

**The Subject Indexing Process:  
an investigation of problems in  
knowledge representation**

**Jens-Erik Mai**

Graduate School of Library and Information Science  
The University of Texas at Austin  
May 2000

**The Subject Indexing Process:  
an investigation of problems in knowledge representation**

by

**Jens-Erik Mai, M.L.I.S.**

**Dissertation**

Presented to the Faculty of the  
Graduate School of Library and Information Science

the University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

**Doctor of Philosophy**

The University of Texas at Austin

May 2000

**The Subject Indexing Process:**  
**an investigation of problems in knowledge representation**

**Approved by  
Dissertation Committee:**

---

Francis L. Miksa, Supervisor

---

Philip Doty

---

E. Glynn Harmon

---

Aloysius P. Martinich

---

Ruth A. Palmquist

Dedicated to Mette without whom this would have been impossible.

## **Acknowledgements**

I wish to thank a number of people and organizations that have been involved in the completion of this dissertation. First and foremost I wish to express my sincere appreciation to Fran Miksa, my supervisor, for his support, encouragement, and belief in this dissertation. My committee has supported and given me excellent feedback during this journey. My thanks therefore go to Phil Doty for sharing his extraordinary breath of view, Glynn Harmon for always bringing a larger perspective to things, Al Martinich for making me understand the basics of philosophy of language, and Ruth Palmquist for always asking the critical questions.

Special thanks also go to my friends and colleges at the Royal School of Library and Information Science, especially to Hanne Albrechtsen, Pia Borlund, Birger Hjørland, and Peter Ingwersen. My good friend Ole Bech-Petersen has improved the readability of the dissertation greatly, for that he deserves a huge thanks.

A number of organizations, schools, and foundations have supported me economically during my doctoral studies. I am grateful to all of them: the

University of Texas at Austin, Graduate School of Library and Information Science; the Royal School of Library and Information Science; the Fulbright Commission; The Danish Research Academy, Aarhus; The Denmark-America Foundation, Copenhagen; Knud Højgaards Fond, Charlottenlund; and Det Obelske Familiefond, Aalborg.

**The Subject Indexing Process:**  
**an investigation of problems in knowledge representation**

Publication No. \_\_\_\_\_

Jens-Erik Mai, Ph.D.

The University of Texas at Austin, 2000

Supervisor: Francis Miksa

The present dissertation is an attempt to explain at least some of the major problems related to subject indexing, that is, the subject representation of documents. In the process of doing so a new approach to the subject indexing process is proposed. This dissertation enters the discussion of how the subject matter of a document is determined from the standpoint that subjects are culturally and socially dependent.

Common approaches to the study of indexing have been to study how indexers index or to prescribe how indexers should index, with the goal of describing the rules of indexing or prescribing a universal and objective way which indexing should be carried out. Here it is argued that it is not possible to

prescribe how indexing should be done and that there are no universal and correct ways of indexing which can be applied to any given situation. Studies of how indexing is done must be based on an understanding of how one arrives at the meaning of a given set of human communicative symbols.

The subject indexing process is ordinarily described as a process that takes a number of steps. Here, however, it is argued that this typical approach characteristically lacks an understanding of what the central nature of the process is. Indexing is not a neutral and objective representation of a document's subject matter but the representation of an interpretation of a document for future use.

Semiotics, based on key ideas found in Charles Sanders Peirce's writings, is offered here as a framework for understanding the "interpretive" nature of the subject indexing process. Peirce's semiotics offers an especially valuable terminology for describing the different kinds of interpretation that takes place throughout the subject indexing process. By placing the subject indexing process within Peirce's semiotic framework of ideas and terminology, a more detailed description of the process is offered here than has been previously possible. While this attempt to show what occurs in the subject indexing process offers a more realistic view of what actually happens in the process, it also suggests strongly that there is no certainty to the result of the subject indexing process.



## Table of Contents

List of Figures .....	xi
1. Introduction.....	1
1.1. To See and to Interpret.....	9
1.2. Structure of the Dissertation .....	15
2. The Subject Indexing Process.....	18
2.1. Determining the Subject of a Document .....	20
2.2. Document and Subject Analysis .....	49
2.3. Stages of Indexer Experience .....	66
2.4. An Interpretive Approach .....	73
2.5. Summary and Conclusion.....	92
3. Representation .....	94
3.1. Determining the Subject Matter.....	97
3.2. Studies on Indexing .....	120
3.3. Relevance Studies.....	171
3.4. Summary and Conclusion.....	181
4. Semiotics .....	197
4.1. Definition of Semiotics.....	200
4.2. C. S. Peirce: A Biographical Sketch.....	202
4.3. The Sign.....	204
4.4. Unlimited Semiosis.....	211

4.5. Phenomenology .....	214
4.6. Categories of Signs .....	219
4.7. Summary and Conclusion .....	249
5. Semiotic Analysis of the Subject Indexing Process .....	251
5.1. The Subject Indexing Process as Unlimited Semiosis .....	253
5.2. The Steps of the Subject Indexing Process .....	260
5.3. Peirce's Categories of Signs and the Four Elements .....	275
5.4. Summary and Conclusion .....	300
6. Summary and Conclusions .....	303
6.1. Approaches to Indexing .....	306
6.2. Indexing as Interpretation .....	309
6.3. Semiotics and Indexing .....	312
6.4. Later Study of Indexing .....	315
6.5. Summary .....	318
Bibliography .....	319
Vita .....	345

## List of Figures

Fig. 1-1. The Necker Cube.....	10
Fig. 1-2. The Jastrow Duck-Rabbit.....	11
Fig. 2-1. The Scope-Matching Process.....	56
Fig. 2-2. The Subject Indexing Process.....	62
Fig. 3-1. Conceptions of Subject Analysis and Indexing.....	130
Fig. 3-2. Variables in Relevance Judgments.....	172
Fig. 3-3. Definitions of Relevance.....	173
Fig. 3-4. Aspects of the Five Conceptions of Indexing .....	189
Fig. 3-5. Five Conceptions of Indexing.....	190
Fig. 4-1. The Semiotic Triangle.....	205
Fig. 4-2. The Y-leg Model.....	208
Fig. 4-3. Unlimited Semiosis.....	212
Fig. 4-4. Triadic Classification of Signs.....	221
Fig. 4-5. Qualisign .....	229
Fig. 4-6. Iconic Sinsign.....	230
Fig. 4-7. Rhematic Indexical Sinsign.....	232
Fig. 4-8. Dicent Sinsign .....	233
Fig. 4-9. Iconic Legisign.....	234
Fig. 4-10. Rhematic Indexical Legisign.....	235
Fig. 4-11. Dicent Indexical Legisign .....	237

Fig. 4-12. Rhematic Symbol .....	238
Fig. 4-13. Dicent Symbol.....	240
Fig. 4-14. Argument .....	242
Fig. 4-15. Ten Categories of Signs .....	244
Fig. 4-16. Signs Embedded in Signs.....	245
Fig. 4-17. The Signs Relations to the Representamen.....	246
Fig. 4-18. The Signs Relations to the Object.....	247
Fig. 4-19. The Signs Relations to the Interpretant.....	248
Fig. 5-1. The Semiotic Model of Indexing .....	255
Fig. 5-2. Document Analysis .....	261
Fig. 5-3. Subject Description.....	266
Fig. 5-4. Subject Analysis.....	270
Fig. 5-5. Matches Between Categories of Signs and Elements .....	276
Fig. 5-6. Category X - Argument.....	279
Fig. 5-7. The Document.....	281
Fig. 5-8. Category IX – Dicent Symbol .....	287
Fig. 5-9. The Subject .....	288
Fig. 5-10. Category VII– Dicent Indexical Legisign.....	292
Fig. 5-11. The Subject Description.....	293

## 1. Introduction

The representation of documents and the knowledge expressed by them is one of the central and unique areas of study within Library and Information Science (LIS) and is commonly referred to as indexing. One might therefore be surprised to learn how little there is actually known about the problems related to the indexing of documents. A common demand in the LIS field is for a set of rules or a prescription for *how to* index. When this demand is raised it is usually based on the assumption that it is possible to explain the intellectual operations in the subject indexing process.<sup>1</sup>

Shaw and Fouchereaux have reported on an investigation that the American Society for Information Science's Research Committee undertook to identify areas within library and information science where research was needed. On the basis of statements of research needs in three volumes of the *Annual Review of*

---

<sup>1</sup> The term "subject indexing process" is here used to separate the problems of representing a document's subject matter from the descriptive part of document representation. This dissertation uses the term synonymously with terms such as "subject cataloging" and "classifying." Due to stylistic concerns, "subject indexing process" is throughout the dissertation referred to as "indexing," "indexing process," or simply "the process."

*Information Science and Technology* (ARIST), they identified a number of areas that need further investigation. One of these areas was “what are the cognitive processes involved in indexing and classification?” (Shaw and Fouchereaux 1993, 25). Milstead has likewise discussed the need for more research in indexing. She noted 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’” (Milstead 1994, 578). Vickery notes, in an investigation of presuppositions in information science, that if the indexing process is carried out by humans--as opposed to some automatic process--it “requires that [the indexer] establishes a *meaning* for the message, so that the choice of the ‘most significant’ elements may be made” (Vickery 1997, 469). Hutchins (1978, 172) has nicely summed 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.”

Statements like these can be found many places in the indexing literature. The present study takes up the challenge and investigates the problems of representing the subject of a document. While it is not claimed that what will be presented in the following pages is the final answer to this fundamental question, at the end the reader should have a clearer conception of the subject indexing process and the problems related to the process.

Since so many, during so many years, have called for more research in this area, a relevant question is why there has been so little research so far. It is beyond the scope of this dissertation to answer that question in detail, but the answer might lie in the research methods commonly used in LIS. As will be discussed at the end of chapter 2, the field favors empirical research methods. But the problems of the subject indexing process reach beyond the process itself, and answers to the problems do not lie in the performance of the process itself. An empirical investigation, therefore, will not reveal much. This assumption is supported by the findings (or lack of findings) of the few empirical studies of indexing which have actually been undertaken (e.g. Chu & O'Brien 1993 and Bertrand & Cellier 1995).

The present dissertation is an attempt to explain at least some of the major problems related to representing the knowledge in documents; more specifically, it attempts to explain the nature of the subject indexing process in a new way.

This study is based on the assumption that it is not possible to make a general prescription of how to index and therefore explores the subject indexing process from the perspective that the process is one of interpretation. The purpose of the dissertation is to explain the nature of the subject indexing process from this perspective and thereby provide an understanding of the subject indexing process that explains why a predictable result can not be expected from the indexing process. It provides an understanding of the subject indexing process that views the process as a number of interpretations which to some degree depend on the specific cultural and social context of the indexer. The aim is not to provide a new and improved method for indexing, but to investigate the nature of the subject indexing process. This investigation is held at a level independent of specific indexing languages and indexing practices.

Because the main problems of the representation of documents are concerned with meaning and language, the subject indexing process is here explored from a philosophy of language perspective. Others have begun with similar assumptions. Fairthorne (1969, 78), 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, vii-viii), in the introduction to his book on language problems in information retrieval, notes that,



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.

In this respect, this study argues that the subject indexing process consists of a number of steps that should be viewed as *interpretations*, not as mental representations or mental rules. Accordingly, the situation requires one to take an *interpretive approach* in order to understand the uncertainty and interpretive nature of the subject indexing process. Benediktsson (1989, 218) has noted the interpretive nature of the process and the need for guidelines that recognize the significance of interpretation.

[A]ny sort of bibliographical description . . . can be considered descriptive. When it comes to interpretation, the question is: Ought not the description to follow a method or standard as any canon, which makes interpretation possible?

The present study will continue the approach to studies of indexing and LIS suggested by Fairthorne, Blair, Benediktsson, and others. In order to gain a better

understanding of the nature of the subject indexing process the present study will seek to answer the following questions:

- What is the distinctive character of the subject indexing process?
- What major problems and discussions have been raised in the LIS literature regarding the subject indexing process?

From the findings based on these questions it is clear that multiple interpretations take place throughout the subject indexing process. If the subject indexing process is a series of interpretations, then a theory that can explain the nature of the process from this perspective is needed. For this purpose the study of signs and semiotics, as discussed in the writings of Charles Sanders Peirce (1839-1914), is suggested as a useful theoretical framework or method for studying and understanding the interpretive nature of the subject indexing process. Peirce's semiotics is useful for this because it includes an explanation of how the meaning of signs is generated, interpreted, and represented.

Peirce provides a general theory of meaning and representation. The theory is general in the sense that it is not limited to language per se, but includes any phenomena that is subject to interpretation and is centered on explaining why a certain sign means one thing for one person and another thing for another person.

