How An Indexer Thinks in Describing Information, In Framing Search Questions, and in Conducting Searches*

By CAROL A. PENN
Jonker Business Machines, Inc., Gaithersburg, Maryland
Received May 18, 1962

Analysis of input to any information system is one of if the not *the* key factor to accurate and adequate retrieval. Nothing determines the quality and nature of output more than the decisions indexers make during the indexing process. It is my intention in this paper to show how analysts index chemical documents and conduct searches —that is, what factors are considered, what decisions are required, and on what bases these decisions should be made. This paper is designed to furnish answers to the question: what are the mental processes required of analysts in indexing documents, in framing search questions, and in actually conducting the searches?

Before discussing indexing in detail, first let me comply with the prescribed procedures for this symposium by discussing the definition of a deep index. I agree with all of the aspects of a deep index that were presented at the start of the symposium, but I wish to discuss two additional points. First, I would insert that the objective of a deep index should be the rapid retrieval of all documents pertinent to any particular question, accompanied by a minimum of non-pertinent references or false retrievals. Second, the depth of treatment should be governed by the questions the users are asking at present and a prediction of the types of questions they are likely to ask in the future. In other words, depth of description is determined by the needs of the users relative to the value of the information.

Since this discussion is based primarily on my previous experience as an indexer, I would like to describe briefly the system with which I was associated and the mechanics of operation of that system. The system in the du Pont Polychemicals Department is based on the unit concept approach as described by J. C. Costello and B. A. Montague in papers previously published in *American Documentation*. Documents are assigned accession numbers as they are issued. They are filed in serial order according to these accession numbers. A system of links and roles was developed to overcome the syntactical problems which would have been particularly troublesome if the deep indexing of chemical information by the coordinate indexing approach had been done without these controls. Links are used to bind or associate together the terms describing each of the separate intellectual components which may be discussed in the document. In retrieval, links prevent the coordination of concepts which, while admittedly present in the complete physical document, are not actually associated in the same intellectual relationship or link. Roles define function or context of each of the terms within each intellectual relationship or link. In searching, roles prevent the retrieval of documents which were described by the search terms functioning in other than the desired contexts.

Accession numbers and terms are stored in an inverted arrangement. The term records on which accession numbers are posted are stored in alphabetical order on magnetic tape by means of term-codes. From this tape, a coordinate index was printed in the form of a dual dictionary which was used for manual searches by comparing the accession numbers and links posted on the term-roles pertinent to the search question. Documents in the system are primarily internally generated research reports and chemical patents issued by the United States Patent Office. The patents were those of one class, the subject matter of which was of interest to the research and development group. The user group consisted primarily of the scientists in that research and development group and the patent attorneys serving the same department. Most of the scientists were chemists and chemical engineers, but the group also included some physicists, mathematicians, and several types of engineers.

The information system personnel responsible for indexing are graduate chemists with the title of scientific literature analyst or indexer. Analysts' responsibilities include participation in editing the vocabulary generated as a result of the indexing and assisting users in conducting searches.

Now that the environment, the user group, and the system serving this group have been defined, let us consider in order, (1) indexing of documents for deep indexes, (2) framing the search questions, and (3) conducting searches. The intellectual considerations that go through indexers' minds as they index chemical documents and which are necessary to assure adequate and accurate input to a deep index evolve from four separate but related lines of thought. These lines of thought are guided by continuously seeking the answers to four groups of questions as each new document is considerd. (1) What information is in this document, how is it organized, and into how many intellectual components is it subdivided? (2) How are the overall document and each of its component subdivisions related to or identified with the current and anticipated needs of the user group? (3) How new, how reusable, or how original is the information in each component? How valuable is it in relation to the current and anticipated

* Presented before the Division of Chemical Literature, ACS National Meeting, Washington, D. C., March 23, 1962.

information needs? (4) How should the information be described?

Now let us consider the first question: "What information is in this document, how is it organized, and into how many intellectual components is it subdivided?" If indexers do not spend time determining what information is in the document, there are certain pitfalls which can result in inadequate and inaccurate input. The first mistake that indexers can make is to pick up a document and start describing it without making general note of the kind of document it is and the nature of the information it contains. This can result in either inadequate depth of description, excessive depth of description, or incorrect description because the intent or purpose of the document was either missed, underestimated, overestimated, or misinterpreted. The second mistake indexers can make is to index words rather than subject intent or subject content. For example, this can result if the indexers select index entries on a basis of the superficial appearance of the document, that is, the words used in the text and an estimate of the frequency of their occurrence. Index entries later serve as locators of information. Hence, selecting the correct locators is a prime determinant of index quality.

