Academic indexing: what’s it all about?

Ross J. Todd

Raises some key issues needing to be addressed by indexing practitioners and scholars in order to deal more effectively with the problems of unacademic indexing. Terminological confusion over the notion of subject and the continuing debate over its permanent or indeterminate nature provide no clear nor consistent explanation of what indexers might actually look for when indexing a document. These problems are highlighted by empirical evidence on measures of inter-indexer consistency suggesting that substantial inconsistency is the rule rather than the exception. Little attention has been paid to how indexers actually determine the subjects of documents or what guides them in establishing the aboutness. While the literature identifies some broad approaches to subject analysis there is little supporting empirical evidence and few attempts to explicate any specifiable procedures. A productive step forward with indexing research would be to begin by examining how indexers actually undertake the process of subject analysis and to explore systematically factors that guide and influence this process. This would shed some light on a theory of subject analysis, clarify some of the central concepts of indexing, and provide an intelligent knowledge-base for effective, academic indexing practice.

John A. Vicker’s paper, ‘Unacademic Indexing’ [The Indexer 18(1), April 1992] highlights two key factors that impact on effective indexing: being able to discriminate between significant and trivial information so as to determine appropriately the subject(s) of documents; and applying indexing principles consistently so that the index is developed as an effective link between the user and the knowledge resource. Examining these problems in the context of the body of knowledge of indexing points the way for future in-depth discussion and research by both the academic community and practitioners, and the hope of solutions to these problems.

Indexing literature asserts that the effectiveness of information retrieval is predicated on the manner in which documents are represented through indexing. It recognizes that the process of indexing consists of two distinct phases: firstly, subject analysis—to analyse, identify and represent the intellectual contents of the document in some way; and secondly, the production of the index representation by translating the outcomes of subject analysis into a specified indexing language used by a particular retrieval system.

However, published accounts of indexing seem to ignore or skate over the mental processes that take place during subject analysis. The focus of the literature is very much on the translation stage, with emphasis on the construction of indexing languages to represent the intellectual contents of documents, the correct formulation of indexing entries, the construction of indexes, and the evaluation of indexes and retrieval systems. There is little about how indexers decide what the subject of a document is, how they determine its aboutness. More interest is currently being shown in automatic indexing than in improving human indexing. Essentially, the subject analysis phase is lamented as an intellectual operation resistant to investigation. This lack of focus on the cognitive processes of indexing—in discussion and in research—has contributed to a number of indexing problems that continue to hound the profession.

What is the subject?

First, what is actually meant by the subject of a document? Rather than clarifying this concept, key writers in the field present considerable terminological confusion.

Cutter defines the subject as the ‘theme or topic of the resource, whether stated in its title or not’. Kaiser clarifies subjects as the ‘things in general, real or imaginary, and the conditions attached to them. We shall call them concretes and processes respectively’. Ranganathan talks about the ‘thought content of a document’—an assumed term; an ‘isolate’. Coates identifies subject as the abstraction of the overall idea embodied in the subject content of a given literary unit. Vickery speaks of ‘the theme on which books, parts of books, articles or parts of articles are written; a complex aggregate of specific aspects; a compound of elementary terms’. Borko and Bernier define subjects as ‘the foci of a work, the central themes toward which the attention and efforts of the author have been directed. They are those aspects of a work that contain novel ideas, explanations, or interpretations. And they should all be indexed.’ And more recent literature appears to use the term aboutness synonymously with the term subject of a document.

Consistency

If we take the time to reflect on these definitions, we would discover that the notion of what a subject actually is is not consistently represented. It seems that at the very outset of the indexing process we are confronted with no consistent or clear explication of what we are looking for in a document in order to represent its intellectual contents as
subjects in some indexing system. Yet it is expected that a subject indexer will select subjects (whatever they are) from among all the concepts, propositions, arguments and examples expressed by the author. This may well be an explanation for the empirical evidence on measures of inter-indexer consistency. Inter-indexer consistency refers to the phenomenon of conflicting indexer decisions: if several different indexers are asked independently and individually to index the same document, a great deal of inconsistency is likely to be apparent in the results. In other words, the judgement and selection of terms to represent the subjects of the document will vary considerably from indexer to indexer. While different formulae for measuring the degree of inter-indexer consistency make generalizations difficult, it can be said as a broad characterization of findings that huge variations have been observed in the levels of inter-indexer consistency arising under different circumstances. In no case have perfect or high levels of consistency been reported: substantial amount of inter-indexer inconsistency is the rule rather than the exception. Zunde and Dexter found that no index term was selected by all indexers, and no two indexers selected the same set of terms for the document in question. Cooper cites some studies with consistency values ranging from 10%-80%, and others with values ranging from 40%-70%. The implication is that documents which are similar may and often would be indexed differently. Their similarity would not be reflected in the sets of indexing terms assigned to these documents. This notion itself introduces a great deal of uncertainty into any information retrieval system.