A number of other philosophers of language have discussed this problem but only Peirce have developed a full explanation of the different kinds of interpretation there are. He did this by constructing a categorization of kinds or types of signs. This categorization suggests that interpretation of signs are of different kinds and since one of the purposes of the present study is to explain the interpretive nature of the subject indexing process, Peirce' semiotics is valuable. By categorizing the interpretive steps in the subject indexing process it is shown exactly what the nature interpretation of these steps are. The following questions are of primary concern:

- What is the distinctive character of the study of signs?
- How can the subject indexing process be analyzed and explained by semiotics?

The object of applying semiotics to the subject indexing process is to demonstrate how fundamentally interpretive the nature of the process is. The semiotic explanation of the subject indexing process presented in this dissertation will elaborate on the multiple-interpretive nature of the process by showing that the individual interpretations that take place throughout the process are of different kinds.

For the purpose of analyzing the indexing process in greater detail, the process is (in chapter 2) described as a number of elements and steps, where each step is an interpretation and each element is that which is interpreted. In the semiotic analysis of the subject indexing process (in chapter 5) it is shown how each step is a sign that is interpreted and that the result of the interpretation is the next element in the process. It is furthermore demonstrated how independent the individual interpretations in the subject indexing process are.

To gain a fuller understanding of the nature of the different interpretations in the indexing process, each element in the process is considered a sign and categorized in Peirce's categories of signs. Peirce defined ten different categories of signs, each of which requires a different kind of interpretation and involves different kinds of uncertainty. By categorizing each element in these categories it is demonstrated that the subject indexing process is highly contingent upon social and cultural contexts rather than being a simple process of converting the document into a representation. Indexing contingency upon social and cultural contexts suggests that empirical studies of indexing need to be concerned with how texts are interpreted rather than with what indexers appears to do when they index.

The value of this study is twofold. First, no one has so far synthesized the literature in indexing of research from the perspective of an interpretive approach.

And no one has created an approach to indexing based on an interpretive understanding. This study will emphasize the main problems related to the subject indexing process and provide an explanation of the nature of the subject indexing process from a hermeneutic standpoint. Second, and perhaps most importantly, the dissertation will reveal that a new direction needs to be taken in the study of indexing. The dissertation will show that the problems of the subject indexing process are buried in philosophical questions and that these questions need to be addressed in a LIS context.

### **1.1. TO SEE AND TO INTERPRET**

Before the problems dealt with in the following pages are introduced, a larger framework of problems related to the topic of this dissertation will be introduced briefly here. A more detailed introduction to the philosophical and methodological framework will be given in section 2.6.

Sometimes two people see the same object, but claim to see two different things. Why is that? Likewise, why are two people who see the same document unable to agree on the subject of the document? These basic, but unanswered, questions which have puzzled psychologists and philosophers for centuries are at the core of this dissertation.

The problem of why people claim to see different things can be explained by illustrations. The first example (see Figure 1-1) is the Necker cube about which Hanson (1958, 8) notes,

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.

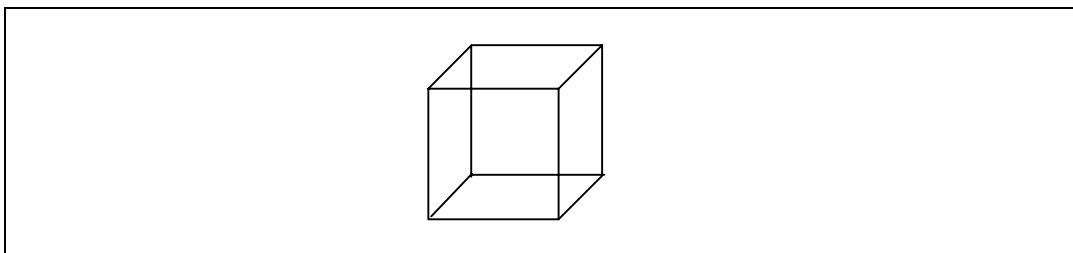


Fig. 1-1. The Necker Cube

Wittgenstein (1958, 193) also discussed the Necker cube. Such an illustration, he said, can be found,

[S]everal places in a book, a text-book for instance. 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.

In other words, it is not the illustration itself that determine what the illustration is about. In this case the text that supplies the illustration will direct how the illustration is understood. But even so, the context does not determine our interpretation. Wittgenstein argues (1958, 193):

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.

The cube itself does not changes and yet we see it differently. The basic question is how such differences can be accounted for.

Another famous example is the Jastrow duck-rabbit, as shown in figure 1-2.

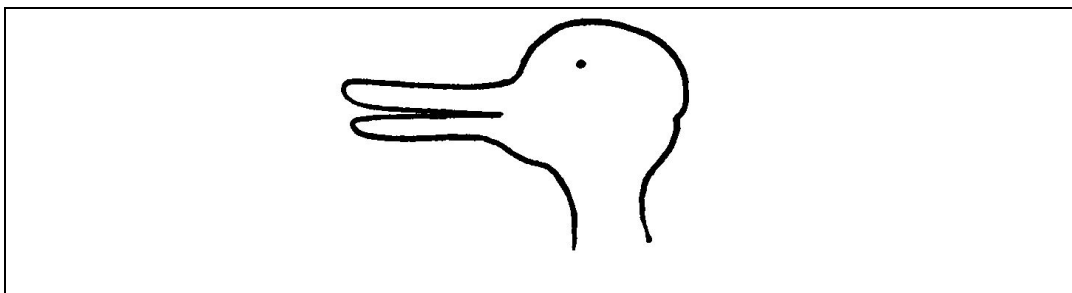


Fig. 1-2. The Jastrow Duck-Rabbit (Wittgenstein 1958, 193)

The illustration can be seen either as a duck's head or as a rabbit's head, but both animals cannot be seen at the same time. What changes when the illustration at one point is seen as a duck's head and the next second as a rabbit's head? The lines are exactly the same. But we would say that the illustrations look different. Wittgenstein (1958, 195) goes on to ask,

But what is different: my impression? my point of view?—Can I say? I *describe* the alteration like a perception; quite as if the object had altered before my eyes.

Wittgenstein (1958, 202) concludes that the difference comes from the way we interpret the drawing,

“And is it really a different impression?”—In order to answer this I should like to ask myself whether there is really something different in me. But how can I find out?—I *describe* what I am seeing differently.

Even though Wittgenstein is tempted to try to determine out what is actually going on in him as he sees the illustration, he argues that this is not viable. He argues that we can learn about what people think or how they interpret only by looking at



what they do. In the case with these drawings we can learn about the different interpretations only by looking at different descriptions of them. These descriptions are of course also subject to interpretation. Thus, Wittgenstein rejects a mental investigation and relies on what is said and done. The reason is that he finds that what is important is what is done, since everything else would merely be a physiological explanation of an experience. Such a physiological explanation will not provide any information about how the drawing is understood. As Wittgenstein (1958, 212) says to interpret is to do something,

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.

Perhaps more importantly, we cannot say, "I believe that it is a picture of a rabbit" if we can see only the rabbit. In that case we must say, "It is a rabbit." In short, we must have taken a view before we acknowledge or recognize other views and argue the view we have. The same holds for the indexing of documents. Only in those cases where the indexer is able to suggest a number of different interpretations of a document will she be able to argue for her choice. If the indexer only sees one possible interpretation of the document she will not be able to argue for that choice, because her choice would then seem as the natural

and obvious choice. As it will be discussed in Chapter 3 empirical studies of indexing have generally assumed that there is one best and correct representation of a document and have therefore not studied whether indexers actually make a choice between several possible representations. Empirical studies have therefore mostly focused on the consistency of the results of the indexing process.

Merrell has also discussed the difference between seeing and interpreting. He states that “people see, not retinas; cameras and eyeballs are blind” (Merrell 1995, 47). The process of seeing the cube or the duck-rabbit entails two distinct operations, namely that of retinal stimulation and of interpretation. Retinal stimulation is an immediate feeling, the might be of a possibility, whereas interpretation is a mediated intellection, “a synthesis of feeling and the effect ... of something ‘out there’ on the sensory organs” (Merrell 1995, 48).

Merely seeing a document does not result in the subject matter of the document being “seen.” The determination of the subject involves an interpretation of the document and the interpretation of a document involves the context in which the document is interpreted.

If the discussions and assumptions mentioned here are correct, how would it ever be possible to represent the subject matter of documents? If all interpretations were equally valid, how would it ever be possible to identify what the subject matter of a document is? This dissertation will argue that this question

can only be answered by understanding the nature of the interpretations that takes place during the subject indexing process. It will be argued that these interpretations are of different kinds and an understanding of them will provide a deeper understanding of the problems of representing documents. This is not to suggest that indexing practice directly will be improved by taking the approach suggested here or that all indexing previously done have not been successful. The suggestion is merely that if one of the central tasks of LIS is to represent documents for retrieval, then the complexity of that task must be understood. An improvement of practice should be guided by a full understanding of the complexity of practice. Practice cannot be improved only by trial and error.

## **1.2. STRUCTURE OF THE DISSERTATION**

This dissertation is divided into six chapters: an introduction, four individual chapters, and a final chapter containing a summary and conclusions.

Chapter Two, *The Subject Indexing Process*, presents and analyzes the nature and characteristics of the subject indexing process in order to show how serious the problem of subject indexing is and how little is known about it. It is shown that existing manuals on indexing do not provide adequate guidance. The chapter further introduces a conceptual framework of the subject indexing process

that will be used in this dissertation. In the last section of the chapter, the *interpretive approach* is introduced, the philosophical and methodological framework used in this dissertation. The main purpose of this chapter is to set the scene for the problem matter of the dissertation.

Chapter Three, *Knowledge Representation*, reviews the literature relevant to representation of documents' subject matter. Due to the limited amount of research in this area, each work will be reviewed rather carefully. This review includes literature, which specifically discusses the subject indexing process, as well as literature on the concept of relevance. The latter is included since this literature contains discussions of the concept of subject.

Chapter Four, *Semiotics*, introduces Charles S. Peirce's semiotics and philosophy. The emphasis is on his concept of sign, his notion of unlimited semiosis, and his categorization of signs. In chapter Five, *Semiotic Analysis Applied to Indexing*, the conceptual framework of the subject indexing process, introduced in Chapter Two, is analyzed semiotically. The goal is to introduce the Semiotic Model of Indexing, and further expand the model by describing the elements in the process in terms of Peirce's ten categories of signs.

The last chapter, *Summary and Conclusions*, sums up the findings and discusses the conclusions and implications drawn from the analysis of the subject indexing process. The main conclusion is that the indexer should be viewed as

having a much more important and central position in the subject indexing process than has been recognized in the past. Indexing needs to be proactive and thought of as a contributing factor in information retrieval, not as a neutral operation. Furthermore, it must be recognized that indexing should be analyzed, studied, and practiced as an act of interpretation and that the knowledge representation should be based in the domain in which both indexer and users work.

## **2. The Subject Indexing Process**

In the preceding chapter it was stated that very little research actually exists on the problems of determining the subject of a document. In this chapter a framework for understanding the indexing process is given. This framework will be used in the semiotic analysis of the subject indexing process in chapter 5. The discussion in this chapter will cover four points.

First, typical approaches to document indexing will be reviewed in section 2.1. A public exchange of views related to PRECIS is recounted in order to set the tone for how little is actually known about the indexing process. After this, two typical examples of manuals and guidelines (DDC and ISO) for classification and indexing will be introduced and discussed. The stated purpose of these guidelines is to give general directions for determining the subject of a document. It will be argued here that such guidelines on indexing are insufficient for this purpose. Second, the issue of indexing “steps” or “stages” will be broached in section 2.2. In the LIS literature it is commonly argued that indexers go through a number of steps when indexing. Here, a full range of steps are elaborated and

discussed, after which it is concluded that the subject indexing process consists of three steps--or sub-processes--and four elements. Third, the indexer's progress from novice to expert is examined in section 2.4. Here, the five stages or steps that Dreyfus and Dreyfus propose (1986) for shifting from the position of novice to expert are examined. On the basis of this approach it is then argued that novice indexers rely more heavily on guidelines than expert indexers and that novice indexers break down the indexing process to sub-processes more often. In other words, novice indexers proceed in a significantly different way than do experts. Studies of how experts index reveals little about how novices should be taught to index, however. Likewise, studies of how novice indexers index reveal little about how experienced indexers index. Lastly, in section 2.6. the interpretive approach will be introduced. This approach to LIS in general and to the representation of documents and knowledge in particular will provide the central justification for how this dissertation proceeds in examining the nature of the subject indexing process.

## **2.1. DETERMINING THE SUBJECT OF A DOCUMENT**

Various scholars have noted in LIS literature how little is actually known about the subject indexing process. This conclusion is highlighted by Farrow (1991, 164), who, after a thorough investigation of the problem, concludes that,

... there is in any case a need for more research into the indexing process itself . . . It is surely remarkable that so little is really known about so basic a professional activity.

