Deconstructing the Indexing Process

Jens-Erik Mai
Royal School of Library and Information Science
Birketinget 6, 2300 Copenhagen S, Denmark

I. Introduction

The representation of knowledge contained in documents is one of the central and unique areas of study within library and information science. One might therefore be surprised to learn how little there is actually known about this area. It is common in the indexing and classification field to demand a set of rules or a prescription for *how* to index. When this demand arises, it is usually based on the assumption that it is possible to explain the intellectual operations that take place in the subject indexing process.

Shaw and Fouchereaux (1993) report on an investigation that the Research Committee of the American Society for Information Science undertook a few years ago. They identified a number of areas within information science which need more research, one of these areas is, "What are the cognitive processes involved in indexing and classification?" Milstead (1994) likewise discusses the need for more research in indexing and notes that "perhaps the most important need for research is one that has never been directly addressed . . . we have no idea of the mental processes involved when an indexer decides what a piece of information is 'about.'" More recently, Vickery (1997) investigated presuppositions in information science. He notes that if the indexing process is carried out by humans, as opposed to automatically, it "requires that [the indexer] establishes a *meaning* for the message, so that the choice of the 'most significant' elements may be made." Hutchins (1978) nicely sums up the situation:

> The literature of indexing and classification contains remarkably little discussion of the processes of indexing and classifying. We find a great deal about the construction of index languages and classification systems, about the principles of classification, about the correct formulation of index entries . . . and about the evaluation of indexes and information systems. But we find very little about how indexers and classifiers decide what the subject of a document is, how they decide what it is "about." (p. 172)

Statements like these can often be found in the indexing literature. This chapter takes up the challenge and makes a further investigation into the

ADVANCES IN LIBRARIANSHIP, VOL. 23

0065-2830/00 $30.00

problems of determining the subject content of a document. This is done through literature review. By critically examining previous understandings, a new understanding of indexing will emerge. It is not claimed that what is presented is the full answer to the fundamental question of how the subject content of a document is determined. What is presented here is the foundation for understanding the problems related to the task of determinig the subject content.

Retrieval of documents relies heavily on the quality of their representation. If the documents are represented poorly or inadequately, the quality of the searching will likewise be poor. This reminds one of the trivial but all too true phrase "garbage in, garbage out." The chief task for a theory of indexing and classification is to explain the problems related to representation and suggest improvements for practice.

By reviewing significant literature surrounding the indexing process this chapter explains some of the problems related to the representation of knowledge and, more specifically, deconstructs the subject-indexing process to explore what indexing is really about. This results in a call to redirect research in the field from striving at making general prescriptions of how to index and explaining the exact actions any indexer should take during the process for the optimal result, to a more holistic and operational approach. It is here suggested that research on indexing should be concerned with the broader questions of what indexing is.

The basic assumption underpinning the present study is that the main problems of the representation of documents is concerned with meaning and language. Similar assumptions have been taken by others; Fairthorne (1969), for instance, noted that "special topics can be treated as isolated topics only at the risk of sterility; therefore some acquaintance with the general problems of language and meaning is essential." Blair (1990), in the introduction to his book on language problems in information retrieval, notes that understanding of language is important. "The central task of information retrieval research is to understand how documents should be represented for effective retrieval. This is primarily a problem of language and meaning. Any theory of document representation . . . must be based on a clear theory of language and meaning" (p. vii–viii). In this respect, this paper argues that the subject-indexing process consists of a number of steps. These steps should be viewed as *interpretations* and not as rules of the mind. The object here, therefore, is to match this reality by taking an *interpretive approach* (Cornelius, 1996). What is needed is an understanding of the uncertainty and interpretative nature of the subject-indexing process in order to be able to properly teach and study indexing.

Benediktsson (1989) has noted the interpretative nature of the indexing process and the need for guidelines that recognize the significance of interpre-

tation. "[A]ny sort of bibliographical description . . . can be considered as descriptive. When it comes to interpretation, the question is: Ought not the description follow a method or standard as any canon which makes interpretation possible?" (p. 218). Before the literature is reviewed, a larger framework of problems related to the topic of this study is introduced. The understanding of how phenomena are perceived is the basic assumption underpinning an interpretative approach to indexing theory.

II. To See and Interpret

Sometimes it happens that two people see the same thing, but claim to see two different things. Why is that? Why is it that when two people see the same document, they cannot agree on the subject matter of the document? These basic, unanswered questions are at the core of the present review; questions that have puzzled psychologists and philosophers for centuries. This chapter does not, however, claim to answer these questions.

The problem of why people claim to see different things is often explained by illustrations. One example is the Necker cube (Fig. 1). Hanson (1958) notes about the Necker cube that it can be seen as representation of many kinds of things. "Do we all see the same thing? Some will see a perspex cube viewed from below. Others will see it from above. Still others will see it as a kind of polygonally-cut gem . . . It may be seen as a block of ice, an aquarium, a wire frame for a kite—or any of a number of other things" (p. 8). Wittgenstein (1958) also discussed the Necker cube. He argued that such an illustration could appear several places in a book and each time be understood differently. "In the relevant text something different is in question every time: here a glass cube, there an inverted open box, there a wire frame of that shape, there three boards forming a solid angle. Each time the text supplies the interpretation of the illustration" (p. 193). It seems that the context in which the illustration is found determines the meaning of the illustration. Even so, Wittgenstein argues that the context does not limit our interpretation.

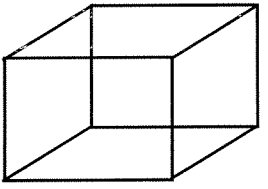

Fig. 1 The Necker cube.

"But we can also *see* the illustration now as one thing now as another. So we interpret it, and *see* it as we *interpret* it" (p. 193). The cube remains the same, but the interpretation of the cube changes. The cube itself does not change, and yet we see it differently. The basic question is how these differences can be accounted for. Although Wittgenstein is tempted to figure out what is actually going on inside his brain as he sees the illustration, he argues that the only way we can learn about the different ways in which the illustration can be seen is by the different descriptions of it. He rejects a mental investigation and relies on what is said and done. The reason for this is that he finds that what is important is what is done. "Do I really see something different each time, or do I only interpret what I see in a different way? I am inclined to say the former. But why?—To interpret is to think, to do something; seeing is a state" (p. 212). Other views must be acknowledged or recognized before a particular view can be argued for. The same holds for the indexing of documents; the indexer must be aware of what he or she picks as the subject entry, among the many subjects contained in the document, before the indexer is able to defend the view the indexer has chosen to index the document from. Indexing is not about seeing a document's text, but about interpreting its content and potentials.

Merrell (1995) likewise discusses this. He states that "people see, not retinas; cameras and eyeballs are blind." The process of seeing the cube entails two distinct operations, namely that of retinal stimulation and that of interpretation. That of retinal stimulation is an immediate feeling, and that of interpretation is a mediated intellectualization, "a synthesis of feeling and the effect . . . of something 'out there' on the sensory organs" (p. 47).

Merely seeing a document does not ensure that the subject matter of the document is "seen." This involves an interpretation of the document. The interpretation of a document involves the context in which the document is interpreted.

If the discussions and assumptions mentioned here are correct and valid, would it then ever be possible to represent the subject matter of documents? If any interpretation is as good and valid as any other, how would it ever be possible to claim that the subject matter of a document is this or that? The argument here is that this is possible, if the different kinds of interpretations involved in the subject-indexing process are known and understood, and the subject-indexing process is well described.

III. The Subject-Indexing Process

Various scholars have noted how little is actually known about the subject-indexing process. Farrow (1991) remarked after a thorough investigation of

the problem that "there is in any case a need for more research into the indexing process itself, if only because the validity of the analogies and parallels contained in the model is untested. It is surely remarkable that so little is really known about so basic a professional activity" (p. 164).

