User-Centered Indexing

Raya Fidel
Graduate School of Library and Information Science, University of Washington, Seattle, WA 98195

Two distinct approaches describe the process of indexing. The document-oriented approach claims that indexing summarizes or represents the content of a document. The user-oriented approach requires that indexing reflect the requests for which a document might be relevant. Most indexing, in practice as well as in theory, subscribe to both, but the document-oriented approach has enjoyed most visibility. While request-oriented indexing is a user-centered approach, it is very difficult to implement with human, *a priori* indexing. Automated indexing with its dynamic and flexible nature is most fit to tailor indexing to requirements of individual users and requests, yet most of current research in the area focuses on the development of global methods. Regardless of the method, user-centered indexing cannot be developed before searching behavior is understood better.

From a most simple viewpoint, "indexing is the procedure that produces entries in an index" (Cleveland & Cleveland, 1983, p. 63). Indexes are necessary to facilitate retrieval of information. Most information systems that are currently available require an index, because sequential or random retrievals to satisfy a request are prohibitively time consuming. While indexes are created for practical uses, the topic of indexing—the actual process of creating an index—has generated much theoretical discussion. As Vickery (1953, p. 48) observes: "The function of a subject index is above all practical . . . Despite this, it is only on the basis of inductively derived principles that a system can be constructed at all."

Some indexes are created mechanically and with no intellectual intervention, such as the computer's internal indexes. For such indexes, the process of indexing is limited to arranging entries in a predefined order. While such a process may rely on computational principles, it requires almost no theoretical consideration. The concept of indexing becomes more complex and theoretically demanding when the process of indexing also involves the creation of entries. Thus, library and information science has focused on the generation of index entries.

The most common types of indexes are name indexes and subject indexes. Name indexes are tailored to the specific needs of the information system and its users, and may include a variety of entities. A name index may include, for instance, names of the authors who wrote the indexed documents, authors cited in these documents, or of people described in them. No matter what entities are represented in these indexes, they usually include names that appear in the stored information. While name indexing is not as simple and straightforward as it might seem, most theoretical work has been done in subject indexing. In fact, most often (as is the case here) "indexing" is used to mean "subject indexing."

In addition, it is common to distinguish between back-of-the-book indexing and database indexing, referring to both printed and machine-readable databases. While both types of indexing have many common features, there are also important differences, and Wellisch (1994) illuminates those later in this issue. Book indexing requires that index entries and the index language both be crafted simultaneously. Thus, the literature on book indexing does not always distinguish between the process of indexing and that of creating an index language. The literature about database indexing, on the other hand, assumes that an index language has already been created and focuses on the process of creating index entries. Because our focal point in this issue is indexing, rather than the construction of index languages, this article draws on the database literature, much of which is relevant to back-of-the-book indexing.

Document-Oriented Indexing

Borko and Bernier (1975, p. 8) explain that "Indexing is the process of analyzing the informational content of records of knowledge and expressing the informational content in the language of the indexing system." Wellisch (1991, p. xxiii) defines indexing according to the International Standard ISO 5127/1 as "An operation intended to represent the results of the analysis of a document by means of a controlled or natural indexing language." Similarly, Rowley (1988, p. 43) elucidates that

"The indexing process creates a description of a document or information, usually in some recognized and accepted style or format." The concept "document" is often construed as the container of information or knowledge; it may take any form and be of any medium, or combinations of media.

The idea that index entries, much like abstracts, represent the contents of a document has led to the notion that indexing is actually the process of creating surrogates for documents, summarizing their contents. It also suggests that when indexed, a document is assigned to classes of similar documents. Using different formulations, this approach to indexing is centered on the document and is guided by two basic ideas:

- The purpose of indexing is to represent (express, describe, or indicate) the content (topic or features) of a document.
- The process of indexing has two components: (a) contents analysis that results in the selection of the concepts to represent the document; and (b) translation, that is, expressing the concepts selected in the index language used by the information system or database.

Each of these components has attracted much attention. Referring to contents analysis, Lancaster (1991, p. 8) explains that "Conceptual analysis, first and foremost, involves deciding what a document is about—that is, what it covers."

