory, the examples provided are not common to all LSS users. It may be best to refrain from using abbreviation except where their meaning is obvious and unambiguous. NUDOCs has attempted to keep an authority files of accepted abbreviations and it is not always up to date or used. The problem will get much worse given the multiple parties contributing to the LSS over long period of time.

Field #9. Edition, Version/Revision

This field will require more detailed instructions to handle:

-- selected pages submitted as Amendment 9 of looseleaf document such as this indexing manual or the application !versus!
-- whole indexing manual or application including revised interfiled pages thru Amendment 9

Shouldn't this field be linked to occurrences in the previous field? Might there not be cases where Rev. 6 to a Sandia report then becomes NRC NUREG-####, which is later supplemented 6 times? Sad but true.

In reference to the use of this field "for describing computer codes and code manuals" , more explanation is required. As one reviewer of these reports stated "I think I know what this means, but surely this needs to be spelled out better so we are all singing from the same hymnal"

Field #10. Author Name

Should concurrences, either by name or organization, be picked up if they appear on the document?

Field #11. Author Organization

How will the authority file rules handle organizational name changes, subsidiaries, reorganizations, etc?

Must develop rules for authors who write in two of more capacities, i.e. letterhead says ACME utility, but author is writing as the head of the utility owners group. OR lawyer works for DEWY, CHEATEM & HOWE but is representing EXXON. OR NMSS staff chairing inter-agency or intra-agency review group? How handled? -- will you pick up both?

In the NRC system, the Affiliations (and the Document Type Codes) have hierarchical scheme to classify document authors and recipients,

NRC AFFILIATION SCHEME

First level - E for external vs N for internal. This would not be appropriate in this system

Second level - type of organization"
i.e. SG = state government
UT = utility
LO = local government
US = Federal agency
LG = legal firm
MV = manufacturer or vendor
etc

DOCUMENT TYPE CODE SCHEME

CLUTN = correspondence/letter/utility to NRC
TRUTIN = text/report/utility inspection report

POINT: These codes can be very powerful in searching, especially along with the Boolean "Not Equal" to narrow scope of searches to their essence. Many times after searching known parameters, the resultant hitlist is still too large to be useful. At this point, the

searcher may not know what he/she wants or be able to positively select a narrowing concept, but he/she knows what he/she does not want. Then, using the 1st (and second) level Affiliation codes [and/or Document Type codes] truncated, he/she can exclude classes of documents by type of author, by type of recipient or by type of document.

Field #12. Recipient Name

Proposed instructions state that this field would include attendees at a meeting as recipients. Some meetings may have a long attached attendance list and it may not be feasible or beneficial to list all of them in this header field. More specific rules must be developed to narrow the scope and intent of this data capture. Consideration: if smaller number of attendees (i.e. less than twelve) are listed at the first of meeting minutes or meeting summary and it was a "participatory"-type meeting, then such persons should be captured. In this case, one could argue that such persons are more "authors" than "recipients. Better yet, have another field for "attendees". The requirement to complete this different field could be triggered for all records having certain document types.

Field #17. Publication Data

Instructions state that an entry is required. From the description of this field, however, it is not clear whether an entry will be appropriate in all instances.

Field #18. Subject Term

Please provide more information as to the intent of this field. The broad nature of the terms may cause this field to be of little value in searching for particular documents. Is it to be used to segment the database? If so, there may be problems because many documents may address several of the listed terms such that the submitter and cataloger would have difficulty in assigning a single term to a document. Also the searchers may take issue with the view of the cataloger. It will be hard to make the segments mutually exclusive by the subject scheme. Page 63 of the CM states that this field will not be used in the Prototype. Therefore, it will be impossible to test the usefulness of this item.

Field #21. Special Class

More discussion is required on this field because it appears that this field and the Document Type fields are being used in combination to "segment" the Adjudicatory Record file for the adjudicatory Boards.

In this field or in the "Project" field, documents related to rulemakings and documents referenced/cited in other documents should be captured.

In the proposed list of "special classes", it is not clear what documents will be encompassed by the following terms: EA-AR (Part of the Environmental Assessment Administrative Record), LA-AR (Part of the License Application Administrative Record), and Lit (Part of EA Siting Litigation). Are these DOE-specific classes? If not, it would be difficult for others to assign such codes. Other parties will have their own "special" codes. The LSS Administrator will have to maintain authority list.

The description of NRC evidence in the special class list should be revised to read "Unit is evidence in an adjudicatory proceeding" because evidence may be oral or written.

Field #22. Abstract.

See comments in section on the Title/Subject field (#2).

Field #25. QA Level Code.

Please provide more information on the scope and usage of this field.

Field #27. Page count.

How will the page count for package records be handled? This has been a sticky issue in the NRC's NUDOCS, especially when an enclosure in the new "package" is a document already indexed earlier and therefore is a "duplicate" which must be tagged to this new package for completeness.

III. PROPOSED ADDITIONS.

The following elements of information were not included as separate fields but may be of value in performing search tasks:

- Date docketed
- HLW adjudicatory document "tag" -- see comments on fields #5, #6 and #21 for more information.
- Concurrence Names
- Reference Affiliation/Organization (use same Controlled Vocabulary as used for Author or Recipient Organization.)
- Referenced Documents and/or regulations (parts of CFR)
- Event Date -- dates of meetings, inspections, "incidents".
- Alternate availability -- other sources of same document, i.e. NTIS, GPO, ORNL and/or location and contact of core samples, data tapes, maps, travel vouchers, etc.