After Austin (1974) presented the theoretical and technical history of the PRECIS system, he simply stated that in a typical sequence of operations the indexer first examines the "document, and mentally formulates a title-like phrase which summarizes its subject content" (p. 49). Moss (1975) wonders why Austin uses only two lines on this matter, and more than forty pages on the rest of the indexing process. Moss points out that "it is apparent that the first—very much taken-for-granted—step is the crucial and vital one in any indexing and classification" (p. 116). Swift (1975) noticed that "Mr. Moss . . . touches on a major blind spot in thinking about indexing, at least for the social sciences. He draws attention to the glossing over . . . of what is involved in the first stage of indexing" (p. 117).

Jones (1976) later sums up the Austin–Moss–Swift discussion, and states that there have been "few attempts to . . . to establish the nature of the indexing *process*. Most studies claiming to be about indexing are, in fact, about indexes" (p. 118). Jones (1983) later supports this conclusion by an investigation into the nature of research into the subject indexing process. He found that "the relationship between text and index is rarely examined" (p. 1); more often, such research is concerned with the technical aspects of indexing and indexing languages.

A. How to Find the Subject of a Document

Manuals on classification and indexing are generally rather uninformative about how to identify the subject of a document. They merely recommend the indexer to examine tables of contents, scan chapter headings, and examine forewords and introductions.

The introduction to the *Dewey Decimal Classification* (DDC; 1996) gives guidelines for the determination of the subject matter of a work. In sum, these guidelines state that the indexer should investigate; (1) the title; (2) the table of contents and/or chapter headings; (3) the preface, the introduction, the foreword, and the book jacket and/or accompanying material; (4) the text itself; (5) bibliographical references and index entries; (6) cataloging copy from centralized cataloging services and the cataloging-in-publication data; and (7) outside sources such as reviews, reference works, and subject experts.

The DDC system's guide to determining the subject of a work merely points the catalogers to obvious places where the subject of a document could be found and states that the subject analysis might already have been done. Little is said, however, about the intellectual processes involved in the deter-

mination of subjects and of how the various pieces of information could be used in the determination of subjects.

The ISO standard (1985) on determining the subject of documents is similarly vague. It merely states that special attention should be paid to certain sources of information, many of which are similar to those found in the DDC's directions.

1. the title
2. the abstract, if provided
3. the list of contents
4. the introduction, the opening chapters and paragraphs, and the conclusion
5. illustrations, diagrams and tables and their captions
6. words or groups of words that are underlined or printed in an unusual typeface. (p. 581)

After the indexer has examined these parts of the document, the ISO instruction states that the indexer should identify the essential concepts that make up the subject of the document. This is done by consulting a checklist on which a number of questions are enumerated, such as the following:

1. Does the document deal with the object affected by the activity?
2. Does the subject contain an active concept (for example, an action, an operation, a process, etc.)?
3. Is the object affected by the activity identified?
4. Does the document deal with the agent of this action?
5. Does it refer to particular means for accomplishing the action (for example, special instruments, techniques, or methods)?
6. Were these factors considered in the context of a particular location environment?
7. Are any dependent or independent variables identified?
8. Was the subject considered from a special viewpoint not normally associated with that field of study (for example, a sociological study of religion)? (p. 581)

These two sets of guidelines, DDC and ISO, do not state anything about *how* the examination of the sources should be done or explicitly what the indexer or classifier should look for. They merely state the potential sources for finding the subject.

Many practitioners follow and use such guidelines, and they constitute the most common way to determine the subject of a document. However, it is quite obvious that these guidelines are vague and that they do not specifically state how to determine the actual subject of a document. Wilson (1968) notes that such manuals often refer to the subject of works, but they "are curiously

uninformative about how one goes about identifying the subject of a writing" (p. 73). Lancaster (1998) argues the ISO standard hardly can be considered a true standard, because "a true standard should be exact. . . and enforceable" (p. 147). Considering that the intention of the ISO standard is to "promote standard practice," the chance of error is great.

Furthermore, neither of these standard guidelines provide help for the cataloger or indexer who wants to determine the subject content based on an analysis of the users' needs or potential needs. These guidelines approach the subject-indexing process as a bottom-up process, where the aim is to determine the *subject* of the document, and not the potential uses of the document.

This suggests that it is almost impossible to formulate guidelines on how to determine the subject of a given document. There is, however, nothing new to this conclusion. Bates (1986), for example, has observed that "it is practically impossible to instruct indexers or catalogers [i] how to find subjects when they examine documents. Indeed, we cataloging instructors usually deal with this essential feature of the skill being taught by saying such vague and inadequate things as 'Look for the main topic of the document'" (p. 360). Cooper (1978) noticed that there has been some investigation into the problems of the process. He states that "it can be said that for one reason or another the findings have not been as enlightening as one could wish on the subject of how an indexer actually index [*sic*]" (p. 107). Cooper also notes that some studies of indexing have the character of investigations of how indexers *do* index, and not how they *should* index. He concludes that "the upshot is that there is as yet no concensus [*sic*] among experts about the answers to even some of the most basic questions of what indexers ought to be told to do or of how an indexer's performance should be evaluated" (p. 107).

Wilson (1968) pointed out four possible methods to determine the subject of a document: (1) the purposive method, (2) the figure–ground method, (3) the constantly referred to method, and (4) the appeal to unity method.

1. The Purposive Method.
   The *purposive method* is author oriented in the sense that the indexer seeks to find out what "the writer is trying to describe, report, narrate, prove, show, question, explain" (Wilson, 1968, p. 78). This is the subject of the document. The indexer should look for clues in the document, such as passages where the author writes explicitly what the purpose of the document is, such as: "I will show that . . . " or "It shall be proved that . . . . " The problem with this approach is that many authors do not state their aims openly.
2. The Figure–Ground Method.
   In the *figure–ground method*, the indexer tries to determine certain

aspects of the document that stand out or are most emphasized. This method relies heavily on the indexers' impression of the document, and this impression may vary from person to person.

3. The Constantly Referred to Method.
   The constantly referred to method (also known as *automatic indexing*) is the most objective method because the subject is determined by counting frequencies of occurrences of words in the document. It is assumed that if the word "Hobbes" occurs many times in the document, then the subject of the document is Hobbes. The problem is that there is not necessarily any correlation between occurrences of words in a document and its content.
4. The Appeal to Unity Method.
   Finally, in the *appeal to unity method*, the indexer tries to determine what makes the document cohesive or what makes the document a whole. The subject is what makes the document complete. Again, this relies heavily on the individual indexer because any two indexers might not agree on the same unity.

It should be noted that Wilson leaves out one major method of indexing that has received much attention lately: namely what could be named the *requirement-oriented method* or *user-oriented method*, which is essentially a variant of the purposive method. In the purposive method, the cataloger or indexer attempts to identify the author's intentions with the document. In the requirement-oriented method, the indexer attempts to identify the users' potential information needs and indexes the document accordingly.

Wilson ends up by saying that any of the methods could be used, and that the use of any of the methods might result in different descriptions of the subject matter of a document. However, none of the methods could be claimed to be better than the others. As Wilson says, "the notion of the subject of a writing is indeterminate" (p. 89).

Wilson follows a tradition within text analysis that typically has stipulated that there are two ideas of interpretation. As Eco (1994) explains, to interpret a text is either to "find out the meaning intended by its original author" or to assume that the text "can be interpreted in infinite ways" independently of the author's intentions (p. 24). If the latter approach is taken, one faces the dilemma of deciding where the meaning of the text then is found. The meaning could be determined by "(i) what the text says by virtue of its textual coherence and of an original underlying signification system or (ii) what the addressees found in it by virtue of their own systems of expectations" (p. 51).