The lack of knowledge about the process that Farrow points out has been discussed from time to time. Such discussions are no better illustrated than in an exchange of comments that followed the publication of the analytic procedure use in the PRECIS system in 1974. PRECIS (Preserved Context Index System) is generally considered one of the most, if not the most, ambitious late twentieth century attempts to create an indexing system from scratch. This indexing system was built on the notion that the meaning we get from a word is dependent on the context of other words (Austin 1974b). When presenting the theoretical and technical history of the PRECIS system, Derek Austin stated (1974a, 1974b) 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” (Austin 1974a, 49).



In a comment on Austin's work, Moss asked why Austin only writes two lines on this matter, but elaborates much more on the remainder of the indexing process. He points out that "it is apparent that the first--very much taken-for-granted--step is the crucial and vital in any indexing and classification" (Moss 1975, 116). Swift agreed with Moss, concluding, "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" (Swift 1975, 117).

Jones (1976, 118) summed up the Austin-Moss-Swift discussion,

... few attempts [have been made] to establish the nature of the indexing *process*. Most studies claiming to be about indexing are, in fact, about indexes.

In other words, studies claiming to be concerned with the steps of the indexing process often solely focus on the last step. Thus, most literature and research on indexing is concerned with the translation of the subject matter into an indexing language. Jones later supported his conclusion by investigating the nature of research about the subject indexing process. He found that "the relationship between text and index is rarely examined" (Jones 1983,1). This kind of discussion is also most often only concerned with the technical aspects of

indexing and indexing languages. In fact, it regularly dwells on the advantages and disadvantages of different kinds of indexing languages and techniques (cf. ex. Rowley 1994 and Bodoff & Kambil 1998).

Even more important than the foregoing general comment about discussions of the overall indexing process is the fact that very little has been done to clarify the intellectual processes concerned with the first or initial step in the indexing process where the subject of a document is identified. A number of scholars have pointed this out. For instance Langridge, who holds that the indexing process consists of four steps, the first of which identifies the subject of a document and the last of which focuses on converting the identified subject into the terminology of an indexing system, states that the first step must be regarded as "the most important and the most difficult part of all classification and indexing" (Langridge 1989, 1). However, he also states that 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, 6). Frohmann, who argues that indexing takes two steps, the first of which focuses on determining the subject of a document and the second of which focuses on converting that identified subject into the language of an indexing system, has pointed out that (Frohmann 1990, 82),

... 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.

Although this emphasizes that studies in indexing typically neglect the initial step in the indexing process where the subject of a document is identified, one might expect some help from sources such as manuals. While guidelines of indexing and classification do not attempt to explain the indexing process theoretically, at least they attempt to outline how the indexer should proceed in indexing. In such guidelines one would reasonably expect some indication of how to determine the subject matter of documents. But in reality they are generally uninformative. They merely recommend examining tables of contents, scanning chapter headings, examining forewords and introductions, etc., and assume that this procedure will make the subject matter of a document evident. But they provide no details or help to determine the subject of a document and appear to assume that the process is essentially intuitive.

#### **2.1.1. The DDC Guidelines**

One example of the above approach is the introduction to the Dewey Decimal Classification system (Mitchell et al. 1996, xxxv), which states that

“classifying a work properly depends upon determining the subject of the work in hand” (Mitchell et al. 1996, xxxv). For this the DDC provides a set of guidelines -- a list of points or sources for determining the subject of a work.

Fourteen possible sources of information are listed as to help identify the subjects of documents. These sources are underlined in the following excerpt from the DDC Introduction. It should be noted that the first ten of the sources are found within or essentially associated with a given document, whereas the last four are found outside the document.

- a) The title is often a clue to the subject, but should never be the sole source of analysis. For example *The Greening of America* is a book about social conditions and social change, not a work on ecology.
- b) The table of contents may list the main topics discussed. Chapter headings may substitute for the absence of a table of contents. Chapter subheadings often prove helpful.
- c) The preface or introduction usually states the author’s purpose. If a foreword is provided, it often indicates the subject of the work and suggests the place of the work in the development of thought on the subject. The book jacket or accompanying material may include the summary of the subject content.
- d) A scan of the text itself may provide further guidance or confirm preliminary subject analysis.
- e) Bibliographical references and index entries are sources of subject information.

- f) Cataloging copy from centralized cataloging services is often helpful by providing subject headings, classification numbers, and notes. Such copy appears on the verso of the title page of many U.S., Australian, British and Canadian books as part of Cataloging-in-Publication (CIP) data. Data from these sources should be verified with the book in hand since cataloging-in-publication is based on prepublication information.
- g) Occasionally, consultation of outside sources such as reviews, reference works, and subject experts may be required to determine the subject of the work.

These guidelines are very vague and weak in the sense that is discussed and demonstrated next.

Point a) directs the indexer to the title of the document. The idea is that the title should give some clue about the subject matter of the document or that the title summarizes the subject matter in a few words. In many cases this might be the case. Take as an example David C. Blair's from 1990m "Language and Representation in Information Retrieval." The book is about language, representation, and information retrieval. However, the indexer must make the connections between 'language,' 'representation' and 'information retrieval.' Even though the title has provided the indexer with some clue to the subject matter, there are still a number of places where the book could be placed in the DDC, including: 004.019 psychological principles (in computer science), 025.48

subject indexing, 025.524 information search and retrieval, 121.68 meaning, interpretation, hermeneutics. 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.

Point b) directs the indexer to the table of contents, or chapter headings in the absence of a table of contents. Whereas the title could summaries of the entire document, the purpose of the table of contents is to name the different parts of the document. The indexer's task is therefore different and the clue much less explicit. The table of contents in Blair's book is four pages long. However, 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." At first these chapter headings might 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 system. Only the fourth chapter seems covered by the title. If the indexer looks at the sub-headings 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 document," "Do we have a complete

typology of illocutionary acts,” “The generic algorithm: relevance,” “Science vs. pseudoscience.” These, and others like these, suggest a different or broader subject matter than simple “information retrieval.” Again, the indexer has been helped some, but the task of summarizing the table of contents into one statement which expresses the subject matter of the document is not simply a task of picking out the subject matter of the book. The table of contents will never directly state the subject matter of the document but simply name some of the parts of the documents. The indexer needs to put the parts together.

Point c) directs the indexer to the preface, introduction, foreword, and book jacket. There is no universal agreement on what these four elements contain. But, some things may be fairly clear. The preface often contains acknowledgments and information about how the document came into existence, but only sometimes deals with the subject matter of the document. The author of the document usually--but not always--writes the preface. The foreword is usually--but not always--written by someone other than the author and does not always contain a statement of the subject of the document. The foreword is sometimes written by an authority in the field or by someone who knows the value of the work of the author. The author almost always writes the introduction, which often contains something about the subject of the book, but not always so. It may for example tell why the document was written, what it hopes to achieve, and what it contains

(a summary of what to expect in each part of the book), but precious little in the way of a summary statement or definition of the document's subject matter. All these places might be helpful to the indexer. However, the indexer still needs to convert the statements into a summary of the overall subject matter. Moreover, statements found at these places might be biased or represent a specific point of view and may not explicitly state the overall subject matter of the document. The indexer might be provided with useful information regarding the subject matter, but she cannot expect to simply pick it out in any of these places. Furthermore, none of these places were created to summarize the subject matter. The preface is usually used to acknowledge various people and say something about the origin of the work. The introduction is used to lead the reader into the work and maybe discuss the subject of the document, but this is often rather vague. The foreword, if written by someone other than the author, says something about the significance of the work. And the book jacket is basically intended to sell the document. Although all of these say something about the document and its content, none of them is created with the intention of summarizing the subject matter.

Point d) directs the indexer to scan the text for further guidance or confirmation of a preliminary analysis. The advice in this point seems to indicate that some actions should be taken before the indexer scans the text. These actions are not outlined. However, it must be assumed that the actions thought of are the



other--or some of the other--points in the DDC guidelines. The idea here seems to be that after the indexer has gone through the other points she scans the text to confirm what she has already found. However, the guidelines do not explicitly state what it means to scan a text. It could mean that the indexer should simply look through the document or maybe even speed-read the entire document. But, it could also mean that the indexer should read selected sections of the text closely, such as the first and last paragraph of each chapter. Nonetheless, it is important to notice that the DDC guidelines recommend that the indexer should not start the examination by reading or scanning the text. The text should be scanned either to finish off the examination of the document and confirm what the indexer has reached elsewhere, or to give further guidance in the analysis. The latter aims at situations where the indexer is stumped in her analysis of the document. Scanning the text might in such situations help the indexer. The crucial point is what it means to scan a text. The guidelines do not discuss this and gives no clue to understand the concept. One common definition is that scanning is a perceptual act<sup>2</sup>. This suggests that the indexer immediately or intuitively recognizes the subject matter of the document.

---

<sup>2</sup> This definition of the term is also used in Farrow's (1994) chapter on indexing in the Encyclopedia of Library and Information Science. Farrow's work will be reviewed later in the dissertation.

Point e) states that bibliographical references and index entries are also sources of information. The first of these--bibliographical references—consists of the bibliography or list of references either at the end of the document or in the footnotes. The idea is that an indexer who knows the literature of a particular field will be able to identify the subject matter of the document on the basis of the literature the author uses or cites. This of course is likely to create some bias in the analysis. The indexer will seldom know all the literature the author refers to, and she is therefore inclined to emphasize the literature she knows. On the other hand, the process of converting knowledge about the literature the author refers to into the subject matter of the document is neither clear nor described. It should be clear that the well-read indexer would gain some knowledge about the document's subject matter by looking at the bibliographical references, but how this knowledge is converted into knowledge about the subject matter is not clear. The other part of the point directs the indexer to the index entries or back-of-the-book-index. The indexer could look through the index for information about the particular subjects that the document treats. The idea is that the subject matter of the document is broken down to a list of index terms and that the indexer could convert these into the document's overall subject matter. The back-of-the-book-index does not weigh the individual subject entries and the indexer will therefore be unable to know the relative weight of the many subjects mentioned in the back-

of-the-book-index. Clearly the indexer needs knowledge about the subject matter of the document before she directs her attention to the back-of-the-book-index. However, this index could serve as a good source of information when the indexer wishes to check whether she has overlooked some major subjects contained in the document.

Points f) and g) direct the indexer to look outside the document--at cataloging copy, reviews, reference works, and subject experts--in order to find summary accounts of the subject matter of the document. This differs from the previous points where the indexer's attention is directed to sources internal to or adjacent to the document itself. In both of these points the guidelines suggest that summaries might well exist in media external to the document.

Point f) directs the indexer to the cataloging-in-publication data. This information usually contains the subject matter of the document expressed in a major indexing language. If the indexer uses the same indexing language (for example Dewey Decimal Classification, Thesaurus of Eric Descriptors, Library of Congress Subject Headings), the indexer's work has in fact been done for her. Of course, the indexer needs to check whether she agrees with the cataloging-in-publication data. If she does not use the same indexing language she might be able to convert the notation from the cataloging-in-publication data into the notation of the indexing language which she uses. This of course works for

indexers in environments where the documents are purchased from publishers who provide cataloging-in-publication data. However, the main problem remains; how will the subject matter of the document be established in the first place.

Point g) tells the indexer to occasionally consult outside sources, such as reviews, reference works, and subject experts. Nothing is said about when outside sources are needed. However, the assumption must be that only in cases where the indexer cannot determine the subject matter of the document herself must she turn to outside sources. Frequently, reviews will summarize the content of a document, but just as often reviews will discuss the contribution and significance of the document. It helps if the reviewer summarizes the content of the document, but sometimes reviews are biased or only discuss a particular aspect of the document. The indexer is also advised to consult reference works. The idea here is that in cases where the indexer knows little about the subject matter of the document she could consult dictionaries or encyclopaedias to learn more about the subject matter. This way the indexer could gain a better understanding of the conceptual context of the document. However, this requires that the indexer know which concepts to look up and is able to understand the explanations. The indexer is not helped much if she is unable to apply this knowledge to the subject matter of the document.

Finally, the indexer is advised to consult subject experts. Since these will have the subject expertise to determine the subject matter of the document, the indexer could either rely on the subject expert's evaluation of the document or the subject expert could help the indexer to understand the content, to be able to determine the document's subject matter. The DDC guidelines are rather implicit in how the indexer should use this information and convert knowledge obtained from these outside sources to the subject matter of the document.

The overall impression of the DDC guidelines is that they point out more or less obvious places where the indexer could look for the subject matter of the document. However, the places pointed seem more likely merely to confirm whether an already determined subject is correct or not. None of the places pointed out seem likely to consistently and explicitly state what the subject matter itself is. The indexer is only given a minimum of help in determining the subject of the document.

A final criticism of the DDC guidelines has to do with its relationship to the DDC as a hierarchically structured classification schema. The ultimate goal of the DDC guidelines is to place the documents in an appropriate place in the hierarchical classificatory structure of the DDC. It therefore seems like a serious

omission that the embedded worldview<sup>3</sup> and structure of the classification scheme is not mentioned as a guide in the process. This could direct the cataloger in her analysis of the document by providing a conceptual framework for the analysis.

