



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

INFORMATION PAPER
ON
ABSTRACTING
IN THE
LICENSING SUPPORT SYSTEM

Office of the Licensing Support System Administrator

September 12, 1990

September 12, 1990

LSSA INFORMATION PAPER ON ABSTRACTING IN THE LICENSING SUPPORT SYSTEM

I. PURPOSE OF THIS PAPER:

At the upcoming October, 1990 meeting of the NRC Licensing Support System Advisory Review Panel (LSSARP), the members are scheduled to continue the discussion on their recommendation to the LSS Administrator (LSSA) on the content of the LSS Header. One open item was the extent to which documents in the LSS should be abstracted. The purpose of this paper is to lay out information about abstracting which the LSSA believes should be taken into consideration by the LSSARP members as they examine this issue.

II. BACKGROUND:

During the March 1990 meeting of the LSSARP, a Technical Working Group was formed to prepare a draft recommendation for the fields for the LSS Bibliographic Header and Full Header. The Working Group met several times and prepared a report to the full LSSARP. The report recommended that abstracts be required only for documents and non-documents that will not be available in searchable full-text (i.e., those with either header only or header and image only). The report further recommended that the abstract field be optional for documents that will be available in searchable full-text. The Technical Working Group determined that the LSSARP should discuss the issue as to which LSS document types or groupings should be abstracted.

During the June 7, 1990 meeting, the LSSARP members agreed that abstracts were required for materials that will not be available in searchable full-text. They then discussed at length the need for an abstract for LSS documents that will be stored in searchable full-text. These discussions centered around cost versus benefit considerations. Differing points were made about:

- the need for any abstract in the header, given availability of full text,
- the sizable cost of abstracting, and
- whether only selected sets of documents might need to be abstracted and, if so, which sets.

No firm recommendation evolved. To focus the issue and to provide more definitive information about the cost implications of alternative abstracting scenarios, the LSSA offered to prepare an issue paper for the members to consider prior to the next LSSARP meeting in October. Since the June LSSARP meeting, the LSSA staff has reviewed existing information science studies related to this issue and gathered industry data on the costs of abstracting. The following is the result of that investigation, including a discussion of abstracting options and some alternatives to abstracting.

III. ABSTRACTING -- WHAT IS IT?

A. TYPES OF ABSTRACTING

In the Library/Information Science discipline, three types of abstracts have evolved. All are based on the human review and summarization of the content of a document. In order of increasing depth and coverage, they are:

- ▶ ANNOTATIVE -- A short description of the document which briefly describes the subject, usually limited to a few lines in length. This type of abstracting can be done by the same staff doing the bibliographic or descriptive cataloging.

- ▶ INDICATIVE -- A longer description than the annotative abstract, giving a more detailed summary of the document scope and content. These abstracts are traditionally about 200 words in length. This type of abstracting is usually done by professional indexers/abstracters having subject matter background and/or experience. The documents are usually reviewed once both for the assignment of subject terms and for the development of the abstract.

- ▶ INFORMATIVE -- The most substantive type of abstracting which includes not only indicative information but also summarizes the findings, answers, or data in the document. Such abstracts often eliminate the need to obtain or read the entire document. The length varies based on depth of document content. As with the indicative abstract, this type of abstracting is also done by professional indexers/abstracters having subject matter background and/or experience.

However, unlike the Indicative Abstracts, this type of abstracting may or may not be done by the same staff that are subject indexing the documents. If not, then another staff resource is required.

It is obviously more expensive as one moves from annotative to informative abstracting because of the additional time and higher level of expertise involved in reviewing the document and composing the abstract. Section IV and Appendix A. contain more information on the cost of abstracting.

B. ABSTRACTING IN THE LSS ENVIRONMENT

Given that the LSS Title/Description field is intended to contain (a) the titles of formal publications or (b) a brief description of less formal or untitled documents, all LSS documents will have annotative-type abstracts. This makes the assumption that titles of publications are somewhat descriptive of content. Therefore, annotative abstracting is not considered from a benefit-costs perspective in this issue paper.