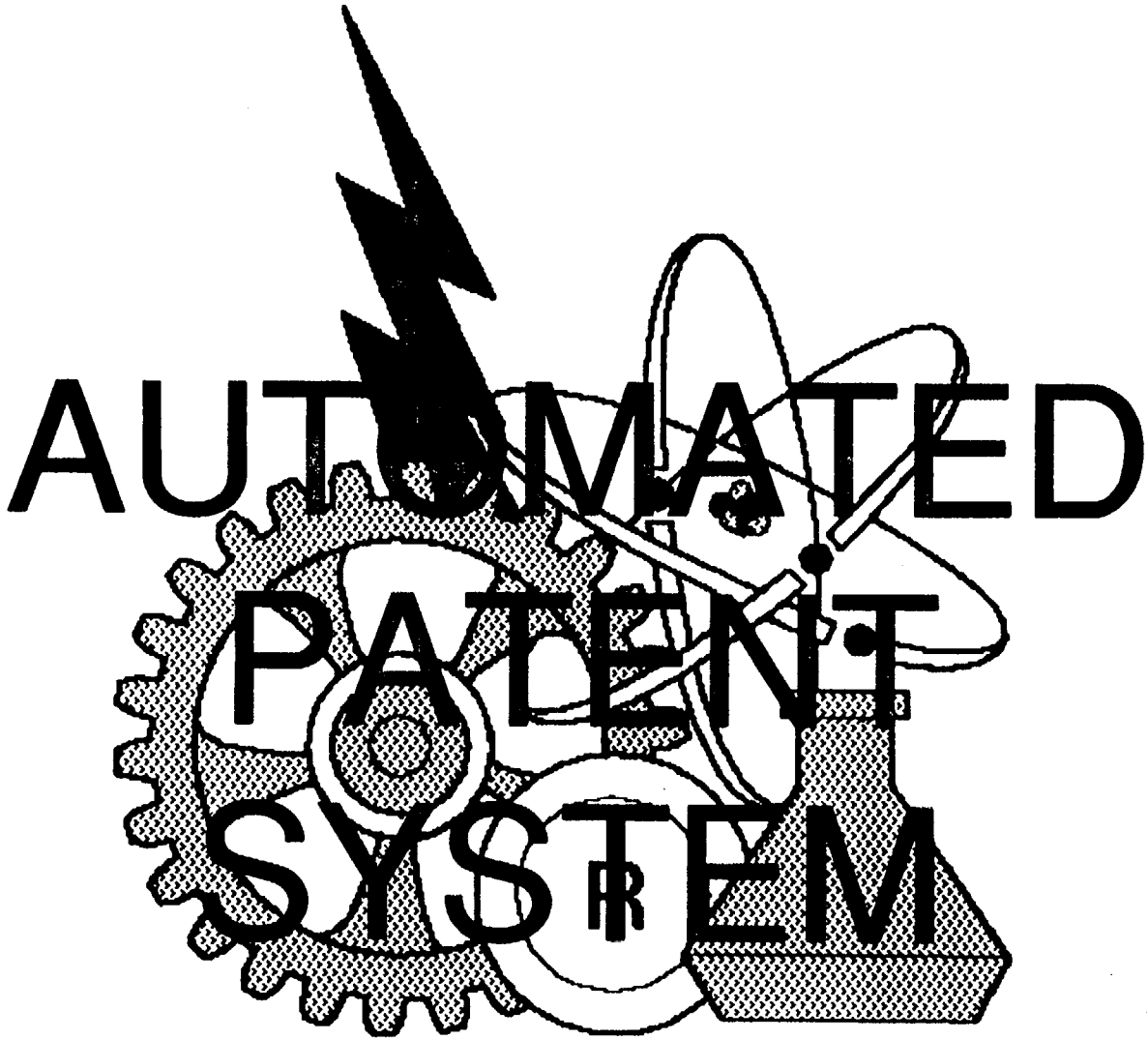
In addition, there are certain elements of information that are captured in more generic fields which might warrant their own specific field:

