Available online at www.sciencedirect.com

SCIENCE DIRECT®

Information Processing and Management 41 (2005) 599–611

INFORMATION PROCESSING & MANAGEMENT

www.elsevier.com/locate/infoproman

Analysis in indexing: document and domain centered approaches

Jens-Erik Mai *

The Information School, University of Washington, Mary Gates Hall, Box 352840, Seattle, WA 98195-2840, USA

Received 1 July 2003; accepted 17 December 2003
Available online 7 February 2004

Abstract

The paper discusses the notion of steps in indexing and reveals that the document-centered approach to indexing is prevalent and argues that the document-centered approach is problematic because it blocks out context-dependent factors in the indexing process. A domain-centered approach to indexing is presented as an alternative and the paper discusses how this approach includes a broader range of analyses and how it requires a new set of actions from using this approach; analysis of the domain, users and indexers. The paper concludes that the two-step procedure to indexing is insufficient to explain the indexing process and suggests that the domain-centered approach offers a guide for indexers that can help them manage the complexity of indexing.

Keywords: Context; Domain analysis; Indexing; Subject analysis

1. Introduction

The purpose of indexing is to determine the subject matter of documents and express the subject matter in index terms (e.g. descriptors, subject headings, call numbers, classification codes, or index terms) to make subject retrieval possible. It is often implicitly assumed that the document will present its subject matter to the indexer and that the indexer can establish the document's subject matter by a simple analysis of the document. An emerging approach to indexing calls for indexers to analyze documents with the users' information needs in mind and argues that the document's subject matter is context-dependent. This paper extends the concept of analysis in indexing by incorporating analyses of the domain in the indexing process, thereby integrating users' perspectives and needs in the indexing process.

Indexing is often divided into two approaches: a document-oriented approach and a user-oriented approach (cf. e.g. Albrechtsen, 1993; Andersen & Christensen, 2001; Fidel, 1994; Hjørland, 1992; Mai, 2000; Soergel, 1985). The document-*oriented* approach focuses on the "entity and its faithful description"

* Tel.: +1-206-616-2541; fax: +1-206-616-3152.
E-mail address: jemai@u.washington.edu (J.-E. Mai).

0306-4573/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2003.12.004

(Soergel, 1985, p. 227) and recommends that the indexer "stick to the text and the author's claims" (Lancaster, 2003, p. 37). The basic idea is that the indexer should establish the subject matter solely based on an analysis of the document itself; the goal is to represent the document as truthfully as possible and ensure the subject representation is valid for a long time. Recent research into indexers' work confirms that some indexers do not "consider the different uses someone could make of a particular book" (Sauperl, 2004, p. 61) and that they select index terms solely based on analyses of the document. A variant of this approach is the document-*centered* approach. In the document-*centered* approach the indexer takes the document as the focal point for analysis, but in selecting the index terms, the indexer bears "in mind the questions, as far as they can be known, which may be put to the information system" (ISO, 1985, Section 6.3.3). These two approaches—document-oriented and document-centered—assume that the subject matter of a document can be determined independently of any particular context or use. They differ in the selection of the index terms; in the document-*centered* approach the index terms are selected in accordance with the users' needs and questions, whereas the document-*oriented* approach does not include an assessment of the users' needs.

The user-oriented approach argues that the indexer should ask, "how should I make this document...visible to potential users? What terms should I use to convey its knowledge to those interested?" (Albrechtsen, 1993, p. 222). The basic idea is that the indexer needs to have the users' information needs and terminology in mind when determining the subject matter of the document as well as when selecting index terms for the document. It suggests that the indexer should have knowledge about the users' needs to determine the subject matter, whereas the document-centered approach merely suggests that the indexer should have the users' needs in mind when selecting index terms.

This paper outlines an approach to indexing that extends the user-oriented tradition by taking the domain as the focal point for analysis. The domain-centered approach to indexing is based on an interpretation of the domain analytic approach to information science suggested by Hjørland and Albrechtsen (1995) and Hjørland (2002). The strength of the domain-centered approach to indexing is that it explicitly bases the analysis of the document on analyses and understandings of the domain and the users. This ensures that the analysis of the document's subject matter is placed within the domain's and users' discourse. The contention is that the indexer should have an understanding of the users' and the domains' goals, purposes, and activities to make an assertion about a document's subject matter.

The paper will first examine current thinking in indexing theory and research by discussing the notion of steps in the indexing process and will then examine the limits of analysis based on documents' attributes. The role of context will be discussed and introduced as a central concept in indexing and the domain will be presented as the part of the context that can be used to guide indexing. The paper will lastly introduce and discuss the domain-centered approach to indexing. This approach to indexing incorporates analyses of the document's context, inserts additional steps in the indexing process, and modifies what the indexer needs to analyze to determine the subject matter of a document.