Eco argues that there has been a change in focus from focusing on what the text says by virtue of its textual coherence, to focusing on what the readers find in the text. The first understanding focuses on textual structure and is

favored by structuralism, the latter focuses on the pragmatic aspect of reading and is favored by social constructivism.

IV. Document and Subject Analysis

A. Steps Toward Indexing

The representation of a document's subject takes, according to varying scholars, either two steps (Frohmann, 1990; Petersen, 1994), three steps (Farrow, 1991; Miksa, 1983; Taylor, 1994), or four steps (Chu and O'Brien, 1993; Langridge, 1989).

The two-step procedure consists of one step in which the subject matter is determined and another step in which the subject is translated and expressed in the indexing language. The three-step procedure inserts a step in which the subject matter is formulated explicitly or implicitly. In the four-step procedure, the translation of the subject matter into the indexing language consists of two steps. The indexer first translates the subject matter from his or her vocabulary into the vocabulary used in the indexing language. Then the indexer constructs the subject entry in the indexing language in the form of index terms, a class mark, or a subject heading.

For experienced indexers and catalogers, all steps, regardless of how many one supposes to be effective, may take place almost simultaneously. However, it is useful to operate with intermediate steps when analyzing the process. The three-step model is used for analysis here.

The first step, the *document analysis process*, is the analysis of the document for its subject. The second step, the *subject description process*, is the formulation of an indexing phrase or subject description. The third step, the *subject analysis process*, is the translation of the subject description into an indexing language.

The three steps link four elements of the process. The first element is the *document* under examination. The second element is the *subject* of the document. This element is only present in the mind of the indexer in a rather informal way. The third element is a *formal written description of the subject.* The fourth is the *subject entry*, which has been constructed in the indexing language and represents the formal description of the subject.

Miksa (1983) introduced a geometrical diagram to represent the subject indexing process, or "scope matching process," which implies that subsequent stages of the process are based on interpretations of earlier stages (Fig. 2). This model implies that the referents of each stage in the process are the cataloger's or indexer's understanding, and hence that the results of the process are rather indeterminate and dependent on the cataloger and indexer who performs the analysis. "Given the document *S*, its subject may be repre-

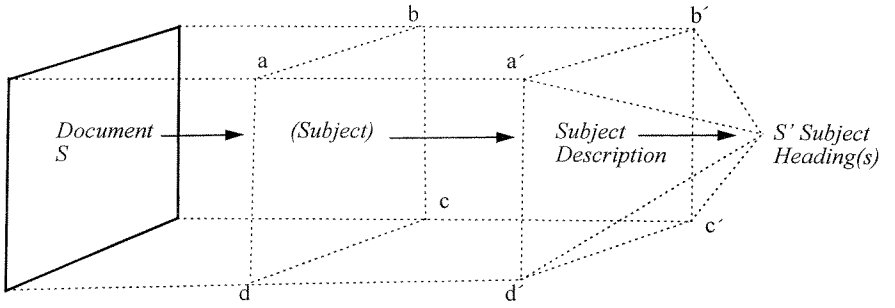

Fig. 2 The scope-matching process.

sented by the arbitrary figure a, b, c, d; the description of the subject by 'a,' 'b,' 'c,' 'd'; and the name of the subject that is useful as a subject heading and that portrays the essential nature of the subject by 'S'" (p. 5). The figure suggests that there is a "certain undefinable substantiality or scope" to the process.

Miksa further discusses what the referent of a subject heading is. He argues that the literature of subject representation tends to answer this question simply by stating that "a subject heading should express or match in some essential way the topical content of a work" (p. 5). He argues that the correspondence between a subject heading and the subject of a document has only a casual correspondence because the subject matter of documents cannot be measured precisely.

The task of a subject cataloger is to move from the document to the subject entry in order to devise a subject name for the subject content of the document. In other words, the task of the subject cataloger is to transform the content of the document into a representation of the document. This process is successful when the subject cataloger has determined a "name or names . . . (which) suggest, represent, fit, match, etc., . . . the supposed substantiality of the topical content" (p. 7).

Miksa's geometrical figure may be modified. The size of each square in Fig. 3 indicates the range of possible referents at each stage. In Miksa's model the three squares are of identical size. In Fig. 3, however, the squares reduce in size during the process. This indicates that the range of possible referents is larger at the beginning of the process than at the end. The idea is that the possible referents are greater for a document than for a subject entry.

B. Novice and Expert Indexers

If the argument that indexing is a set of closely related interpretations is valid, then an investigation of the subject-indexing process most likely reveals

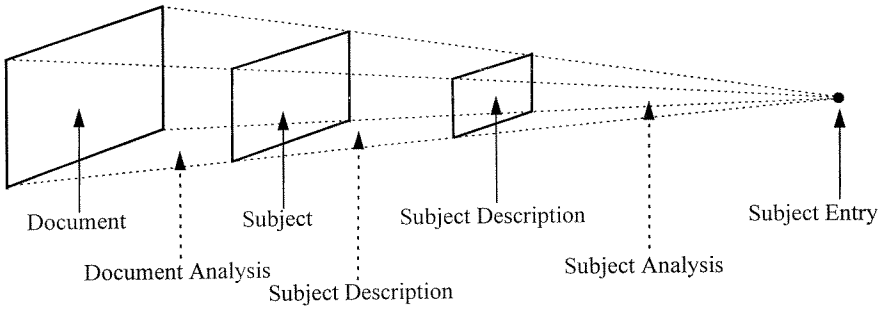

Fig. 3 The subject-indexing process.

that novice indexers often break down the indexing process to individual processes and that expert indexers seldom think about indexing as consisting of more steps. Dreyfus and Dreyfus (1986) developed this idea and identified five stages from being a novice to becoming an expert. These stages are introduced here and related to the skills an indexer has at the different stages.

Stage 1: Novice. "The novice learns to recognize various objective facts and features relevant to the skill and acquires rules for determining actions based upon those facts and features" (Dreyfus and Dreyfus, 1986, p. 21). The novice is only able to apply the learned rules in familiar situations. The learned rules are therefore *context-free rules.* The novice applies the learned rules in a particular situation, if the situation matches the learned proto-situation. The novice is not able to accommodate for the given context of the situation. The novice indexer has typically read manuals and guidelines and applies these strictly to every indexing situation. True novice indexers are probably only found in introductory classes in library schools.

Stage 2: Advanced beginner. As the novice becomes familiar with real situations, they advance "through practical experience in concrete situations with meaningful elements, which neither an instructor nor the learner can define in terms of objectively recognizable context-free features" (Dreyfus and Dreyfus, 1986, p. 22). Thus, the advanced beginner uses both context-free rules and "situational" elements in performing a task. The advanced beginner has learned the basic rules and indexed some documents and is able to engage in discussions of the indexing of particular documents in a particular system.

Stage 3: Competence. The competent performer approaches the situation with a plan to organize the situation in mind. The situation is then examined according to the plan, and only those factors that are important to the

given plan are given consideration. The competent performer "feels responsible for, and thus emotionally involved in the product of choice" (Dreyfus and Dreyfus, 1986, p. 26). The competent indexer looks for certain information in the document and, from that information, indexes the document.

Stage 4: Proficiency. The proficient performer relies on intuition and not on rules when performing a task. Intuition here means the "sort of ability we all use all the time as we go about our everyday tasks" (Dreyfus and Dreyfus, 1986, p. 29). It is unlikely that the proficient performer is able to explain in detail how she solved the problem. The proficient indexer examines the document at hand without manuals and guidelines in mind. She follows intuition without being able to articulate the exact actions taken. When the proficient performer has decided on a subject entry, she most likely is not able to explain how the result was reached.

