Title
AN SDI SYSTEM BASED ON NSA MAGNETIC TAPES. USER PROFILING AND THE IMPLICATIONS OF DECENTRALIZED INDEXING

Authors
Smith, G.L.
Herr, J.J.
Wakerling, R.K.

Publication Date
1969-12-01

For Symposium on the Handling of Nuclear Information, Vienna, Austria
February 16-20, 1970

UCRL-19290
Preprint
IAEA-SM-128/33
c.2

AN SDI SYSTEM BASED ON NSA MAGNETIC TAPES USER PROFILING AND THE IMPLICATIONS OF DECENTRALIZED INDEXING

G. L. Smith, J. J. Herr, and R. K. Wakerling

December 1969

AEC Contract No. W-7405-eng-48

LAWRENCE RADIATION LABORATORY
UNIVERSITY of CALIFORNIA BERKELEY

DISCLAIMER

AN SDI SYSTEM BASED ON NSA MAGNETIC TAPES. USER PROFILING AND THE IMPLICATIONS OF DECENTRALIZED INDEXING*

G. L. Smith, J. J. Herr, and R. K. Wakerling

Lawrence Radiation Laboratory
University of California, Berkeley, California 94720

INTRODUCTION

Our first work on mechanized selective dissemination of information began in mid-1963 and was based on a keyword index of report titles that was being produced for a semimonthly reports-acquisition list. The keywords were computer-selected from titles, with modifying keywords added by indexers when the titles were completely uninformative. A small but representative group of participants from among the research groups at LRL was invited to participate by providing data upon which system design and evaluation would be based. Information about each participant's subject interests was gathered (a) from his written statement, (b) from responses to questions asked in a structured interview, and (c) from words he selected from a word list of about 8700 terms machine-selected from document titles. For each participant, interest profiles were prepared, three from the data collected by the three methods and a fourth from the combined data. The profiles were matched by an IBM-1401 computer program.

The first phase of the experiment was completed in the spring of 1964 and led to the conclusion that to achieve a desirable level of recall with an acceptably low level of invalid retrievals both language control and some form of coordinate search strategy were necessary.

Fortunately, in the Fall of 1963 discussions between EURATOM and the AEC were begun looking toward a cooperative agreement whereby the material prepared for Nuclear Science Abstracts would be indexed with descriptors from the EURATOM Thesaurus. We began in the spring of 1964 to plan for the use of current NSA input indexed with EURATOM descriptors in an automated selective dissemination system as soon as it became available. The work mentioned above was brought to a close because we believed that the fundamental difficulties presented by the free language of document titles made further pursuit of that path unprofitable.

Before NSA tapes of satisfactory quality could be produced it was necessary for DTIE Oak Ridge to solve a variety of man and machine problems. The difficulties experienced in developing the capability at Oak Ridge to do satisfactory EURATOM indexing, in addition to the regular subject indexing done for the printed NSA, were formidable; they offer valuable insights into some of the problems of decentralized indexing soon to be faced by INIS. Dual indexing of NSA material by DTIE began with Vol. 18 No. 19 (Oct. 15, 1964). The indexing information was keypunched and the punched cards were loaded onto magnetic tape in the EURATOM tape format at the Argonne National Laboratory until Oak Ridge could acquire the necessary facilities. By the Sept. 15, 1965 issue (Vol. 19 No. 17) DTIE was able to take over the production of the NSA tapes at Oak Ridge.

Simultaneously with the work at Oak Ridge we at LRL developed programs for the IBM-1401 computer to make use of the NSA tapes in an experimental SDI system. A search strategy based upon Boolean combinations of descriptors was adopted, and subject profiles were structured accordingly. The programs were first tested in February 1966 in a pilot operation involving ten participants. On the basis of responses from the participants on several subsequent test runs, the profiles were improved to produce output that better matched the wishes of the participants.

In July 1966 the AEC began issuing the NSA tapes in a new format. We had decided to prepare new programs for a large computer (the CDC-6600) rather than to try adapting the IBM-1401 programs. Greater speed of search, increased flexibility and ease of operation resulted. By September 1966 the new programs had been tested and we were prepared to offer SDI service on a regular basis to a small group of users. This experimental operation led us to make some revisions to the programs in the spring of 1967 to allow them to operate faster and more economically. Also, several special programs were devised to provide statistics on operating costs, and data on descriptor usage and category assignment, for use in question formulation and for monitoring the indexing done at DTIE.

All parts of the system were operating satisfactorily, so that regular pilot operation, including gathering of comprehensive statistics, was begun in April of 1968. Procedures based on earlier experience were formalized for routine use in construction and refinement of user profiles. The number of users has been increased gradually to the current total of more than 70.

DESCRIPTION OF THE LRL SDI SYSTEM

The NSA tape for each issue is divided into two parts: The Entry File gives the descriptive cataloging information for each item in the issue in abstract number sequence, the Keyword File contains the EURATOM descriptors (also called selectors) assigned to the items. There are about 10 to 12 descriptors per item. The abstract number is the link between these two files. The bibliographic elements on the tape are described in report TID-4577 (Rev. 3) [1], and the tape format has been described by O'Connor [2].

Our first program in the system converts the NSA tape for an issue to a binary search tape, suitable for use on the CDC-6600 computer, in which all the information on a particular document is combined into one record. It also produces a library printout consisting of the complete bibliographical information and an author index for the issue. Because the tape is customarily available a month in advance of the corresponding issue of NSA, the library printout is a valuable interim library reference tool. (A part of the work done by the conversion program would not be necessary if we were to use an IBM-360 computer.)