2. Indexing and guidelines for indexing

Indexing is often described as a process that takes a number of steps. The simplest description of the process states that indexing is done in two steps. First, the indexer analyzes the document to determine its subject matter. Second, the indexer translates the subject matter into index terms.

There has been some research aimed at understanding the procedure that indexers use to determine the subject matter of documents (cf. e.g. Bertrand & Cellier, 1995; Chu & O'Brien, 1993; Sauperl, 2002); however, the exact procedure or steps indexers take are not well understood (Albrechtsen, 1993; Bates, 1986; Beghtol, 1986; Cooper, 1978; Farrow, 1991; Hutchins, 1978; Jones, 1976; Jacob & Shaw, 1998; Mai, 2001; Milstead, 1994). The first step in indexing, that of determining the subject matter, is the least well understood even though it is "the most important and the most difficult part" (Langridge, 1989, p. 1): the most important because it is the foundation for the subject representation of the document, and the most difficult because little or no guidance can be given as to how the indexer should establish the subject matter. Yet, while "there is...plentiful literature on the [last step in] indexing, that of making an appropriate entry...[the first step] has been badly neglected" (Langridge, 1989, p. 6). As Frohmann argues, "most research focuses on the...[last] step, while the first continues to be lamented as an intellectual operation both fundamental to indexing yet so far resistant to analysis" (Frohmann, 1990, p. 82). This lack of understanding of the indexing process creates problems when instructing indexers how to index; Bates notes that "it is practically impossible to instruct indexers or catalogers [on] 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'" (Bates, 1986, p. 360). The second step in the indexing process, the translation of the subject matter into an indexing language, is explained much better; each indexing language has a particular set of rules and guidelines for how to express the subject matter using the indexing language.

The exact relationship between the two steps remains to be investigated in greater detail. It could be argued that the first step "relates to the document and not to the system" (Langridge, 1989, p. 7) and that the indexer should not be concerned with the particular system in the first step. However, one could reasonably expect that the use of controlled vocabularies—especially subject-specific controlled vocabularies—guide indexers' analysis of the document and directs the indexing process, resulting in greater consistency among indexers (Sievert & Andrews, 1991). The relationship between the first and second steps is, however, beyond the scope of this paper and the focus here will be on the nature of the first step in indexing.

One possible guide for indexers is standards or guidelines for indexing. Standards can roughly be divided into two types: "(a) Compatibility standards, the kind of standard that is essential if a 3.5 in. high-density disk will actually function in your hard drive, and (b) advisory standards, advice from experts on how best to carry out some function or produce some product" (Anderson, 1994, p. 628). Where compatibility standards prescribe how a certain entity should be designed, advisory standards merely give advice on how to do something.

Indexing standards fall somewhere between these two types of standards. The goal of the ISO (1985) standard, for instance, is to "promote standard practice (a) within an agency or network of agencies; (b) between different indexing agencies, especially those which exchange bibliographic records" (ISO, 1985, Section 1.4). The indexing process as a whole cannot be standardized in the same sense that disk drives can be standardized. Nor, however, can advisory standards ensure standard practice, since each agency can choose to implement indexing as they want. Indexing guidelines cannot prescribe a standard practice in the determination of the subject matter; they can, however, standardize the expression of the subject matter in the indexing language. Thus, the second step in the indexing process can be standardized to a certain degree in terms of ensuring that a given concept is expressed a in particular way. It is questionable whether the first step can be standardized. The value of indexing guidelines is, therefore, in prescribing the technical aspects in the construction of index terms.

In guiding indexers on how to establish the subject matter of documents, guidelines for indexing, such as the ISO standard, usually recommend examining the table of contents, scanning chapter headings, examining forewords and introductions, etc., and assume that this procedure will make the document's subject matter evident. While not informative about the process itself, this listing of sources for subject information is nevertheless valuable since it directs all the indexers or catalogers within an indexing agency to the same sources. Moreover, ranking such sources indicates to the indexers which information has the greatest weight. However, guidelines for indexing "are curiously uninformative about how one goes about identifying the subject" (Wilson, 1968, p. 73) because they focus on which document attributes the indexer should pay attention to without explaining how the indexer should extract subject information from them. The guidelines are "fairly document-centered" (Hjørland, 1997, p. 44) because they assume that the indexer can determine the subject matter from the document attributes. The document-centered approach to indexing and guidelines for indexing are alike in their reliance on the document itself to reveal its subject matter and do not consider context as a component for analysis in the indexing process.

The following section illustrates the problems in taking a document-centered approach to indexing and shows that the indexer always has to rely on an interpretation of attributes in order to place the attributes in context.