Stage 5: Expertise. "An expert generally knows what to do based on mature and practiced understanding" (Dreyfus and Dreyfus, 1986, p. 30). The expert does not see problems in a detached way; she is so much a part of the situation that she does not need to be aware of the situation to handle it. When things work as they normally do, the expert is able to deliberate before acting. However, this deliberation is not based on calculative problem solving, but on critically reflecting on her intuition. As the performer becomes an expert, her performance becomes fluid. The expert is not able to legitimize decisions, but seldom makes wrong decisions. The expert indexer typically has years of experience within the same environment. Unlike the lower-level performers, the expert indexer is able to index the same document using different approaches. In other words, the expert indexer is able to choose to index a document according to an entity-oriented or according to a user-oriented conception (these conceptions are discussed later in the chapter).

The action taken by indexers with different levels of experience might not be exactly the same, but the steps and elements of the subject-indexing process are considered fundamental for any indexing process. Whether this assumption actually is the case is beyond the scope of the present investigation.

Dreyfus and Dreyfus' idea of the development from novice to expert is helpful in explaining the different stages indexers go through.

V. Studies of Indexing

This section reviews and discusses a portion of the literature on the problem of how to determine the subject matter of documents. The aim is to show

the range of topics discussed in the field and argue that the higher aim of these studies basically has been to find the best way to index.

First, however, a terminological clarification. The phrase "subject analysis" is used in at least two different ways in the library and information science literature: (1) to denote the area of study concerned with the construction and use of indexing language and classification systems, and (2) to denote analysis of the topical content of a document. The first area has received most attention. It has been reviewed by, for example, Lancaster, Elliker, and Connell (1989), Liston and Howder (1977), Schwartz and Eisenmann (1986), and Travis and Fidel (1982). The concern here is with the latter definition of the phrase—namely, that of determining the subject content of a document. In this tradition of study, *subject analysis* is sometimes used to denote the last step of the process (Chan, Richmond, and Svenonius, 1985) and sometimes to denote the first step (Langridge 1989). Hjørland (1997) notices that other terms have been used; for instance, content analysis, conceptual analysis, information analysis, aboutness analysis, and text analysis.

To clarify the vocabulary in the area, as discussed previously, it is suggested that the phrase "document analysis" is used for the first step in the process and "subject analysis" for the last step (see Fig. 3).

A. Perceptual and Conceptual Indexing

Farrow (1991, 1994, 1995) examined the subject indexing process on the basis of speed-reading studies. Farrow argues that the process could be analyzed as a top-down process, using information that is not contained in the document, but is part of the indexer's world knowledge; or as a bottom-up process, using information that is contained in the document for indexing. Farrow calls the first process (top-down) *conceptual indexing*, and the last process (bottom-up) *perceptual indexing*.

Perceptual indexing takes the form of scanning the text for cues. This might be long words, words that are italicized or underlined, or words in headings or otherwise emphasized. The focus of the scanning could also be areas of the texts such as the introduction, conclusion, or lead sentences, such as: "In this paper, we . . . ," "We will, in this paper, prove that . . . ." The indexer simply picks out words or phrases from the text.

Conceptual indexing, however, relies on the knowledge of the indexer. The indexer's knowledge of the subject matter, as well as the structure of the text itself, influences the quality of indexing. Farrow adds three more factors: indexers' knowledge about (1) the system they are using, (2) the users of the systems, and (3) the background of general world knowledge.

Farrow provides a model for indexing, abstracting, and classification, based on a comprehension model that is simple and highly structured. How-

ever, he fails to mention the importance of the knowledge on which the indexing process relies: namely, system knowledge, user knowledge, and world-knowledge. He further argues that this framework of understanding pinpoints the "causes of inadequate or inaccurate indexing" (p. 163).

B. Conceptions of Indexing

Albrechtsen (1992, 1993) argues that there are three general conceptions or viewpoints of subject analysis and indexing (Fig. 4). She argues that these relate to the type of information that constitutes the subject and which method of indexing is being used. The model incorporates three conceptions of subject analysis and indexing: the simplistic conception (subjects as objective entities), the content-oriented conception (implicit subjects), and the requirement-oriented conception (focusing on users' needs).

Albrechtsen has in a very simple but effective way drawn the boundaries for discussing subject analysis and indexing. She provides a very useful framework for investigating and studying the limits and problems in subject analysis and indexing. More specifically, she argues for a pragmatic approach to indexing in which the subject of the document is determined by the use of the document, and not merely by the text in the document.

C. Information Needs and Indexing

Weinberg (1988) discusses why traditional indexing fails to satisfy the needs of researchers. She argues that scholars and researchers are interested in "ideas and theories, and want to know whether specific ideas have previously been expressed in the literature" (p. 3). Traditional indexes do not provide this kind of information, and indexers are not trained to derive this kind of information from the documents. Weinberg concludes that researchers and scholars have to rely on extensive reading and prodigious memory.

Vickery (1968) discusses analysis of information, by which he means "deriving from a document a set of words that serves as a condensed represen-

Conceptions of subject analysis and indexing	Type of subject information	Indexing method
Simplistic conception	Explicit information	Extraction
Content-oriented conception	Implicit information	Assignment
Requirement-oriented conception	Pragmatic information, contextual potentials	

Fig. 4 Conceptions of subject analysis and indexing.

tation of it" (p. 355). He notes that any text could be used for multiple purposes, depending on the needs of the user. Simultaneously, each of these purposes could represent a point of view from which the subject analysis could be made. He makes the crucial observation that "the subject analysis we make will depend on the users we expect to serve" (pp. 359–360). Vickery holds that any analysis must have the potential users of the documents in mind, and the analysis should be based on the needs of these users.

He notes that there are two types of indexers—namely, those who fully understand or attempt to fully understand the subject content of the document under investigation, and those who do not. The first are capable of expressing the subject content in words other than the authors'. The second merely pick out words that the authors have emphasized, such as title section headings, conclusions, and summary. Vickery notes that very little is known of how it is possible for indexers to achieve such an understanding.

Cooper (1978) developed an indexing method, Gedanken Indexing—thought experiment indexing—that is based on users' probable *utility* of one index term versus another. Cooper argues that the indexer should perform the following task to decide whether to assign a given index term:

1. Predict the relative propositions of future users who will derive negative and positive utilities from their experiences with the card (index term).
2. Predict the average negative utility.
3. Predict similarly the average positive utility.
4. Compute the predicted average utility by adding the results in (2) and (3) weighted by the propositions determined in (1). (p. 114)

The idea is that the indexer should try to picture what a sample of her library users might look for under each of the terms under consideration. The task is to "picture their precise information need" (Cooper, 1978, p. 112).

The method requires that a list of suitable index terms actually be available. The method then assumes that the user population is fixed and well known by the indexer. For indexing in large scale databases in which the user population is not well known this method does not seem feasible.

It is evident that any representation of documents must take the users', or potential users', information needs into account when determining the subject of the document. The problem is that it is difficult, perhaps even impossible, to prescribe how to do this.

D. Summarization vs. Depth Analysis

Taylor (1994) defines subject analysis as that part of cataloging that deals with the conceptual analysis of an item, the translation of the analysis into the conceptual framework of the classification system, and the translating

of the conceptual framework into specific classificatory symbols or specific terminology. A conceptual analysis of an item deals with determining what "the intellectual content of an item is 'about' and/or determining what an item 'is' " (p. 101).