The determination of what a document is "about" can be subjective, however. Thus, the concept of "aboutness" generated much attention. Various writers pointed to humans' limited ability to objectively determine aboutness (e.g., Maron, 1977; Swift, Winn & Bramer, 1978; Wilson, 1978). Others see aboutness as a twofold attribute: part of it is inherent to a document and thus objective, and the other part is determined by the reason or purpose for which the document is being used (e.g., Beghtol, 1986; Fairthorne, 1969; Hutchins, 1978). Svenonius (1994) and Shatford Layne (1994) explain later in this issue that the concept of aboutness requires special consideration when indexing nontextual information. On the practical side, strong evidence to the subjective nature of human indexing has been collected with tests of indexer consistency. Most of these tests have revealed that indexers are likely to be in poor agreement among themselves about how to index a document, and that they frequently index the same document differently in different times.

The second component of indexing, translation, requires rules for its implementation. Thus, indexing policies have guided and directed the process of expressing the surrogate concepts in the database's index language. Indexing policies may address a variety of issues, and are usually determined by the purpose of the database and its users. However, the cumulative experience of indexers has generated a core of issues that should be addressed by an indexing policy:

- *Sources of index terms.* Which vocabulary sources can the indexer use for the selection of index terms? The policy may limit the indexer to the descriptors in a database's thesaurus, or it may allow the assignment of additional index terms using natural language.
- *Specificity.* How specific should the indexer be in translating a concept into index term(s)? That is, should the index term selected be as specific as the surrogate concept, or should it be broader in meaning?
- *Weights.* Can the indexer express the relative weight of a concept in a document? Weighted indexing reflects the importance of a concept to the document; central concepts receive higher weights than less important ones.
- *Accuracy.* How accurate should the translation be? Should the indexer index under related terms, and how should the indexer translate a concept when it has no equivalent descriptor?
- *Degree of precombination.* Should the indexer assign elemental index terms or their combination? For example, should an indexer assign the index term "health education," or the index terms "health" and "education," to represent the concept "health education?"
- *User language.* Can the indexer assign index terms in the language most likely to approximate the user's? Some indexing policies describe the intended users (e.g., whether they are professionals or lay people) to guide indexers in the selection of index terms most suitable for users.

A few issues in indexing policies address contents analysis:

- *Exhaustivity.* How comprehensive, or exhaustive, a representation should the indexer provide—that is, the degree to which various aspects are represented in the indexing? For instance, should an article describing a health education project be indexed under "women" and "Asian Americans" if most of the participants happened to be Asian-American women?
- *Indexable matter.* What intellectual parts of a document should the indexer consider for representation in the indexing? For example, should negative results, implications, possible uses, or suggestion for future research be represented in the indexing?

Some policy issues gained prominence; later in this issue Anderson (1994) explains why they are covered by the indexing standard, and Soergel (1994) discusses their effect on retrieval performance.

Request-Oriented Indexing

Because indexing is performed to facilitate information retrieval, it is necessary to consider the requirements put to a database by user needs. According to the document-oriented approach, indexing can be done with no knowledge or consideration of users or their needs. However, no developer or supporter of this approach ever recommended ignoring these important partners. Lancaster

(1991, p. 8), for instance, emphasizes that "Effective subject indexing involves deciding not only what a document is about but also why it is likely to be of interest to a particular group of users." Thus, an article about the use of isotopes in hydrology should be indexed under "hydrology" for a database that covers scientific literature about isotopes, but under "isotopes" for a database on hydrology.

Although the document-oriented approach does not exclude users, and may even require that users are described in indexing policies, users as individual seekers of information do not play a central role. The centrality of the user, however, is not a new notion. A few examples illustrate this point.

Calvin Mooers, who introduced the concept "descriptor" in the early 1950s, explains that "The particular scope of meaning for a descriptor is *assigned* in such a way that the descriptor will be most useful for retrieving information in a specified collection" (Brenner & Mooers, 1958, p. 347). Consequently, his method of developing an index language was an "empirical process." A panel of users examined a collection of documents, and for each answered the question: Why would any one of our users be interested in this document? The terms generated as answers became the descriptors in the index language. Mooers proposed a similar method for indexing, which he calls a "filtering" technique. The indexer checks each descriptor against the indexed document and asks: Would any one of our users who is interested in the content of this document use this descriptor as part of the query formulation?