The search tape produced is next processed by a "matching" program, which selects from it any documents that satisfy user profiles. The profiles are in the form of search questions (several for each profile) formulated in coded Boolean statements. Query formulation is discussed in the next section. For economy of search time the actual matching is done on the selector I. D. numbers. The search program prints user notifications and accumulates statistical data on the results of the run. A typical notification to a user is shown in Fig. 1. For each item selected the bibliographical data, including the NSA abstract number, and the list of descriptors are given. Each descriptor that caused the item to be selected is marked with a + sign.

Several special-purpose programs are available in addition to the basic conversion and search programs. A previous paper [3] describes the programs involved in the LRL-SDI system. The programs have been made available to several CDC-6600 users, and are currently being employed at the Westinghouse Bettis Laboratory. Tape copies of the LRL programs are always accompanied by the LRL Procedures Manual [4], which outlines the handling and disposition of the system's tapes and printed output.

We are accumulating the search tapes and using them to do retrospective searches on demand. The file extends back to July 1966.

QUERY FORMULATION FOR SEARCHING

Search questions, whether for user profiles or for retrospective searches, consist of exact terms chosen from the Thesaurus to describe the query, grouped together in logical combinations by the operators AND, OR, and NOT. In addition to searching on descriptors and descriptor combinations, we can also search to any level of specificity within sections and subsections of NSA, which is a powerful and useful search aid. Other elements that can be searched are language of the original paper, country of affiliation, corporate code, and journal title (CODEN).

The specific method used for formulating subject searches will be explained by an example. Suppose that one of our SDI users is interested in radiation effects on human bones and tissues. An examination of the EURATOM Thesaurus shows that the descriptors "radiations," "radiation effects," "radiation injuries," "man," "tissues," and "bones" are acceptable terms, so it is permissible to look for documents in which the terms "radiations" or "radiation effects" or "radiation injuries" are associated with the terms "man" or "tissues" or "bone". The statement can be displayed as: (radiations OR radiation effects OR radiation injuries) AND (man OR tissues OR bone) or symbolically as

$$(A_1 + A_2 + A_3) * (B_1 + B_2 + B_3).$$

Furthermore, we may want to reject documents that relate to radiation effects on plants or insects. This can be done by adding the statement

NSA/SDI NOTIFICATION

LISTED BELOW ARE THE DOCUMENTS SELECTED FOR YOU BY SDI.
KEYWORDS PRECEDED BY (+) ARE THOSE YOU HAVE CHOSEN TO SELECT
DOCUMENTS. PLEASE FILL IN THE LAST PAGE OF THIS NOTIFICATION.

99 LEFOG, LEROY L. BLDG 508 RM 4206 X6308
NSA 23(21) NOVEMBER 15, 1969

43506 NSA 23(21) JOURNAL

EFFECTS AND PROTECTION OF RADIATION FROM ATOMIC FACILITIES. 2. DISPOSAL OF RADIOACTIVE WASTES TO OCEAN.
HIYAMA, YOSHIO- SHIMIZU, MAKOTO (TOKYO UNIV.). GENSHIRYOKU KOGYO, 15- NO. 3, 9-13(MAR. 1969). (IN JAPANESE).

CAT. 24 ENGINEERING / 70 RADIOACTIVE MATERIAL HANDLING

ASIA	FISH
+MAN	MONITORING
+RADIATION EFFECTS	RADIATION PROTECTION
RADIOACTIVITY	SAFETY
SEA	WASTE DISPOSAL
WATER	JAPAN
RADIOACTIVE WASTES	

43612 NSA 23(21) JOURNAL

GLASS DOSIMETER FOR MEASURING THE ABSORBED DOSE IN CRITICAL ORGANS.
YOKOTA, RYOSUKE- MUTO, YUHEI (TOKYO SHIBAURA ELECTRIC CO.). HOKEN BUTSURI, 4- 497-501(JUNE 1969). (IN JAPANESE).

CAT. 26 INSTRUMENTATION / 20 RADIATION DOSIMETERS

ABSORPTION	BODY
DOSEMETERS	GLASS
LUMINESCENCE	+RADIATIONS
+TISSUES	ORGANS
PHOTOLUMINESCENCE	

43673 NSA 23(21) BOOK / THESIS

MATERIALY PO TOKSIKOLOGII RADIOAKTIVNYKH VESHCHESTV SERA-35, KAL'TSII-45, FOSFOR-32. VYPUSK 6. (MATERIALS ON THE TOXICOLOGY OF RADIOACTIVE MATTER (/SUP 35/S, /SUP 45/CA, /SUP 32/P). NUMBER 6).
LETAVET, A. A. (ED.). MOSCOW, IZDATEL'STVO MEDITSINA, 1968. 168P.

CAT. 28 LIFE SCIENCES / 13 BIOCHEM., ETC. / METABOLISM, PHYSIOL., + TOXIC.

ALBUMINS	ANIMALS
+BLOOD	BLOOD FORMATION
BONE MARROW	+BONES
BRAIN	CANCER
EYES	GLANDS
GONADS	INJECTION
LEUCOCYTES	METABOLISM
NUCLEIC ACIDS	RADIATION DOSES
+RADIATION INJURIES	RADIATION SICKNESS
TIME	TOXICITY
CALCIUM 45	PHOSPHORUS 32
SULFUR 35	PITUITARY GLAND
SARCOMAS	TESTES
DOSE RATES	PHAGOCYTOSIS

Fig. 1. Notification to a user.