About aboutness

A second problem relates to the notion of a permanent aboutness that appears to underpin indexing practice. The current practice of subject analysis, leading to assigning index terms or subject headings as significant indicators of the intellectual contents of documents, makes the assumption that documents have a relatively permanent aboutness, and that indexers are able to state what the document is about by formulating expressions which indicate the subjects and its contents. Subjects are perceived as static, repeatable and objectifiable. The low measures of inter-indexer consistency do not clearly support this view, and currently there is some debate over the permanent or indeterminate nature of aboutness.

Fairthorne distinguishes between extensional aboutness as the inherent subject of a document, and intensional aboutness, the reason or purpose for which it has been acquired by the information agency or requested by the user. Boyce makes a similar distinction using the term topicality to define aboutness based on a view of subject matter treated in documents as being self-evident, and informativeness to express its subjective meaning as ascribed by users. Van Dijk makes the assumption that documents have an intrinsic subject, an aboutness that is at least to some extent independent of the temporary usage to which an individual might put one or more of its meanings. Beghtol takes the general position that documents of all kinds have a relatively permanent aboutness but a variable number of messages or meanings, each different according to exact use made of the document by users.

Wilson challenges the assumption that for any document a subject can be found that is co-extensive with its aboutness. He argues that the nature of the subject is essentially indeterminate, that nothing definite can be expected of the things found at any given position. Being on a given subject means 'being the sort of writing which our methods of assigning single locations assign to the position with such and such a name'. Wilson suggests that subject analysis is very much a subjective process, and that an indexer’s view of what a subject actually is shapes this process. Maron also recognizes that aboutness can be interpreted from several different points of view, and that each different viewpoint gives rise to different interpretations of the meaning of about. Oliver similarly argues that the same document can have different meanings for the same reader or indexer at different times, and it can have an unlimited number of meanings according to the perceptions of different uses that will be made of the documents at certain times.

The purpose of indexing

Such debate also raises a more fundamental question: What is indexing? Is indexing to make documents retrievable to satisfy information requests, or is indexing intended to convey the information conveyed in a document? It also raises the question of the relationship between indexing consistency and effectiveness of information retrieval systems, the extent to which increasing inter-indexer consistency will improve indexing (and the implication that consistent indexing is good and inconsistent indexing is bad).

Determination

These issues also raise another key problem that needs to be addressed. How does an indexer go about determining the subjects of a document? While it is accepted that an indexer performs an intellectual operation of some kind to determine the subject and represent it by means of index terms, what kind of operation is actually undertaken? The current literature of indexing does not address how the indexer is expected to be able to undertake the process of subject analysis. There is a need to understand the rules, procedures that govern the mental activities of subject analysis, not merely to establish causes to reduce the notorious inconsistency of indexers, but also to broaden the theoretical understanding of what is a fundamental operation of information practice. How are indexers able to write down the subjects of documents? What internal guidance systems determine their actions? What is the cognitive process of identifying the intrinsic aboutness of documents, that is the cognitive process by which an indexer describes the intrinsic subject or aboutness of a document? The literature indicates that there is no consensus about what or how indexers ought to do. Yet it seems that there must be some rules
guiding the mental activities of indexers, for otherwise it becomes impossible to explain how they are able to determine and write down statements of aboutness. It is essential that these questions set the direction of indexing research into the 21st century.

Several broad and somewhat overlapping approaches to subject analysis are presented in the literature, and these perhaps may form the starting point for future investigation. These broad approaches, however, provide little insight into the mental processes of subject analysis. Rather, they are more indicative of the notion of subject. Borko and Bernier identify a paraphrase approach that results in a short verbalization of the subject. Jones and Wilson suggest a purposive approach or textual framework approach whereby the predominant tendency, aim or purpose of the document is considered and the mental framework of the author identified. Wilson also suggests a number of additional approaches: a 'cast of characters' approach whereby the predominant ideas, names, and places are listed; an appeal to unity approach to identify the unifying concept that explains the presence of the rest and leaves no questions unanswered; and the figure-round approach to identify the idea occupying the centre of attention or that gives the impression of dominance. The term-significance approach of Coates is similar to the figure-round approach—the words or phrases that evoke the clearest mental image of a document.

Hutchins presents an approach based on linguistic analysis, particularly the thematic organization of text. He argues that text structure is based on themes: i.e., the given elements which the author assumes the reader already knows, and rhemes: i.e., elements expressing new or otherwise unpredictable ideas. The natural progression of given and new information enables the author to convey the message. This basic notion of semantic progression suggests a user-oriented process: identifying assumed or pre-supposed knowledge, that which is taken for granted by readers—i.e., certain linguistic competence, levels of general knowledge, cultural and educational backgrounds, interest or inquisitiveness, temperament; and that which is new to improve the present state of knowledge. In other words, an approach to subject analysis formulated in terms of knowledge presupposed by authors and that which is new.