Taylor argues that one must have a clear idea of which approach to exhaustivity will be applied. Depth indexing identifies all the main concepts dealt with in the document and recognizes many subtopics and subthemes. Summarization identifies only a dominant, overall subject, and recognizes only concepts embodied in the main theme. Taylor argues that library cataloging has traditionally been concerned only with the summarization level, which is the attempt to identify one overall concept that would encompass the whole document and has been justified by extensive back-of-the-book indexes. The users would retrieve a large class of books based on the summarization of the books in catalogs and pick the relevant ones on the basis of the back-of-the-book indexes.

Taylor defines the difference between *document retrieval* and *information retrieval* as the difference between summarization and in-depth indexing. Summarization allows the user to retrieve documents (document retrieval), after which the user will use the document's (book's) internal subject index to retrieve the relevant information (information retrieval). Taylor's differentiation between information retrieval and document retrieval is less useful. There has been much research and development in document retrieval in which the level of analysis is in-depth analysis, but what is retrieved is still documents and not information. It might be more useful to simply say that information retrieval is the same as document retrieval.

E. Empirical Investigation

Chu and O'Brien (1993) argue that the first step of the subject indexing process has been neglected. The authors therefore set up an investigation of how well indexers perform the first critical step. To emphasize that they were only interested in the first step, they asked experimental participants to condense a text into a single sentence.

The authors asked a total of 104 students from the departments of library and information studies at the University of California at Los Angeles and Loughborough University of Technology to read three short, popular articles and subject analyze the documents. One article selected represented science, one represented social science, and one represented the humanities.

The participants were asked to fill out a questionnaire in which they were asked about the subject of the article and what they would consider the primary and secondary subjects and their relative importance. They were further asked how easy it had been to determine the subject and whether the

layout (defined as the bibliographic apparatus, e.g., title, abstract, first and last paragraph, keywords, illustrations, and the physical presentation) of a work had helped the analysis.

The authors concluded that it is possible for novice indexers to determine the main subject of a short text, but added that there seems to be "a serious problem when participants are required to isolate primary and secondary topics" and that "the bibliographical apparatus is a major factor in determining the general subject matter of a text" (Chu and O'Brien, 1993, p. 453).

It does not seem possible to generalize the results of Chu and O'Brien's study to the indexing of longer articles or books. It is inevitable that longer and more complex texts will make it more difficult to determine the subject and the potential uses of the document. It should also be noted that Chu and O'Brien used novice library and information science students as participants, and used popular articles that required very little (or no) specialized knowledge.

F. Mentalism

Frohmann (1990) criticizes what he calls "mentalism in indexing" as it has been promoted by Farradane (1979, 1980), Anderson (1985), and Beghtol (1986). The intention of these studies is to find the rules by which one should index.

Frohmann argues that "there must be some rules guiding the mental activities of indexers, for otherwise it becomes impossible to explain *how* they are able to utter or to write down an indexing phrase for the text. The problem is to discover the precise form of these rules" (p. 84).

The mentalistic-oriented conception of indexing assumes that it is possible to uncover the intellectual processes within the process and that when these rules are uncovered it would be possible to prescribe the best way to index. This assumption is closely connected to the cognitive viewpoint in information science that became popular through the 1980s. This approach was, to a certain degree, based on the cognitive science movement, as it, for instance, was defined by Gardner (1985), who argues that "it is necessary to speak about mental representations and to posit a level of analysis wholly separate from the biological or neurological, on the one hand, and the sociological or cultural, on the other" (p. 6). The idea is that the indexers derive the subjects of documents by invoking mental rules, and thereby automatically generate the mental representation of the subject matter from the mental representation of the documents.

Frohmann argues, in the words of Wittgenstein, that to follow a rule is a practice, technique, or custom, and hence rules are embedded in social life. To understand a rule therefore requires an understanding of the practice.

Rules and practice are of the same kind and cannot be understood separately. He argues that mentalism in indexing research conceals "fruitful directions of inquiry in at least the following five areas" (p. 94):

1. Mentalism focuses on *discovering* the rules of the indexer when focus should be directed toward *constructing* the rules. It is false to believe that tacitly known rules are followed unconsciously.
2. Mentalism favors rule systems that take their ground in the rules of the mind, and thereby conceal other rule systems.
3. Mentalism conceals the text. The text is not regarded independently of the object of study; namely, the mental representations.
4. Mentalism conceals intertextuality by focusing on single-text processing.
5. By focusing on the processes that occur in the mind, mentalism conceals the crucial social context of rules.

Common for much of the literature on indexing is that the authors are searching for an answer to *how* to index. They are driven by the need to find the rules that an indexer must follow to represent knowledge properly. Frohmann challenges this common assumption. He argues that such an approach *conceals* very important aspects of the process.

Frohmann argues that an attempt to find the rules of indexing is bound to fail. There are, in principle, two reasons for this; first, the *application* of rules always take place in a social praxis, and this praxis determines whether the rule is correctly used; and second, the *creation* of rules always take place in a social praxis, and the meaning of the rules are determined by the social praxis. The social praxis, the creation of rules, and the application of rules are closely connected. The mentalistic approach attempts to create context free rules that can be applied in any social context. This is not possible.

VI. Five Conceptions of Indexing

There is a trend in the indexing literature to discuss the advantages and disadvantages of different approaches to indexing. Such discussions have often led to dual conceptions of indexing; Soergel (1985) discusses the difference between entity- and request-oriented indexing; Fidel (1994) discusses the difference between a document- and a user-oriented approach; and, as shown here, Albrechtsen (1992, 1993) discusses the difference between user- and requirement-oriented indexing. These discussions have led to the belief that there are only two approaches or conceptions of indexing, and that these are opposed. A closer look at the conceptions reveal that it is possible to enumerate at least five conceptions. These conceptions are syntheses of understandings

discussed in this chapter, but they are furthermore argued to be based in different epistemological positions.

Hjørland (1997) argues that there are four basic epistemological positions that are of interest to knowledge representation. He defines these as (1) empiricism or positivism—the view that knowledge is given *a priori*, that knowledge is modular, that individual sensations are the basis for obtaining knowledge, and that knowledge only can be obtained empirically; (2) rationalism—the view that it is possible to formulate basic principles for obtaining knowledge, that the primary source of knowledge is reason, and that a good analysis will lead to the truth; (3) historicism—the view that the principles for obtaining knowledge develops historically, that it is not possible to define exhaustive and fundamental principles for obtaining knowledge, that knowledge is defined by cultural and historical contexts and not by individual sensations or rationalizing, and that knowledge is not modular; and (4) pragmatism—the view that knowledge develops primarily from praxis, that the future use determines the context for knowledge, and that knowledge is context dependent. In the following discussion, the relationship between different conceptions of indexing and different epistemological positions are brought together.

The five basic conceptions of indexing are therefore defined and argued as follows:

1. The *simplistic conception* of indexing. This conception is similar to Wilson's (1968) constantly referred to method and Albrechtsen's (1992, 1993) "simplistic conception," and focuses solely on automatic extraction and statistical manipulation of words. The idea is that the sum of the words in the document constitutes the subject matter; therefore, this conception of indexing is linked to empiricism.
2. The *document-oriented* conception focuses on the information that is present in the document. This relates to Farrow's (1991, 1994, 1995) definition of perceptual indexing and to Wilson's (1968) figure–ground and purposive methods. In this conception, the indexer investigates specific parts of the document, and the importance of the information is determined by the indexer. The document-oriented conception is closely related to the rationalist position—that by pure reasoning it is possible to objectively determine the subject matter of documents. Such an approach to indexing defines some fundamental principles by which it is possible to determine the subject matter objectively.
3. The *content-oriented* conception attempts to describe the content of the document as fully as possible. This is an objectivist conception that, in the extreme, would claim that there is only one correct analysis

of a given document. As with the historic epistemological position, the content-oriented approach assumes that by a careful investigation into the different interpretations of the document, it is possible to determine the subject of the document. This conception determines the basic subject matter of the origin of the document; that is, the context in which the document is produced. In other words, it is historical and cultural circumstances that determine the subject matter of the documents.