"and NOT (plants OR insects)." The query would then be symbolized as

$$[(A_1 + A_2 + A_3) * (B_1 + B_2 + B_3)] - (C_1 + C_2).$$

This statement when properly formatted for machine search would appear as follows in the user's profile

<u>Group 1</u>
Radiation effects
Radiation injuries
Radiations

<u>Group 2</u>
Bones
Man
Tissues

<u>Group 6</u>
Insects
Plants

The program provides for 15 groups for one question. Groups 1-5 can be combined with AND in a positive request; groups 6-10 are available for negation; and groups 11-15 can be used to add simple term combinations in order to save computer running time. The number of terms within a group is essentially unlimited. A profile may contain as many as 99 questions.

Groups 11-15 are not often used. They are reserved for terms commonly found together as "liquid" and "nitrogen," or "nuclear" and "cross sections." Use of two of these groups is illustrated as follows. Let us extend the example by supposing that the user is also interested in the use of tracer techniques in studying human blood. In recognition that some indexers may consider "labeling" as synonymous with "tracers," we could add two more questions to the profile:

<u>Group 1</u>	<u>Group 1</u>
Labeled compounds	Tracer techniques
<u>Group 2</u>	<u>Group 2</u>
Blood	Blood
<u>Group 3</u>	<u>Group 3</u>
Man	Man

These questions are closely related to the first one, so in the interest of saving computer processing time we can combine them all into a single statement which would be symbolized as

$$\{[(A_1 + A_2 + A_3) * (B_1 + B_2 + B_3)] - (C_1 + C_2)\} + [(D_1 * E_1 * B_2) + (F_1 * E_1 * B_2)]$$

The computer printout of the above complex question is shown as Fig. 2.

Statistics on the frequency of use of descriptors in NSA indexing are a valuable aid in preparing search questions. Heavily used terms must be combined with others to avoid a useless flood of output, while infrequently used descriptors can be used in single-term searches, as in question 2 on Fig. 2.

As mentioned above, searches can be done on elements provided on the NSA tapes besides subject descriptors. This is illustrated by the profiles (Fig. 3) synthesized for two hypothetical SDI users Mr. Doe and Mr. Moe.

```
        PROFILE 59       99   LEFUG, LEROY L.    BLDG  50A RM  4206 X6309                          99000
         LANGUAGE ALL                                                                         99001

                          WORD                                   I.D. NO.     TYPE     COUNT      AVE.
QUESTION    1 HAS   14 TERMS
------------

                    GROUP  1          3 TERMS
                         RADIATION EFFECTS                          3925        1       5357       223
                         RADIATION INJURIES                         3926        1       1167        49
                         RADIATIONS                                 3930        1       1508        63

                    GROUP  2          3 TERMS
                         BONES                                       599        1        443        18
                         MAN                                        2757        1       2311        96
                         TISSUES                                    5059        1       1079        45

                    GROUP  6          2 TERMS
                         INSECTS                                    2256        1        252        10
                         PLANTS                                     3559        1        605        25

                    GROUP 11          3 TERMS
                         LABELLED COMPOUNDS                         2476        1        586        24
                         BLOOD                                       582        1        326        14
                         MAN                                        2757        1       2311        96

                    GROUP 12          3 TERMS
                         TRACER TECHNIQUES                          5105        1        663        28
                         BLOOD                                       582        1        326        14
                         MAN                                        2757        1       2311        96
QUESTION    2 HAS    4 TERMS
------------

                    GROUP  1          4 TERMS
                         ACETYLCHOLINE                              5676        9         10      LT 1
                         ACETYLCHOLINESTERASE                      14584        9          0      LT 1
                         CHOLINE                                   11249        9         10      LT 1
                         CHOLINESTERASE                            10842        9         12      LT 1
```

Fig. 2. Example of a user profile.

```
        PROFILE 57       80   DOE, J.Q.          BLDG  50B RM 4206   X6368                         90000
         LANGUAGE ALL

                          WORD                                   I.D. NO.     TYPE     COUNT      AVE.
QUESTION    1 HAS    4 TERMS
------------

                    GROUP  1          2 TERMS
                         ACCELERATORS                                  3        1        458        19
                         34 PHYSICS (HI-ENG.) / 60 PARTICLE ACCELERATORS                                 3460

                    GROUP  2          1 TERMS
                         AFFILIATION  UR                           32522

                    GROUP  6          1 TERMS
                         CODEN --- PRTEA                      2022240501
QUESTION    2 HAS    2 TERMS
------------

                    GROUP  1          1 TERMS
                         28 LIFE SCIENCES / 62 RADIATION EFFECTS ON ANIMALS / VERTEBRATES              2862

                    GROUP  2          1 TERMS
                         CORP. CODE ---     639000
```

```
        PROFILE 58       90   MOE, I.R.          BLDG  50B RM 4206   X6368                         90000
         LANGUAGE ITALIAN   RUSSIAN                                                          90010

                          WORD                                   I.D. NO.     TYPE     COUNT      AVE.
QUESTION    1 HAS    6 TERMS
------------

                    GROUP  1          2 TERMS
                         ACCELERATORS                                  3        1        458        19
                         34 PHYSICS (HI-ENG.) / 60 PARTICLE ACCELERATORS                                 3460
                    GROUP  2          2 TERMS
                         PROTON BEAMS                               3884        1        848        35
                         SYNCHROTRONS                               4685        1        312        13
                    GROUP 11          2 TERMS
                         PROTONS                                    3885        1       2524       105
                         BEAMS                                       444        1        700        29
QUESTION    2 HAS    3 TERMS
------------

                    GROUP  1          1 TERMS
                         BIBLIOGRAPHY                                523        1        640        27
                    GROUP  2          1 TERMS
                         REACTOR SAFETY                             4025        1        927        39
                    GROUP  3          1 TERMS
                         CORP. CODE ---    6171000
```

Fig. 3. Profiles for two hypothetical users.

