Development of an Index for In-House Research and Development Technical Records*

By J. M. McILVAIN and L. N. LEUM
Technical Information Section, Research Division, Research and Development Department,
The Atlantic Refining Company, Glenolden, Pennsylvania
Received December 9, 1963

The Atlantic Refining Company is an integrated company in the petroleum industry, engaged in the search for and production of crude petroleum, the transportation of crude petroleum and its products, and the manufacture and sale of a full range of petroleum products including many chemical derivatives. Research and development work in these areas is carried out in our Dallas, Texas laboratories for crude exploration and production, and in the Philadelphia and Glenolden laboratories for the other listed functions, plus studies in other areas of possible diversification. This paper discusses one segment of the technical information activities of our Pennsylvania research and development laboratories. Activities based in the Dallas laboratories are the subject of a separate paper.[1]

The Technical Information Section in our Philadelphia and Glenolden Research and Development Laboratories maintains a complete file of the project reports, technical memoranda, and certain other technical records coming out of the work of these laboratories. These records date back to 1924. The information in them is an asset obtained by major expenditures and affords sound justification for providing a means whereby the information is made readily available to our staff. Accordingly, we initiated a project in 1962 to develop an indexing and retrieval system for Atlantic Research and Development Department's technical information.

To determine what kind of indexing system would best suit our needs, we reviewed our own experience, accelerated an intensive study of the recent literature, and availed ourselves of opportunities to hear and talk with authorities in this field. We also listed the features which we considered desirable in the retrieval system to be developed. These were:

1. The index should be capable of covering all of Atlantic's technical research reports.

2. The index should be adaptable to including other technical material of interest to us, if so decided later. Our own experience has convinced us that a consolidated index of technical information is much more likely to be understood, maintained, and used than a multiplicity of indexes probably differing in principles, equipment, and accessibility. With suitable attention to its structure, one soundly based and properly maintained index would provide access to desired information in the shortest time and with the assurance that no indexed information was overlooked by the inadvertent failure to search other potentially pertinent indexes. Since all coverage would be of technical information pertinent to Atlantic Research and Development interests, the indexing vocabulary and the indexing procedures could be coordinated.

3. The output from the indexing and retrieval system should be document numbers rather than copies of the referenced documents. All of the reports to be indexed are available in centralized files, easily accessible for duplication if desired.

4. The index should be constructed to provide retrieval by concept coordination rather than from alphabetical or classification listings. The pros and cons of coordinate indexes have been discussed by numerous authors. Costello has recently provided a concise statement of the pros.[2] Our choice was based largely on the desire for (a) sufficient depth of indexing, (b) quick and easy retrievals (intellectual effort expended once, in the indexing, rather than at each retrieval), (c) ease of adding (or subtracting) blocks of indexed matter when new fields of interest are entered or old ones discontinued, (d) minimizing clerical effort (typing, posting, etc.) and its attendant requirement of training personnel and proofreading their work with possibilities of undetected errors or omissions (coordinate indexing adapts readily to machine processing the results of the intellectual activity), (e) minimizing the size of the index itself (as compared to an alphabetical index of equivalent scope and depth).

5. Posting in the index should be document numbers on terms to provide searching by scanning only relevant parts of the index, as opposed to posting terms on document numbers which would require sequential scanning of all or much of the index for each search.

6. Vocabulary control should be established by developing and maintaining a thesaurus of indexing terms, collected from the material to be indexed, to provide specificity and to minimize the number of terms used.

7. Indexing consistency should be established and maintained by free discussion among the indexers, a manual of indexing rules and conventions, and occasional review of each other's indexing.

8. The indexing system should provide for searching (a) by machine for comprehensive or complex searches, and (b) by a manual method for most occasions.

9. Indexing should use links (since many of the reports to be indexed deal with more than one major topic).

10. Indexing should use role indicators to promote consistency and speed in indexing and to minimize false drops in retrieval.

11. The indexing system should be as compatible as feasible with that being developed by the American

* Presented before the Division of Chemical Literature, 145th National Meeting of the American Chemical Society, New York, N. Y., Sept. 10, 1963.

Petroleum Institute, since it might be desired to merge the two for our use.

With these preliminary objectives established we looked into our needs for reaching them. These were (a) to index enough documents to accumulate the beginnings of a thesaurus, (b) to edit the accumulated terms, (c) to establish the format of a thesaurus and to publish the first edition, and (d) to select the kind of manual search index to be used and to prepare the same.