### **2.1.2. The ISO Guidelines**

The second set of guidelines to be examined here are found in the International Standards Organization's publication entitled *Documentation—Methods for Examining Documents, Determining their Subjects, and Selecting Indexing Terms* (ISO 1985). These guidelines are much more detailed than the DDC guidelines about the exact actions the indexer should take throughout the indexing process. These guidelines limit themselves to “any agency in which human indexers analyze the subjects of documents and express these subjects in indexing terms” (ISO 1985, 1.2). However, the guidelines focus only on how to determine the subject matter; they do not deal with the practices of any particular indexing system. The intention of the ISO guidelines is to (ISO 1985, 1.4),

[P]romote standard practice

a) within an agency or network of agencies;

---

<sup>3</sup> The idea here referred to is that any classification scheme or thesaurus expresses a particular view of the world. The indexer needs be aware of this and may even use it actively in her work. See for instance the work of Sanford Berman (1993), Hope Olson (Olson 1996; 1998, Olson & Ward 1997), and Bernd Frohmann (1994b).

- b) between different indexing agencies, especially those which exchange bibliographic records.

The guidelines state that the indexing process consists of three stages, although these tend to overlap in practice (ISO 1985, 4.3),

- a) examining the document and establishing its subject content;
- b) identifying the principal concepts present in the subject;
- c) expressing these concepts in the terms of the indexing language<sup>4</sup>.

The ISO guidelines are divided into a number sections, and each of the three stages above is dealt with in its own section.

In the first stage of the indexing process the document is examined and its subject content established. The guidelines note that a complete reading of the document often is impracticable, but that “the indexer should ensure that no useful information has been overlooked” (ISO 1985, 5.2). The guidelines thereafter mention a number of places which the indexer should give particular attention, (ISO 1985, 5.2),

- a) the title;
- b) the abstract, if provided;
- c) the list of contents;
- d) the introduction, the opening chapters and paragraphs, and the conclusion;
- e) illustrations, diagrams, tables and their captions;
- f) words or groups of words which are underlined or printed in an unusual typeface.

For performing the first task in the indexing process, examining the document and establishing its subject content, the ISO follows the DDC guidelines in stating that the indexer should look for hints about the subject of the document by examining elements of the document. The ISO guidelines are not as thorough as the DDC guidelines, however. The DDC guidelines also do not include sources outside the document such as reviews, reference works, cataloging copy, and subject experts. Its list of elements, while somewhat similar to the DDC list, contains important differences, including, abstracts, the opening chapters and paragraphs, the conclusion, illustrations, diagrams, tables and their captions, and underlined words or words in an unusual typeface.

---

<sup>4</sup> The ISO standard is an anonymous work, however, the ideas presented in it are the same as those which Austin expressed in the PRECIS system, although the wording differs a bit. Austin (1974b, 28) identified three stages in the indexing process: “a) a subject is first broken down into separate concepts defined in terms of their syntactical roles and hence their relationships with the other components of the subject; b) operators which express these roles are then assigned to the terms which represent concepts; c) finally, index terms are organised into a context-dependent string according to the filing nature of the operators.”



It seems obvious that the ISO guidelines and the DDC guidelines differ in one major way. The DDC guidelines appear to be aimed at books and the ISO at articles in proceedings of societies and academic journals. These kinds of articles are much shorter and more focused than book-length treatises. In some cases they have abstracts whereas books commonly do not. There is little provision here for works of art and of the imagination. There is little provision here for documents that are philosophical or artistic in presentation. In contrast, the DDC guidelines do not seem very adaptable to articles.

Given that the ISO guidelines differ from the DDC guidelines in their specific focus on journal articles, they also point the indexer to different sources of information within the document. Even though journal articles are shorter than books, to read them completely is also often impracticable. The indexer is therefore advised to consider all the above parts of the document. The indexer should be able to learn about the content of the document from the title and the abstract. However the ISO guidelines warn that titles could be misleading and the abstract “should not be regarded as a satisfactory substitute for an examination of the text” (ISO 1985, 5.2). The guidelines do not describe or discuss in any detail what the indexer should do to determine the subject besides that all the above “elements should be scanned and assessed by the indexer” (ISO 1985, 5.2). After

the indexer has examined the document she should be able to establish its subject content. The indexer thereafter proceeds to the next stage in the indexing process.

In the second stages of the indexing process the indexer identifies the principals concepts in the subject. The second stage is laid over the first stage in the sense that the indexer should not go back to the document to look for concepts. Rather, the indexer should look for concepts within the findings of the first step; that is the natural language representations of the subject content. The ISO guidelines suggest that the indexer should “follow a systematic approach” (ISO 1985, 6.1) to determine a document’s concepts and the indexer should be helped and guided in identifying the concepts. For that purpose the ISO guidelines suggest that the indexing agencies “establish check lists of those factors which are recognized as important in the field covered by the index” (ISO 1985, 6.1). The questions below illustrate the general factors that such checklists should establish (ISO 1985, 6.1),

- a) Does the document deal with the object affected by the activity?
- b) Does the subject contain an active concept (for example an action, an operation, a process, etc.)?
- c) Is the object affected by the activity identified?
- d) Does the document deal with the agent of this action?
- e) Does it refer to particular means for accomplishing the action (for example special instruments, techniques or methods)?

- f) Were these factors considered in the context of a particular location environment?
- g) Are any dependent or independent variables identified?
- h) Was the subject considered from a special viewpoint not normally associated with that field of study (for example a sociological study of religion)?

The idea is that the indexer should pose questions like these to the principal concepts present in the subject identified in the first stage. The questions are designed to ensure that the indexer does not overlook important aspects of the document.

The above questions are only examples of “general factors which are likely to apply in any subject field” (ISO 1985, 6.1). Each indexing agency needs to formulate questions which are specific to the particular disciplines covered. This again suggests that the creators of the ISO thought of indexing of scientific articles in specific domains when they formulated these guidelines. Although it may be possible to formulate such questions when indexing documents in a specific scientific discipline, in contrast, it might be difficult to formulate such questions for humanistic literature or indexing done at public libraries for the public at large.

The ISO guidelines stress that the concepts that the indexer selects or rejects should depend on the purpose of the indexing. However, what the indexer chooses to represent is mostly affected by the levels of exhaustivity and specificity which the indexer chooses. Exhaustivity is defined as the number of factors discovered through using the questions in the checklist. Specificity is defined as “the extent to which a particular concept which occurs in a document is specified exactly in the indexing language” (ISO 1985, 6.4).

By using the checklist procedure it should be possible to “identify all the concepts in a document which have potential value for the users of an information system” (ISO 1985, 6.3.1). However, the indexer is advised to bear in mind that the index could be used by more than one specific group of users and therefore should not interpret the subject matter of the document too narrowly. The main criterion for selecting concepts, however, “should always be the potential value of a concept” (ISO 1985, 6.3.3). When the indexer has to choose between a number of concepts, “the indexer should bear in mind the questions, as far as they can be known, which may be put to the information system” (ISO 1985, 6.3.3).

Lastly, it is noted that no arbitrary limits should be set to “the number of terms or descriptors which can be assigned to a document” (ISO 1985, 6.3.4). Such a limitation would likely “lead to some loss of objectivity in indexing” (ISO 1985, 6.3.4).

The idea is that a checklist of questions should help the indexer think of all of the important aspects that make up the document's subject matter. In other words, it is the belief that the indexer will be able to exhaust a document for its entire subject matter.

The other aspect which the ISO standard considers important is the level of specificity. Specificity is defined as "the extent to which a particular concept which occurs in a document is specified exactly in the indexing language" (ISO 1985, 6.4). Indexers are advised to represent concepts as specifically as possible. However, a general level might be chosen in certain circumstances, for instance, if "the indexer considers that over-specificity might adversely affect the performance of the indexing system" (ISO 1985, 6.4). One occasion for this might be if the document refers to specific concepts which "only occur in the fringe area of the subject field covered by the index" (ISO 1985, 6.4). Another circumstance where the indexer is advised to keep a more general level is when an idea presented is not fully developed, and when the author only casually refers to an idea.

During the first two stages the indexer has established the subject content of the document and identified the principal concepts in the subject. The indexer is hereafter ready to translate the concepts into the indexing language, that is the third stage in the indexing process.

The indexer should be familiar with the particular indexing language and the specific rules and mechanisms of the indexing language. If the concepts that the indexer has identified during the second stage are present in the indexing language the indexer should translate the concept into her preferred terms. If the concepts are not present in the indexing language, the indexer should check terms in dictionaries, encyclopedias, thesauri, and classification systems for accuracy and acceptability and add them to the vocabulary. The indexer is moreover advised that “subject experts, especially those with some knowledge of indexing or documentation” (ISO 1985, 7.1) could be consulted at this stage in the indexing process.

At this point in the indexing process the indexer should be aware that indexing languages “may impose certain constraints” (ISO 1985, 7.1) in expressing the concepts. If the indexer uses a controlled indexing language, this “may not permit the exact representation of a concept encountered in a document” (ISO 1985, 7.1). The concern is that the concepts that the indexer identified during the second stage of the indexing process might not be present in the indexing language. The indexer is then forced either to choose a term that does not express exactly the same concept or add a new term to the vocabulary to represent the concept.

The quality of the indexing depends on factors such as (ISO 1985, 8.1),

- a) the qualifications and expertise of the indexer;
- b) the quality of the indexing tools.

But in an ideal situation the terms assigned to a particular document and the level of exhaustivity “attained during indexing should be consistently the same regardless of the indexer employed” (ISO 1985, 7.1). In other words, it should be possible for two indexers to establish the same subject for a document and to identify the same concepts in that subject. Whether this occurs depends on the qualifications of the indexers and the quality of the indexing tools. It is however, noted that such consistency is not always possible in practice, “but the goal of consistency, and hence of predictability, is an important factor in the performance of the indexing system” (ISO 1985, 8.1). In other words, the ISO standard assumes that a high degree of consistency leads to a better performance in retrieving documents. Therefore the goal is a high degree of consistency.

Consistency can be achieved by “complete impartiality on the part of the indexer” (ISO 1985, 8.2); the indexer should refrain from “subjective judgment in the identification of concepts and the choice of indexing terms” (ISO 1985, 8.2). The indexer should furthermore have “adequate knowledge of the field covered by the documents he is indexing” (ISO 1985, 8.3) and the indexer should “have direct contact with users” (ISO 1985, 8.4).

The governing idea of the ISO guidelines is that indexing should be neutral, objective, and independent of the particular indexer's subjective judgment. The goal of the guidelines is to "promote standard practice" (ISO 1985, 1.4). However, this goal is at odds with the indexing process because that process requires decisions which cannot avoid being subjective. Therefore, such decisions could hardly reach the complete impartiality that the guidelines seem to require (ISO 1985, 8.2). For instance, the indexer should determine the purpose of the index and select the concepts that should be represented (ISO 1985, 6.2). But, the indexer should also determine what interests the document covers (ISO 1985, 6.3.2), and should base the selection of concepts on the users' potential (future) needs (ISO 1985, 6.3.3). All such decisions are unavoidably subjective, and they play an important part in the indexing process. Even the ISO guidelines infer this dependence on subjective judgment by stating that the quality of the indexing depends on the indexers' qualifications (ISO 1985, 8.1). At the same time, however, the guidelines do not discuss what these qualifications are other than stating that the indexer needs knowledge about the indexing tools and adequate knowledge of the subject field. Due to the individual indexers' degree of involvement in the outcome of the process, nothing definite can be said about the outcome and standardization of practice is an unrealistic goal.



One of the most important suggestions of the ISO guidelines is to include the potential users of the information system as a factor in the indexing process. The guidelines suggest that the indexer should “bear in mind the questions, as far as these can be known, which may be put to the information system” (ISO 1985, 6.3.3) and that the indexer should “have direct contact with the users” (ISO 1985, 8.4). An indexer who has contact with the users might better be able to represent the documents in accordance with how the users think. The idea is that the indexer should not necessarily attempt to determine *the* subject of the document but take the users’ questions and information needs into account when indexing. This might help the indexer when a document contains multiple concepts, in such instances the indexer could choose to represent only those concepts which are regarded as “most appropriate by a given community of users” (ISO 1985, 6.3.3.a).

Although the ISO standard stresses the use of users’ questions as a factor, and this appears to enhance the effectiveness of the indexing, the ISO guidelines do not state how its emphasis on users’ questions correlates with the statements about consistency and exchange of bibliographic records. It seems obvious that if an indexing service pays attention to “a given community of users” (ISO 1985, 6.3.3.a), this indexing service might index a document differently than another indexing service with a different user community. This would make consistency

among the indexing services less important and make exchange of records more difficult. This again points to the inherent contradiction in some of the suggestions given by the ISO standard.