3. Analysis based on document attributes

A document-centered approach to indexing relies on the indexer's analysis of a number of document attributes as the basis for establishing the document's subject matter. The exact attributes that indexing guidelines or textbooks recommend that the indexer examine varies, but archetypical examples are: the title, the abstract, the table of contents, chapter headings, chapter subheadings, preface, introduction, foreword, the text itself, bibliographical references, index entries, illustrations, diagrams, and tables and their captions. The exact recommendations vary according to the type of document that is being indexed (monographs vs. periodical articles, for instance).

While it makes sense that the indexer examines all or some of the aforementioned attributes, the exact nature of the indexer's analysis is seldom addressed in these guidelines and textbooks. The indexer is often left with "vague and inadequate [recommendations such] as 'Look for the main topic of the document'" (Bates, 1986, p. 360) and the exact use of document attributes for indexing is left open to interpretation. However, the indexer will inevitable draw on experience and knowledge about the collection, users, and potential use of the document when deciding on a document's subject matter.

The title is a document attribute that in many cases can give clues about the subject matter of the document. However, even with titles that are fairly descriptive the indexer needs to make decisions that reach beyond the document and the title. Take as an example David Blair's book "Language and Representation in Information Retrieval" (Blair, 1990); the title states fairly precisely that the book is about language, representation, and information retrieval. However, the indexer must establish the connections between 'language,' 'representation,' and 'information retrieval.' Even though the title has provided the indexer with some clues as to the subject matter, there are still a number of places where the book could be placed in the Dewey Decimal Classification (DDC), including: 004.019 Psychological principles (in computer science); 025.48 Subject indexing; 025.524 Information search and retrieval; and 121.68 Meaning, interpretation, hermeneutics. All of these representations match the subject expressed in the title fairly well. However, the exact choice of classification code can only be made once the indexer considers how the users would potentially use and search for the book. So, even if the title of the book is fairly precise, as in this case, the indexer is only provided with a clue, not the actual subject matter. The indexer will have to make a decision about the actual representation of the document based on more than the information gathered from the title.

Another example of a document attribute is the table of contents. The purpose of the table of contents is to name the different parts of the document, and clues for the indexer are less explicit and require a different analysis than that of the title. The table of contents in Blair's book is four pages long. The headings of the six main chapters are: "The nature of information retrieval," "The principal formal models of information retrieval," "The evaluation of information retrieval systems," "Language and representation: The central problem in information retrieval," "Communication and adaptation," and "Information retrieval and the nature of scientific theory." The first chapter headings suggest that the book is mainly about information retrieval, however, the last two chapters deal with communication and scientific theory and the third chapter with evaluation of information retrieval systems. Only the fourth chapter (“Language and representation: The central problem in information retrieval”) seems covered by the book’s title. If the indexer looks at the sub-headings he/she will find entries like, “The goals of document retrieval: Prediction criteria and futility points,” “Consequences of the lack of reliable recall/precision studies,” “Semiotics,” “Ludwig Wittgenstein,” “The nature of documents,” “Do we have a complete typology of illocutionary acts,” “The generic algorithm: Relevance,” and “Science vs. pseudoscience.” These, and others like these, suggest a different or much broader subject matter than simply “information retrieval.”

The table of contents will never directly state the subject matter of the document but simply name some of the parts of the document, since the purpose of the table of contents is to guide the reader through the book. The indexer needs to put the parts together, and, as such, the subject description is as much a product of the indexer as of the document. The indexer needs to make a decision about how to represent the document and, again, this decision needs to be based on information other than that which the indexer can gather from the document.

Even if the indexer scans the text, he/she would still need to apply contextual knowledge to determine the document’s subject matter. Scanning the text could mean that the indexer simply looks through the document or maybe even speed-reads the entire document. It could also mean that the indexer reads selected sections of the text closely, such as the first and last paragraph of each chapter. Farrow (1994) distinguishes between scanning as a perceptual act as opposed to scanning as a conceptual act. Farrow defines perceptual indexing to mean that the indexer immediately or intuitively recognizes the subject matter of the document from the text itself, and conceptual indexing to mean that the indexer uses some kind of world knowledge to determine the subject matter. No matter which approach one takes to scanning, the indexer’s task is to determine the subject matter by identifying the dominant aspect of the document. But since it cannot be expected “that everyone’s attention will be dominated by the same things” (Wilson, 1968, p. 83), each indexer’s analysis of the subject matter will depend as much on his/her knowledge as on the text.

Indexing according to document attributes has especially serious implications for social scientists (Swift, Winn, & Bramer, 1979), humanists (Langridge, 1976; Tibbo, 1994), and interdisciplinary scholars (Bates, 1996) because the exact understanding and use of documents in these fields is context dependent and might have only a slight relation to the specific object under study in the document. Blair’s book, for instance, might be used by people who only have a marginal interest in information retrieval but have an interest in, for instance, the philosophy of language or scientific communication.