Since we had no experience in the techniques of indexing for retrieval by concept coordination, a contract was executed with Jonker Business Machines, Inc., for consulting service to train our staff in indexing and vocabulary editing. This proved to be most helpful to us, not only in the organization and execution of our own indexing and editing activities, but also by keeping us in touch with outside trends, such as the American Institute of Chemical Engineers (AIChE) and subsequent Engineers Joint Council (EJC) programs.

In order to ensure an adequate sampling of report topics and vocabulary, and to afford training to the six professional members of our Technical Information Division staff, some 919 reports of various kinds were indexed, perhaps 50% more than the minimum needed for this initial phase. However, the more generous sample gave us more information on the rate of vocabulary buildup and the effect on vocabulary of going to different interest areas in the indexing of our reports.

Indexers entered directly on their work sheets the accession number of the report being indexed and (a) report title, (b) report type and serial number if any, (c) project number of the work reported, (d) author(s) of the report, and (e) each combination of link, term, and role. A separate IBM card was punched for each of the items a–e together with the accession number. The resulting cards were machine listed for proofreading and erroneous cards replaced. All bibliographic data cards (items a–d) were sorted out and set aside, in order of the accession numbers. These cards averaged six per accession number.

The remaining cards (items e) were then sorted alphabetically by terms and within each term successively by document number, link, and role. The machine listing of the resultant deck provided the raw vocabulary for our first editing, which we are completing at the time of writing this paper.

The most time-consuming phase of the editing has been that concerning chemicals, where we generally have conformed each chemical term to the nomenclature and indexing principles of *Chemical Abstracts*.[3] Our raw vocabulary listing totaled an estimated 33,000 lines, generated by indexing the 919-document sample. Many of these lines were repetitious with regard to terms, since each term was separately listed for each combination of term, accession number, link, and role. Our alphabetical sort brought all occurrences of each term together, so that each term needed to be edited only once, no matter how many times used (in the same sense).

We expect that after editing our sample vocabulary will consist of about 6,300 terms, of which 2,600 will be names of chemicals. These figures do not include the bibliographic data for the 919 documents.

As the final step in the editing, we plan to convert the role indicators for the non-bibliographic indexing terms to the definitions and usage of the EJC.[4] These definitions are the end product of an evolution embodying the creative thinking of workers in this field over a decade and can be expected to be widely used in the publications of the engineering societies.

The format of the thesaurus toward which we are working is broadly similar to that of the AIChE.[5] It will be divided into two parts. One part will be an alphabetical listing of the names of chemical structures. The other part will be an alphabetical listing of all other non-bibliographical indexing terms, including trade names. We believe that segregating the names of chemical structures in one list will make easier, quicker, and more certain the selection of terms of this category for both indexing and searching.

Each alphabetic list of terms will display indented under each lead term, its seen-from, broader, narrower, and related terms. Since machine-search tapes are to be a component of the system, each chemical structure and other term in the thesaurus will have alongside it its serial identifying number. Use of these numbers is expected to save time in indexing and manual retrieval, in addition to being almost mandatory for preparation and use of machine-search tapes.

We have given considerable thought to the question of how many occurrences of an indexing term will permit it to be included in the thesaurus. Pending further experience, we have decided to include each term as it comes up, feeling that this will avoid separate lists to be consulted or overlooked, and the necessity for reconsideration when (if ever) the term recurs.

The production of the thesurus will be by photoduplication from computer print-out.

We anticipate that 80 per cent or more of our searches will be by manual coordination of terms and their associated document numbers. Because of this heavy usage, it is highly desirable that the coordinations be attainable quickly, easily, and with a minimum requirement of preliminary training on the equipment. We also are interested in keeping the clerical aspects of posting items on terms at a minimum skill level and expenditure of time. The TERMATREX equipment of the Jonker Business Machines, Inc., meets all these requirements and was selected. It has the additional attractive feature that the Jonker Company will transfer postings between TERMATREX cards and IBM punched cards on a service basis, if this should become desirable.

Since the simplest TERMATREX system will accomodate 10,000 postings before going into an additional deck of term cards, it was of interest to estimate how far one deck would take us. A survey of our current areas of interest and a count of report items by years in those areas indicated that one deck could be expected to cover about the most recent seven years. Since in any particular area of current interest the reports thin out as one goes back year by year, a second deck should be adequate for something more than an additional seven-year period.