There is no agreement on these broad approaches, little attempt to explicate a specifiable procedure for deciding aboutness, and little supporting empirical evidence. Yet underpinning all these broad approaches is the notion of summarization, the abstraction of the subject by some nebulous process of reduction or some kind of semantic condensation. Summarization is seen as a process of identifying and selecting subjects which are significant indicators of the contents and which together sum up the message of the documents. The inadequacies of present indexing procedures, problems of low inter-indexer consistency and inadequate explications of aboutness analysis, suggest that research might fruitfully begin by focusing on the mental processes involved in summarization.

Van Dijk and Kintsch perhaps come the closest to providing a potentially useful theory of summarization. Their descriptive model is based on the notion that information is semantic and hence propositional, and that any content analysis must yield a sequence of propositions. They claim that determining a document's subjects results from an ability to reduce the information in a document; it is an automatic reductive cognitive process of summarization. In this analytic and synthesizing process, sequences of sentences with their underlying logical propositions are cognitively compressed and summarized so as to produce a hierarchically governed sequence of macropropositions that express the sum of the meanings of the propositions subsumed under them. The end product of this continued reduction is the subject or aboutness. This is the highest specific macroproposition that is produced during cognitive reduction. Thus the level at which no further cognitive propositional reductions can be fruitfully made is the subject of the text because it hierarchically organizes all the detailed textual propositions into the most general macroproposition that meaningfully expresses the aboutness of the text.

Van Dijk suggests that these cognitive actions are governed by macrorules or macrostrategies or operations, and accordingly postulates five rules of subject analysis:

- **Rule 1. Weak deletion rule:** each proposition that is not an indirect interpretation of a subsequent proposition may be deleted;
- **Rule 2. Strong deletion rule:** each proposition that is not a direct interpretation of a subsequent proposition may be deleted;
- **Rule 3. Zero rule:** admit a proposition directly into the macrostructure with no reduction or summarization in short statements where no information is judged irrelevant;
- **Rule 4. Generalization rule:** each sequence of propositions may be substituted by the general proposition denoting an immediate superset;
- **Rule 5. Construction rule:** each sequence of propositions may be substituted by a proposition denoting a global fact of which the facts denoted by the microstructure propositions are normal conditions, components or consequences.

What is created is a statement of the document aboutness in the surface verbal elements available in the particular natural language. The model represents a specific formulation of how people mentally formulate the subject of a document. It is a bottom-up data driven approach, processing information actually contained in the text. The model has some support in empirical research done by Van Dijk in the area of recalling and summarizing stories and complex discourse, but it has not been tested by indexers. The notion of cognitive processes is simply assumed to operate during the subject analysis of a document for the purpose of indexing it by means of a particular indexing vocabulary.
Systematic effects

Any attempt at testing such a model in an indexing context must also generate discussion of and research into other factors that might also be brought to bear on the process of subject analysis. Beghtol suggests that the indexing system may influence aboutness identification. The indexer, knowing that the highest appropriate macroproposition beyond that associated with normal fluent reading, may unknowingly fail to analyse the document independently from the knowledge of the indexing system. This possibly means that documents with similar aboutness may be indexed differently by different indexers or by the same indexer at different times, depending on the indexing language used. Farrow suggests that indexers work under conditions quite different to the psychological experiments measuring readers' rates of comprehension of texts—they work under time constraints, scan documents rather than reading at the normal rate, and comprehend documents for the purpose of indexing them, taking no further interest in the text once the task is completed. Also, indexers work within the narrow range of text types and subject fields, suggesting that repetitive element may lead to automatic processing beyond that associated with normal fluent reading. Jones and Hutchins present some empirical evidence to suggest that subject matter, word frequency, semantic nets and structural features shape aboutness. Blair identifies subject area and indexer levels of training and experience as contributing factors.

Van Dijk also recognizes that extra-linguistic reality is a further ingredient that must be accounted for in a theoretical description of the indexing process. In the reductive process the reader is continuously and simultaneously undertaking a complementary process of inserting pre-existing knowledge or frame knowledge (conventional knowledge, expectations for both the content and structure of different kinds of documents, their interest, norms, purpose, attitudes) to complete the summarization process. This top-down approach, using information not contained in the document, is also recognized by Hutchins, who argues that readers possess an appropriate detailed level of frame knowledge and that authors write assuming this presupposed knowledge.

A productive step forward with research would be to begin investigating indexing practice that focuses on the first stage of indexing, subject analysis, and examine how indexers in practice actually undertake this process, and to identify and examine some of the factors that guide them in subject analysis. This would begin to shed some light on a theory of subject analysis, to clarify further some of the central concepts that underpin indexing practice, and provide an intelligent knowledge base for effective indexing practice. There is an urgent need for the profession to develop a theoretical knowledge base of indexing, and this knowledge, as a basis for practice, will contribute to more academic indexing.

References


Bibliography

Blair, D. Indeterminacy in the subject access to documents. Information Processing & Management 22(2) 1986, 229–41.

Ross Todd is lecturer in the School of Information Studies. University of Technology, Sydney, Australia.