The examples and discussions above demonstrate that an indexer always needs to rely on information that is not provided in the document itself when determining the subject of a document. The subject matter does not exist in the document, ready to be discovered and pulled out by the indexer. The indexer will necessarily include knowledge of or guesses about the document’s potential use and the users’ needs in his/her analysis of the document. In other words, a true document-centered approach is problematic, and indexing will necessarily involve context in the indexer’s final decisions about subject matter and index terms.

4. The role of context

The document-centered approach is based on the notion that the subject matter of documents can be established independently of any particular context or use, or in other words, that the indexer can base the determination of the subject matter on document attributes alone. This notion is closely tied to the position that text can be analyzed in itself and the meaning of text can be established independent of context.

The last few decades have seen a change in textual analysis theory from viewing a textual structure as something that should "be analyzed in itself and for the sake of itself" (Eco, 1994, p. 44) to focusing on the "pragmatic aspect of reading" (Eco, 1994, p. 44). The latter position argues that the function and meaning of a text only can be explained by taking the reader of the text into account. In this tradition, the meaning of a text depends on the interpretive choices the reader makes:

> [I]f meaning is embedded in the text, the reader's responsibilities are limited to the job of getting it out; but if meaning develops, and if it develops in a dynamic relationship with the reader's expectations, projections, conclusions, judgments, and assumptions, these activities (the things the reader *does*) are not merely instrumental, or mechanical, but essential, and the act of description must both begin and end with them (Fish, 1980, pp. 2–3).

In this understanding of texts and their interpretation, a text does not have meaning in and by itself. The meaning of the text is created by the reader as the text is read and used. A reader does not *respond to* the meaning of a text. The reader's response *is* the meaning of the text. But even though the reader's interaction with the text is personal, the meaning of the text is not entirely a personal and private construct. The meaning of words and the correct use of language is closely tied to the community in which the words and the language are used. Even though words come from an individual person and are perceived by an individual person, language is not the product of these individual persons. Language belongs to the community in which it is used. It is the community and its activities that defines and determines the meaning of the words used. Words, therefore, do not have objective and true meanings, but neither are words' meanings fluid and individual. Introna (1998, pp. 8–9) gives the example of how one needs to start with the community's "already there language," even if one wants to disagree with the community:

> I cannot stand up in a conference on the philosophy of language and propose that the audience somehow entirely 'forget'—if this is possible at all—the already there tradition of philosophical discourse on language that emerged over thousands of years. Even if I want to disagree with it entirely, or use concepts in totally different ways, I will still have to draw on this tradition—of linguistic distinction—to say how, or in what way, my use of this language will be different.

Meaning, language, cognition, and practice are interwoven in the social sphere and cannot be captured separately. When reading a text, a reader will construct an understanding of the text and that understanding is tied to the reader's social context. The reader cannot reasonably interpret the text to mean just anything. The context limits and guides the interpretation and understanding of the text.

Following this interpretative tradition, an indexer will ask questions such as, "Who will use this document?", "To what domain do the users of the document belong?", and "Who wrote this text?" to determine the subject matter of a document. Only if the indexer were instructed to, say, derive nouns from the title—or perform some similarly mechanical sort of indexing—would it be possible for the indexer to base indexing on the document attributes alone, without any interpretation or judgment about the potential use of the document. Any indexing that allows the indexer to make choices will force the indexer to make an interpretation of and judgment about the use of the document.

In sum, the indexer cannot determine and represent the subject matter of a document without some understanding of the future use of the document. Furthermore, the indexer cannot understand potential usage of the document without an understanding of the context in which the document is used. The determination and representation of the subject matter of documents is connected to discourse and activities in a context, and the indexer needs to have an understanding of this discourse and these activities.

5. The notion of domain

Context is often defined as that which surrounds a specific word, passage, event, or situation. For the purposes of guiding indexers in the indexing process, it is necessary to operate with a more precise notion of context.

From the information needs and seeking research perspective, Talja, Keso, and Pietilainen (1999) offer a discussion and clarification of the concept of context. They distinguish between an objectified approach to context in which "context is evoked and described" (p. 752), and an interpretive approach in which context "is not understood as an independent entity, but as a carrier of meaning" (p. 752). The idea is that context and the study of context can be approached as something that can be captured, described, and made concrete in order to control the context in certain situations. Talja, Keso, and Pietilainen call this an objectified approach. On the other hand, context can be seen as everything that surrounds a given phenomenon and gives meaning to that phenomenon. In this sense, a context is not a set of entities that can be identified and represented, but is rather those intangible notions that create meaning and understanding. Talja, Keso, and Pietilainen call this an interpretive approach.