PROFILING

By profiling we mean the gathering of information about the user's subject interests in the technical literature, and the preparation of search questions that will select from the data base the items that are pertinent to these interests. Good subject profiling is the key to satisfactory SDI service. Its importance has not been stressed adequately in the literature on SDI services. If the user's SDI profile is not good he will not be satisfied for long, no matter how fast the system operates or how beautiful the output looks. Because the documents in the NSA data base are indexed by subject experts using a well tested and controlled indexing vocabulary, we believe that it should be possible by careful profiling to produce high quality output for the users of the service.

There are four important steps in profiling;

a. Gathering the information on the user's interests.
b. Structuring this information into search questions.
c. Gathering and evaluating user response.
d. Refining the profile by use of the information from the response.

Steps b, c, and d can be recycled until the user is satisfied with the quality and quantity of the output he gets from the system. It must also be recognized that profiles are not static: changes in user interests must be reflected in corresponding changes in their profiles.

The first step, that of data gathering, may be carried out in a variety of ways. We have used written statements from the users, structured interviews, questionnaires, selection of Thesaurus keywords jointly by user and profiler, and combinations of these. We believe that a well designed interview technique is the best. At the time of the interview the general features of the SDI system are described to the user, and he is given information about the content of the NSA tapes. The interviewer points out that the documents on the tapes are indexed by experienced subject specialists on the basis of terms from the EURATOM Thesaurus, not on words from titles or abstracts. It is emphasized that his profile search questions will be framed in terms selected from this same indexing vocabulary. He is shown sample profiles and SDI notifications. A number of questions are then asked to get information on the user's subject interests, his use of the literature, what secondary sources he uses, documents he has written recently, etc. A more detailed description of the procedure followed is given in Appendix A.

The second phase in profiling is the structuring of the information gathered from the user into search questions. It is very important that the profiler be completely familiar with the EURATOM Thesaurus and its use, and with the NSA categorization scheme. Experience as an indexer is very helpful. In addition to the material gathered at the interview or from other sources, the profile makes use of the EURATOM Terminology Charts and statistical data on the frequency of use of indexing terms in NSA. We have evolved the procedure for profile structuring that is given in Appendix B.

After the draft profile has been put into form for machine search it is run against an NSA tape, and the notifications produced are examined. The results are used to make any obvious improvements in the profile. The improved profile is run and the resulting selection from the sample NSA tape and the profile are reviewed with the user. His response is used as the basis for further profile refinement. Subsequent to making the profile changes based upon this first feedback from the user, the user is added to the regular notifications service system. Henceforth he is sent semimonthly notifications routinely by mail. An evaluation form regularly accompanies the notifications. Evaluation forms returned by the user are employed in making further refinements in his profile. At our Laboratory the average user profile has five or six questions with a total of about 35 terms.

We have developed a procedure for profile refinement upon the basis of our experience with local users. Most problems fall into five categories.

a. Too many citations (over 50) are selected by the profile. One first checks the frequency count list to determine the high-frequency terms and tries to limit their effect, either by separating them into different groups, or by replacing them by low-frequency terms, or by adding terms or categories as restricting measures. Also, one looks for citations not in the user's field of interest and identifies the terms that produced these citations. These terms can then be eliminated or replaced by other terms, category restrictions can be added, etc.

b. Too few citations are selected by the profile. In this event one can remove limiting or restricting terms and categories, add categories as single-term searches, or break up combinations of terms into single-term searches.

c. Questions are redundant--i. e., the same citation is selected by more than one profile question. This problem can usually be solved by compressing the various term combinations into a single profile question.

d. Citations are selected by only a portion of the profile questions. The remedy is to treat the low-producing profile questions by the procedures given under item b above.

e. There are too many "no interest" evaluations, as indicated by evaluation sheets from the user. Our approach in this case is to make a tabulation of the no-interest citations and examine the reasons for their selection. The index terms in the profile that produced these citations are studied with a view toward either eliminating them replacing them by others, or combining them with other terms.

The refinement process may be recycled as many times as required. Follow-up interviews may be necessary to help in eliminating particularly difficult snags.

All users are requested to notify the SDI system operators of changes in their interests that would necessitate changes in their interest profiles.

ECONOMICS OF THE SYSTEM

Cost information is gathered on the operation of the system. For example, we record the computer time required to prepare the search tape from the semimonthly NSA tape supplied by the AEC. This is one of our largest items of cost because of the amount of processing required to make a tape usable on our CDC-6600 computer from the tape prepared on an IBM-360 computer at Oak Ridge. The cost of preparing the search tape depends on the number of items in the corresponding issue of NSA, but is independent of the number of users in our SDI system. It averages about $42 per issue, based on our computer charge rate of $155 per hour. The library author index is prepared at the same time, but the cost is small and so has not been separated out.