In summary, the ISO standard is not really a standard but a set of guidelines, which can help the indexer or an indexing unit to develop some standard approach to indexing. As such, the ISO guidelines are helpful and useful. However, the guidelines stated intention--to promote standard practice--is questionable, and the parts of the guidelines that focus on this makes the use of the guidelines indistinct. Moreover, the philosophical underpinnings of the guidelines are questionable<sup>5</sup>.

---

<sup>5</sup> One way to interpret the underpinning ideas of the ISO standard is to focus on their definition of some central notions. The ISO standard defines a concept as “a unit of thought. The semantic content of a concept can be re-expressed by a combination of other and different concepts, which may vary from one language or culture to another” (ISO 1985, 3.2). A subject is defined as “any concept or combinations of concepts representing a theme in a document” (ISO 1985, 3.3). An indexing term is defined as “the representation of a concept in the form of either a term derived from natural language . . . or a classification symbol” (ISO 1985, 3.4). Although central, the term ‘theme’ is not defined. This could mean that concepts exist before communication or language and can be expressed in various ways through language. The better specific concepts are expressed in language the better the communication flows. A subject is then any concept or combination of concepts which is expressed in the document. The readers’ task is to interpret the words and sentences in the document in order to understand the concepts. Whether a reader understands a document depends on how precisely the author expresses the concepts she refers to and whether the reader is aware of the concepts the author expresses. The basic idea is that the concepts exist before the author writes the document and the reader reads the document.

Similarly, the indexer’s task is to identify concepts in the document and re-express these in indexing terms. This is done first by establishing the subject content, or in other words the content of concepts in the document. Thereafter the principal concept presented in the subject content is identified, and finally, the concepts are re-expressed in the indexing language. The indexing is successful when the document and the indexing term express the same concepts.

Although the ISO standard does not explicitly state its philosophical foundation the above analysis shows that the guideline does in fact make a clear assumption about the relation between reality, words, and meaning. In other words, the guideline is based on a certain theoretical tradition.

### **2.1.3. Summary and a Suggestion for a Change in Focus**

Guidelines like the ones put forth by DDC and ISO could potentially be of high value, if they could state exactly the actions an indexer or classifier should take in representing the subject matter of a document. However, it is questionable whether such guidelines are able to do so due the very nature of the indexing process.

Lancaster has argued that guidelines such as the ISO can hardly be considered true standards because a “true standard should be exact . . . and enforceable” (Lancaster 1998, 147). However, the indexing process is neither exact nor enforceable. Wilson noted as early as 1968 that manuals of indexing and classification often refer to the subjects of documents, but they “are curiously uninformative about how one goes about identifying the subject of a writing” (Wilson 1968, 73). The point is that attempts to explicitly prescribe how an indexer or classifier should determine the subject matter of a document are insufficient.

This might suggest that studies of indexing simply have not been thorough enough and that much more research along the same line of thought is needed in order to describe precisely what indexers do when they index. Or, it might suggest that it is not possible to describe precisely what one does when indexing

because the process is imprecise by its very nature and that research in the area therefore needs to shift focus. The present dissertation argues the latter.

Much previous research in the field has centered on attempts to describe how indexing is done, but the present work is an attempt to describe what indexing is about. The underlying assumption of this study is that it is not possible to make exact prescriptions of how to index and that it is more valuable to discuss the difficulties of indexing. Studies of indexing have a practical end, namely to enhance the quality of indexing. However, studies of how indexing is done must be based on an understanding of something deeper--they must be based on the underlying process of how one arrives at the meaning of a given set of human communicative symbols. This aspect is absent in most studies of indexing and in guidelines of indexing and classification. Furthermore, when one investigates this more fundamental activity of mind, which though a semiotic reading involves interpretations, it becomes clear that it is not possible to pin down meaning exactly. And if this is the case, it is also not possible to make exact prescriptions of how to index. Although studying the indexing process cannot make the process less vague--because it is inherently vague or at least has inherently vague elements--it can indicate to the indexer at just what points vagueness sets in. The assumption underpinning this dissertation is that if one knows at just what points vagueness sets in, measures may be taken on the one

hand to understand the limitations of the process, and on the other hand, hopefully to keep vagueness to an acceptable minimum.

In other words, studies of indexing need to shift focus from deductive-nomological explanations of the process to idiosyncratic studies and theoretical investigations of the process. From the latter some general principles of indexing should emerge. These principles will not be descriptions of how to index, but descriptions of the nature of indexing.

## **2.2. DOCUMENT AND SUBJECT ANALYSIS**

The previous section tried to show that typical discussions of the indexing process, especially those found in two significant sets of guidelines, do not really tell how to identify the subject of a document. The section ended with a call for a new focus which, while accepting the fundamental imprecision of the process, calls for investigations of the central nature of the indexing process simply to gain a greater understanding of it. Here this effort begins through the establishment of an intellectual framework for looking at the indexing process.

### **2.2.1. Indexing Viewed in Terms of ‘Steps’**

The foregoing discussion of the ISO guidelines suggested that the indexing process could be viewed in terms of a number of steps. This accords with the way other commentators have written about the subject indexing process. In fact, the process is portrayed variously as involving two steps, three steps, or sometimes even four steps.

The two-step approach (cf. e.g. Frohmann 1992; Petersen 1994) consists of one step in which the subject matter is determined and a second step in which the subject is translated into and expressed in an indexing language.

The three-step approach (cf. e.g. Farrow 1991; Miksa 1983; Taylor 1994) adds one more step to the process. The subject is still determined in the first step. However, a second step is then included in which the subject matter found in step one is reformulated in more formal language than is associated with the first step. Thereafter, in a third step the subject as more formally stated in the second step is further translated into the explicit terminology of an indexing language.

The four-step approach (cf. e.g. Chu and O'Brien 1993; Langridge 1989) is similar to the three-step approach in the first two points. The first step determines the document's subject matter more or less informally. In the second step the indexer then summarizes the subject matter of the document more formally, usually in her own vocabulary and in the form of a more compressed statement.

From this point forward, this approach differs from the three-step approach. Here the translation of the subject matter into an indexing language consists of two steps rather than a single step. In a third step the indexer translates the sentences into the vocabulary used in the indexing language. And in a fourth step the indexer constructs one or more subject entries in the indexing language--in the form of index terms, class marks, or subject headings--with respect to their syntax and relationships.

It should be noted that the idea of “steps” as recounted here has to do chiefly with the logic of the indexing process, not necessarily with the actual sequence of mental and physical operations. It may well be that some indexers, particularly those who are beginners in such work, may accomplish their indexing “by the numbers,” ticking off the steps as they go. However, this is less likely as experience is gained. In reality, experienced indexers and catalogers may not be conscious of the various steps at all, and all steps, regardless how many one supposes are most accurate, may well take place almost simultaneously. In short, an experienced indexer, one who we might call an expert (the question of expertise in indexing will be dealt with ahead), will perform the indexing process in just one complex action. However, it is useful to operate with the idea of steps when analyzing the process because breaking down the process into its individual logical parts will allow one to examine the process in greater detail.

Accordingly, this dissertation is based on the assumption that indexers actually do proceed through a number of steps, at least logically. This section will introduce a conceptual framework for viewing that process. This framework is important and will become essential later in this study when the indexing process is analyzed semiotically.

Furthermore, not only will the idea of steps be followed here but a three-step approach will actually also be applied. A three-step approach is chosen here for several reasons. The two step model is too simplified in its conception of the subject indexing process. In fact, the two step approach appears to be used chiefly as a device to separate two distinct activities in the subject indexing process—determining the subject of a document and converting that subject to the terminology of an indexing language. It is seldom used to discuss the details of the process itself. In contrast, the four-step approach appears to add an unnecessary complication to the final part of the process. The final part of the process consists of the activity of translating the subject of a document into the terminology of an indexing language. The four-step approach breaks that final part of the process into two parts. But it is the conclusion here that it is not useful to split the translation activity into two separate steps. There is really no essential difference between these two steps but only a difference of general versus specific activity. In the first of these two final steps, the subject of a document is said to



be translated into the language of a given subject access vocabulary, whereas the next step only translates the results into indexing terms or strings of terms (i.e., the syntax) in the system. One might just as well combine these two statements and state simply that one will translate the subjects into the language of a system and, more specifically, into the syntax used by that language. In sum, therefore, one can only conclude that separating these two points is unnecessary from any logical point of view. Certainly, they do not add substantially to the understanding of the overall process.

### **2.2.2. An Enhanced Three Step Approach to the Indexing Process**

The subject indexing process is presented here as a three-step process. However, the view here does not merely focus on the steps themselves, but rather presents an enhanced rendition of the steps. More accurately, the view of the indexing process presented here consists of *four elements and three steps*, where an element consists of an object that is acted upon and a step is the action taken upon the object. The sequence of the elements and steps are as follows: Element 1 - Step 1 - Element 2 - Step 2 - Element 3 - Step 3 - Element 4.

The first element consists of the *document* under examination. As an object upon which action is focused, this document is a given. Its presence causes the indexing process to swing into action.

The first step, called the *document analysis process*, occurs in response to the presence of the document. It consists of the act of examining the document (i.e., looking at the kinds of sources outlined in the DDC and ISO guidelines—the title, the table of contents, the abstract, if there is one, the back of the book index, reviews of the item, and so on.) in order to identify its subject.

The second element is the product of the first step. It consists of some mental sense of the subject of the document on the part of the indexer. It could be called the subject of the document as it exists initially in the mind of the indexer and includes a relatively unordered mass of mental impressions, phrases, terms, etc., which have been collected in the process. These ideas have been generated from the sources that form the basis of the examination process in the first step.

The second step is the indexer's response to the second element. Named *the subject description process*, it consists of the act of attempting to create a cohesive formulation of the subject of the document in language. In short, the product of the first step is a relatively unordered mass of mental impressions, phrases, terms, etc., which have been collected in the document examination process. The

product of the second step is the result of a concerted effort to give those various impressions, phrases, terms, etc., some sort of order and structure.

The third element is the product of the second step. It consists of the more or less cohesive formulation of the subject of the document in language. The second element consists primarily of a mental product, a sort of a running mental tab of the various candidate terms, ideas, concepts, and so on that one collected in examining a document. This third element represents an attempt to compress all of those various terms, ideas, concepts, and so on into something that in a relatively cohesive way summarizes the subject of the document.

The third step, which is now prompted by the presence of the third element, that is, of a relatively cohesive summary of the subject of the document in language, and which is named here *the subject analysis process*, consists of translating the product of the third element into a formal statement of the same thing, only this time in terms of the language of the appropriate subject access system. In short, it means converting one's language statement into, for example, subject headings or subject heading strings of terms if one is using the *Library of Congress Subject Heading* system, or into a class number, if one is using the *Dewey Decimal Classification* system. In this activity, one must of course be aware of all of the various rules, conventions, proscriptions, and so on that any system uses.

The fourth element, the terminus of the process, is simply the product of the third step. It consists of the completed subject indexing term or terms from a given system that the indexer has finally chosen to represent the subject of the document.

### 2.2.3. Miksa's Approach as the Source of the Approach Used Here

Miksa (1983) has introduced a geometrical diagram to represent the subject indexing process or “scope matching process” as he calls it. His diagram and subsequent discussion provide a basis for the four element-three step process described here. This diagram is reproduced in Figure 2-1 and is taken from his work *The Subject in the Dictionary Catalog from Cutter to the Present* (Miksa 1983, 6).

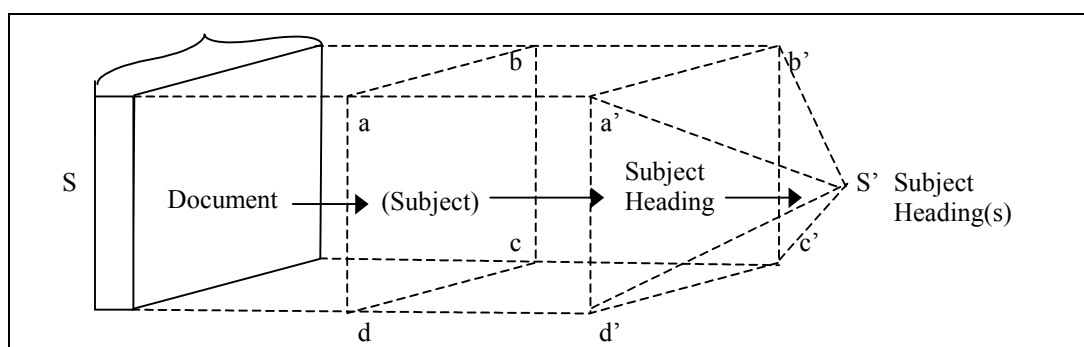


Fig. 2-1. The Scope-Matching Process

His approach and diagram explicitly portray four elements in the indexing process. Miksa (1983, 5) comments on the diagram in the following way,

Given the document S, its subject may be represented 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'.