4. The *user-oriented* conception focuses on users. Focus is directed toward the user's general knowledge level or toward the user's work or research domain. The indexer pays special attention to the knowledge level of the users. Users at a public library may require a different kind of indexing than users at a university library, even for the same document. If the indexer works in an organization that serves users with a particular interest or work domain, the indexer should pay special attention to the work domain, and not to the individual users. The user-oriented conception is clearly based on the potential future use of the document, and is therefore based in a pragmatic epistemological position. The analysis of the document is shaped by the potential user group's work in a domain. In such an approach, the subject matter of documents cannot be determined objectively. It changes as the members of the user group changes, as the interests and tasks of the user group changes, and as the domain changes.
5. In the *requirement-oriented* conception, the indexers have knowledge about the users' individual information needs and work tasks. A requirement-oriented conception is only useful in smaller organizations—for example, a consulting firm that employs 15 consultants, in which it is possible for the librarian to know the needs and tasks of each individual. As with the user-oriented conception, the requirement-oriented conception is also based in the pragmatic epistemological position. The requirement-oriented conception is more narrow in focus because it bases the document analysis on specific information needs of specific persons. However, common to both is that they base the analysis on the potential future use of the documents. Indexing done according to the requirement-oriented conception, like the user-oriented conception, changes over time.

The conceptions cannot be seen in isolation. There is probably no indexing unit that bases their indexing solely on one of the conceptions mentioned here. They are related to each other, as on a continuum. An indexing unit will most likely choose a combination of two of the five different conceptions. A consulting firm might choose to combine the user-oriented conception

with the requirement-oriented conception to ensure that the indexing is of some lasting value. University and public libraries might choose to combine the user-oriented conception with a content-oriented conception, again to ensure lasting value of the indexing.

The upshot is that it is impossible to claim that an indexing unit is based in only one conception. It is most likely using a combination, and it is impossible to define one conception of indexing that works better than others—it all depends on the particular situation of the indexing unit. Moreover, as is shown here, each conception is closely tied to different epistemological positions.

VII. Relevance Studies

Measuring the performance of an information retrieval (IR) system is a central aspect of information retrieval theory. Relevance is used to measure how well a particular IR system or technique retrieves documents on the basis of a request or information need. The underlying assumption for this chapter is that researchers of the concept of relevance must operate (explicitly or implicitly) with a concept of subject in order to measure how well a system or technique retrieves documents on a particular subject. Thus, if they operate with a particular concept of a subject, that concept might have bearing on the conception of the subject-indexing process.

Cooper (1971) distinguishes between two uses of relevance; namely, relevance as a logical relationship between the topic sought and the document found, and relevance as usefulness for the user. Cooper argues that it seems viable to argue that relevance is a relation between portions of stored information in documents and something called an *information need.* Cooper does not attempt to define what an information need is, but merely states that it is a psychological state that is not directly observable. He further acknowledges that the words stated by the user are *not* the same as the information need. The aim, therefore, is to define relevance as the relationship between the stored information and the user's information need as stated.

Cooper argues that because the utility of documents is relative to different users it is not possible to measure such a concept of relevance. He believes that relevance should be defined as the relation between topics of the stated need and the retrieved documents. Cooper's definition of relevance also requires a clear concept of subject.

Swanson (1977) defines two frameworks for understanding relevance. Within the first framework relevance is understood as something that users create or construct from whatever new knowledge they derive from the document. In this sense, relevance is not something that is present in the

document that can be measured or judged. The relevance judgment is solely based on the requester's use of the information in the document. Within the second framework of relevance, "relevant" is defined as "being on the same topic." In this sense, a document is relevant if it is on the same topic as the question that the user enters into the system. The judgment of relevance in this sense can be made by people other than the user. It is important to note that even though a document provided and a stated request are on the same topic, the document might not necessarily fulfill the user's underlying need for information.

Buckland (1983) distinguishes between three uses of the term *relevance*. He defines these as (1) responsiveness: the measurement of the system's ability to retrieve correct data on the basis of the attributes used as the basis for retrieval; (2) pertinence: a narrower use of responsiveness—namely, when the attribute used for retrieval is the subject matter; and (3) beneficially: the degree to which the user of the system can utilize the retrieved data.

Buckland argues that relevance in the third sense, the degree to which the document is beneficial to the user, cannot properly be used to evaluate the performance of retrieval processes as retrieval processes. Relevance should only be used in the narrower definition, in which it is the ability of the system to respond to an inquirer's formulated inquiry.

Buckland argues that measurements of relevance need to leave out the user's subjective and individual information needs because it is not possible to objectively measure the effectiveness of an IR system on the basis of how users utilize data retrieved from the system. However, a measurement of an IR system based on the *pertinent* sense of relevance, in which relevance is measured against the subject of the document, and the query requires a theory of how to derive the subject matter of a document.

Saracevic (1970) clarifies confusing terminology used in literature on relevance. On the basis of earlier definitions of relevance, he develops an algorithm in which he joins these definitions:

> Relevance is the __A__ of __B__ existing between __C__ and __D__ as judged by __E__.
>
> where: A represents the gage of measure;
> B represents the aspect of relevance;
> C represents the object upon which relevance is measured;
> D represents the context within which relevance is measured;
> E represents the assessor. (p. 120)

With the appropriate words inserted, many statements could be formulated (Fig. 5). By combining the terms from the five categories, different connotations of a definition of relevance may be given, but the basic problem remains the same: A judgment of relevance is a subjective measurement. Two people with the same information need will not have the same perception of the

Relevance is the
a. measure
b. degree
c. extent
d. judgment
e. estimate
f. appraisal
g. relation
h. quantity
i. dimension

of
a. utility
b. matching
c. informativeness
d. satisfaction
e. appropriateness
f. usefulness
g. correspondence
h. importance
i. connection
j. fit
k. similarity
l. applicability
m. closeness

between a
a. document
b. document representation
c. reference
d. textual form
e. information provided
f. fact
g. article

and a
a. question
b. question representation
c. research stage
d. information need
e. information used
f. point of view
g. request

as judged by a
a. requester
b. intermediary
c. expert
d. user
e. person
f. judge
g. information specialist
h. librarian
i. delegate

Fig. 5 Definition of relevance.

relevance of a document. Saracevic's definition of relevance clearly shows the many variables the notion has and how problematic the concept is.

Green (1995) argues that other relations than topicality might influence the relevance judgment of documents. Green distinguishes between strong and weak relevance. Strong relevance is the relationship between the user's need for information and the documents, whereas weak relevance is the relationship between the topic of the user's request and the topic of the documents. Green states that the ideal system must be one that considers relevance as "the property of a text's being potentially helpful to a user in the resolution of a need" (p. 647).

Thus, the measurement of the system should rely on the system's ability to provide the user with useful information. It is not enough simply to measure the system's ability to retrieve topical relevant information.

Barry (1994) argues that there are two important implications for the evaluation of IR systems. First, there are factors other than topical appropriateness that influence users' judgment of relevance. Second, users are apparently able to recognize and discuss nontopical aspects of documents that influence their judgments. Barry argues that because "users approach information retrieval systems in hopes of finding information that has some *meaning* for them" (p. 151; author's emphasis), retrieval mechanisms based primarily

on topical matching may be failing to address the needs that users bring to the systems. Barry, however, advocates that it must be accepted that IR systems can only retrieve documents based on topicality and should thus be evaluated on that basis alone.