Also, in the opinion of the LSSA, the LSS should not attempt under any scenario to provide informative abstracts because (1) the costs are excessively high and (2) such treatment of LSS documents is unwarranted given the availability of the document text on-line. The LSS abstract would only be intended as a search aid, not as a surrogate for the document itself, which is often the case with systems providing informative abstracts.

Therefore, in discussing the pros and cons of abstracts in the LSS environment, this paper assumes that any abstracts would be of the indicative type.

C. BENEFITS OF INDICATIVE ABSTRACTS

The following is a list of the potential or reputed benefits of having an abstract field in a full-text database. Where applicable, we have included a summary of the information gained from relevant research studies. It should be noted that no specifically applicable research has been found that directly speaks to the benefits/costs of abstracts in a full-text database having keyterms and header data, such as will be the case with the LSS.

1. IMPROVED PRECISION -- The presence and use of abstracts may improve the precision of subject/content searches because it is assumed that if a word or phrase is in the abstract, then it is probably a primary topic of the document. This

precision is gained by limiting word/phrase searches to the abstract field, either initially or after retrieving a document set via search of full-text or other parameters.

There is a current on-going debate in the information science literature about the benefits and power of full-text database software as compared to traditional systems that have only bibliographic (fielded) data, subject indexing, and abstracting. Most of this debate centers around the balance of "recall" versus "precision" capabilities. The attached articles are representative of the discussions and data surrounding this debate (see Attachments #1 through #5).

It is known that in striving to achieve the greatest recall (retrieval of all relevant documents), the precision (retrieval of only relevant documents) of search results suffers. This axiom is applicable to all types of information systems, ranging from bibliographic only to full-text systems. However, the degradation of precision to assure greatest recall is magnified in large full-text systems, especially for collections on a narrow and/or homogeneous topic, such as the HLW LSS. This problem will be further exacerbated in the LSS environment of decision support and litigation support where knowledge of all relevant materials appears more to be essential.

In a 1986 article (Attachment #1), Gerald Salton summarizes the results of several related studies. Simplistically presented, the precision/recall performance of different access methods can be drawn from two of the studies. These data support the belief that searching the abstracts can significantly improve recall (as compared to searching the full-text alone without) a significant loss in precision.

	<u>Recall Ratios*</u>	<u>Precision Ratios*</u>
Searching the:		
a. Text of Abstract	0.78	0.63
b. Controlled Descriptors Subject Indexing	0.56	0.74
c. Full Document Text	0.20	0.75

* Recall Ratio is number of retrieved relevant documents as percentage of all of the relevant documents in the database.

Precision Ratio is the number of retrieved relevant documents as percentage of all retrieved documents

As indicated in line b. above, the recall ratios are better if one has controlled subject terms to search as well as the full-text, without any significant loss of precision. Subject indexing will be done in the LSS.

2. RELEVANCY REVIEW -- Abstracts provide a summary of the entire document. Therefore, browsing the abstracts of a retrieved set of documents can aid in determining the usefulness of the document and the context in which the subject is treated without having to roam around in the text.

Also, abstracts can be very helpful when reviewing document listings or bibliographies in hardcopy away from the LSS workstation. This would be the case when LSS search specialists or intermediaries, e.g. librarians, research assistants, and paralegals, are performing searches in response to "client" requests. In one study, the presence of an abstract reduced the number of "missed documents" -- documents judged as not relevant by a review of the titles only, but which were subsequently determined as relevant after a review of the abstracts (Attachment #6).

3. COST SAVINGS -- Abstracts can potentially reduce the need for printing hardcopy of documents if a review of the abstract is sufficient for the searcher to determine the relevancy of the document for his/her needs.
4. TIME SAVINGS -- Abstracts can reduce on-line time if, as above, review of the abstracts negates the need to browse/read the full-text.

D. LIMITATIONS:

1. Abstracts are only as good as the abstracter. They are subjective, whether it be the author's characterization of his/her work or the abstracter's interpretation of the author's work.
2. Abstracts do not improve recall of subject/content searches in a full-text database if the abstract does not contain different terminology from the text. Different terminology that could improve recall might be more generic, more specific, synonyms, or the translation of jargon.
3. Abstracting only certain document types/categories places a burden on the user to know when abstracting was done and when it was not. Otherwise, users could unknowingly formulate search strategies that would provide false results. For

example, if all documents in a collection are not abstracted, then searches limited to the abstract field will automatically exclude non-abstracted documents and thereby possibly exclude relevant materials from the resulting hitlist.