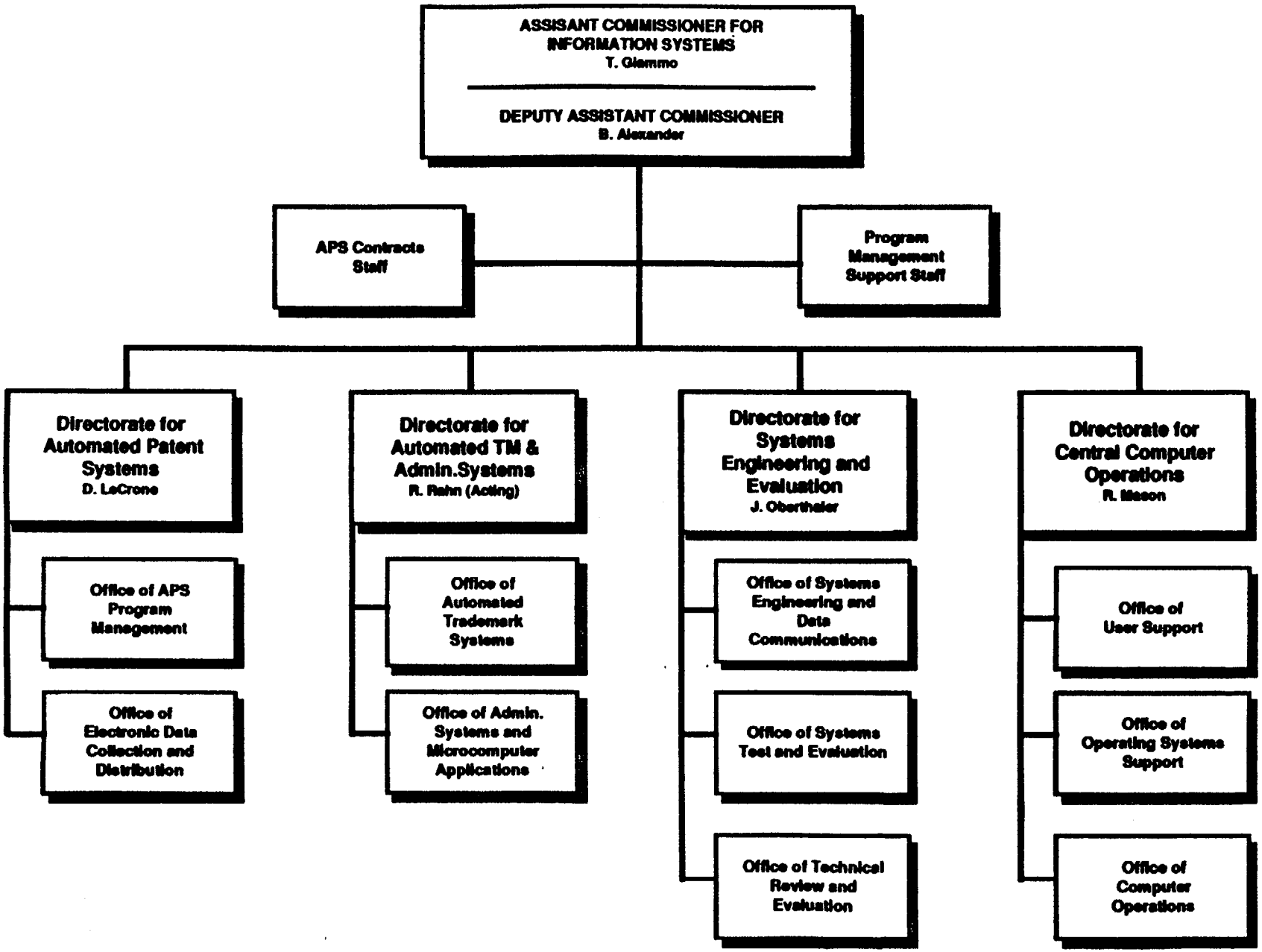
- Witnesses & Speakers (currently in 'author field')
- Attendees (currently in 'recipient field. If kept as part of more generic field, I could debate that attendees should go in 'author field', especially for small meetings - less than ten people.)
- Contract numbers (currently in the 'Report field')

There is an argument which states that some of the above listed information could be found by searching the full-text. Also, some of this information could be loaded into more general fields. However consistency of capture and format would argue for a specific field. The existence of such fields would trigger indexers to capture the information in a standard format. This would relieve the burden on the searchers.

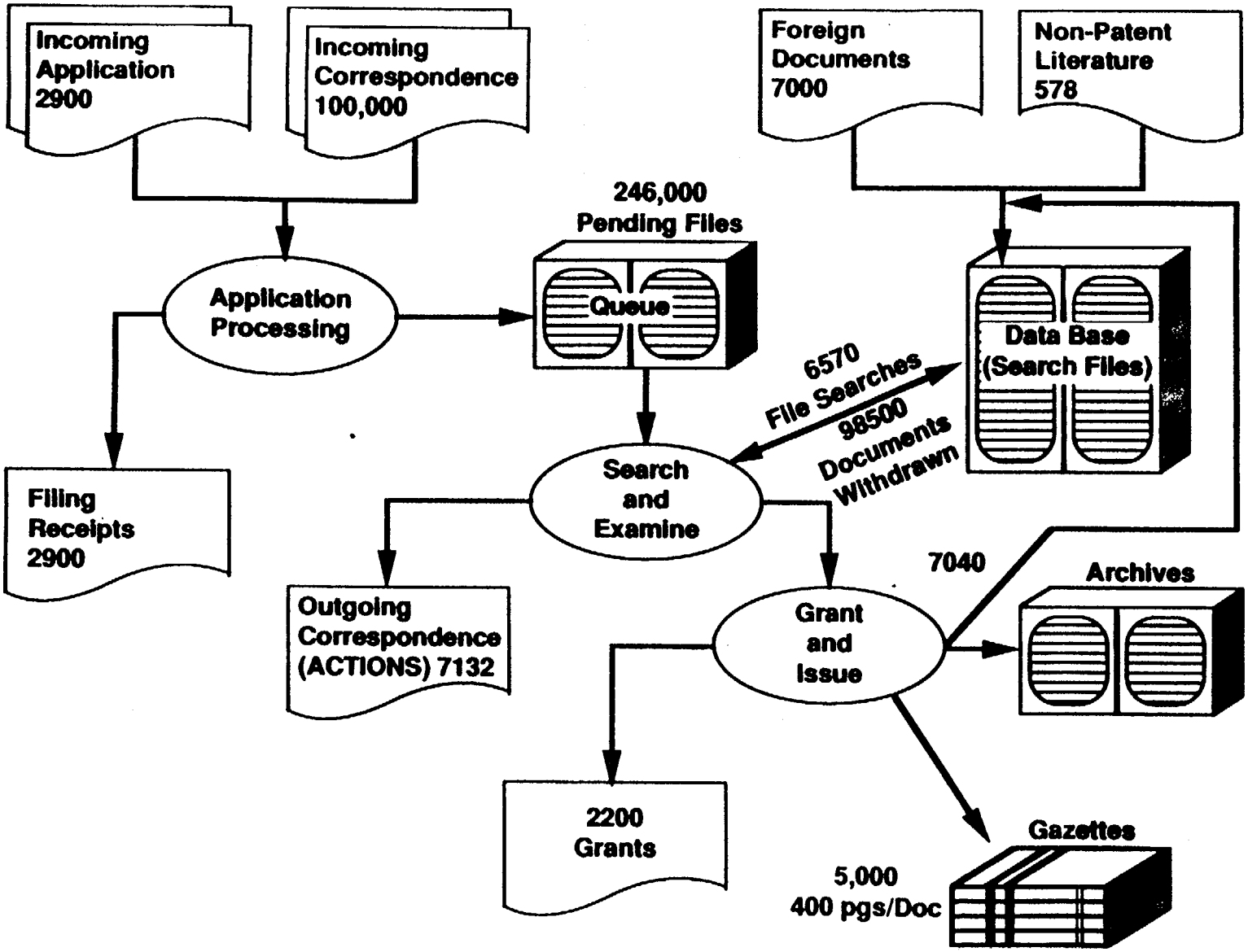
A paragraph or so explaining and justifying the exclusion of such data capture and alternate retrieval methods should be provided. It would be helpful to those of us who follow (advisory committees) and wonder "why not?". Further, weren't there other fields proposed or discussed during the negotiations? If so, what was their disposition? Those fields that were considered but not included should be listed and discussed <u>somewhere</u>.