Park (1994) argues that topical relevance is context free and is based on fixed assumptions about the relationship between the topic of a document and a search question, ignoring an individual's particular context and state of needs. Measurements of relevance therefore must take into account the users' individual information needs and base the evaluation on these.

The best summary of the problem of defining the concept of relevance is probably given by Rees and Saracevic (1966), who incorporated findings of the early research on relevance measurement into basic characteristics of relevance.

There is a sharp distinction between relevance to a question and relevance to the underlying information need.
Only the user can judge whether a document is relevant (i.e., the relevance judgment is subjective).
Relevance judgments are not constant; they will change over time.
Documents found relevant by one user will not necessarily be found relevant by another user with the same question (i.e., the underlying information need may differ for users with the same question. (p. 229)

VIII. Information Needs, Relevance, and Indexing

The essence of the relevance debate seems to be about what the relevance judgment should be matched against, and therefore which level of information need documents should be represented to satisfy. None of the reviewed authors specifically deals with the subject-indexing process and they only touch on the concept of subject—because their focus is somewhere else. Furthermore, there is a trend in the newer literature to focus solely on the users' utility of the document and not on the relation between representation of the documents and retrieval of the documents.

Nearly every researcher who is concerned with the relevance measurement problem seems to agree that relevance is a relationship between an information need and information contained in documents, and that relevance judgments rely on much more than a match in topicality. The major disagreement is on the degree to which these subjective judgments should be used as the basis for evaluation of IR systems/techniques. In other words, they disagree about the level of the information need the information contained in documents should be measured against.

Taylor (1968) has defined four stages in the development of the information need:

1. The visceral need.
2. The conscious need.
3. The formalized need.
4. The compromised need.

As the information need moves from stage 1 toward stage 4, it becomes expressed in words, narrower, and more measurable. The essential problem in relevance measurements is to agree on which stage the document should be related to. If the first two stages are used, the definition of relevance is based on the degree of usefulness the document has for the user, whereas if one of the later stages are used, the measurement of relevance is based on the degree to which the *topic* of the request and the *topic* of the retrieved documents match. The first type of relevance measurement is usually defined as *user-oriented relevance* and the latter is usually defined as *topical relevance.*

When testing an IR system's ability to retrieve relevant documents, it makes a difference which of the two kinds of needs it is being tested against.

Suppose a researcher is writing a paper on nurses who in the 1840s traveled around in Texas and visited poor farm villages to care for the ill, and that these nurses brought a number of books with them that they sold to wealthier people. The researcher is writing a book about this and wants some more knowledge about this. This is the writer's information need. She comes to the IR system with this need and reformulates it into a number of demands for information, such as "I want literature about farm villages in Texas in the 1840s"; "I want literature about nurses' working conditions in the 1840s"; "I want literature on the reading habits of wealthy people in Texas in the 1840s." The librarian might help her reformulate her information need into these requests for literature. The IR system might be able to retrieve a number of relevant documents that meets each of the researcher's demand for literature, although it does not solve her problem that generated the information need. Was the IR system successful? Depending on whether it is assumed that the IR system should be tested on its ability to retrieve *useful* information for the users (user-oriented relevance), or the IR system's ability to retrieve *topical* relevant information (topical relevance), the answers differ.

Researchers of IR evaluation have focused their concern mainly on defining the kind of information need that the documents should be related to. They assume that as documents are indexed, this side of the problem has been solved. The assumption is that only the user side of the problem is relative, the subject of the document can be objectively defined.

A subject representation of a document has two primary functions: (1) it must represent the subject content of a document, and (2) it should

help the users of the system find documents on a particular subject. The challenge is to find a balance between these two functions. As discussed in the previous section on the five conceptions of indexing, this is a balance because indexing should not solely be based on the users' requirements or on the document's text.

If one focuses solely on the representation aspect and ignores future users, one might risk representing documents in a way that would be of no use for the users. An indexer who does not pay much attention to the users might choose to represent subjects of documents that are of no interest to the users, or might use a different vocabulary from the users, or might represent the subject on a level that is too broad or too narrow for the users. However, if the indexer pays too much attention to the users of the system, the indexer might represent documents in such a way that the subject representation of the documents only serves the current users and those current information needs.

In relevance studies, the balance between the two functions is highly important. Evaluation of relevance that tests only the users' ability to find documents that will solve, or be useful for, their particular information needs tends to forget the difficulties and possibilities of subject representation. A study that solely discusses and investigates a group of users' information needs and how they solve these needs by using a particular information retrieval system reveals little about how well the subjects of the documents are represented.

IX. Summary and Conclusion

In the late nineteenth century, Charles A. Cutter (1904) defined a set of universal objects of a catalog. These much cited objects state that one of the central objects of a catalog is to provide access to documents on the basis of their subject matter.

Although efforts have been made to describe the process of indexing, it seems quite obvious that it is poorly documented as well as poorly explained. That there has been no adequate theory to explain the indexing process, nor a recognition that such a theory should focus on what indexing is, at the expense of determining how to index. The existing manuals and guidelines for determining the subject matter of documents are insufficient for their purpose. Attempts to describe actual methods of indexing have likewise been unsuccessful. The indexing process can be deconstructed to show three steps (document analysis process, subject description, and subject analysis process) and four elements (document, subject, subject description, subject entry). A general analysis of these steps and elements indicates that they can be viewed

as interpretations. An indexer goes through a number of stages from being a novice indexer to becoming an expert indexer.

Much of the literature on the subject-indexing process is concerned with finding the rules of indexing. The one major goal that has guided research in the field has been to find out what it is that indexers do when they index. The goal of this kind of research has been to be able to prescribe *how* to index. This has been challenged; it has been argued that it will not be possible to find these rules. Even if such rules are found, they will not be of much use. By using Wittgenstein's argumentation, Frohmann shows that a rule has meaning only in a social context.

A common discussion in the indexing literature is to define two distinct conceptions of indexing, one that takes its basis in the document itself, and another that takes as its basis the users' needs or potential needs for information. This paper suggests that these two conceptions should be expanded to five conceptions of indexing, and that in practice, a combination of these should be used.

A central concept in the library and information science field is relevance. This might even be the most important and most central concept of all. Some literature on this concept was reviewed in order to search for conceptions of the subject-indexing process and the concept of subject. It was concluded that most researchers do not explain the concept of subject they use, and none of them discusses the subject-indexing process *per se*. There is also a trend to investigate the users' utility of the documents at the expense of investigations into the relations between representation of the documents and the retrieval of the documents.

Investigations into, studies of, and the teaching of indexing should be taught as and thought of as a number of choices or interpretations rather than an exact skill to be learned. No matter how much time is spent searching for the rules of indexing, they will never be found, and it will never be possible to determine the exact actions an indexer or cataloger should take during the process. The focus should therefore be shifted toward explaining what indexing is.

Indexing is often defined as the process of creating entries in an index. Most of the studies of indexing have been concerned with technical aspects of translating the subject description into the vocabulary of the indexing language—the first step in determining the subject matter has largely been ignored. The reason for this could be that indexing has been approached as an exact skill that can be mastered, such that inter-indexer inconsistency can be eliminated. Instead, indexing must be thought of as a humanist interpretative art.

Studies of indexing must be tied more closely to studies of relevance and evaluation of information retrieval systems and the practice in which the

indexing and retrieval is conducted. Indexing cannot be studied separately from the social context and domain in which the indexing is done and the users operate. Studies of evaluation must be tied to the representation process and not solely to the retrieval process. There needs to be a shift from objective scientific methods to hermeneutic qualitative methods, and more holistic and operational approach to research in indexing.