Although Talja, Keso, and Pietilainen present these two approaches as distinct, they could more fruitfully be considered as layers. The interpretive approach acknowledges that we as actors carry a particular point of view and bias as we conduct activities. The objectified approach attempts to articulate and point out the biases and particular instances in the context that facilitate the point of view. For actors who acknowledge that they carry a particular point of view, the objectified approach is merely an attempt to work with that fact by selecting among the carriers of meaning those considered most significant. As such, the objectified approach is an attempt to articulate parts of the context, since articulating the whole and full context will not be possible.

The idea that the organization and representation of information should start with an analysis of the context, discourse, and activities, is central to *domain analysis*, introduced by Hjørland and Albrechtsen in 1995. The focus of domain analysis is to understand the activities of a particular domain:

> The domain-analytic paradigm in information science (IS) states that the best way to understand information in IS is to study the knowledge-domains as thought or discourse communities, which are parts of society's division of labor. Knowledge organization, structure, co-operation patterns, language and communication forms, information systems, and relevance criteria are reflections of the objects of the work of these communities and their role in society. The individual person's psychology, knowledge, information needs, and subjective relevance criteria should be seen in this perspective (Hjørland & Albrechtsen, 1995, p. 400).

The core of the domain analytic approach is to study the activities and products of domains to gain insights into "already there" structures and meanings. The assumption is that domains produce artifacts that can be used to study them. It can furthermore be argued that an understanding and analysis presumes and includes an understanding and analysis of the users' information needs and information usages. An understanding of the domain informs decisions made about information related activities, e.g. design of information systems, construction of controlled vocabularies, strategies for evaluation, and indexing of documents.

Hjørland and Albrechtsen do not clearly define what they mean by "domain" and it might not be possible, or even desirable, to define the concept in an exclusive manner. Domain is an evolving and open concept that will develop as the concept is used and applied in research and practice. In this paper, however, the concept is used to refer to *a group of people who share common goals*. A domain could, for instance, be an area of expertise, a body of literature, or a group of people working together in an organization.

This activity-centered use of the notion of domain helps to determine the context of the indexing task. While some may limit the notion of domains to disciplines, the concepts of disciplines and specialties are not sufficient to frame activities in knowledge organization and indexing because of their lack of stability.

A discipline is defined as "a cluster of specialties" (Dogan, 2001, p. 14851) that represents "historical evolutionary aggregates of shared scholarly interests" (Chubin, Porter, & Rossini, 1986, p. 4) and activities are organized into disciplines mainly because they share historical commonalities. The boundaries of disciplines are unclear in the sense that what to one person might be *sociology* will to another be *politics* or *history*; people are collaborating across disciplines, publishing outside their own disciplines, and pulling in knowledge from various disciplines.

This rather vague definition of discipline suggests the notion that "a scholarly discipline contains a fundamental ambiguity" (Birnbaum, 1986, p. 53) in the sense that it is difficult to determine the commonalities between the specialties within a discipline and thereby define what the discipline is. Today, boundaries between disciplines are blurred, and work within disciplines is typically highly specialized.

Disciplines consist of multiple specialties that may have little in common. Sociology, for instance, consists of more than 50 specialties (International Sociological Association, www.ucm.es/info/isa/rc.htm) among many of which there is little communication and exchange of information (Dogan, 2001, p. 14852). It can furthermore be observed that scholarly and scientific work advances by intersecting and relating knowledge from multiple areas of work; this is evident from citation studies that show how researchers, to a large extent, use literature from outside their own discipline (cf. e.g. Dogan, 2001, p. 14854).

Specialties are constructed "along substantive, epistemological, methodological, theoretical, and ideological lines" (Dogan, 2001, p. 14852) but a single issue may belong to several disciplines (Klein, 1996b, p. 142):

> The problem of "poverty" for example, appears simultaneously in economics, policy studies, sociology, and women's studies. Similarly, the problem of "disease" appears in social medicine, anatomy, gerontology, and a host of medical specialties. Yet, "poverty" and "disease" are constructed differently in each disciplinary domain. Boundaries are drawn along particular disciplinary, professional, and interdisciplinary lines.

The crossovers have given rise to interdisciplinary fields (Klein, 1996a) and recombination of specialties (Dogan & Pahre, 1990). As a result, disciplines and specialties by themselves do not give sufficient context to the indexing task to place a particular document in a discourse. The concept of domain cuts across formal definitions and boundaries and focuses on people's activities, collaboration, and common goals when placing documents in discourse communities.

The exact boundaries and composition of particular domains is determined through an analysis of the domain, focusing on establishing the structures, ontology, and communication patterns present in the domain, that is, the activities that take place, the circumstance under which they can take place, and the constrains imposed by paradigms and current research fronts (Tennis, in press).