Enclosure 8

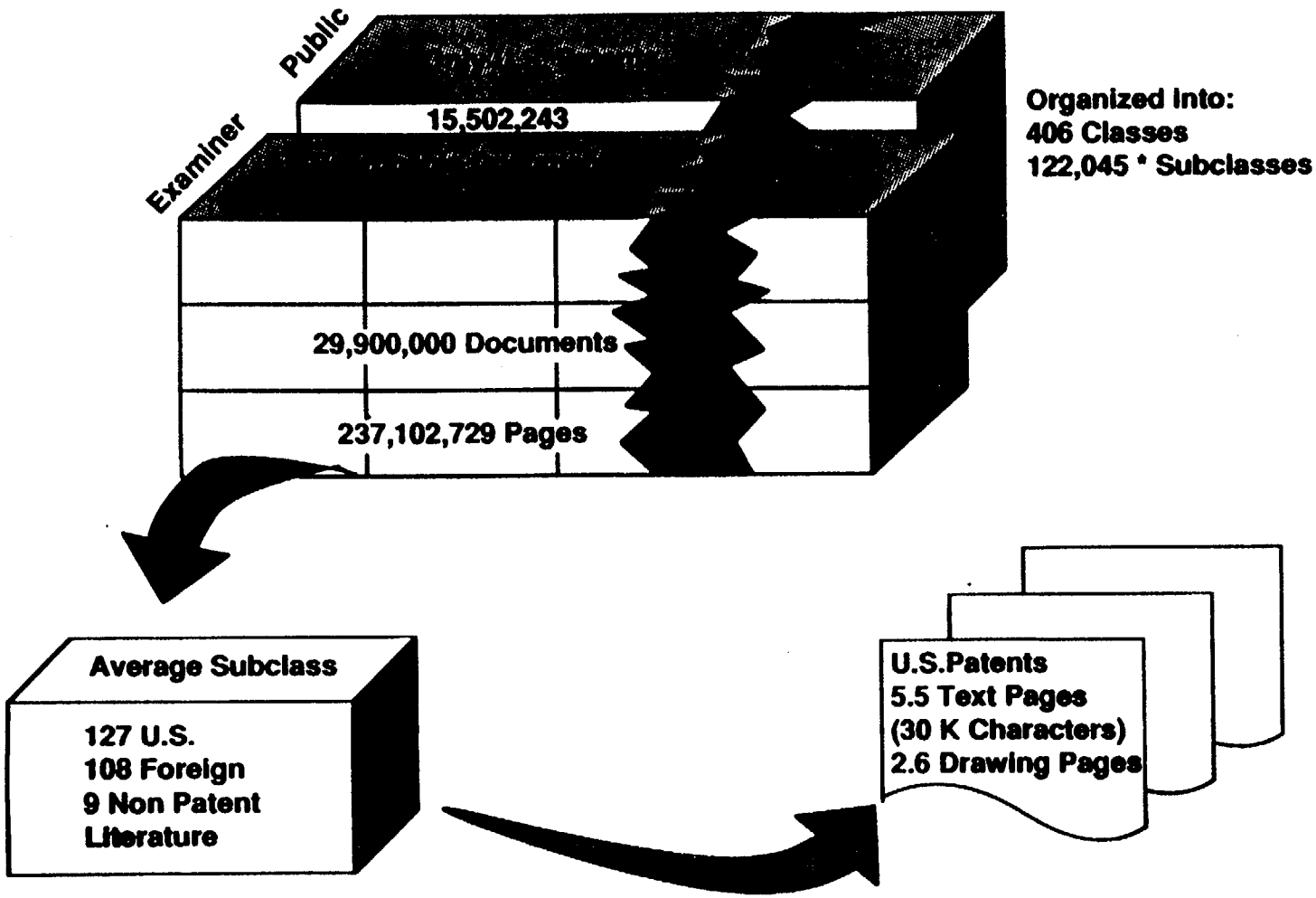



ASSISANT COMMISSIONER FOR INFORMATION SYSTEMS
T. Glemmo
DEPUTY ASSISTANT COMMISSIONER
B. Alexander
APS Contracts Staff
Program Management Support Staff
Directorate for Automated Patent Systems
D. LeCrone
Office of APS Program Management
Office of Electronic Data Collection and Distribution
Directorate for Automated TM & Admin.Systems
R. Rahn (Acting)
Office of Automated Trademark Systems
Office of Admin. Systems and Microcomputer Applications
Directorate for Systems Engineering and Evaluation
J. Oberthaler
Office of Systems Engineering and Data Communications