The first prerequisite in preventing this type of superficial analysis is that indexers be technically competent in the subject matter of the document collection. Regardless of how elaborate the device may be which manipulates the index for retrieval purposes, the quality of output can be only as high as the quality of analysis at the time of input. Indexers analyzing chemical information should be graduate chemists or chemical engineers who have a sound, general understanding of the subject matter of the field. They start with this basic understanding; from experience in indexing and by seeking assistance of user group for explanation of more specialized, highly technical information, they become better qualified generalists. In a system serving a research and development group, indexers are likely to be working with information of a highly specialized nature. It is probable that graduate chemist-indexers might have some difficulty in understanding the work of an infrared spectroscopist, a nuclear chemist, or a theorist in viscoelastic flow. In such cases, indexers have at their disposal numerous references to which they can refer. Among these may be, hopefully, the author, whom they can consult personally; or if he is not available, other authorities may be, who can discuss the work with them. However, if they lack a fundamental understanding of the field, neither references nor the author nor-available authorities can be of much assistance in preventing a substandard quality of indexing.

Assuming then that the indexers do have knowledge of the subject matter, they undertake the analysis of each document in an organized fashion. They familiarize themselves initially with the nature of the document by an examination of the title, introduction, and perhaps the table of contents. This creates an awareness of the intent or purpose of the document and the main topics considered in it. For example, a document might consist primarily of a detailed procedure for the chemical analysis of a particular product. From this first examination, the indexer can tell whether the research was done on the procedure itself and its feasibility for possible use, using the particular product as an example, or whether the research was aimed at the product and an accurate means for identifying its components. The indexer here determines the nature of the work distinguishing, for example, between basic research, quality control, and product development. Knowledge of the author's field of specialization and the identity of the research team of which he is a member also provide clues to the nature of the work recorded. In indexing patents, the indexer's first familiarization consists of examining the title and the first few paragraphs of the patent which disclose the state of the art and the objects of the invention.

After general familiarization, then the indexer examines each component of the document in detail. This includes abstract, summary and conclusions, the text or body of the document, and appended material, considering in order the most condensed statements first and proceeding to the most narrative. In this fine-line analysis, the indexer is actually interpreting the words of the document into a comprehensive picture in his mind. By visualizing each separate intellectual idea and the proper relationships between these ideas, the indexer is creating his own flow chart of the document. Each separate intellectual relationship discussed is a unit; when all the units are properly related, the result is a central theme accompanied by a number of supporting themes.

In the analysis of patents, the indexer first reads the claim or claims. Keeping the object of the invention in mind helps to overcome the difficulties in understanding the legal terminology of the claims. Then the indexer examines the disclosure of the patent. By the time he completes his analysis of the claims and specifications, the indexer is able to relate the inventive feature to the information needs of the user group.

Having determined the information content, then the second question is considered: "How are the overall document and each of its component subdivisions related to or identified with the current and anticipated activities of the users?" At this point the indexers establish a relationship between the concepts discussed and the objectives of the system. This phase of indexing is a kind of mental reworking of the information as it is written in the document to relate it to the environment of the users. The indexers identify the information from the users' viewpoint as well as from the author's viewpoint. They interpret and relate the author's terminology to the accepted terminology of the user group. Identification is most vital when the documents originate outside of the user group.

When the indexer has analyzed the document so that he has a complete understanding of its information content and has identified the information with the users' needs, then he asks himself the third question: "How new, how reusable, or how original is the information in each component?" How valuable is it in relation to the current and anticipated information needs? Knowing the objectives and limitations of the system, the indexers decide what elements of the information are valuable and the depth to which locators should be provided. The hazard in this phase of analysis is that of misjudgment on the part of the indexers. If indexers cannot anticipate the types of information users will require of the system, or if they are unable to relate the content of the document to the types of questions being asked, then the depth of

indexing may be too great or too shallow. Since the ultimate decision as to whether or not to include locators to certain information is based on the reuse value of the information itself and the needs of the user group, indexers should keep themselves informed of the activities of that group in addition to maintaining an active interest in the advances being made in the field as a whole. They should be able to relate the contributions of the user group to the state of the art in general. Maintaining close association with the activities of the user group can be a valuable asset since by so doing the indexers can increase their professional competence and their proficiency as indexers. This association also has the inevitable results of increasing the users' confidence in the system and its operation.