References

Albrechtsen, H. (1992). PRESS: A thesaurus-based information system for software reuse. In *Classification Research for Knowledge Representation and Organization.* Elsevier, New York.

Albrechtsen, H. (1993). Subject analysis and indexing: From automated indexing to domain analysis. *The Indexer* **18**(4), 219–224.

Anderson, J. D. (1985). Indexing systems: Extensions of the mind's organizing power. *Information and Behavior* **1**, 287–323.

Austin, D. (1974). The development of PRECIS: A theoretical and technical history. *Journal of Documentation* **30**(1), 47–102.

Barry, C. L. (1994). User-defined relevance criteria: An exploratory study. *Journal of the American Society for Information Science* **45**(3), 149–159.

Bates, M. (1986). Subject access in online catalogs: A design model. *Journal of the American Society for Information Science* **37**(6), 357–376.

Beghtol, C. (1986). Bibliographic classification theory and text linguistics: Aboutness analysis, intertextuality and the cognitive act of classifying documents. *Journal of Documentation* **42**(2), 84–113.

Benediktsson, D. (1989). Hermeneutics: Dimensions toward LIS thinking. *Library and Information Science Research* **11**, 201–234.

Blair, D. (1990). *Language and Representation in Information Retrieval.* Elsevier Science Publisher, New York.

Buckland, M. (1983). Relatedness, relevance and responsiveness in retrieval systems. *Information Processing and Management* **19**(3), 237–241.

Chan, L. M., Richmond, P. A., and Svenonius, E. (1985). *Theory of Subject Analysis: A Sourcebook.* Libraries Unlimited, Littleton.

Chu, C. M., and O'Brien, A. (1993). Subject analysis: The first critical stages in indexing. *Journal of Information Science* **19**, 439–454.

Cooper, W. S. (1971). A definition of relevance for information retrieval. *Information Storage and Retrieval* **7**(1), 19–37.

Cooper, W. S. (1978). Indexing documents by gedanken experimentation. *Journal of the American Society for Information Science* **29**, 107–119.

Cornelius, I. (1996). *Meaning and Method in Information Studies.* Ablex Publishing, Norwood, New Jersey.

Cutter, C. A. (1904). *Rules for a Dictionary Catalog* (4th ed.). Government Printing Office, Washington, DC.

DDC (1996). *Dewey Decimal Classification and Relative Index.* Forest Press, Albany.

Dreyfus, H. L., and Dreyfus, S. E. (1986). *Mind Over Machine: The Power of Human Intuition and Expertise in the Era of the Computer.* Basil, London.

Eco, U. (1994). *The Limits of Interpretation.* Indiana University Press, Bloomington.

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

Farradane, J. (1980). Knowledge, information and information science. *Journal of Documentation* **2**, 75–80.
Farradane, J. (1979). The nature of information. *Journal of Information Science* **1**, 13–17.
Farrow, J. F. (1995). All in the mind: Concept analysis in indexing. *The Indexer* **19**(4), 243–247.
Farrow, J. F. (1991). A cognitive process model of document indexing. *Journal of Documentation* **47**(2), 149–166.
Farrow, J. F. (1994). Indexing as a cognitive process. *Encyclopedia of Library and Information Science* **53**(16), 155–171.
Fidel, R. (1994). User-centered indexing. *Journal of the American Society for Information Science* **45**(8), 572–576.
Frohmann, B. (1990). Rules of indexing: A critique of mentalism in information retrieval theory. *Journal of Documentation* **46**(2), 81–101.
Gardner, H. (1985). *The Mind's New Science: A History of the Cognitive Revolution.* Basic Books, New York.
Green, R. (1995). Topical relevance relationships. I. Why topic matching fails. *Journal of the American Society for Information Science* **46**(9), 646–653.
Hanson, N. R. (1958). *Patterns of Discovery.* Cambridge University Press, Cambridge.
Hjørland, B. (1997). *Information Seeking and Subject Representation: An Activity–Theoretical Approach to Information Science.* Greenwood Press, Westport, Connecticut.
Hutchins, W. J. (1978). The concept of 'aboutness' in subject indexing. *Aslib Proceedings* **30**(5), 172–181.
ISO (1985). *Documentation—Methods for Examining Documents, Determining Their Subjects and Selecting Indexing Terms.* International Organization for Standardization.
Jones, K. (1983). How do we index? A report of some ASLIB informatics group activities. *Journal of Documentation* **39**(1), 1–23.
Jones, K. (1976). Towards a theory of indexing. *Journal of Documentation* **32**(2), 118–125.
Lancaster, F. W. (1998). *Indexing and Abstracting in Theory and Practice.* University of Illinois, Champaign, Illinois.
Lancaster, F. W., Elliker, C., and Connell, T. H. (1989). Subject analysis. *Annual Review of Information Science and Technology* **24**, 35–84.
Langridge, D. W. (1989). *Subject Analysis: Principles and Procedures.* Bowker-Saur, London.
Liston D. M., Jr., and Howder, M. L. (1977). Subject analysis. *Annual Review of Information Science and Technology* **12**, 81–118.
Merrell, F. (1995). *Semiosis in the Postmodern Age.* Purdue University Press, West Lafayette, Indiana.
Miksa, F. (1983). *The Subject in the Dictionary Catalog from Cutter to the Present.* American Library Association, Chicago.
Milstead, J. L. (1994). Needs for research in indexing. *Journal of the American Society for Information Science* **45**(8), 577–582.
Moss, R. (1975). PRECIS (letter). *Journal of Documentation* **31**(2), 116–117.
Park, T. K. (1994). Toward a theory of user-based relevance: A call for a new paradigm of inquiry. *Journal of the American Society for Information Science* **45**(3), 135–141.
Petersen, T. (1994). Introduction. In *Guide to Indexing and Cataloging with the Arts and Architecture Thesaurus.* Oxford University Press, New York.
Rees, A. M., and Saracevic, T. (1966). The measurability of relevance. *Proceedings of the American Documentation Institute* **3**, 225–234.
Saracevic, T. (1970). The concept of relevance in information science: A historical note. In *Introduction to Information Science.* R. R. Bowker, New York.
Schwartz, C., and Eisenmann, L. M. (1986). Subject analysis. *Annual Review of Information Science and Technology* **21**, 37–61.

Shaw, D., and Fouchereaux, K. (1993). Research needs in information science. *Bulletin of the American Society for Information Science* **19**(3), 25.

Soergel, D. (1985). *Organizing Information: Principles of Data Base and Retrieval Systems.* Academic Press, San Diego.

Swanson, D. R. (1977). Information retrieval as a trial and error process. *Library Quarterly* **47**(2), 128–148.

Swift, D. F. (1975). PRECIS (letter). *Journal of Documentation* **31**(2), 117–118.

Taylor, A. G. (1994). Books and other bibliographic material. In *Guide to Indexing and Cataloging with the Arts and Architecture Thesaurus.* Oxford University Press, New York.

Taylor, R. S. (1968). Question-negotiation and information seeking in libraries. *College and Research Libraries* **29**, 178–194.

Travis, I. L., and Fidel, R. (1982). Subject analysis. *Annual Review of Information Science and Technology* **17**, 123–157.

Vickery, B. (1968). Analysis of information. *Encyclopedia of Library and Information Science* **1**, 355–384.

Vickery, B. (1997). Metatheory and information science. *Journal of Documentation* **53**(5), 457–476.

Weinberg, B. H. (1988). Why indexing fails the researcher. *The Indexer* **16**(1), 3–6.

Wilson, P. (1968). *Two Kinds of Power: An Essay on Bibliographic Control.* University of California Press, Berkeley.

Wittgenstein, L. (1958). *Philosophical Investigations.* Macmillan Publishing, New York.