Office of Systems Test and Evaluation
Office of Technical Review and Evaluation
Directorate for Central Computer Operations
R. Mason
Office of User Support
Office of Operating Systems Support
Office of Computer Operations

Patent Operations Weekly Work Volumes

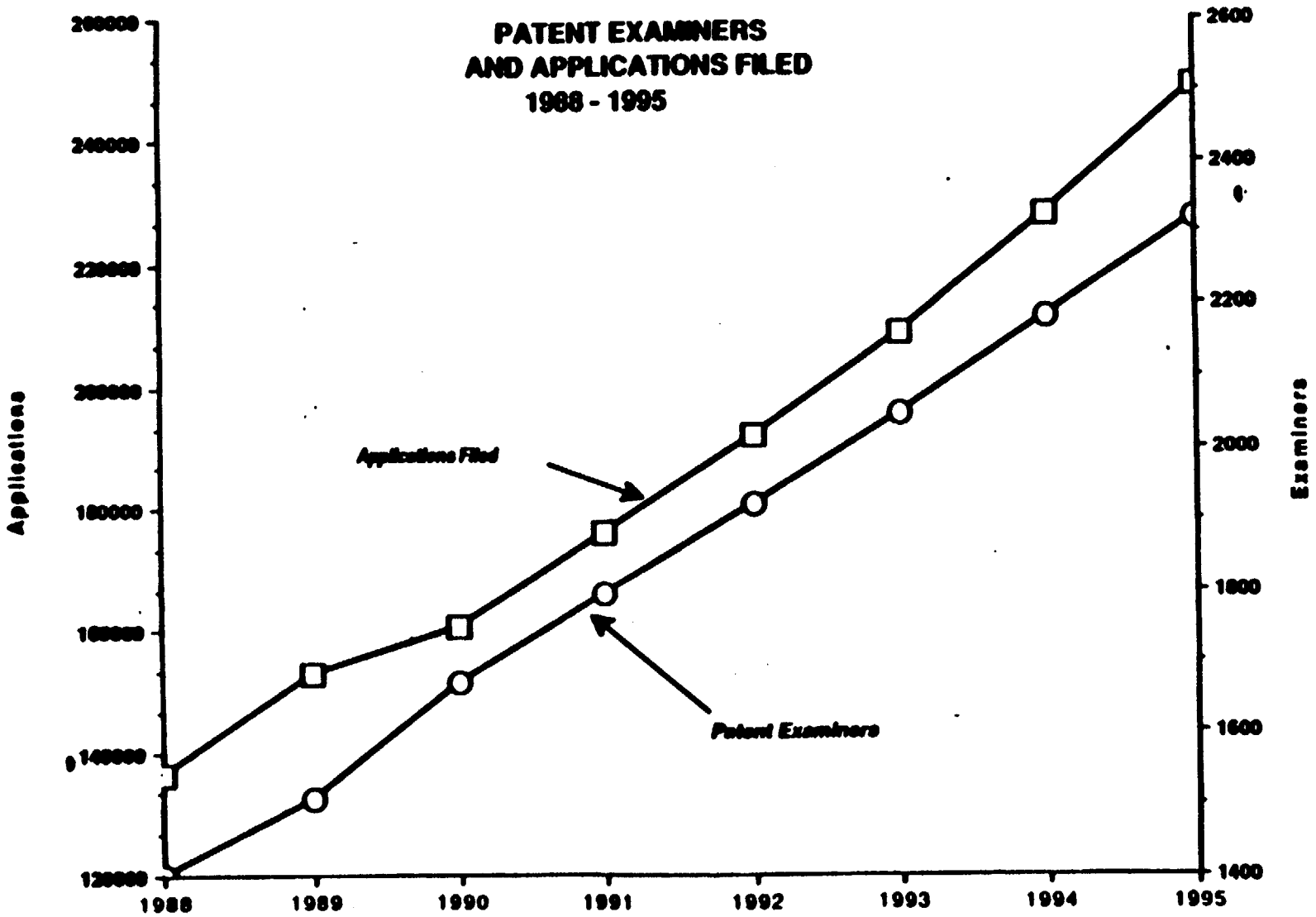


Document Data Base
(Search File)



* 3% Reclassified Annually

Figure 2


PATENT EXAMINERS
AND APPLICATIONS FILED
1988 - 1995
Applications
Examiners
Applications Filed
Patent Examiners
240000
220000
200000
180000
160000
140000
120000
2600
2400
2200
2000
1800
1600
1400
1988
1989
1990
1991
1992
1993
1994
1995