He does not discuss either the individual elements or the individual steps that connect them. This is because his concern is not with details of the process, but only with the beginning and end of the process. He is concerned chiefly with explaining how modern subject heading practice conceived of the relationship of a subject heading to a document. He expressed this concern in terms of a basic question that asked what the “referent” of a subject heading is in modern practice. His diagram was simply his way of showing that ultimate relationship. In his attempt to show that relationship, he had to explain not simply how the subject cataloger moved from a document to a formal subject heading, but also to show something of the vagueness of the process. He explicitly denies that the process could ever be fully represented by a precise geometrical figure, but concludes

nonetheless that the figure is useful in showing that there is a "certain undefinable substantiality or scope to" the process (Miksa 1983, 7).

Miksa's diagram of what the referent of a subject heading is, provides a way of illustrating his argument. He argues that the literature of subject representation tends to answer his question simply by stating that "a subject heading should express or match in some essential way the topical content of a work" (Miksa, 1983, 5). In the light of his portrayal of the referent of a subject heading, Miksa finally concludes that the task of a subject cataloger is to move from the document to the subject heading 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" (Miksa, 1983, 7).

Miksa's approach to the subject representation process has three aspects of significance for the explanation of the subject indexing process. First, his use of the idea of "referent"--that a product of the indexing process will inevitably refer to something that preceded it--although limited in his discussion chiefly to the relationship between the beginning and the end of the process, provides a useful way of viewing the entire process. In short, given a series of objects upon which a

series of matching and intervening actions operate and in some cases which such actions produce, it is legitimate to portray each pair of objects within the entire process as that of referent to a later object. From this standpoint, each referent is related to the object that succeeds it. The entire procedure is then conceptualized as a series of referents and objects. It is this serial relationship that forms the basis of viewing the indexing process in the form of element – step – element – step – element – step - element, where the first element or object is the document itself, and the final element is the formalized subject term representation of the document.

Second, Miksa's inclusion of implied internal steps that take the subject cataloger from document to the final subject heading term, although not otherwise explained in his discussion, provides a basis for delineating the steps in the present explanation. In particular, his idea of a subject that more or less exists independently of the document itself is used here. And his use of a subject description as an attempt to formalize that sense of a subject somewhat correspond to the products respectively of the first and second steps in the present dissertation.

Third and finally, Miksa's insistence that the process deals with objects or elements that have only a "certain undefined substantiality" (Miksa 1983, 7) and are at best casual and cannot be defined precisely, suggests a need to identify more

explicitly the source of such indefiniteness. Here, it is concluded that the indefiniteness of the steps and their corresponding products arises from the fact that interpretation on the part of the indexer is involved. In short, it is concluded here that the steps constitute interpretations of the elements. In other words, the indexer interprets an element and through this interpretation she creates the following element. Therefore latter elements are only connected to preceding elements through an interpretation.

In sum, Miksa's approach to the subject indexing process is the basis for the view presented here, with the understanding that it has been greatly extended beyond his use. Specifically, his idea of a referent as a way to connect the first and last element in the subject indexing process is adopted as a way to speak of all of the internal relationships between each pair of elements. Again, his suggestion of internal steps and elements is important in helping to delineate the nature of the steps and elements in the process here. And finally his suggestion of the indefiniteness of the entire process has provided basis for looking for a cause of the indefiniteness, and this has ultimately been identified with a process of interpretation that makes up the steps in the process. To this basis, however, it should be noted that one other factor has been added. That is, the idea that referents are progressively more extensive the farther back one delves into the process and progressively more constricted the closer one gets to the end of the



subject indexing process. The reasoning behind this is based more or less on a common sense view of the objects that are found or produced at each stage in the process. When one begins, the document contains all of the words, paragraphs, chapters, and so on that the author used to express what amounts to the subject content of the document. In contrast, by the time a formal subject indexing term has been produced at the end of the process, the content of the representation is reduced to very few words in some carefully drawn syntax. The extent of each of these is obviously very different. It is surmised that the intervening steps show a continuous narrowing of the extent of the representation of the subject. The first step yields a cluster of terms, ideas, and so on, but the second step narrows even this down to a more compressed linguistic statement of the subject. Finally, even the compressed linguistic statement produced by step two is compressed by its conversion to the terms and syntax available in step three, or so this would appear in many if not most instances.

An alternative model of the indexing model, which incorporates the foregoing idea of progressive compression of the representation of the subject, is presented in figure 2-2.

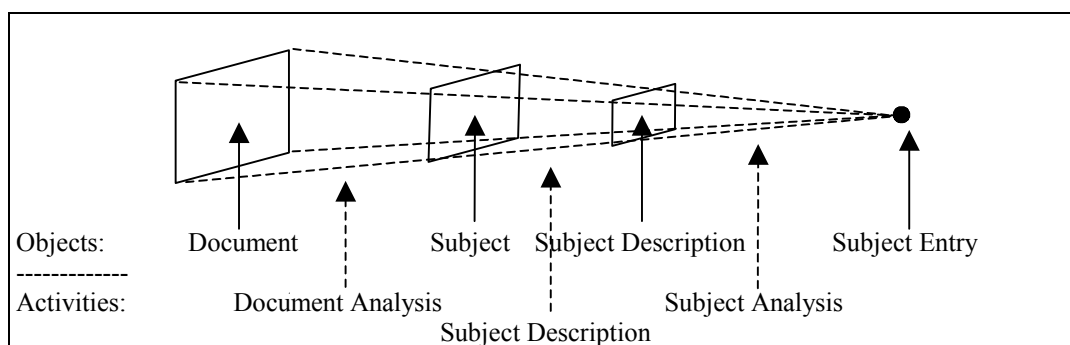


Fig. 2-2. The Subject Indexing Process

In Miksa's model, figure 2-1, the three squares are of identical size. In figure 2-2, however, the squares diminish 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 represented by a document itself are larger than that which is represented by a subject entry itself. A fuller explanation of how this occurs will be discussed in detail in the semiotic analysis of the subject indexing process in chapter 5.

One final matter should be noted before summarizing the argument as it has now proceeded. A close comparison of the four element-three step process as it has been portrayed here with the subject indexing process as found in the ISO guidelines suggests that there is some likeness between them. Both have three steps and four elements, the elements made explicit in the view provided here, but

only implied in the ISO guidelines. Each begins with a document, draws out of the document materials useful to a more formal description, and uses those materials to formulate a statement of the subject of the document which is afterwards converted to the terminology of a specific system. Where they differ is the nature of the product of the second step. Whereas the ISO guidelines argue that the second step produces a formalization of concepts, the framework presented here argues that the second step produces a compressed language statement of the subject. In this sense the first step produces unorganized material that can be used for a subject description, and the second step compresses and organizes this material (and obviously excludes some or even much of it) into a compressed statement.

#### **2.2.4. Summary**

In summary, this chapter so far has focused on describing the seriousness of the problem faced by indexers in attempting to represent the subjects of documents and has taken the first steps towards an alternative approach. Section 2.1 has shown that discussions of indexing not only offer little help in understanding the overall nature of the indexing process, but are especially silent on the initial step in that process, where the subject of a document is identified

and summarized. Instead, such discussions typically emphasize the end of the process, the point at which the conversion of an identified subject into the terminology of an indexing language is accomplished. Further, it has also been shown that even manuals of indexing procedure are relatively uninformative. In the cases of two important sets of guidelines examined here--the DDC and the ISO guidelines--various procedures for how to go about the subject analysis of documents are presented, but ultimately fail to explain how such procedures accomplish the task.

Section 2.2 has focused on the beginning of an alternative understanding of the same process. It began with the recognition that the indexing process can be analyzed in terms of its logical steps and showed that various commentators have viewed the process in terms of two, three, and four steps. Here, a three step approach has been adopted, but it has been enhanced by including the objects or elements that surround the steps and upon which the steps as actions operate. The source of this enhanced approach to the indexing process has been shown to be the work of Miksa, but his work, limited in its scope and purpose, has been expanded by focusing on all the elements and steps in the process. By focusing on the relationship between the steps as one of decreasing fullness, and by emphasizing the interpretive nature of the steps, a fuller picture of the process was given. In the end, therefore, this dissertation focuses not only on the first step but

on the coherence between the steps and elements in the process. It should be noted that this view of the indexing process does not claim that the steps that are portrayed here (document analysis, subject description and subject analysis) constitute the only valid way of subdividing the overall indexing process. There may well be other ways to portray what takes place. But for the purpose of developing a model of the subject indexing process it is assumed that the subject indexing process consists of these four elements and three steps. This assumption is based on previous descriptions of the subject indexing process in the indexing literature.

Only one thing remains to be discussed at this point, the relationship of the various elements and steps to the actual practice of indexing. Earlier it was suggested that novice catalogers may proceed in their work somewhat “by the number” in terms of steps in the indexing process, whereas experienced catalogers or indexers may compress all the steps into a single complex action. The difference between these two extremes appears not to be a function of the formal logic of the process, but rather of how people develop expertise in any given process. In order to view this reality in greater detail, a discussion of the stages of indexer experience will form the final part of this chapter.

### 2.3. STAGES OF INDEXER EXPERIENCE

An investigation of the steps taken in the subject indexing process will most likely reveal that novice indexers break down the indexing process into individual sub-processes more often than expert indexers.

Dreyfus and Dreyfus (1986) have identified five stages of development from being a novice at some task to becoming an expert. The idea is that when we acquire a certain skill we go through five stages of competence. For instance, when a child learns to ride a bike, the child first needs to be told exactly what to do. But as the child becomes a more skilled bicyclist she need not be aware of these instructions. When the child totally masters the skill she does not even think about what she is doing. At this stage it also becomes difficult for her to explain others how to bike. Dreyfus and Dreyfus's idea of the stages will be related to the skills of an indexer at 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, 21). The novice will only be able to apply the learned rules in contexts and situations with which the novice is familiar. The learned rules are therefore *context free rules*. The rules are context free in the sense that the novice has learned the rules objectively and will apply the rules to the particular situations with no regard for the situation

itself. The novice will apply the learned rules in situations that match the learned proto-situation and will not be able to accommodate for the given context of a new and different situation.

The novice indexer will typically have read manuals and guidelines about indexing and will apply these strictly to every situation. Such indexers will do exactly as instructed. This means attempting to determine the subject of a document by looking at the title, table of content, abstract, etc. If the indexer is taught to index using a particular indexing system, the indexer will follow the exact rules of that system. In other words the novice indexer will follow the indexing guidelines step by step and do exactly as prescribed.

True novice indexers are probably only found in introductory classes of indexing and classification in library schools where students learn how to use particular indexing systems. But as the indexer becomes more experienced, manuals and guidelines about indexing must be put aside. The novice indexer proceeds to the next stage.

Stage 2. Advanced beginner. As the novice becomes familiar with the skill, she advances to a marginally acceptable level of performance. This is done “through practical experience in concrete situations with meaningful elements, which neither the instructor nor the learner can define in terms of objectively recognizable context-free features” (Dreyfus and Dreyfus 1986, 22). The

advanced beginner begins to recognize and apply previously experienced situations to new situations. Hence the advanced beginner uses both context-free rules and “situational” elements in performing a task.

The advanced beginner will typically know the basic rules and techniques of indexing, will be familiar with a fair amount of the problems of indexing, and will have indexed a significant number of documents. The advanced beginner indexer is typically fairly new in a particular indexing service. She has therefore had some experience with real user needs and also has gained some knowledge about the particular domain in which she works. The indexer therefore starts to index from experience and not solely from manuals and guidelines. The advanced beginner will be able to engage in a somewhat detailed discussion about the indexing of particular documents in a particular system.

Stage 3. Competence. Whereas the novice and the advanced beginner merely followed a set of rules in coping with a situation, the competent performer sees a situation as a set of facts. 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 which are important to the given plan are given consideration. Since the competent performer acts according to a plan which she has chosen herself, she “feels responsible for, and thus emotionally involved in the product of choice” (Dreyfus and Dreyfus 1986, 26).



The novice and the advanced beginner, on the other hand, act according to specified elements or rules, and thus feel that an unfortunate outcome is the result of inadequately specified elements or rules.

The competent indexer has a conception of what to do in mind then looks for certain information in the document, and from there indexes the document. In other words, she now actively uses her knowledge of the needs of the users at her information service, she knows the indexing language very well, and she is able to represent documents accordingly. The indexer therefore approaches each document with a particular plan; namely to determine how the document is best represented to serve her community of users. For this she uses her intuition as much as her reasoning.