Looking at indexing from a different angle, William Cooper (1969) examined the validity of the widely held assertion that consistency among indexers is indicative of the quality of indexing, and that an increase in the level of consistency improves retrieval effectiveness. Through a mathematical analysis of a counterexample he concludes that indexer consistency cannot be used as a measure of indexing quality. Further, he illustrates a hypothetical case where an increase in indexer consistency might result in a decrease in retrieval effectiveness. This is likely to happen if indexers consistently exclude a term that is used in a request for which the indexed document is relevant. He introduces, therefore, a new concept: "Evidently there is another kind of consistency beside interindexer consistency which is important—namely, *indexer–requester consistency.* If indexers tend to assign a given index term to a document to the same extent to which there is a tendency for the term to appear in requests to which the document is relevant, we say the indexer–requester consistency is high; otherwise, it is low" (p. 270).

On a more general level, Robert Fugmann (1984) has developed a theory of information supply and indexing based on five axioms: Definability, Order, Sufficient Degree of Order, Predictability, and Fidelity. Although part of a theory of indexing, three of these axioms relate directly to searching. The Axiom of Definability states: "The compilation of information relevant to a topic can be delegated only to the extent to which an inquirer can define the topic in terms of concepts and concept relations" (p. 118). The last two axioms address factors affecting performance, that is, precision and recall ratios. The Axiom of Predictability states that performance depends on the degree to which the query formulation predicts the modes in which concepts and their relations are expressed in the database, and the Axiom of Fidelity states that performance depends on the fidelity with which concepts and their relations are expressed in the database.

Based on Mooers's ideas, Dagobert Soergel (1985, chap. 13) proposes a comprehensive user-centered approach to indexing: the request-oriented approach. He reminds us that indexing would not be necessary if databases were very small. Theoretically, the best way to find an answer to a request is to check each document in a database and determine its usefulness for the request. This method of searching is very time consuming, both for the searcher and the user who has to wait for the information (and obviously impossible for real-life databases). Batching requests may save some time, as the searcher can examine each document for a batch of requests. However, while it saves much of the searcher's time, it makes the user's wait even longer.

The best way to save time (i.e., to make searching for information realistic) is to *anticipate* user requests and check each document when it is entered into the database against a list of anticipated requests. When a match occurs (i.e., a document is likely to satisfy a request on the list), the document is indexed under the request. According to this approach, indexing means representing each document in the database in terms of anticipated requests. The list of anticipated requests, or of their components, forms the index language. In indexing, the indexer asks: what requests (or components of requests) would the document satisfy? The index language "informs" the indexer about possible requests.

In indexing, then, each document is checked against every descriptor. This method is called the *checklist technique* of indexing:

- First, the indexer gets familiar with the subject and the structure of the document.
- Next, for each descriptor, the indexer asks: Will a person looking for material about (the descriptor) be interested in seeing this document?

Some issues in indexing policy do not apply to the checklist technique. For example, indexers cannot choose sources of terms because they use only descriptors. In addition, in order to index a document, an indexer may need to read different sections of a document, and even use additional sources. For instance, in an attempt to index exhaustively an article about a health ed-

ucation project in a certain city section, one may need to find out the ethnic composition of the section, even if it is not described in the article.

The checklist technique is likely to improve retrieval performance, even more so if supplemented with document-oriented indexing, and thus increase user satisfaction. But it is costly. Moreover, it is very difficult to implement when the index language is large, as is often the case. A proper display of the index language might help, however. When the index language is displayed in a classified structure, with relationships among descriptors, the indexer can check the class first and only then decide whether to check its individual members. In practice, several databases have selected a subset of their index language to be checklist descriptors. Indexers then check each document against this limited list of descriptors that are particularly important. The check tags used to index Medline are an example of such a list.

Automated Indexing

While request-oriented indexing derives from a user-centered approach, it can never predict all the requests which a document could satisfy. In addition, intellectual indexing, whether document oriented or request oriented, is always done *a priori,* before users search the database. Thus, if it is discovered in some way that indexing "failed" to predict certain elements of a request, nothing could be done to "improve" the indexing for an individual request. This static nature of intellectual indexing requires users to interact with retrieval systems to improve the results of their searches. But even the most skillful searcher may not be able to get satisfactory results if indexing does not address the searcher's requirements. Therefore, indexing processes that are dynamic are more promising as user-centered indexing methods because they may tailor indexing to the requirements of each request. Such dynamic methods would require the use of computers, that is, they would employ automated indexing.