PTO's Automation Objectives

- Provide Automated Searching Services to Patent and Trademark Examiners
- Create Electronic Data Bases Containing U.S. and Foreign Patents and U.S. Trademarks
- Broad Dissemination of Patent Information in Electronic Form

PTO's Automation Objectives

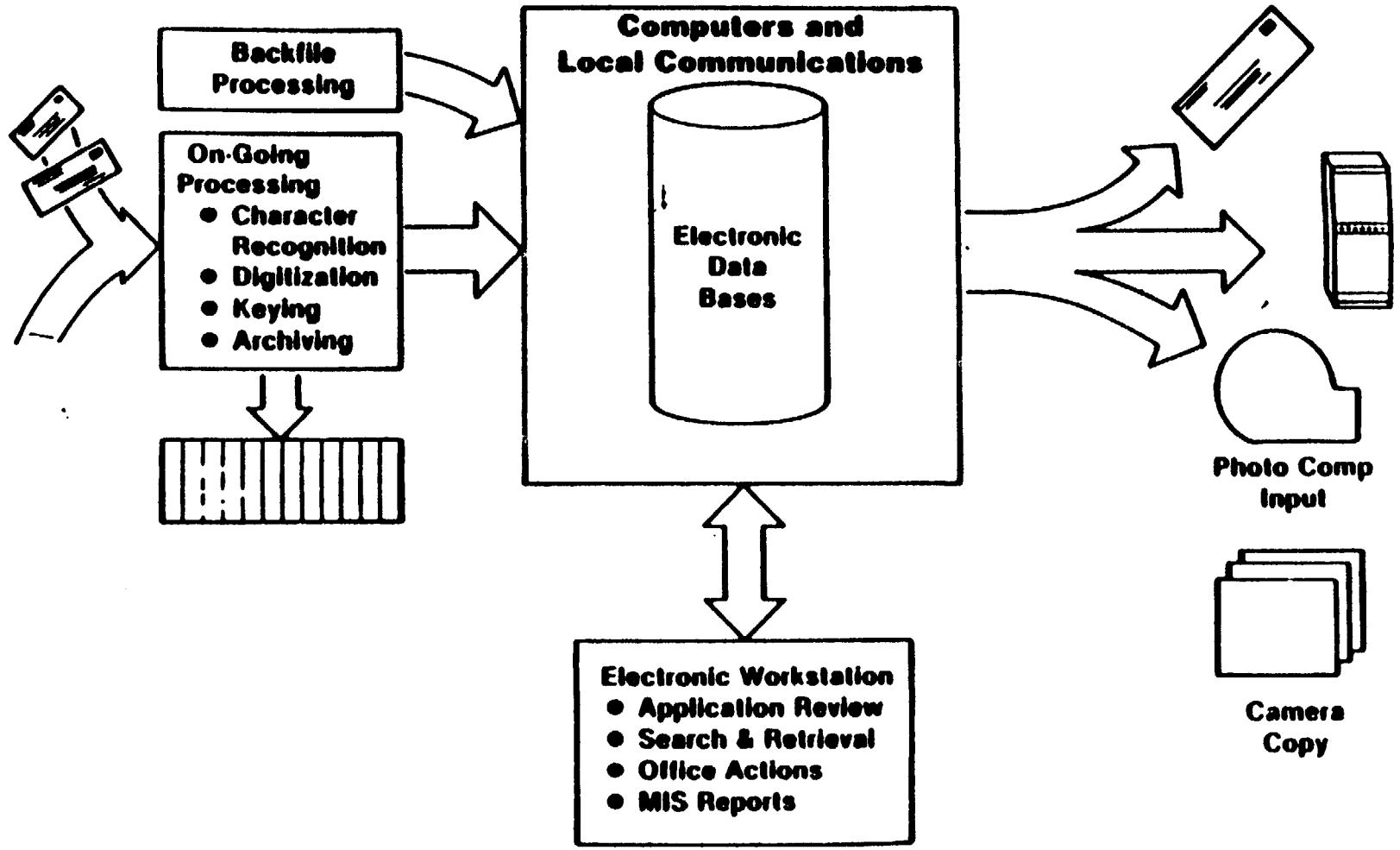
(continued)

- Permit Filing of Applications in Electronic Form
- Enhance all Patent and Trademark Processes through Automation

System Characteristics

- Large Mainframes for Text Search
- Sophisticated Workstations for Digital Image Searching
- Massive Data Base on Optical Disks
- High Speed Communications Network

Automated Patent System
General Concept



APS Development Strategy

- Production System
- Operational Testbed - Group 220
- Group 220 Composition
  - Small Number of Examiners
  - All Technologies
    - Electrical
    - Mechanical
    - Chemical
- Long-Term Optional Quantity Contracts for Deployment
- Modular Architecture to Allow for Technology Enhancements
- Conversion of Complete U.S. Data Bases
- Exchanges - European and Japanese Data Bases

Fully Deployed APS

- Number of U.S. Patents:
  - 5 million
- Number of Foreign Patents:
  - 7-10 million
- Total Optical Data Base Size:
  - 32 terabytes
- Number of Image Workstations:
  - 1000
- Projected Capacity of Communications Network:
  - 400-500 megabits/secon