Stage 4. Proficiency. Unlike the previous performers, the proficient performer does not rely on rules in performing a task. At this stage she merely uses her intuition. This should not be confused with either irrational conformity or guessing. To guess is to reach a conclusion when one does not have sufficient knowledge or experience. Intuition here means the “sort of ability we all use all the time as we go about our everyday tasks” (Dreyfus and Dreyfus 1986, 29). The proficient performer is characterized by her ability almost immediately without conscious effort to follow a plan after encountering a problematic situation. Yet it

is unlikely that the proficient performer will be able to explain in detail how she solved the situation.

The proficient indexer will examine the document at hand without manuals and guidelines in mind. She will follow her intuition without being able to articulate the exact actions taken. Likewise, when the proficient indexer has decided on a subject entry, she will most likely not be able to explain exactly 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, 30). The expert does not see problems in some 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 are proceeding normally, experts don’t solve problems and don’t make decisions; they do what normally works” (Dreyfus and Dreyfus 1986, 30-31). When things work as they normally do, the expert is able to deliberate before acting. But this deliberation is not based on calculative problem solving. Rather, it is based on critically reflecting on her intuitions. Most importantly, as the performer becomes an expert her performance becomes fluid. The expert seldom needs to consider what she does, but when she needs to she is able to do so critically. And, like the proficient performer, the expert will not be able to

articulate her exact actions. The expert will likewise not be able to legitimize her decisions, but even so she will seldom make wrong decisions.

The expert indexer will typically have years of experience in the same environment, but this will not limit her performance. An expert indexer will perform the task as if it only consisted of one single operation and will not break down the indexing process into sub-processes. Therefore she will not be conscious of the task she performs and will not be able to explain in detail which parts of the document she pays attention to or what questions she asks herself to determine the subject. The expert will think about the indexing process as representing the subject matter of a particular document. She will consider the particular document as part of the collection and as a whole and a means for users' problem solving. Due to the experts' lack of consciousness about their actions, an investigation of what they do and how they index will reveal little.

### **2.3.1. Summary**

When a person learns a skill she develops her expertise through five stages. At the first stage she becomes familiar with a technique, but by the last the stage she masters the skill to perfection. Learning how to index most likely develops in the same way. An inexperienced indexer needs to learn the skill--that is how to

index--just as the inexperienced chess player needs to learn the basic rules of chess to be able to play. This means that what is what inexperienced indexers or chess players do is not necessarily the same as that which expert indexers or chess players do. Although someone at a particular stage will be able to imitate the performance of a higher stage, the person will perform poorly because she lacks practice and experience.

Dreyfus and Dreyfus' idea of the development from novice to expert is helpful in explaining the different stages indexers go through. It is also helpful to remember that the novice indexer's task is not to learn to mimic the performance of expert indexers, but to learn how to index. They will only become masters of that skill through practice, experience and awareness of the key problems related to the skill.

The exact actions taken by indexers with different levels of experience might not be identical, but the previously described steps and elements of the subject indexing process are considered basic in any indexing process. The present dissertation will elaborate on the actions taken by indexers in the light of those steps and elements.

## 2.4. AN INTERPRETIVE APPROACH

The present study is a philosophical and theoretical discussion of the subject indexing process. Philosophical discussions do not always explicitly discuss their theoretical or methodological underpinnings. In this case, however, it is necessary to discuss methodology because the method used here is quite different from typical approaches to methodology in the LIS field. Furthermore, the method used here needs clarification because given its direct dependence on interpretation, some might suggest that the method is invalid for an LIS study. The following will present what has been named the interpretive approach to LIS (Cornelius 1996).

This dissertation presents the subject indexing process from the standpoint of *an interpretive approach* to LIS. This section will present the philosophical justification for the present dissertation. It will become apparent that the chosen philosophical approach entails research methods and techniques that rely heavily on the arguments of the researcher.

This section will first discuss the philosophical underpinnings of quantitative and qualitative research models and argue that the present dissertation uses a qualitative approach to research. Thereafter it is argued that LIS for a large part has had a positivistic approach to research. It is argued that a more humanistic oriented approach is desirable. In such an approach, the core objects

of study are the humans at the expense of the artifacts used. It is suggested that Wittgenstein's (Wittgenstein 1958; Wittgenstein 1969) concepts of 'form of life' and 'world pictures' could be used as frameworks for an approach to theories on the representation of knowledge.

#### **2.4.1. Quantitative vs. Qualitative Research Models**

An interpretive approach is part of the qualitative research paradigm, and it is therefore be useful to begin by discussing methods and qualitative research in particular.

There are three distinct levels of analysis in the research process; a) philosophy and theory, b) methods and techniques, and c) data. Philosophical theory addresses the epistemological assumptions of the research and should also articulate the researcher's agenda and hidden assumptions about the research project. Methods and techniques are the procedures which the researcher will use to put her theories to test. In the end such tests will lead to new insights and new theories. Data is the direct outcome of the theories and the methods. Data is presented as the tangible results of the research process (Sutton 1998).

Fidel (Fidel 1993), in discussing the use of qualitative research methods in LIS, found that there is a growing interest in these methods in the LIS community.

But since there is no “universally agreed-upon definition” (Fidel 1993, 220) of qualitative research methods it is difficult to claim that a study uses qualitative research methods. The present dissertation operates in this borderland.

Qualitative research is sometimes defined as the opposite of quantitative research or simply as essentially different from quantitative research (Fidel 1993, 220). Bradley and Sutton (1993) define the philosophical assumptions in both quantitative research methods and qualitative research methods. They find that qualitative research method paradigm (Bradley & Sutton 1993, 406),

... has origins in hermeneutics (the view that observation is an interpretive process), relativism (the theory that truth is contingent on the observer and the time and place of the observation), and idealism (the view that reality is essentially a property of the mind rather than a phenomenon itself).

These basic philosophical traditions condition qualitative research methodology in that the methods and techniques used arise in the context of these assumptions.

This approach to research is sometimes referred to “as the naturalistic, constructivistic, or interpretive paradigm” (Bradley & Sutton 1993, 406). These philosophical underpinnings of qualitative research are seldom discussed, however. More often than not, qualitative research is defined by the actual

research practice use, typically by such techniques as “participant observation or unstructured interviewing” (Bradley & Sutton 1993, 406).

Quantitative research approaches, on the other hand, are often associated with some variation of positivism and are often linked to philosophical traditions, such as (Bradley & Sutton 1993, 406-407),

... realism (the view that objects of sense impressions exist independently of the observer), logical positivism (the view that sense impressions alone are the most valid source of knowledge and that the factuality of propositions is based on those impressions that can be verified), and empiricism (the claim that experience is the only process for gaining knowledge).

Associated with these traditions are the views that there is one tangible reality, that it is possible to separate the observer from what it is observed, and that this approach is a value-free methodology (Bradley & Sutton 1993, 407).

Fidel finds that “the qualitative approach offers the best methods for *exploring* human behavior” (Fidel 1993, 221). Qualitative research methods offer the best solutions for investigating a complex but little understood problems. This is the case because whereas quantitative research method emphasizes the general applicability of the results of the research, qualitative research is more concerned with the subject matter of the research at the expense of the method. Qualitative



researchers therefore generally have a more pluralistic approach towards methodology and find that the problem to be investigated and its specific conditions should determine the relevant method (Fidel 1993).

Through a large literature review Sutton (Sutton 1998) has shown how LIS research has focused on quantitative research methods in LIS at least since the 1950s. This focus on quantitative research methods was based on the idea of objectivity. To be acceptable for research the object of investigation had to be distinctive from the researcher. A number of researchers even criticized qualitative research methods and argued that these methods instead of focusing on the truth tend to focus on verbal expressions of the seemingly likely and plausible. It is simply concluded here that such arguments limit the scope of LIS research.

#### **2.4.2. Positivism in LIS**

A few scholars have argued against such positions, and insisted that the field needs to allow for a much broader range of methods. Some have even suggested abandoning positivistic research methods altogether. The common ground for these criticisms is that “that the modern conception of the library has been grounded in a positivistic commitment to neutrality and objectivity” (Sutton 1998, 268-269) and that the LIS field has to give up this commitment. The field should

recognize that new knowledge could be created under different circumstances than in the natural sciences. In other words, the field should recognize that the problem itself should determine the method used, and not vice versa. As Bradley and Sutton (1993, 408) have noted, “no research perspective is best for all questions.”

Harris (Harris 1986a; Harris 1986b) has investigated the historical roots of the commitment to positivism in LIS. He finds that it can be traced back to the Carnegie Cooperation’s establishment of the Graduate School of Library Science at the University of Chicago in 1928. Harris argues that the first faculty at the school defined and established what he at one point calls “a ‘psychosociological’ research program” (Harris 1986a, 517) and at another point calls “a positivistic epistemology” (Harris 1986b, 218). The school further succeeded in forming a conception in the LIS field over the next several generations that ‘good research’ is founded in such an approach (Harris 1986b, 218). A positivistic approach to LIS can be characterized as based on three assumptions (Harris 1986a, 518),

1. Library science is a genuine, albeit young, natural science. It follows then that the methodological procedures of natural science are applicable to library science; that quantitative measurements and numeration are intrinsic to the scientific method; that epistemological issues are best treated with respect to specific research questions; and that complex

phenomena can best be understood by reducing them to their essential elements and examining the ways in which they interact.

2. The library (broadly defined) must be viewed as a complex of facts governed by general laws. The discovery of these laws and theories is the principal objective of research.
3. The relation of these laws and theories to practice is essentially instrumental. That is, once the laws and theories are in place, we will be able to explain, predict, and control—i.e., produce a desired state of affairs by simply applying theoretical knowledge.
4. The library scientist can and should maintain a strict "value-neutrality" in his or her work.

Harris further argues that the belief in the possibility and desirability of a science of librarianship recalls the situation of the social sciences in the 1950s (Harris 1986a, 528). Most social science theorists have now generally dismantled this hope or belief. It is now more or less accepted that the social sciences never will become 'sciences of society', but they have to establish their own epistemological foundation. Harris (1986b, 220-221) summarizes the case against positivism in the social science in three points,

1. the social sciences have never been able to generate "scientific paradigms" that might govern research in ways analogous to the hard sciences;

2. the social sciences simply cannot sustain the essential division between the subject and the object of research so central to the positivist epistemology;
3. the value free pretensions of the social sciences have been proven to be mystifications designed to camouflage the extent to which the social scientist is governed by prejudgments and domain assumptions.

LIS must separate itself from a positivistic epistemology in the same way that it has been necessary in the social sciences, even though a positivistic or quasi-positivistic approach remains the most widespread approach to research and thinking in the LIS field. This is seen by the emphasis on objectivity, generalization of findings, empirical research, quantification, etc. Most importantly such an approach to the “field has severely limited the range of questions that can be investigated and has rigidly defined the characteristics of relevant answers” (Harris 1986b, 221-222).

The literature of and research on the subject indexing process has likewise been influenced by this dominating approach to research in LIS. This means that the majority of research has concentrated on finding the laws or rules that govern the subject indexing process. The hope seems to have been that if it were possible to actually find these laws or rules then they could be applied in practice and subsequently would ensure consistency in indexing. The problem is that it seems

unlikely that such laws or rules actually exist, and that even if they did exist their application would be very limited.

#### **2.4.3. A New Focus**

There has in recent years been a growing dissatisfaction with the dominating philosophy within information science, which more or less explicitly argues for an approach that establishes the field on an empirical scientific basis. This movement advocates an approach to the field that is based in the humanistic aspects of the field. A few scholars within library and information science have recently argued for such a reorientation of the approach to the field (Budd 1995; Dick 1995; Cornelius 1996; Hjørland 1997). A humanistic approach shifts the research focus towards the interpretive aspects of the collection-organization-retrieval-evaluation process, which typically is used to define the scope of the field.

Epistemological discussions have very often been centered on attempts to define the concept of information (Frohmann 1992) and this has resulted in an emphasis on artifacts such as indexes, documents and abstracts, at the expense of humans such as authors, indexers, abstractors, and users (Dick 1995). This is the most important change in focus for an interpretive approach. The object of study

for the field has to shift focus from the artifacts used in the collecting-organizing-retrieving-evaluating process to the people in the process. The implication is that the field should base its methodology on an approach where the human interactions and processes are stressed and less on empirical studies and other objective approaches. The field should be studied as a human study – a study wherein the pursuit of laws is a pedantic affection (Natoli 1982).

A humanistic approach to the field might be rejected, since it may tend toward epistemological beliefs which argue in favor of a subjective idealism or even solipsism. However, library and information science in general and representation of knowledge in particular can be placed in a humanistic tradition without becoming relativistic. A view that focuses on subjective understandings of the world might not at first seem applicable as a basis for organization of knowledge. But an interpretive approach such as the one advocated here is ultimately based on a hermeneutical-phenomenological tradition which focuses in turn on social context as the determining factor for meaning. The humanistic interpretive approach is, therefore, based in realism.