IV. COSTS OF ABSTRACTING

A. AVERAGE COST PER ABSTRACT

The LSSA collected abstracting cost and productivity information from six companies that perform abstracting services. The information provided by respondents varied in terms of assumptions, such as variations in the size of documents, the QC reviewers/supervision ratios, and scope of abstracting. It was therefore difficult to normalize the data. However, there was not such a disparity in the data that some useful figures could not be compiled. The assumptions used for this paper are listed in the Table below and Appendix A.

Data was also provided by SAIC, based on their experience in the LSS prototype cataloging efforts. Their data show abstracting times of about seven (7) minutes per document based on a sample of 47 documents, each averaging 48 pages. Unfortunately, the SAIC timing estimates did not include a quality control review. Also, it was uncertain whether these times consistently included the actual review and analysis of the document scope and content before the composition and keying of the abstract.

B. ESTIMATED COSTS IN THE LSS

The following table presents the estimated costs of abstracting LSS documents by document type. The figures on the number of documents are extrapolations from recent SAIC re-evaluations of the size of the LSS database (see Attachment #7). The estimated number of pages in this SAIC report was divided by nine (9) to develop an estimated number of documents. The figure of nine (9) pages per document was selected because this was the size of the average document in the DOE Nevada RIS collection, which will contribute the vast majority of documents to the LSS.

The distribution of the estimated number of documents by major document types is based on recent figures from the three major HLW document collection: DOE's RIS systems in Las Vegas and at DOE Headquarters and the NRC's NUDOCS system.

Even though the figures in the table below are just gross estimates and may differ from the actual volume/costs experienced in the future; these figures are based on the best available data. For the purposes of this paper, they do provide the LSSARP members with a significantly improved basis for decision making.

Table 1. ESTIMATED COSTS OF ABSTRACTING IN THE LSS
(Numbers of Documents & Dollars in thousands)

Cumulative Document Counts and Costs by Specified Year

LSS DOCUMENT COLLECTION BY DOCUMENT TYPE	BY 1995		BY 2000		BY 2005	
	NO. OF DOCMNTS	EST. COSTS	NO. OF DOCMNTS	EST. COSTS	NO. OF DOCMNTS	EST. COSTS
TOTAL	1,278	\$33,179	2,296	\$59,595	3,759	\$97,581
CORRESPONDENCE (64%) 3 doc/hour	818	\$17,996	1,469	\$32,318	2,406	\$52,932
PUBLICATIONS/ REPORTS (23%) 2 doc/hour	294	\$9,700	528	\$17,427	864	\$28,512
LEGAL & OTHER DOCUMENTS (13%) 2 doc/hour	166	\$5,483	299	\$9,850	489	\$16,137

Assumptions:

1. A fully loaded rate of \$66.00 per hour. This includes the costs of labor (abstracters, quality control reviewers, and supervisors), G&A, overhead, and fee. Abstracting work activities include reading documents, composing abstracts, keying in the abstracts, and performing quality control and supervision.

2. A production rate of two abstracts developed and reviewed per hour (\$66.00 divided by 2 = \$33/abstract) was used for the Publications/Reports and Legal/Other Document categories. This is the production figure used by the National Federation of Indexers and Abstracters for 200 word indicative abstracts. For correspondence with typically fewer pages than the other two categories, a production rate of three per hour was used (\$66.00 divided by 3 = \$22/abstract).

3. While it is acknowledged that a portion of the LSS documents, particularly formal publications, will have an abstract or summary within the body of the document, no cost reduction was factored into this table. This decision was based on responses of the surveyed abstracting companies. They were reluctant to reduce estimates even if documents contained abstracts, due to the time required to verify the quality of the existing abstract and to edit as required for consistency of coverage with other abstracts. This decision was also supported in the timing tests performed by SAIC in their prototype. Also, no adjustment was made to acknowledge that some documents, such as transmittal correspondence, would not warrant abstracting, given that an annotative summary would be contained in the Title/Description field.