Database Development for A P S

> **Full Text of U.S. Patents from Printing Process**

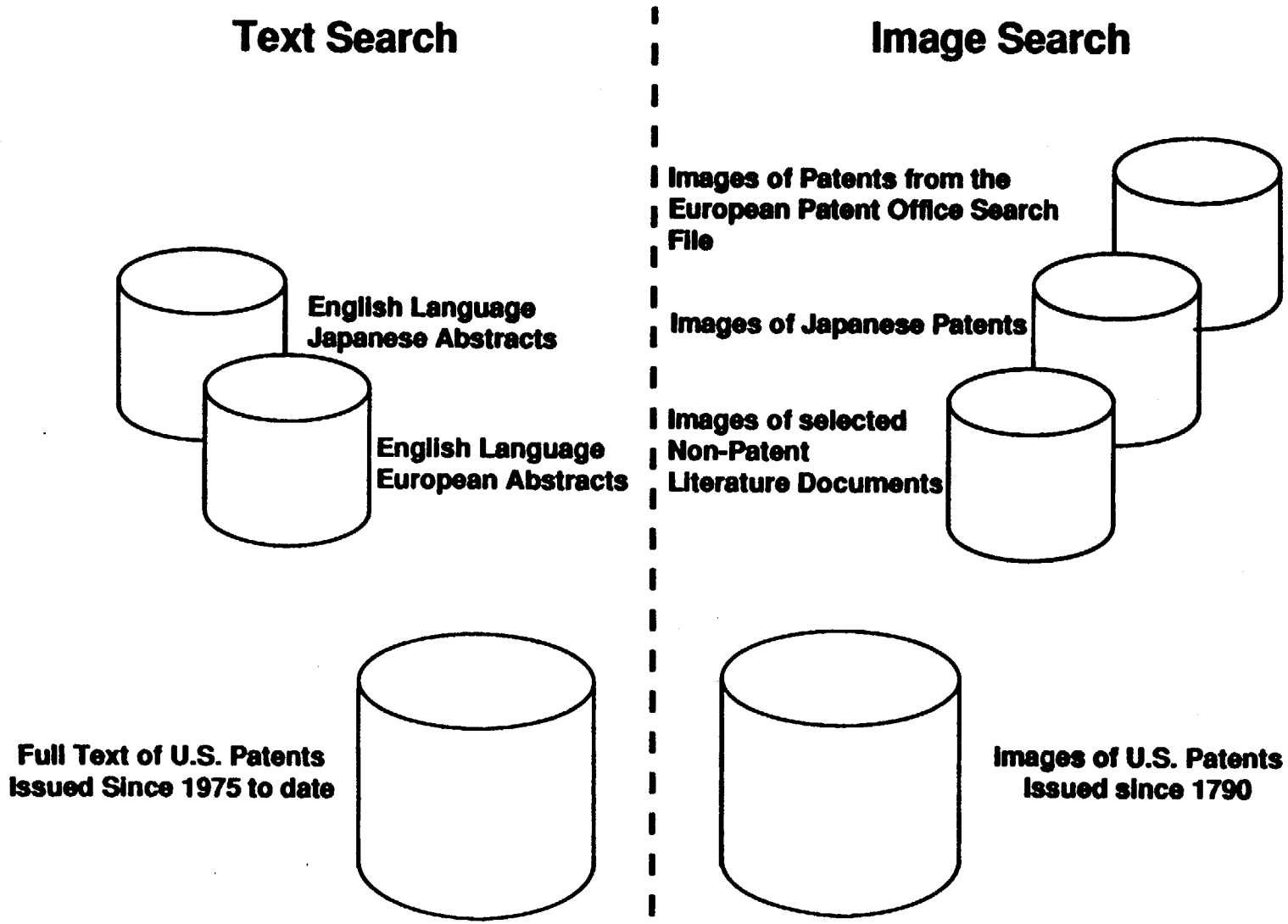
> **Capture Digitized Images of U.S. Patents**
- **Scan Patents at 300 DPI**
- **Write to Optical Discs**
- **Image Capture Complete in First Quarter of FY-89**
- **Display Images at 150 DPI**

> **Images of Foreign Patents**
- **Develop Trilateral Image Standards**
- **Via Trilateral Agreement with European Patent Office and Japanese Patent Office - Exchange Images**

> **Load Images and Install Discs on A P S as Needed**

> **Access to Commercial Data Bases**

Automated Patent System
Searchable Databases



BENEFITS

- High Quality Patents (More Comprehensive Search)
- Ability to Meet Ever Increasing Workloads
- Dissemination of Technology to the Public
- Access to Comprehensive Database of Foreign Patents.

GROUP 220 STATUS

- Group 220 examiners like the system and are using it full time operationally.
- Examiners are using the system with advanced and sophisticated search strategies.
- We believe the system to be productivity-neutral at this time - with improved quality.
- Release #4 (requested by users) provided improved functional capabilities and up to 30% improvement in systems performance.
- New display screens made by Techtronics are currently being tested in the Group cluster room.

Public Search Room

- Four APS Text Terminals Installed
- Over 600 Public Users Trained
- Public Use of the System is High
- Public User Fee is $40 Hour
- Image Workstations will be Added Soon

BOTTOM LINE

- Text Search is Operational and Deployed. Modest Evolutionary Enhancements Continuing.
- Image Search Software is Mature (i.e. Near End of Development. User Requirements for Additional Enhancements Identified and Programmed for Implementation). (Release #5 and #6)
- Hardware Improvements Identified and Scheduled for Reprocurement.
- Accelerated Deployment of Image Search (from the Schedule in the October 1988 Plan is Achievable and Can be Justified).
- Need to Reinitiate Developmental Stages of the Electronic File Wrapper, PALM, Patent Copy Sales, Classification Data Systems and Photocomposition.