Furthermore, the concept of domain has the strength of being applicable to areas outside the scholarly and scientific communities, such as a pharmaceutical company (Nielsen, 2001) or film archives (Albrechtsen, Pejtersen, & Cleal, 2002). In these cases the domain is defined and delimited by an analysis of the goals and objectives of the organizations for which the indexing is done. This analysis provides a description of the organization's goals, objectives, and tasks that are used when formulating indexing strategies. Users of public libraries could be defined as a domain in the sense that public library users typically have information needs that are more general in nature than subject experts' information needs.

When defining a domain as *a group of people who share common goals* it is made clear that the concept of domain is closely tied to human activities. Disciplines and organizational structures are often based on formalities and might not reflect the activities that actually take place in them. The notion of domain brings together work level perspectives and formal structures, and provides a strong concept for analysis of human–information interactions.

6. Domain-centered indexing

Given our understanding of the concepts of context and domain, we can then reexamine the document-centered approach. The document-centered approach starts with an analysis of the document to establish "its subject content" (ISO, 1985, Section 1.4). The subject is then expressed according to the needs and use in the domain. The document-centered approach moves from an analysis of the document to the use in the domain (see Fig. 1). In the document-centered approach the indexer will primarily ask: "What is the subject matter of the document?" The focal point is to analyze the document to determine what the document is about. The indexer might move on and ask: "In which parts of the subjects matter will the users be interested?" However, the aim of the indexer is to establish the subject matter independently of any particular use and domain and only place the document in a particular use after the analysis of the document and the establishment of the subject matter has taken place. The basic assumption in the document-centered approach is that it is possible to establish the document's subject matter and select the subjects that are appropriate for a given domain. The focus of the document-centered approach is on the documents' intrinsic meaning and subject matter and on the translation of the subject matter into the users' needs.

The main arguments in favor of the document-centered approach are that: (1) the representation of the document will be based on stable characteristics of the document that are not likely to change over time and that the indexing therefore will be of lasting value; and (2) it is difficult to predict how the document will be used in the future and it is therefore better to index the document according to what it actually contains. However, the document-centered approach does not address the inherent subjective interpretative nature of indexing and fails to show how the indexer might establish the subject matter independently of any context or use. Section 3 of this paper demonstrates that an analysis of a

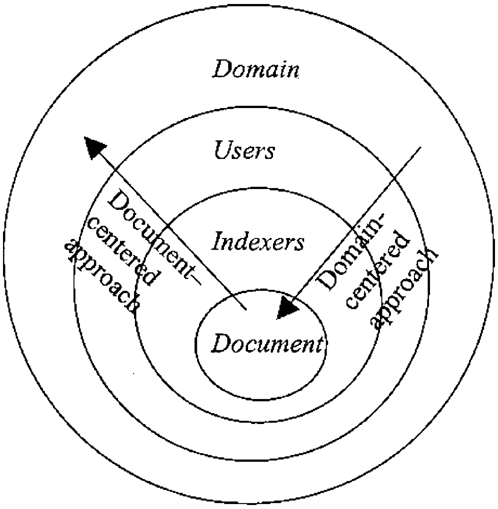

Fig. 1. Approaches in indexing: document-centered and domain-centered.

document will necessarily draw on the indexers' and documents' contexts and thereby unavoidably represent a particular understanding of the document. The indexers' perceptions of the context cannot be blocked.

The domain-centered approach to indexing takes a different approach to indexing than the document-centered approach; the domain-centered approach takes the domain as the focal point of analysis. Domain-centered indexing starts with an analysis of the domain, then moves on to analyze the needs of the users, determine the indexers' perspectives and roles, and lastly analyze the document in the context of the domain and the users' needs (see Fig. 1). The assumption in the domain-centered approach to indexing is that the subject matter and meaning of the *documents* can be determined only in the context of an understanding of the *domain*. In other words, a domain and the users' needs frame the meaning and subject matter of documents through the users' discourse in the domain. It is, therefore, of utmost importance that indexers understand the domain, the users' roles and interests in the domain, and critically analyze their own roles as indexers before a document is analyzed for its subject matter.

The domain-centered approach will ask a series of questions, starting with questions about the domain, moving on to questions about the users, then to questions about the indexers, and finally will pose questions about the document to be indexed. The indexer will address questions about the domain, users, and indexers in preparation for analyzing a document and ultimately indexing it. An indexer need not ask questions about the domain, users, and indexers for every single document that is indexed; these analyses are only performed as needed on a regular basis.

The specific type of questions and the amount of questioning that is needed varies from domain to domain. The following discussion serves only as an overview and guideline.

The subject matter of the document is closely tied to the domain in which the document will be used. The indexer, therefore, needs a substantial understanding about the nature of the domain, including its history, schools of thought, paradigms, research fronts, activities, goals, and objectives, to understand the roles a particular document will play in the domain. The indexer can gain this knowledge by studying the activities and discourse of the domain. Hjørland (2002) presents a number of approaches that might be taken in the analysis of the domain. The main concern is that the indexer identifies the domain's overall goals and purposes, its historical development, and its current research traditions. The ultimate goal of this analysis is to gain an understanding of the domain that enables the indexer to position an argument within the domain.