In automated indexing the computer indexes documents. The full text of documents is the input and indexing is the output. At the beginning of its development, the main selling point of automated indexing was twofold. First, in contrast to human indexers who are subjective and inconsistent, computers are both objective and consistent. Second, in time it would become much cheaper to use computers than human indexers. Thus, over the last 30 years, developers of automated-indexing techniques tested the retrieval performance of their systems to show that they perform at least as well as systems employing intellectual indexing, and at times even better.

After being limited to the experimental realm for almost three decades, automated indexing is beginning to be used in systems that are publicly available, such as online public access catalogs (OPACs) and textual and bibliographic databases. There are several types of approach to automated indexing. In the statistically based approaches, the text of the documents is subjected to statistical manipulations of quantities such as frequency of term occurrence or co-occurrence in a document or in the whole database. Dictionary/rule-based approaches use external knowledge such as machine-readable, Roget-like thesauri and rules related to the process of indexing. In addition, automated indexing can be based on syntactic analysis as well as on semantic analysis. While these are distinct approaches, most systems employ a combination of approaches, and may even integrate citation indexing, Boolean searching, and links to intellectual indexing that use controlled vocabularies.

Automated indexing is clearly document oriented, because indexing is based only on the stored text. At the same time, its dynamic nature and flexibility make it a promising approach to user-centered indexing. In fact, automated indexing is already considered to be more user-friendly than intellectual indexing for four reasons[1]:

- Most automated systems accept requests in natural language. Users, therefore, do not have to express their queries in Boolean formulations, as they are required in systems that employ intellectual indexing.
- Such systems can provide relevance feedback, where the retrieval is improved after the user marks relevant documents that were retrieved previously by the system.
- Most systems provide ranked output in which the retrieved documents are listed according to their relevance to the request, most relevant first.
- Automated indexing and retrieval provide the capabilities for query expansion. That is, terms can be added automatically to the original query formulation to improve retrieval.

Further, from its very early stages, automated indexing promoted the idea that indexing and searching are two sides of the same coin, and that both are dynamic and interactive processes. Another fact indicates the user-centered nature of automated indexing: It has never addressed the issue of aboutness; the major quest has always been to find the technique that results in the "best" retrieval, not the one that represents documents best.

Thus, current research in automated indexing reflects a contradiction. On the one hand, it is the most user-centered approach because of its dynamic, helpful, and flexible nature. On the other hand, indexing is based solely on the text stored and is completely immune to the particular group of users and their queries. Even the most user-oriented features, relevance feedback, and query expansion, use request characteristics to modify *query formulations,* while indexing remains unchanged.

[1] Although these user-oriented facilities could be employed in databases with human indexing, in reality, they have been tested and implemented only for systems with machine indexing.

The idea that indexing can be tailored to each request approximates Soergel's suggestion mentioned earlier that a database be searched for each request. This notion, however, is foreign to automated indexing, where the purpose is to develop powerful methods that would always perform well. This might not be a useful target, however. Tibbo's (1994) article in this issue, for example, shows that subject domain is an important factor to consider, as information seeking in the humanities makes special requirements on indexing. Moreover, in theory, automated indexing has the best tools to respond to specific or momentary information needs; why should it then be limited to developing global techniques?

Soergel (1975) demonstrates the potential of automated indexing when he suggests that it follows the checklist technique. Rather than deriving index terms from the text, algorithms should be developed to determine the expected relevance of a document to each of the concepts in an index language that is composed of anticipated requests. Another example may illustrate this potential. Suppose it is discovered that users, when looking for information about a topic, at times prefer documents that are general in nature, at other times want to get all the specifics, and the rest of the times do not make this distinction. To accommodate for such requirements, indexing can be done on two levels. The first is *a priori*—whether intellectual, automated, or both—to reflect topics and other predetermined features. The second is *ad hoc,* tailored to specific requirements that are not satisfied by the first level, such as the level of treatment, whether general or specific. Current methods of automated indexing can retrieve articles about topics, but cannot discriminate among levels of treatment, even though it is probably possible to develop an indexing technique that does just that. This is because issues that are typical of individual requests and situational factors have not been addressed by automated methods.