Planned APS Enhancements

- Expand Test Bed to Second Examining Group
- Load Images of All U.S. Patents on Optical Disks
- Expand Text Search Data Base

Enclosure 9

LSS ADVISORY REVIEW - PLANNING AGENDA

	1990	1991	1992	1993
LSSARP Agenda	October (Tentative) Review of Revised Topical Guidelines Review of ARP Subcommittee Recommendation on Header Review of SAIC Design Documentation Discussion of Priority Documents Production Schedule Presentation on Access to Technical Data Presentation on Compliance Evaluation Program	First Meeting Second Meeting		
LSS/HLW Milestones	February: SAIC Prototype Report March: SAIC Capture System Design Document; SAIC LSS Thesaurus (Draft) August: SAIC Image System Design Document; SAIC Workstation Hardware Configuration Design September: SAIC Search System Design Document; SAIC Controlled Vocabulary	January: Surface Investigations Begin August: Final RFP for LSS Contract	April: Award LSS Contract November: Exploratory Shaft	January: First LSS Station Operational

U.S. Patent and Trademark Office

Office of Information Systems - Automation

Automation in the Patent & Trademark Office

In 1980 the Patent & Trademark Office (PTO) began its current automation efforts by congressional mandate through Public Law 96-517, section 9, whereby the Commissioner was charged with preparing a plan to fully automate the operations of the agency.

In preparing, the PTO identified its current systems as corner stones for the future systems. A comprehensive plan was drawn up to cover all operations of the agency.

In 1982, the PTO submitted to the Congress a plan to improve the quality of patents and trademarks through automation. Congress approved the plan's concepts and instructed the office to go ahead with the implementation of its plan.

TRADEMARK AUTOMATION

Since 1982, the entire Trademark Examination Operation has been automated and is now using a search and retrieval system with a data base of over 600,000 active Trademarks. Thirty-five percent of these trademarks contain picture images, stored electronically as digitized images, of the design elements in one data base, and 100 percent of the textual information in ASCII form stored in another. Text and Image searches are undertaken by trademark examining attorneys to accomplish the examination of applications for Trademark Registration. This search system is called T-Search. The T-Search software is a modified version of a commercial software package called ORBIT available from Maxwell Online Inc.. This system is operated in a conventional IBM mainframe computer configuration connected to workstations. Both image information and text information are stored on magnetic disc media. The search software allows unique searching capabilities for both text and image in a combined search statement or separately as text or image searches. The capability for simple text, phonetic, syllabic and numeric searches either separately or in combination are also possible. In a text search, both left and right hand word truncation operations may be performed in a single search statement. The workstation used by the trademark attorneys is a Burroughs B-22 microcomputer.

PATENT AUTOMATION

The Automated Patent System (APS) is being implemented in response to a need to improve patent quality . This system provides improved access for the prior art search performed by examiners as a preliminary to patentability decisions. The first step towards automation is the availability of full text searching of all U.S. Patents which have issued since January of 1975 and English language abstracts of Japanese patents. All 1600 patent examiners have been trained to use full text search which is available through the use of text terminals connected through the APS.

Eventually, full electronic search as depicted in the attached system architecture chart will be available to the patent examiner at their high resolution, dual screen workstations. This system has already been installed as a production system in one of the 16 patent examining groups (Group 220).

The APS is not a conventional architecture as can be seen from the attached architecture chart. Both Image and Text type searches may be conducted, but unlike the T-Search system they may not both be searched in a single search statement. The Search software used for the Text Search portion of APS is also a commercially available package which has been augmented for Patent Full Text Search it is called Messenger and is a Chemical Abstracts Services product. Image Search has been created for the Patent and Trademark Office by its contractors, Planning Research Corporation and Chemical Abstracts Service. The use of both Image and Text search is made possible through the use of a highly sophisticated workstation allowing what the PTO refers to as Full Electronic Search. Full Electronic Search capability via a High Resolution dual screen workstation, allows the Group 220 examiner to search picture images of all of the U.S. Patents (Over 163,000) assigned to areas of technology assigned to Group 220. In addition to the images the examiner may also conduct a full text search of every word of over 1,000,000 U.S. Patents issued since January,1975 and every word of over 1,170,000 English language abstracts of Japanese patents and over 6,000 English language abstracts of Published Chinese patent applications. Examiners in Group 220 and all of the other examining groups may also access from their text terminals or workstations certain commercial data bases.

Approximately 4.6 million of the 4.8 million U.S.Patents have been scanned as digitized images; 163,000 of these patents (over one million images) are loaded on the APS for retrieval by the examiners from optical disks and can be displayed at the workstation ten seconds from the request command. Each page of each retrieved document may be seen at a rate slightly over one second per page if desired by the examiner. High resolution screens allow the examiner to view the printed text and complex drawings at somewhat over 20% of the actual printed page size. Each workstation is equipped with a laser printer allowing the user to walk away with a very high quality paper copy of the patent documents retrieved from the optical disk system. This optical disc system segment of the APS makes the PTO the largest government installation of optical disk technology. Both single disk drive rapid access devices and multi-drive high density, optical juke box devices make up the optical disc system.

In the late fall of 1987 a blue ribbon Industry Review Panel was established by the Secretary of Commerce, headed by