The average cost of running the search-sort routines is about $0.90 per profile per issue of NSA for a user group of 70. Because the system is experimental we have placed no limit on the number of questions or the number of terms in profiles. The largest profile contains 15 questions with a total of 1084 terms: the smallest profile consists of one question with a total of one term. The issues of NSA also vary in size from about 1800 to about 3000 items, averaging about 2200 items.

The total cost for the first 100 users is currently averaging about $1.70 per user per issue of NSA, or $3.40 per month. The total cost per user decreases with the number of users because the cost of preparing the search tape is spread over a wider base. There is no cost included for the input tapes because they are provided free of charge by the AEC.

There are several approaches one might take to reduce the cost of SDI service. For example, a group profile could be used to replace the individual profiles of users working in small, tightly knit research groups. Or limitations could be placed on the number of questions or the number of terms in a profile, or both. Another possibility is to have several options available so that the user can choose which quality of service fits his information needs and his pocketbook. NASA management came to the conclusion that their centralized SDI service was too costly and consequently decided to offer NASA SCAN in its place. We believe that some of the alternatives mentioned above in a decentralized system would reduce costs and produce a better service than SCAN. We plan to investigate this matter.

We have only approximate costs for the other operations, such as profile data gathering, profile formulation, and analysis of user response, because the system is experimental and some of these operations have been mixed with the development of procedures and the study of the indexing quality. However, we can give some approximate figures for the amount of time required for a documentalist to perform these operations.

Interview and data gathering on user's interests	2 hours
Preparation of the draft profile	4 hours
Testing and refinement of the profile	4 hours

As yet we do not have much data on the cost of doing retrospective searches with the same programs. The questions can vary so widely in such respects as their complexity and the time span to be searched that it is difficult to determine average or typical search costs.

DECENTRALIZED INDEXING IN RELATION TO SDI SERVICE

The knowledge and skill of the information scientist who prepares SDI profiles and retrospective search questions contribute much to the quality of the search results. We believe that an important benefit of the decentralized input plan of INIS is that trained indexers will be available in many places to assist with profiling and question formulation for machine searching.

Our experimental work on decentralized input to NSA has convinced us of the value to an SDI system of having the information scientists who are responsible for the preparation of input to the data base also be concerned with the retrieval of information. They have a thorough knowledge of the indexing vocabulary and guidelines for input to the data base. Their experience in the analysis of the subject content of scientific documents is directly applicable in analyzing the data on SDI user's scientific interests, and in formulating SDI profile search questions with terms chosen from the indexing vocabulary.

The indexer also profits from this association with the information user. He sees first hand how his indexing influences the results of searches. He can get the direct reaction of the user on such matters as the quality of the input and output in terms of the scope of coverage, the timeliness, and the adequacy of the subject analysis and categorization.

In the course of our study of problems of decentralized input to NSA we are working on aids to indexers. Some of them can also be of assistance to the searcher. One such aid is a collection of subject-centered vocabularies to supplement the Thesaurus. Subject-centered vocabularies are concentrated lists of 400 to 600 Thesaurus terms that form the core vocabularies for various subject areas. Because the terms useful for a given subject are extracted from the alphabetically arranged 12 000-term, multidisciplinary EURATOM Thesaurus, the specific terms needed for a given document are more available to the indexer or profiler than they are in the full thesaurus. In addition to their usefulness to indexers and profilers, these concentrated vocabularies can be formatted to make possible the use of mark-sensing readers as input devices for keyword indexing, thereby reducing input cost. Because only the positions of the marks are important in mark sensing, the vocabulary lists could be in any language.

A variety of forms can be used for subject-centered vocabularies. For profile preparation, we have employed a straight alphabetical listing, from which easily remembered nonconcept terms, such as chemicals, have been removed. When the hierarchical posting is added to the alphabetical lists, microthesauri are produced as a further refinement. Perhaps the most valuable are the categorized lists, in which similar terms are clustered. An example of such a categorized list, in a form suitable for use as an indexing work-sheet for the Atomic and Molecular Physics subsection, is shown in Fig. 4. On this form the hierarchical posting has not been included: such a task is easily done by computer. If properly prepared, this form can make it easier for the indexer and profiler to find the appropriate index terms.

Computer programs have been written to generate subject-centered vocabularies from our file of NSA tapes. In addition to the lists of terms, the programs give the frequencies of term use. Several forms of output are available. The programs have been used to produce 17 vocabularies, some of which have been applied to preparation of search profiles. All lists of terms studied have small enough core group of terms that it is possible to make up subject-centered vocabulary check lists that are compact enough to be useful. Furthermore, the vocabularies change slowly enough that revisions need be made only rather infrequently.

The advantages to SDI service of decentralized input to the data base are realized irrespective of whether the actual machine searching is carried out locally or centrally, as long as the formulation and refinement of the profiles are decentralized.