V. ALTERNATIVES TO ABSTRACTING

Section III.C presented the potential benefits of having abstracts in the LSS. This section highlights some of the LSS features currently specified in the SAIC draft LSS Search and Image Design Document which will provide some of the same benefits of abstracting without the continuing costs of abstracting. These software features, if not part of the off-the-shelf database package, can be developed at a finite, one time cost. This section also discusses some other features that could increase precision and recall.

A. CURRENT DOE LSS DESIGN FEATURES

1. Header Field Analysis: After a searcher has developed a hitlist of documents based on his/her search statement, this optional feature, if invoked, would present to the user a computed table of the frequency of occurrences of values for any specified Controlled Vocabulary Header Field. This shows the distribution of Descriptors, Sponsoring Organizations, Author Organizations, etc. within their hitlist.

For example, given the best known search strategy, the user creates a hitlist of 230 documents on boreholes and volcanic rocks. The user then requests the Header Analysis feature, using the Descriptor field. The LSS system would then present a listing of all Descriptors used to describe the 230 and show the number of documents having each descriptor, in decreasing frequency order. The table would look something like:

This query found 230 units.
Header Analysis on Descriptor Field:

<u>Descriptors</u>	<u>Frequency</u>
Fractures	47
Fractures (Geologic)	43
Topopah Springs Member	39
Boreholes	36
Drill Cores	30
Stratigraphy	25
:	
:	
:	
Volcanic Rocks	11
Structural Geology	10
Strain (Geology)	4

The user could use this information about their hitlist to select parameters of greatest or least interest to refine the search statement and create a query with greater precision. For example,

the searcher might now want to broaden the search to include all documents on Topopah Springs Member while also excluding documents on Stratigraphy and Strain.

2. Ranking Retrieved Documents Based on Selected Term Frequency: This LSS feature will allow the user to rank and display the documents in his/her hitlist in decreasing order according to density of selected ASCII-text words in the text. Density is defined as the number of times a relevant words or phrases appear in the document as a percentage of the total number of words in the document. For example, the words abstracts, abstracted, abstracting, and abstracters are repeated about 140 times in this 4,000 word paper. This represents 3.5% of all words in this paper. The percentage would be even greater if "stop" words (such as a, the, were, most, in, etc.) were excluded from the total word count. This process will present the hitlist in an order which provides the most **relevant** documents first on the assumption that if the specified words are repeated frequently in the document, that is a major topic covered in the document.

B. POTENTIAL LSS DESIGN FEATURES

The following are search and retrieval software features that are not currently in the DOE design. These features may warrant further investigation, given the costs of abstracting, the concern of excessively large hitlists, and the problems of low recall and low precision in large text databases.

1.a. Automatic Abstracting -- There are current software packages that purport to scan existing text and present the contents into an abstract-like summary. Such a software feature could be used to add a summary to the LSS header record for presentation to searchers and reviewers of bibliographies to enhance their determination of the relevance of documents retrieved. This would potentially provide the benefits of: (a) reducing the orders for non-relevant documents or (b) finding relevant documents that might have judged non-relevant upon review of the bibliographic information only.

1.b. Optional Extensive Bibliography Format -- LSS users could have the option of ordering the "first" ASCII page of each document in their hitlist to be printed along with a header bibliographic listing. Such a feature would have the same benefits as Automatic Abstracting, described above.

2. Sophisticated Ranking Algorithms -- Over the past several years, the information science literature has contained many articles about research to improve text search results using a variety of statistical and lexical analysis methods. Basically, these are centered on the clustering of related or synonymous terms

and word patterns. Attachments #4 and #8 are examples of such techniques. The capabilities of such software enhancements to improve recall and precision will be carefully monitored. As features become proven, they could be incorporated into the LSS design over the life of the system.

VI. PROS & CONS OF DIFFERENT OPTIONS FOR ABSTRACTING:

A. ALL DOCUMENTS

- PROS: ▶ Consistency and simplicity
- CONS: ▶ Prohibitively Expensive
- ▶ Not warranted for traditional 'correspondence' given:
- ▶ use of Title/Description Field which will provide short annotative summary for relevancy review.
 - ▶ full-text search capability
 - ▶ multiple other access points in the header fields for content/subject searches of all documents, such as descriptors, identifier, project/special class fields etc.