An indexer should position the users within the domain. The indexer analyzes the users' roles in the domain, how and where they fit into the domain, and establishes their needs and special interests within the domain. This analysis provides the indexer with knowledge about users in the context of the domain. The indexer asks where the users fit into the larger domain to identify special topics that are of potential interest to the users, or if there is a specific bias in which the users are interested, or if there is a particular level of specificity that the users need. The indexer needs this information to identify the specific subject matter of the document that might interest or be useful to the users. The indexer needs to place the users within the domain to fully understand the potential uses of a document. Usefulness and relevance to the users extend the subject matter of the document; the indexer needs to identify not only the document's subject matter but also how the document will be used.

An important and often overlooked aspect in indexing is the indexers' implicit or explicit interests, perhaps since it is expected that the indexers will assume a neutral role or no role at all. However, since analysis in indexing is a context-dependent interpretation and decision making process, the indexer unavoidably assumes a certain position. An analysis of the indexer's role is needed to make it explicit. The indexer could, for instance, assume an active, "exploitative" (Wilson, 1968, p. 25) role in which the indexer makes clear judgments in the representation of the documents; alternately, the indexer could assume a more "descriptive" (Wilson, 1968, p. 51) role where the indexer attempts to represent the document as neutrally as possible. These choices cannot be made at the global level of indexing standards; the exact role of the indexer is determined locally.

The three aforementioned analyses of the domain, the users, and the indexers should be performed on a regular basis to keep the indexers up-to-date on any changes and developments in the domain, user population, and role of the indexers. These analyses are used as the basis and frame of reference for the indexers' daily work.

Having performed an analysis of the domain, users, and indexers, the analysis of the document's subject matter will than start by asking which domain discourse the document is part of and how the users will use the document. Asking these questions will ensure that the document is put into the context of the domain and the users. The answers to these questions will be based on the previous analyses of the domain and the users' needs and interests. Only after the indexer has put the document into context will the indexer identify the subject matter that should be represented and translate the subject matter into the indexing language.

In situations where the indexers serve a retrieval system with a general collection and users (e.g., a university library or public library), the analysis of a document's subject matter can either be placed in multiple domains as applicable or the users can be viewed as a specific domain. The principle is that the document's subject matter is tied to the domain and that the identification of the subject matter has to be done within the context of domains (Hjørland, 1997, p. 50, note 2).

The domain-centered approach to indexing takes the domain as the focal point of analysis and uses knowledge about the domain and the users to determine the subject matter of documents. The benefit of this approach is that indexers have a clear frame of reference for making decisions when indexing, it ensures that indexing is consistent with users' use of information, and it provides effective results.

7. Conclusion

The concept of analysis in indexing needs to be broadened to cover more than just documents. An analysis of a document's subject matter evidently draws on the context of the document and the potential use of the document. The indexer brings knowledge, impressions, and experience to the analysis of the document and the domain-centered approach makes explicit what knowledge the indexer should bring and how the indexer should use that knowledge.

Indexing is often explained as a two step procedure whereby indexers first analyze the document to determine its subject matter and then translate the subject matter into index terms. In the domain-centered approach, these two steps are the last two steps in a series of analyses that begins with an analysis of the domain and includes an analysis of the users' information needs as well as the indexers' roles. Although the two step procedure is beneficial in separating the analysis of the document's subject matter and the creation and selection of index terms, it is insufficient in laying out the steps that an indexer must go through when indexing a document. The analysis of documents to establish their subject matter is a complex phenomenon and it requires complex explanations to guide the indexers. The domain-centered approach to indexing offers a framework that manages the complexity and guides the indexer through the range of required analyses.

References

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

Albrechtsen, H., Pejtersen, A. M., & Cleal, B. (2002). Empirical work analysis of collaborative film indexing. In H. Bruce, R. Fidel, P. Ingwersen, & P. Vakkari (Eds.), *Emerging frameworks and methods. Proceedings of the fourth international conference on conceptions of library and information science* (pp. 85–108). Greenwood Village: Libraries Unlimited.

Andersen, J., & Christensen, F. S. (2001). Wittgenstein and indexing theory. In H. Albrechtsen & J.-E. Mai (Eds.), *Advances in classification research. Proceedings of the 10th ASIS SIG/CR classification research workshop* (vol. 10, pp. 1–21). Medford, NJ: Information Today.

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

Bates, M. J. (1996). Learning about the information seeking of interdisciplinary scholars and students. *Library Trends, 44*, 155–164.

Bates, M. J. (1986). Subject access in online catalogs: a design model. *Journal of the American Society for Information Science, 37*, 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*, 84–113.