Document Identification Number ____________

Atomic and Molecular Physics Vocabulary Check List (Subsection 32.2)

Categories:

- ____ 32.21 Beams--properties of and interactions with solids
- ____ 32.22 Properties (including spectroscopic) of atoms and molecules--measurement and calculation of
- ____ 32.23 Mesic and muonic atoms and molecules; positronium and muonium
- ____ 32.24 Collision phenomena; gas-phase reactions
- ____ 32.25 Atomic and molecular theory--general theory

COMMON MATERIALS

- ALKALI METALS
- ALUMINUM
- AMMONIA
- ARGON
- BERYLLIUM
- BORON
- CARBON
- CARBON DIOXIDE
- CESIUM
- DEUTERIUM
- DEUTERIUM COMPOUNDS (to indicate deuterium substitution)
- HELIUM
- HELIUM 4
- HYDROGEN
- KRYPTON
- LITHIUM
- MERCURY
- METHANE
- NEON
- NITROGEN
- NITROGEN OXIDES
- OXYGEN
- POTASSIUM
- RARE EARTHS
- RARE GASES
- RUBIDIUM
- SODIUM
- WATER
- XENON

PARTICLES, ATOMS, IONS, MOLECULES

- ANIONS
- ATOMS
- BETA PARTICLES
- CATIONS
- CHARGED PARTICLES
- DEUTERONS
- ELECTRONS
- HEAVY IONS
- IONS
- MESIC or MUONIC ATOMS
- MESIC or MUONIC MOLECULES
- MUONIUM
- PHOTONS
- POSITRONIUM
- POSITRONS
- PROTONS

BEAMS (use for beams and for incoming particles)

- ATOMIC BEAMS
- DEUTERON BEAMS
- ELECTRON BEAMS
- ION BEAMS
- MOLECULAR BEAMS
- MUON BEAMS
- PROTON BEAMS
- POSITRON BEAMS

ENERGY RANGES

- EV RANGE
- KEV RANGE
- MEV RANGE
- MILLI EV RANGE

STATES OF MATTER

- CRYSTALS (list structure if possible)
- FOILS
- MONOCRYSTALS
- POWDERS
- SOLIDS
- SURFACES
- LIQUIDS
- GASES
- VAPORS
- AFTERGLOW
- ELECTRIC DISCHARGES
- GLOW DISCHARGES
- PLASMA
- POSITIVE COLUMN
- ADSORPTION (to indicate adsorbed materials)
- MIXTURES
- SOLUTIONS

RADIATIONS

- BREMSSTRAHLUNG
- GAMMA RADIATION
- INFRARED RADIATION
- LIGHT (visible radiation)
- MICROWAVES
- RADIO WAVES
- ULTRAVIOLET RADIATION
- X RADIATION

FIELDS

- ELECTRIC CHARGES
- ELECTRIC FIELDS
- ELECTROMAGNETIC FIELDS
- ELECTROSTATIC FIELDS
- MAGNETIC FIELDS

ENERGY LEVELS AND SPECTRA

- EXCITED STATES
- METASTABLE STATES
- P-STATES
- ROTATIONAL STATE
- S-STATES
- VIBRATIONAL STATE
- COUPLING
- ENERGY LEVEL MIXING
- INTERACTIONS
- LEVEL CROSSING
- LEVEL WIDTHS
- LIFETIME
- ABSORPTION
- DIPOLE TRANSITIONS
- EMISSION (photons or electrons)
- EXCITATION
- FLUORESCENCE
- QUADRUPOLE TRANSITIONS
- SPECTRA
- TRANSITIONS
- AUGER EFFECT (autoionization)
- BALMER LINES
- COSTER-KRONIG TRANSITIONS
- LYMAN LINES
- RYDBERG EQUATION

- EINSTEIN COEFFICIENTS
- FRANCK-CONDON PRINCIPLE
- OSCILLATOR STRENGTH
- SELECTION RULES
- FINE STRUCTURE
- HYPERFINE STRUCTURE
- ISOTOPE SHIFT
- LAMB SHIFT
- SPECTRAL SHIFT
- STARK EFFECT
- ZEEMAN EFFECT
- DOPPLER EFFECT
- LINE BROADENING
- WIDTH (line width)

ELECTRONIC STRUCTURE

- ATOMIC MODELS
- BORN-MAYER EQUATION
- BORN-OPPENHEIMER APPROXIMATION
- CONFIGURATION INTERACTION
- CORRELATION ENERGY
- DENSITY MATRIX
- ELECTRON CORRELATION
- GAUSSIAN ORBITALS
- GOLDSTONE DIAGRAMS
- HARTREE-FOCK METHOD
- HEITLER-LONDON THEORY
- HYLLERAAS COORDINATES
- L-S COUPLING
- MULLIKEN ANALYSIS
- OCCUPATION NUMBER
- ORBITALS
- QUANTUM NUMBER
- RESONANCE
- SELF-CONSISTENT FIELD
- SLATER DETERMINANT
- SLATER INTEGRALS
- SLATER ORBITALS
- SPIN-ORBIT INTERACTION
- SPIN-SPIN INTERACTION
- THOMAS-FERMI MODEL
- WAVE FUNCTIONS

ATOMIC AND MOLECULAR PROPERTIES

- ATOMIC NUMBER
- BOND ANGLE
- BOND LENGTHS
- CORIOLIS FORCE
- DIPOLE MOMENTS
- DISSOCIATION ENERGY
- ELECTRIC DIPOLE MOMENTS
- ELECTRON AFFINITY
- ELECTRON DENSITY (also plasma)
- INTERATOMIC DISTANCES
- IONIZATION POTENTIAL
- MAGNETIC MOMENTS
- NUCLEAR ALIGNMENT
- NUCLEAR MAGNETIC MOMENTS
- NUCLEAR SCREENING
- POLARIZABILITY
- QUADRUPOLE MOMENTS
- WORK FUNCTIONS