As mentioned earlier, our indexing will relate term–role–link combinations to document numbers. However, terms only (not term–role combinations) will appear in the thesaurus and on the TERMATREX term cards. Furthermore, only document numbers (not document number–link combinations) will be posted on the term cards.

Manual retrieval therefore will be by coordinating terms (not term-roles) and will yield document numbers (not document number-link combinations).

The usefulness of the links and roles is threefold. First, they materially help the indexer to analyze the document he is indexing, leading to more effective indexing. Second, we plan to maintain an "Accession Register" which will list all indexed documents in the order of their document numbers, together with the complete bibliographic data and all link-role-term combinations for each document. Third, the document number-link combinations (rather than document number only) will be entered on our search tape in relation to the corresponding term-role combinations.

Thus, in a manual search, terms will be coordinated to give document numbers. Location of these document numbers in the Accession Register will provide telegraphic abstracts of all the documents. In most instances this will obviate any need for reference to the original documents.

In a machine search, the read-out could give the document number-link combinations for reference to the Accession Register. Or, if the volume of tape searches warrants, it would be simple enough to have tape search results reported as print-outs of the complete information—bibliographic and telegraphic abstract—on each referenced document, from a tape master of the Accession Register.

Manual searches and the preparation of requests for machine searches will be provided by our Technical Information Section as a service to our technical staff. However, we intend to encourage use of the manual search equipment by all who are interested in it as a quick access to specific information in past reports.

Acknowledgments.—The development of this index has been a group activity to which important contributions have been made by Jane A. Bennett, E. Ray Birkhimer, Joyce R. Crossley, Louise U. Matternas, and Irene H. Sachs. Donald Van H. Harrison of Atlantic's Systems and Data Processing Department provided helpful computer know-how and John C. Costello, Jr. (then of Jonker Business Machines, Inc.), invaluable instruction. Finally, without the encouragement and support of James E. Connor, Jr., the work would never have been done.

REFERENCES

(1) J. R. Bilhartz, "Experiences in Information System Design," presented before the Division of Chemical Literature, 145th National Meeting of the American Chemical Society, New York, N. Y., Sept. 10, 1963.

(2) J. C. Costello, Jr., in "Information Handling: First Principles," P. W. Howerton, Ed., Spartan Books, Washington, D. C., 1963, Chapter 3.

(3) "The Naming and Indexing of Chemical Compounds by *Chemical Abstracts*," Introduction to the Subject Index of Vol. 56, (Jan.-June, 1962) of *Chemical Abstracts*.

(4) "The Engineers Joint Council System of Roles-Meanings-Examples-Explanations-Exclusions," Battelle Memorial Institute, Columbus, Ohio.

(5) "Chemical Engineering Thesaurus, A Wordbook for Use with the Concept Coordination System of Information Storage and Retrieval," American Institute of Chemical Engineers, New York, N. Y., 1961.

Retrieval of Analytical Research Information by a Coordinate Indexing System*

By BARBARA A. MONTAGUE

Research and Development Division, Plastics Department, E. I. du Pont de Nemours and Company, Inc., Wilmington, Delaware

Received April 3, 1964

The objective of this paper is to describe the storage and retrieval of recorded analytical research information by concept coordination using links and roles. Analytical information is retrieved by indexing the object of the analysis, the matrix in which the analysis was performed, the technique, and the reagents employed.

The results of research efforts in the field of analytical chemistry in the Plastics Department of the Du Pont Company are recorded in notebooks, correspondence, and finally in formal reports. This knowledge may comprise the main topic of an analytical report or lie buried in a report written by a scouting or product group for whom the research was performed. The passage of time and transfer of personnel weakens the link with the past, and such recorded information becomes the prime source for the acquisition of knowledge by others. In 1959 the Plastics Department recognized a need for improved methods for retrieval of all their internal research information and initiated a program utilizing concept coordination with the expectation of obtaining faster, more selective access than was provided with shallow indexing by classification used at that time.

An information system was designed[1-3] using deep indexing (40 terms per document) by concept coordination with links and roles to reduce irrelevant retrieval. An example of the indexing of a technical report with links and roles is presented in Fig. 1, and the definition of the roles is given in Fig. 2. This report was indexed in one

* Presented at the Fisher Award Symposium honoring John Mitchell, Jr., before the Division of Analytical Chemistry, 147th National Meeting of the American Chemical Society, Philadelphia, Pa., April 9, 1964.