In separating the reusable from the non-reusable, there are generally two types which should not be included—that which is basic common knowledge in the field and that which is already included in the system. Common knowledge is valueless since it will likely never be requested. An exception to this would be patent information in a system designed to serve patent attorneys, for example, who for infringement purposes might want to retrieve patents pertaining to a particular process for large scale synthesis of a specific compound. The compound and its synthesis may have been recorded in text books since the beginnings of the science of chemistry, but since the patent has definite legal implications, the synthesis must be included along with all the particulars which the patent covers. In such cases, the basic ground rules for system coverage will define how these situations should be handled.

If a particular idea is indexed in sufficient depth at the time of input, then there is no reason to index it into the system again. For instance, consider progress reports and final reports. Controls and checks should be developed so that duplicate information in later documents is not included. With respect to this principle, patents provided another exception. There are many patents covering the polymerization of ethylene, and they may differ from each other by only one variable. In order to assure the patent attorneys complete coverage from the legal viewpoint, locators for the reaction must be entered into the system again and again, despite the fact that there may be no substantial difference other than the one variable.

The decision to include or exclude, as the case may be, is based on the needs of the users and an evaluation of the relative quality of the information in reference to those needs. Those needs are not as difficult to judge for the immediate future as they are for the distant future. Since trends and objectives of research organizations are continuously changing, indexers must predict the types of questions that must be answered in the future. They must estimate the likelihood that a project will be undertaken which would have need for the information in question. If they can visualize no situation where the information might have reuse value, then there is no reason to include locators for that information. On the other hand, if indexers cannot immediately conclude that the information has no reuse value, then they include it, since any hesitation would seem to indicate that it does have possible reuse value. Here the difference between possibility and probability is the deciding factor.

The type of document helps in evaluating the reusability of information. If the document is an internally generated research report, it is doubtful that there will be any question as to the value of the information content in its own environment. Its purpose is generally to record procedures, results, and conclusions. If the document is a literature review of what others have recorded in previous documents, then it will likely be necessary to index only the original documents. Again, the most important factor in deciding to include or exclude is the needs of the user group.

When there is any question, doubt, or hesitancy about including locators to information, the general principle that indexers should follow is—if you don't index it, you can't get it back. It is not much more expensive to include several additional index entries on the work sheet, since identification and evaluation, the expensive and time-consuming operations, have already been accomplished. It requires considerable time and effort to find an item of information if it was not indexed at the time of input.

Once indexers have decided to include information, they apply the principle of evaluation in deciding the depth to which that information should be indexed. Since depth of description depends not only upon the number of locators, but also upon the level of specificity of those locators, determination of depth also is based on the needs of the users. Usually indexers will not have to decide on the depth of description. They will let the specificity of the discussion in the document govern the specificity of description. If the information is worthwhile enough to be included in the first place, then it should be included at the level at which it was discussed, subject only to evaluation in light of its originality and uniqueness of treatment.

The fourth question which indexers consider in indexing documents is: "How should the information be described?" In order to describe information adequately, indexers should have complete understanding of the principles upon which the system is built. A training program for indexers should include not only drill in the basic techniques of indexing, but should also provide an understanding of the problems encountered in searching for information and how these problems can be overcome. With this understanding, indexers are better able to know how information should be described in order that it may be retrieved. The competence of indexers in the technical field is an important prerequisite to the adequate description of information. Both recognition of the technical concepts and a knowledge of the correct terminology for those concepts are necessary for maintaining a high quality of input.

Describing the information in a document is an anticipation of how the information might be used or in what way it might be requested in the future. Indexers must realize that different users in different situations may have use for the same piece of information for different purposes. In a sense indexers are actually presearching the information they are indexing by framing search questions to fit the information. The index entries noted by indexers are their predictions of how search questions might be posed to recall this information.

In attempting to anticipate the different ways that certain information might be requested, indexers are being redundant in their indexing. A certain degree of redundancy is necessary in a system which utilizes links and roles in order to anticipate different viewpoints which may exist in the user group. For instance, the inventive feature of a patent which claims the process for the reaction of chemical A and chemical B may be unusually low pressure or unusually low temperature. The research chemist may have an interest in this patent because of the reaction variables. On the other hand, the patent lawyer must be able to retrieve this patent as a domination of the process for the reaction of chemical A and chemical B. Indexers must describe information from as many viewpoints as necessary to satisfy the needs of all potential users.