B. ALL NON-CORRESPONDENCE-TYPE DOCUMENTS - "everything but .." Exclude letters, memos, telephone conversation reports...

B.1 Abstract all non-correspondence regardless of how long or short the document.

PROS: ▶ Less expensive than Option VI.A.

CONS: ▶ Somewhat wasteful given that some "short" documents do not warrant such treatment.

B.2 Abstract only non-correspondence over a certain page count.

PROS: ▶ Less expensive than VI.B.1.

- ▶ Increased benefits of relevancy review and precision
- CONS:
- ▶ Selection of document size cutoff is arbitrary and subject to debate.
 - ▶ Searchers are very unlikely to keep this arbitrary rule in mind. Therefore, if they limit their searches to the Abstract Field for precision, then they could unknowingly exclude whole sets of documents and get erroneous search results.

C. ABSTRACT ONLY SPECIFIC DOCUMENT TYPES.

C.1 For All Documents Coded as Specified Document Types -- Pick up Abstracts/Summaries as available within documents or compose and add if not.

- PROS:
- ▶ Less Subjective or arbitrary in the selected universe than VI.B.2.
 - ▶ Much less expensive because of smaller universe of documents to be abstracted.
 - ▶ Most understandable alternative to most, if not all, searchers. Therefore least likely to be misused in searching.
- CONS:
- ▶ Still somewhat subjective in that the assignment of Document Type codes is somewhat subjective.
 - ▶ Inconsistent treatment of abstracts and therefore varying quality if abstracts drawn from the text are not strictly reviewed for consistency with LSS abstracting standards.

C.2 Only Store Abstracts in Headers for Documents which have author-generated Abstracts/Summaries available in the text which can be "grabbed" and put in header as searchable full-text.

- PROS:
- ▶ The least expensive alternative while still allowing searching of this text because submitter's preparation staff and/or LSSA staff do not have to compose and enter the abstract.

- ▶ The abstract listed in bibliographies will assist the reviewer in determining the potential relevance of documents retrieved.
- CONS:
- ▶ Universe of documents which contain abstracts for searching and for presentation is totally random. This does not appear to be a viable option because searchers could not use these randomly existing abstracts with any reliability for identifying relevant documents.
 - ▶ Subjective in determining if document contains text which could be used as an abstract.
 - ▶ Inconsistent treatment of abstracts and therefore varying quality if abstracts drawn from the text are not strictly reviewed for consistency.

C.3 Only Store Abstracts in Headers for Documents which have author-generated Abstracts/Summaries available in the text which can be "grabbed" and put in header but not allow this Abstract field to be searchable.

- PROS:
- ▶ The least expensive alternative. A minimal cost to transfer and store the pre-existing text in the header in a non-searchable field.
 - ▶ The abstract listed in bibliographies will assist the reviewer in determining the potential relevance of documents retrieved.
 - ▶ By not allowing searches to be limited to Abstract Field in this option, it prevents users from unknowingly eliminating potentially relevant sets of documents.
- CONS:
- ▶ This option presents a design issue to be solved because the abstracts in LSS header records that describe documents or data that are not stored in searchable full-text would have to be made searchable.

VII. CURRENT LSSA STAFF VIEW:

The LSSA staff believes strongly that manually prepared abstracts should not be created for inclusion in the Licensing Support System

in searchable text for those documents that are already stored in searchable full-text due to the substantial costs projected for abstracting in comparison to the benefits. Although there is the potential for low recall and precision ratios in large text databases, abstracting is not the only remedy. The other access points in the LSS header fields and the software features specified in the current LSS design will greatly enhance to searchers ability to create useful sets of documents. Also, the LSSA staff will continue to work with DOE in investigating additional software tools to increase performance and will recommend the development of such software if it is a cost-effective approach.