Bertrand, A., & Cellier, J. (1995). Psychological approach to indexing: effects of the operator's expertise upon indexing behaviour. *Journal of Information Science, 21*, 459–472.

Birnbaum, N. (1986). The arbitrary disciplines. In D. E. Chubin, A. L. Porter, F. A. Rossini, & T. Connolly (Eds.), *Interdisciplinary analysis and research* (pp. 53–66). Mt Airy: Lomond.

Blair, D. (1990). *Language and representation in information retrieval*. New York: Elsevier.

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

Chubin, D. E., Porter, A. L., & Rossini, F. A. (1986). Interdisciplinary research: The why and the how. In D. E. Chubin, A. L. Porter, F. A. Rossini, & T. Connolly (Eds.), *Interdisciplinary analysis and research* (pp. 3–10). Mt Airy: Lomond.

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

Dogan, M. (2001). Specialization and recombination of specialties in the social sciences. In N. J. Smelser & P. B. Balters (Eds.), *International encyclopedia of social and behavioral sciences* (pp. 14851–14855). New York: Elsevier.

Dogan, M., & Pahre, R. (1990). *Creative marginality: innovations at the intersection of social sciences*. Boulder: Westwiew Press.

Eco, U. (1994). *The limits of interpretation*. Bloomington: Indiana University Press.

Farrow, J. F. (1991). A cognitive process model of document indexing. *Journal of Documentation, 47*, 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-oriented indexing. *Journal of the American Society for Information Science, 45*, 572–576.

Fish, S. (1980). *Is there a text in this class: the authority of interpretive communities*. Cambridge, MA: Harvard University Press.

Frohmann, B. (1990). Rules of indexing: a critique of mentalism in information retrieval theory. *Journal of Documentation, 46*, 81–101.

Hjørland, B. (1992). The concept of "subject" in information science. *Journal of Documentation, 48*, 172–200.

Hjørland, B. (1997). *Information seeking and subject representation: an activity–theoretical approach to information science*. Westport, CT: Greenwood Press.

Hjørland, B. (2002). Domain analysis in information science: eleven approaches—traditional as well as innovative. *Journal of Documentation, 58*, 422–462.

Hjørland, B., & Albrechtsen, H. (1995). Toward a new horizon in information science: domain-analysis. *Journal of the American Society for Information Science, 46*, 400–425.

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

Introna, L. D. (1998). Language and social autopoiesis. *Cybernetics and Human Knowing, 5*, 3–17.

ISO (1985). Documentation—Methods for examining documents, determining their subjects and selecting indexing terms. International Organization for Standardization, ISO 5963-1985.

Jacob, E. K., & Shaw, D. (1998). Sociocognitive perspectives on representation. *Annual Review of Information Science and Technology, 33*, 131–185.

Jones, K. (1976). Towards a theory of indexing. *Journal of Documentation, 32*, 118–125.

Klein, J. T. (1996a). *Crossing boundaries: knowledge, disciplinarities, and interdisciplinarities*. Charlottesville: University of Virginia Press.

Klein, J. T. (1996b). Interdisciplinary needs: the current context. *Library Trends, 44*, 134–154.

Lancaster, F. W. (2003). *Indexing and abstracting in theory and practice* (3rd ed.). Champaign: Graduate School of Library and Information Science, University of Illinois.

Langridge, D. W. (1989). *Subject analysis: principles and procedures*. London: Bowker-Saur.

Langridge, D. W. (1976). *Classification and Indexing in the humanities*. London: Butterworths.

Mai, J.-E. (2000). Deconstructing the indexing process. *Advances in Librarianship, 23*, 269–298.

Mai, J.-E. (2001). Semiotics and indexing: an analysis of the subject indexing process. *Journal of Documentation, 57*, 591–622.

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

Nielsen, M. L. (2001). A framework for work task based thesaurus design. *Journal of Documentation, 57*, 774–797.

Sauperl, A. (2004). Catalogers' common ground and shared knowledge. *Journal of the American Society for Information Science and Technology, 55*, 55–63.

Sauperl, A. (2002). *Subject determination during the cataloging process*. Lanham, MD: Scarecrow Press.

Sievert, M. C., & Andrews, M. J. (1991). Indexing consistency in information science abstracts. *Journal of the American Society for Information Science, 42*, 1–6.

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

Swift, D. F., Winn, V. A., & Bramer, D. A. (1979). A sociological approach to the design of information systems. *Journal of the American Society for Information Science, 30*, 215–223.

Talja, S., Keso, H., & Pietilainen, T. (1999). The production of context in information seeking research: a metatheoretical view. *Information Processing and Management, 35*, 751–763.

Tennis, J. (in press). Two axes of domains for domain analysis. *Knowledge Organization*.

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

Wilson, P. (1968). *Two kinds of power: an essay on bibliographic control*. Berkeley: University of California Press.