In a system using links and roles, indexers separate each relationship of concepts into separate links in much the same way that an author separates his writing into sentences or paragraphs. When an indexer encounters a new intellectual relationship, he indexes that in a separate link to distinguish it from other intellectual relationships in the same document. Sometimes, the ideas are not completely distinct and separable; in such cases, the indexer's decision regarding the use of links is governed by his prediction of false retrievals which might result. In applying roles to the index entries, the definitions of the roles provide their own rules for use. The indexer interprets the mental picture he has into terms and roles, roles defining context or function for the terms. If the first three phases of the indexer's analysis have been accomplished adequately, then the fourth phase is an operation of converting thoughts to words on paper, using synthetic grammatical devices to portray the proper association and context.

Framing a search question involves procedures not unlike those used in indexing. First the indexer identifies what information is desired. Then he relates the search question to the limitations and coverage of the system. Finally he describes the search question in the language of the system. Negotiating the question with the user is the best technique that the indexer can use to identify the true nature of the information sought. To better understand the question, the indexer inquires about the nature of the project on which the scientist is working. Most users ask questions on a much more general level than that of the information they need. By inquiring into the nature of the user's work and the effect that the information will have on his course of action, the indexer can frame the search question on a level of specificity compatible with his needs. As an example, one user may ask for information on the reactions of chemical A and chemical B, fully anticipating that there may be no references to so specific a question. Another may ask the same type of question, such as for information on reactions of chemical C and chemical D, when actually he wants information on the reaction of Generic Class C compounds with Generic Class D compounds. These searches should be made on different levels of specificity, the levels being determined by negotiating the questions with the users.

After obtaining a clear definition of the information that the scientist needs, the indexer must relate the question to the types of information in the system and to its index terminology. Knowing the kinds of information which are included in the system, the indexer can predict the probability of locating that which is needed. Translating the search question into the appropriate concepts or locators includes assigning the pertinent roles according to the context of the concepts in the search question. Framing a search question is the assignment of term-roles to the information to be retrieved, just as indexing is the assignment of term-roles to information to be stored.

In framing search questions, indexers apply a certain degree of redundancy in searching. They now recall the different contexts in which certain concepts may have been indexed, and they search on all the contexts which relate to the user's request. For instance, a property of a particular product could have been studied for its effect on another variable. On the other hand that same property in another document may have been discussed as a dependent variable, that is, some other variable might have had an effect on it. Depending upon the viewpoint of the scientist, the indexer decides on the contexts or roles on which to search. Whenever indexers have any doubt about the level of specificity of the search, they always search on the more general level to assure that the user gets all pertinent information. Indexers should let the ultimate rejection of information as not pertinent rest with the user. Comparable to the indexing process, the framing of searches is carried out with the overall objective of the system in mind—that of retrieving all pertinent documents with a minimum of false retrievals.

Once the search question has been framed, the process of conducting the search is a more or less mechanical process of comparing item numbers and links, which appear on the selected term-roles, assuming that a manually operated device is used. In a search involving logical product, the answer to any question can never be more document numbers than the number of postings on the least densely posted term. In these searches, indexers compare the postings on the least densely posted term with the postings on the next least densely posted term. This minimizes the time required in manual operations. Comparable observations could be made for systems which depend on the collating of cards or the comparison of term records on magnetic tapes or discs.

When the analyst has obtained the document numbers that satisfy the search requirements, he checks the validity of the references by examining abstracts or the title list. This allows him to be more informative in his reply to the user, and it enables him to identify any non-pertinent references. If the search produces only a few references, he checks the title and/or abstract for each one. If the search produces a long list of references, then an examination of a few randomly selected abstracts gives him an indication of the relative number of false references and perhaps some clue as to the cause. This will enable him to conduct a revised or reworded search if necessary.

The rules that govern these processes of input and output are best summarized by stating that decisions should always be based on the needs of the users. In instances where these needs are not easily determined or not quite clear, the rule should always be to make a decision which will enable the user to determine the pertinence of a reference for himself. If the usability of

a particular item of information is questionable, the decision should always be to index it. In searching, if there is any doubt about including a particular aspect of the question, then the decision should always be to exclude that aspect rather than narrow the scope of the search by using it. An indexer provides more effective service if he lets the user decide for himself whether or not the document has any value. When indexers judge the value of information too strictly, users will have justifiable cause for questioning the efficiency and adequacy of the information system.