GENERAL PROPERTIES

- ANGULAR MOMENTUM
- ENERGY
- ELECTRON TEMPERATURE
- FREQUENCY
- KINETIC ENERGY
- MASS
- MOMENTUM
- TRAJECTORIES
- VELOCITY
- ANGULAR CORRELATION
- ANGULAR DISTRIBUTION
- ANISOTROPY
- ASYMMETRY
- COUPLING
- COUPLING CONSTANTS
- DEPOLARIZATION
- DISTANCE
- DISTRIBUTION
- EFFICIENCY
- MOTION
- ORIENTATION
- POLARIZATION
- RESONANCE
- TIME DEPENDENCE
- BINDING ENERGY
- DIPOLE MOMENTS
- ELECTRIC DIPOLE MOMENTS
- GYROMAGNETIC RATIO
- LANDE FACTOR (g-factor)
- MAGNETIC MOMENTS
- MULTIPLICITY
- OCCUPATION NUMBER
- QUADRUPOLE MOMENTS
- QUANTUM NUMBER
- SPIN
- WAVE FUNCTIONS

CONDITIONS

- ABUNDANCE (concentration)
- HIGH TEMPERATURE
- LOW TEMPERATURE
- PRESSURE
- TEMPERATURE
- THICKNESS

COLLISION PROCESSES

- ELASTIC SCATTERING
- GLORY EFFECT
- INELASTIC SCATTERING
- RAINBOW SCATTERING
- RAMSAUER EFFECT
- RESONANCE SCATTERING
- SCATTERING
- SMALL ANGLE SCATTERING
- CHEMICAL REACTIONS (rearrangement reactions)
- DECOMPOSITION
- ISOTOPIC EXCHANGE
- AUGER EFFECT (autoionization)
- CHARGE EXCHANGE (electron transfer)
- ELECTRON ATTACHMENT
- ELECTRON DETACHMENT
- ION-ELECTRON RECOMBINATION
- ION-ION RECOMBINATION
- IONIZATION
- PENNING EFFECT
- PHOTOIONIZATION
- ABSORPTION
- ANNIHILATION
- CAPTURE
- EMISSION (electrons or photons)
- ENERGY TRANSFER
- EXCITATION
- FINAL-STATE INTERACTIONS
- INTERACTIONS
- RELAXATION
- TRANSITIONS
- COMPTON EFFECT
- RAYLEIGH SCATTERING
- BOMBARDMENT
- SECONDARY EMISSION
- SLOWDOWN
- SPUTTERING

WAVE PROPAGATION

- DIFFRACTION
- INTERFERENCE
- MODULATION
- OSCILLATIONS
- REFRACTION

COLLISION PROPERTIES

- CROSS SECTIONS
- DIFFERENTIAL CROSS SECTIONS
- EXCITATION FUNCTION
- TOTAL CROSS SECTIONS
- ACTIVATION ENERGY
- ENERGY LOSSES
- Q-VALUE
- THRESHOLD ENERGY
- ISOTOPE EFFECTS
- PROBABILITY
- PRODUCTION
- REACTION KINETICS (rate constants, order of reaction)
- YIELD
- IMPACT PARAMETER
- PHASE SHIFT
- SCATTERING AMPLITUDE
- SCATTERING LENGTH
- BOUND STATE
- TRANSIENTS
- TRANSITION STATE

INTERACTION POTENTIALS

- CENTRAL POTENTIAL
- HARD-SPHERE POTENTIAL
- HARMONIC POTENTIAL
- INTERMOLECULAR FORCES
- LENNARD-JONES POTENTIAL
- LONDON FORCES
- MORSE POTENTIAL
- NONCENTRAL FORCES
- VAN DER WAALS FORCES

COLLISION THEORY

- BORN APPROXIMATION
- DWBA APPROXIMATION
- IMPULSE APPROXIMATION
- WKB APPROXIMATION

GENERAL THEORY

- ADIABATIC APPROXIMATION
- CORRESPONDENCE PRINCIPLE
- EXCHANGE INTERACTIONS
- EXPECTATION VALUE
- FIELD THEORY
- HAMILTONIAN FUNCTION
- MANY BODY PROBLEM
- NEWTON MECHANICS (classical mechanics)
- P-WAVE
- PARTIAL WAVES
- PERTURBATION THEORY
- QUANTUM MECHANICS
- RAYLEIGH-SCHROEDINGER FORMULA
- RELATIVITY THEORY
- S-WAVE
- SCHROEDINGER EQUATION
- T INVARIANCE
- TUNNEL EFFECT
- VARIATIONAL METHOD

MATHEMATICS

- ANALYTICAL SOLUTION
- ASYMPTOTIC SOLUTIONS
- BOUNDARY CONDITIONS
- CORRECTIONS
- CORRELATION FUNCTION
- EIGENFUNCTIONS
- EIGENVALUES
- EIGENVECTORS
- EQUATIONS (only for new equations)
- GAUSS FUNCTION
- GREEN FUNCTION
- GROUP THEORY (list specific groups if possible)
- HERMITIAN OPERATORS
- HILBERT SPACE
- INTEGRALS
- ITERATIVE METHODS
- LAGUERRE POLYNOMIALS
- LIE GROUPS
- MATRICES
- NUMERICAL SOLUTION
- OPERATORS
- POISSON EQUATIONS
- PROBABILITY
- RACAH COEFFICIENTS
- ROTATION GROUP
- SERIES EXPANSION
- SPHERICAL HARMONICS
- WHITTAKER FUNCTIONS
- YOUNG DIAGRAM

DEVICES

- DRIFT TUBES
- ELECTRON SOURCES
- ION SOURCES
- LASERS
- MAGNETOMETERS
- MASS SPECTROMETERS
- SPECTROMETERS

TECHNIQUES

- COINCIDENCE METHODS
- DETECTION
- MEASUREMENT
- MONITORING
- OPTICAL PUMPING
- SPECTROSCOPY

GENERAL TERMS

- BIBLIOGRAPHY (also review)
- DESIGN
- OPERATION
- TABLES

Fig. 4. Typical indexing work sheet.