Before we can develop indexing methods that are truly user-centered, however, we need to find out the specific requirements that users put to databases. Barry (1993), for example, discovered that faculty and students used 20 quality categories to determine the relevance of documents, such as: The extent to which information is presented in a clear or readable manner; the extent to which information is recent or up-to-date; the extent of knowledge with which the user approaches the document; and the extent to which a source of the document is well known or reputable.

While these results are a promising start, the study of online searching behavior of end-users is still in its infancy. New insights into this behavior are sorely needed, and not only for automated techniques. Milstead (1994), for instance, reports, in this issue, that human indexers have already expressed their need to understand searching behavior better. Whether to support intellectual or automated indexing, studying information seeking and searching behavior is necessary for user-centered indexing to develop. At present, information science is ready to consider a fully developed user-centered approach to indexing, but the tools and the understanding that are required for implementing this approach are still in a very early stage of development.

References

Anderson, J. D. (1994). Standards for indexing: Revising the American National Standard Guidelines Z39.4. *Journal of the American Society for Information Science, 45,* 628–636.

Barry, C. L. (1993). A preliminary examination of clues to relevance criteria within document representations. *ASIS '93: Proceedings of the 56th ASIS Annual Meeting, 30,* 81–86.

Beghtol, C. (1986). Bibliographic classification theory and text linguistics: Aboutness analysis, intertextuality and the cognitive act of classifying documents. *Journal of Documentation, 42,* 84–113.

Borko, H., & Bernier, C. L. (1978). *Indexing concepts and methods.* New York: Academic Press.

Brenner, C. W., & Mooers, C. N. (1958). A case history of a Zatocoding information retrieval system. In R. S. Casey, J. W. Perry, M. M. Berry, & A. Kent (Eds.), *Punched cards* (pp. 340–356). New York: Reinhold.

Cleveland, D. B., & Cleveland, A. D. (1983). *Introduction to indexing and abstracting.* Littleton, CO: Libraries Unlimited.

Cooper, W. S. (1969). Is interindexer consistency a hobgoblin? *American Documentation, 20,* 268–278.

Fairthorne, R. A. (1969). Content analysis, specification, and control. *Annual Review of Information Science and Technology, 4,* 73–109.

Fugmann, R. (1984). The five-axiom theory of indexing and information supply. *Journal of the American Society for Information Science, 36,* 116–129.

Hutchins, W. J. (1978). The concept of 'aboutness' in subject indexing. *Aslib Proceedings, 30,* 172–181.

Lancaster, F. W. (1991). *Indexing and abstracting in theory and practice.* London: The Library Association.

Maron, M. E. (1977). On indexing, retrieval and the meaning of about. *Journal of the American Society for Information Science, 28,* 38–43.

Milstead, J. L. (1994). Needs for research in indexing. *Journal of the American Society for Information Science, 45,* 577–582.

Rowley, J. E. (1988). *Abstracting and indexing* (2nd ed.). London: Clive Bingley.

Shatford Layne, S. (1994). Some issues in the indexing of images. *Journal of the American Society for Information Science, 45,* 583–588.

Soergel, D. (1975). Theoretical problems of thesaurus construction with particular reference to concept formation. In R. Petoefi (Ed.), *Fachsparache-Umgangssprache* (pp. 355–381). Kronberg, Germany: Scriptor.

Soergel, D. (1985). *Organizing information: Principles of data base and retrieval systems.* Orlando, FL: Academic Press.

Soergel, D. (1994). Indexing and retrieval performance. *Journal of the American Society for Information Science, 45,* 589–599.

Svenonius, E. (1994). Access to nonbook materials: The limits of subject indexing for visual and aural languages. *Journal of the American Society for Information Science, 45,* 600–606.

Swift, D. F., Winn, V., & Bramer, D. (1978). 'Aboutness' as a strategy in the social sciences. *Aslib Proceedings, 30,* 182–187.

Tibbo, H. R. (1994). Indexing for the humanities. *Journal of the American Society for Information Science, 45,* 607–619.

Vickery, B. C. (1953). Systematic subject indexing. *Journal of Documentation, 9,* 48–57.

Wellisch, H. H. (1991). *Indexing from A to Z.* Bronx, NY: H. W. Wilson.

Wellisch, H. H. (1994). Book and periodical indexing. *Journal of the American Society for Information Science, 45,* 620–627.

Wilson, P. (1978). Some fundamental concepts of information retrieval. *Drexel Library Quarterly, 14,* 10–24.