The LSSA staff does believe that the text of abstracts that already exist in documents should be captured in the Full LSS Header. This would be in a non-searchable field to be used for presentation and relevance review only, (Option C.3) above. This assumes the design issue can be solved related to the need to search abstracts for those documents/data not stored in searchable text.

SUMMARY OF INDUSTRY SURVEY OF ABSTRACTING COSTS

APPENDIX A

DIRECT HOURLY LABOR RATES	COMPANY A	COMPANY B	COMPANY C	COMPANY D	COMPANY E	COMPANY F	NFAIS
ABSTRACTERS	\$13.50 - 18.00	\$25.00	\$10.00 - 15.00	Unit Charge	nr	\$12.00	\$13.50
QUALITY CONTROL REVIEWERS	nr	\$25.00	nr	"	nr	nr	nr
SUPERVISORS	\$30.00	\$25.00	nr	"	nr	nr	nr
<hr/>							
RATIO OF QC PERSONNEL TO ABSTRACTERS	1:2	1:5	nr	1:3	1:4	nr	1:4
RATIO OF SUPERVISORS TO ABSTRACTERS	1:20	1:15	nr	1:15	Same Person as QC	nr	nr
<hr/>							
UNIT CHARGE PER ABSTRACT	nr	\$58.50	nr	\$33.29	\$16.77	nr	nr
TIME TO PRODUCE AN INDICATIVE ABSTRACT	20 Pages of doc. per hour	135 mins/ document	nr	49 mins/ 35 page document	37 mins/ 12.5 page document	nr	30 mins/ document

NOTES: nr = not reported
NFAIS = National Federation of Abstracters and Indexers

CALCULATIONS OF FULLY LOADED HOURLY RATEAverage Direct Hourly Rate:

Abstracters	=	\$15.75
QC Personnel	=	20.00
Supervisors	=	27.00

Ratio of QC Personnel to
Abstracters = 1:3.5

Ratio of Supervisors to
Abstracters = 1:15

<u>Abstractor's hourly rate</u>	\$15.75	
+ portion of QC rate	<u>5.71</u>	(\$20 hourly rate for QC personnel divided by 3.5)
	\$21.46	
+ portion of Sup.rate	<u>1.80</u>	(\$27 hourly rate for Supervisors divided by 15)
	\$23.26	
+ Overhead (120%)	<u>27.91</u>	
	\$51.17	
+ G & A (20%)	<u>10.23</u>	
	\$61.40	
+ Fee/profit (8%)	<u>4.91</u>	
	\$66.31	=== Fully loaded hourly rate for abstracting services.

ATTACHMENTS

- #1 Salton, Gerald. "Another Look at Automatic Text-Retrieval Systems." Communication of the ACM. 29(7). 648-656. July 1986.
- #2 Blair, David C. and M.E. Maron. "An Evaluation of Retrieval Effectiveness for a Full-Text Document-Retrieval System." Communications of the ACM 28(3). 289-299. March 1985.
- #3 Tenopir, Carol. "Contributions of Value Added Fields and Full-Text Searching in Full-Text Databases." Proceedings of the National On-Line Meeting - 1985. Medford NJ: Learned Information, Inc., 1985. pp. 463-470.
- #4 Ro, Jung Soon. "An Evaluation of the Applicability of Ranking Algorithms to Improve the Effectiveness of Full-Text Retrieval. I. On the Effectiveness of Full-Text Retrieval." Journal of the American Society for Information Science. 39 (2), 73-78. 1988.
- #5 A. Jordan, John S. Letter to the Editor, Journal of the American Society for Information Science (JASIS) 40(3), 362-363. 1989
- B. Lancaster, F.W. Letter to the Editor, JASIS 40(3), 362. 1989.
- #6 Saracevic, Tefko. "Comparative Effects of Titles, Abstracts, and Full Texts on Relevance Judgements." Proceedings of the American Society for Information Science. Vol. 6 Oct.1-4, 1969. pp. 293-299.
- #7 Science Applications International Corporation. Licensing Support System, Revised Data Scope Analysis. Draft. dated August 28, 1990.
- #8 Deerwater, Scott et. al. "Indexing by Latent Semantic Analysis" Journal of the American Society for Information Science 41(6): 391-407. 1990.