CONCLUSIONS

An information research project in SDI and decentralized indexing at the Lawrence Radiation Laboratory in Berkeley has resulted in the following developments:

1. SDI programs have been developed and are in use with input tapes of Nuclear Science Abstracts to select and disseminate printed references of interest to Laboratory scientists.
2. Techniques have been produced and tested for developing user profiles and for subsequently refining and updating them.
3. Current computer (CDC-6600)—costs are around $40 per year per SDI user.
4. Knowledge and experience gained in the process of indexing for the NSA tape system are being applied effectively to SDI retrieval. Subject-centered vocabularies promise valuable aid to indexers and searchers alike.

APPENDIX A. PROCEDURE FOR INITIAL SDI INTERVIEW

1. Describe the SDI system to the prospective user and show him some sample output.

2. Explain a sample profile to him in detail. He should understand how it works before going to the next step. Emphasize that the documents on the NSA tapes are indexed on the basis of terms from the EURATOM Thesaurus, not on words from titles or abstracts, and that the SDI selection process is based on this same indexing vocabulary.

3. Describe the NSA category system, and explained to the user that categories can be used in addition to or in combination with index terms for searching. Choose category limitations for the user's profile.

4. Explore the user's familiarity with NSA. If he uses it, does he scan it for current awareness, or does he use it for retrospective searching only? If he scans it, does he restrict himself to certain specific subject categories? How many documents of interest does he usually find? How many machine selections will he consider reasonable to scan? This information, though not vital to the profile, gives guidance on what is a suitable amount of output for that user.

5. Ask the user whether he plans to share the SDI output with members of the group he works in, or to use it personally only. In the first case explore the size and general responsibility of his group, and his particular work within the group.

6. Next seek detailed information on the user's subject interest. To begin with, he is requested to describe his work, i.e., projects or programs he is working on. The interviewer may ask him to elaborate on each topic or point, and may ask questions such as the following;

a) What subject fields or discipline are pertinent to your work?

b) What specific materials do you work with?

c) What methods or processes do you work with?

d) What experimental environments are important to your work?

e) What aspects of each topic are you <u>not</u> interested in?

In recording the information the interviewer takes particular note of all words the user employs that may be index terms.

7. The user is asked to show any journal articles, reports, and books he has read recently that represent the kinds of documents he would like to be informed of.

8. What areas of information of interest to the user are, in his opinion, inadequately covered in the literature?

9. Information about any articles, reports, or books the user is in the process of writing is asked for.

10. The interviewer asks the user to thumb through an issue of NSA and pick out documents of interest in each of the subject areas he has described. If the reason for the choice is not clear, the matter is discussed.

11. The special search elements available are explained and the user is asked to specify any requirements on language of the original documents, corporate authors, countries of affiliation, or journal titles.

APPENDIX B. PROCEDURE FOR PROFILE CONSTRUCTION

1. List each topic or subtopic mentioned by the user.

2. Start with the first topic and select the term or terms from the Thesaurus that best describe this topic. Avoid extremely general terms and favor terms of narrower scope.

3. Look up each term in the frequency list for its frequency count and identification number. If the frequency of occurrence is low (less than 10 per issue), consider using the term in a one-group question--i.e., not pairing it with other terms. If the frequency is high (more than 10 per issue), it should be paired in a logical product with another term.

4. Group the terms into Boolean statements, preferably two groups at first, thus forming the first question.

5. Select additional terms which can be grouped with the above terms as alternative selections in the Boolean statements.

6. Make several separate questions rather than one complex one; they can be combined later. It is easier to detect errors or the need for additional terms in shorter questions.

7. "NOT" terms should be handled very cautiously at first. They can be added later as necessary.

8. Repeat the above steps for each remaining topic of interest to the user to complete the first draft of his profile.

9. Look up the indexing for any documents the user picked from NSA to discover any terms that should be added to the profile to retrieve these pertinent documents.

REFERENCES

[1] Division of Technical Information, USAEC, Descriptive Cataloging Guide, Report TID-4577 (Rev. 3), Dec. 1968.

[2] O'CONNOR, Joel S., "AEC Division of Technical Information Tape Distribution System," presented at the Annual Meeting of the American Documentation Institute, New York, Oct. 1967.

[3] SMITH, Gloria L., "AEC Tapes -- User Experiences," presented at the Annual Conference of the Special Libraries Association, Los Angeles, June 1968.

[4] SMITH, Gloria L., RAYMOND, Marcus R., HEALEY, Robert N., Procedures Manual of SDI Programs for Processing Nuclear Science Abstracts Tapes on a CDC-6600 Computer, Lawrence Radiation Laboratory, Berkeley, California, Report UCRL-19249, Oct. 1969.

TECHNICAL INFORMATION DIVISION
LAWRENCE RADIATION LABORATORY
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720