Indexers are the necessary components of an information system who create the index, a device to bridge the gap between the originators of information, on the one hand, who do not know when or where the information might be needed, or by whom, and on the other hand, the users of information who do not know when, where, by whom, or under what circumstances the information might have been written. In creating the index as a device to solve one problem, another problem has been created. Indexers, not usually specialists, are surrounded by specialization and are working with information originated by specialists. This problem of having generalists analyzing specialized information can be greatly reduced if indexers plan for the continual growth of their technical competence by associating and familiarizing themselves with the intellectual activities of the scientists, the type of information they generate, and the type of information they use. Attendance at group research planning sessions and research reviews is most helpful.

In addition to being competent as scientists and as information analysts, indexers must in a sense be public relations people to provide liaison between the system and the scientific group they serve.

Indexers serving any one group function as a team. They assist each other in making decisions regarding indexing and searching techniques. They discuss among themselves such things as highly technical concepts and the reuse value of information. As decisions are made, ground rules are created which serve as standards for future decisions and which provide guides for new indexers. Thus the system is built on and maintained by a schedule of written rules which evolve from practice.

I would like to expand briefly on the concept of "user group" which I have mentioned frequently in this discussion. There are many instances where the user group and its needs cannot be defined clearly or specifically. Such a group would be all the members of a professional organization such as the American Chemical Society. Since the information needs of all the members of such a group are virtually limitless, a system which serves this group must be geared to satisfy the requirements of as many of the specialized sub-groups as is economically justifiable. In this type of situation, the more diverse the needs of the user group and the broader the information coverage of the system, the greater the need for description on the level of specificity dealt with in the documents. When the user group is well-defined, as in individual corporate organizations and governmental agencies, the analysis must be adequate to meet the needs not only of personnel who have been members of the research team for a number of years, but also of the new employees who have just completed their formal education.

In summary, analysis of documents for input to an information system requires that analysts or indexers be capable of understanding the content of the document, evaluating that information for its reuse value, and describing it so that it can be located quickly and completely with a minimum of non-pertinent material. Locating information requires these same considerations relative to the information requested. Whereas these principles of analysis have been discussed with particular reference to a system utilizing links and roles, the same considerations are essential for a system that does not use these or comparable syntactical controls.

It is obvious from this discussion of the human judgments necessary for high-level information analysis that I do not believe that machines are capable of producing indexes except on a crude basis of statistical count or frequency of use of words.

BIBLIOGRAPHY

(1) Bernier, C. L., "Correlative Indexes II: Correlative Trope Indexes," *Am. Document.* **8**, 47–50 (1957).

(2) Costello, J. C., Jr., "Uniterm Indexing Principles, Problems and Solutions," *Am. Document.* **12**, 20–26 (1961).

(3) Costello, J. C., Jr., "Storage and Retrieval of Chemical Research and Patent Information by Links and Roles in du Pont," *Am. Document.* **12**, 111–210 (1961).

(4) Costello, J. C., Jr., "Some Solutions to Operational Problems in Concept Coordination," *Am. Document.* **12**, 191–197 (1961).

(5) Documentation Incorporated, "The Uniterm System of Indexing, Operation Manual," Washington, D. C., 1955.

(6) Holm, B. E., "Information Retrieval in Industry," paper presented at the American Management Association Data Processing Conference, New York, March 6, 1961.

(7) Luhn, H. P., "Keyword-in-context Index for Technical Literature (KWIC-Index)," paper presented at the 136th Meeting of the American Chemical Society, Division of Chemical Literature, Atlantic City, N. J. September 14, 1959.

(8) Montague, B.A., "Patent Indexing in du Pont by Concept Coordination Using Links and Roles," *Am. Document.* **13**, 104–111 (1962).

(9) Schultz, C. K., "Intellectual Problems in the Organization of Information: Storing Information for Machine Retrieval," paper presented at the first American Documentation Institute Meeting-in-Miniature, Philadelphis, Pa., February 10, 1961.

(10) Wall, E., "A Practical System for Documenting Building Research," from the proceedings of the 1959 Fall Conferences of the Building Research Institute, Division of Engineering and Industrial Research, Publication No. 791, Nat. Acad. Sci.-Nat. Res. Council, Washington, D. C., 1960.