#### 2.4.4. The Core Object of Study

In a recent article, Budd (1995) calls for a clarification of what he calls the “ontology of the library.” The ontology of the library is the core of the library’s being, the reason for the library’s existence. Budd states that the reason and core of the library’s existence is “to collect, organize and provide access to information” (Budd 1995, 306). This is a very general description, which is nonetheless valuable in the sense that it clearly and simply states the basics of the field. But Budd's definition is weak in the sense that he seems to suggest that *information* is the central concept of the field. However, in reality the sentence “to collect, organize and provide access to information” contains two different notions, namely 1) the *process* of collecting, storing and retrieving and 2) the *object* of the process, namely information. The first notion in reality requires an agent, and the focus for research should be on the agent and not on the process itself. In short, it is the *people* in the process of collecting-storing-retrieving information who should be the focal point of research. The people as agent, not information, should be the object and, therefore, the focal point of research.

The LIS field must be based on a philosophy that takes into account the people who are involved in the process. The study of library and information science is the study of *humans* and their relations to libraries, information systems, etc.

Information must in this sense be regarded as a subjective construct, not as an objective object. A subjective construct is used here in the sense of personal knowledge as defined by Polanyi (Polanyi 1958), not as the more cognitive oriented way the term recently has been used in the LIS field by for instance Cole (Cole 1994; Cole 1997). The most important feature of this redefinition of the object of study for LIS is that it requires a change in research methods.

Benediktsson (1989, 226) has argued that,

There should be less emphasis on empirical methods, which is only possible if it is acknowledged that information or information needs are not measurable or prescriptive.

Wright (Wright 1976; Wright 1978) has argued that the object of study for LIS and the research cannot be separated from the researcher. He therefore suggests that LIS should be concerned with metaphysics rather than physics. In other words, LIS should concern itself with the humanistic aspects of the information collecting-organizing-retrieving-evaluating process before or instead of the technological aspects.

A field of study with a subjective construct as its object of study calls for other methods than those used in sciences with 'objective' objects of study.

Natoli (1982, 164) has argued that,



It is my contention that research in librarianship cannot be modeled to any great extent on an objective approach, not without producing conclusions far removed from the reality of librarianship as readily perceived by every librarian in the field.

Natoli argues that the LIS field must adopt methodologies which do not require an objective object of study. He further argues (Natoli 1982, 163),

The goal of research in a human study is to recreate the human condition of the object of study in the mind of the reader by utilizing the reader's natural propensity to both experience and understanding.

Natoli holds that the LIS field should be regarded as a *human science*, and that a human science must use methodologies where the *interpretation* of the object of study becomes clear for the reader.

LIS's focus of study is the communication process involved with the generation, transfer, organization, storage, retrieval, and use of information. In this view, new methodologies are needed, as pointed out by Benediktsson (1989, 205),

Therefore, quantitative statistical methods can be used only in those areas in which the human perception of a situation is not a factor. The presence and validity of human perception is a clear indication of phenomenological-hermeneutical methods.

Since LIS in general and representation of knowledge in particular revolve around human perception and human interpretation, a phenomenological-hermeneutical approach seems reasonable.

#### **2.4.5. Knowledge in a Wittgensteinian View**

Wittgenstein (Wittgenstein 1969) discussed the concept of knowledge in his work, *On Certainty*. His discussions and conclusions can be used as the basis for the LIS field in general and knowledge representation in particular. It should be noted that some of the discussions in *On Certainty* can also be found in his better known work, *Philosophical Investigations* (Wittgenstein 1958), and that claims such as “Wittgenstein defined...,” “according to Wittgenstein...,” etc. are stated with some reservations. A large part of Wittgenstein’s work is formed as discussions between himself and some imaginary person, and statements about various ideas tend to change over time.

Wittgenstein argued that to understand a statement is be part of praxis. The fact that humans constitute discourse communities, in which it is possible to communicate fairly well, rests on the premise that humans--through a good amount of conscious and unconscious training--learn to use words in the same manner. Wittgenstein compares the use of words with the use of tools. There are various uses of a hammer – but somehow we have agreed on the few correct ways a hammer could be used. The functions of words are as diverse as the functions of tools. The meaning of words is defined through *language games*. In language games the meaning of words (and language) are not defined by what they refer to, or signify, but by their use.

In Wittgenstein's terminology discourse communities are named 'forms of life.' It is within these that words are used and defined. To speak a language is part of an activity. What we do besides using language cannot be separated from our use of language; language does not refer to some meaning that is outside language itself. Judgments of right and wrong, true and false, and meaning in general are based on these forms of life. It is within these forms of life that meaning is determined--that it is judged whether a language game is understood correctly. The same could be said of representation of knowledge. Any judgment of the representation and use of knowledge rests on a form of life.

In *On Certainty*, much of the discussion revolves around what it is possible to *know* and how it is possible to claim to *know* something. Wittgenstein argued that the claim to *know something* only makes sense within a certain unquestioned context. This context is called a ‘world picture’. A world picture is not something that we can acquire, change, or replace as it pleases us. A world picture is fundamental to the way we understand the world, hence also to the way we organize and represent knowledge. A world picture is a system or structure of--explicit and implicit--assumptions about the world, which is not easily removed or changed. There are four points which define the concept of a world picture:

1. Our world picture is closely related to our praxis.
2. Our world picture rests neither on empirical knowledge nor on verifications of hypotheses.
3. Our world picture is not easily changed due to empirical information which contradicts our world picture.
4. A shift in world picture will be similar to a conversion—a fundamental change in the view of the world.

The upshot is that we regard our world picture as true and correct, but whether it correlates with reality is beyond discussion. Since our world picture is such a

fundamental part of our approach to the world, it is not possible to investigate and discuss, because the world picture itself would form the basis for such an investigation. Thus, a discussion of how to organize and represent knowledge is limited and shaped by our world picture. If we wanted to explore whether a particular representation and organization of knowledge were correct or true, we would not have any thing to hold the organization against.

Even though this sounds like relativism, that was not Wittgenstein's mission. He argues that we must have the openness and ability to understand and judge others' world pictures. If someone for instance claims that Canada is a member of the European Union, we know that the person is wrong. We would have knowledge which the person does not have and we would know this. Wittgenstein's way out of relativism is a call to common sense. Common sense is shaped in language games. He is back at the social context.

Therefore, as individual as the concept of world pictures may sound they are relatively stable. They are shaped through language games and as such are part of a social praxis, and consequently more or less alike for all human beings in a particular social praxis.

#### **2.4.6. Summary**

This chapter has described the subject indexing process in detail. It stressed that the process contains a number of interpretations. This section introduced a general view and theory for the LIS field and outlined a framework to study the interpretive nature of processes of the collecting-storing-organizing-retrieving process.

The general state of theory and method in the LIS field was first discussed. It was noted that there is an ongoing discussion in the field over the use of different research methods. It was argued that the discussion would benefit from a general shift from discussing particular methods to the philosophical underpinnings of research. This would focus on the epistemological commitments that the researcher takes.

It was argued that a foundation of the field should be based in a humanistic oriented approach, where the people in the collection-organization-retrieval-evaluation process are in focus. It was further argued that the social praxis is the determining factor for meaning, information, and knowledge. Lastly it was argued that Wittgenstein's concepts world pictures could be used as frameworks to study the field.

Wittgenstein's notion of form of life and world picture was introduced as a framework for defining the scope of the LIS field. In such an understanding the

focus is on the praxis and the meaning producing community in which documents are used.

The implication for research in knowledge representation is that the immense focus on being as *user-oriented* as possible changes meaning. User oriented indexing generally means ‘to be able to represent documents according to the users’ needs and request for information;’ in short, to represent knowledge according to the users’ world pictures. But, if Wittgenstein’s argumentation is followed, this would clearly be impossible and to attempt to do so would be absurd.

Representation--and organization--of knowledge takes place within a particular social practice. This praxis, therefore, is the determining factor for the meaning of the document and therefore also for the representation and organization of knowledge within that practice. If the principles and values for the organization and representation were made clear and explicit, the use of the representation and organization would be easier and hence user oriented. The users of a particular information service must understand the form of life and world picture which have shaped that service. In other words, users have the responsibility for obtaining the best search results, but the information service has the responsibility for the best organization and representation of knowledge.

## **2.5. SUMMARY AND CONCLUSION**

This chapter has laid the ground for the problem of this dissertation. It seems obvious that the subject indexing process has been poorly documented and explained. It was shown that guidelines for classification and indexing are insufficient for their purpose of explaining the subject indexing process in such detail that they give the indexer enough information to determine the subject matter of a particular document and index or classify it. The conclusion here is that the chief reason for this is that there has been no adequate theory to explain the process. As a first step in providing such a theory the subject indexing process was described as consisting of four elements (document, subject, subject description, subject entry) and three steps (document analysis, subject description, and subject analysis). A general analysis of these elements and steps indicated that they become less full as they progress and that they can be viewed as a series of interpretations. The chapter further used Dreyfus and Dreyfus to show how an indexer goes through a number of stages from novice indexer to expert indexer. This was done to begin to show how an indexer might make use of the new view of the subject indexing process that will be discussed in succeeding chapters. In the last section of the chapter the philosophical and methodological framework used in the dissertation was introduced and discussed.



The next chapter will take a step back and review literature on indexing before presenting semiotics and returning to the framework presented in this chapter. However, the literature will be reviewed from the standpoint that was presented in this chapter.

### **3. Representation**

Representation of knowledge or documents is used in this dissertation to denote the representation of the subject matter/subject content/topic/aboutness of documents in a broad sense. This process is only one part of representing documents for retrieval. A distinction is usually made between the descriptive part of document representation or cataloging on one hand and subject representation on the other. The first deals with the description of “the physical make-up of an item and identifies the responsibility for intellectual contents” (Wynar 1992, 18). The latter, subject representation, deals with “determining what subject concept or concepts are covered by the intellectual content of a work” (Wynar 1992, 19). Sometimes a distinction is also made between classification, subject cataloging and indexing. However, the fundamental intellectual processes involved in classification, subject cataloging and indexing, as discussed in the previous chapter, are the same. Lancaster (1998, 16) notes that the confusion over the use of different terms for the process arises from a failure

to distinguish between step 1, document analysis, and step 3, subject analysis. The first two steps (i.e. step 1: document analysis and step 2: subject description) are identical no matter what indexing language or system the indexer will use in the third step; it is not before the third step that a particular indexing language comes into the process. A terminological distinction based on the kind of indexing language or system used is “artificial, misleading and inconsistent” and “only serve to cause confusion” (Lancaster 1998, 16). The process described and discussed in the previous chapter is the same, no matter how the documents’ subject matter are represented and no matter whether only portions of the document are represented or the whole document is represented. Lancaster (1998, 17) stresses this point,

The process of deciding what some item is about and of giving it a label to represent this decision is conceptually the same whether the label assigned is drawn from a classification scheme, a thesaurus, or a list of subject headings, whether the item is a complete bibliographic entity or a portion of it, whether the label is subsequently filed alphabetically or in some other sequence (or, in fact, not filed at all), and whether the object of the exercise is to organize items on shelves or records in catalogs, printed indexes, or electronic databases.

Representation of knowledge or documents will in this dissertation be used in a broad sense to cover classification, subject cataloging and indexing.

Representation of knowledge or documents is further used synonymously with subject indexing.

This chapter will review literature on the problems of how to determine the subject matter of documents. The literature will be examined in detail due to the relative limited amount of research specifically on this topic. Various items in the literature are chosen either because the authors are concerned with the *entire* subject indexing process or because they are concerned with an aspect of the first step in the process. Besides this literature, writings on the concept of relevance will also be reviewed. However, due to the large corpus of literature on this central concept, only a relatively small but highly significant number of writings will be covered. The purpose of this relatively limited review of the literature on relevance is solely to investigate the degree to which researchers who have reported on the concept of relevance have related the problems in that area to the subject indexing process.

Before beginning, however, a terminological clarification must be made. The phrase *subject analysis* is used in at least two different ways in the LIS literature, 1) to denote analysis of the topical content of a document and 2) to denote the area of study concerned with the construction of indexing language and

classification systems. The latter has received most attention and has been reviewed extensively, by for example, Liston and Howder (Liston Jr. and Howder 1977), Travis and Fidel (Travis and Fidel 1982), Schwartz and Eisenmann (Schwartz and Eisenmann 1986), and Lancaster, Elliker and Connell (Lancaster, Elliker, and Connell 1989). In contrast, the concern here is with the first definition of the phrase, namely that of determining the subject contents of a document. In this tradition *subject analysis* is sometimes used to denote the last step of the process described in chapter 2 (Chan, Richmond, and Svenonius 1985) and sometimes also to denote the first step (Langridge 1989). In addition, Hjørland (Hjørland 1997) has noticed that a few other terms also have been used, for instance content analysis, conceptual analysis, information analysis, aboutness analysis, text analysis, etc. In order to be consistent in the use of terms and as already noted in chapter 2, the phrase “document analysis” will be used here for the first step in a three-step process and “subject analysis” for the last step in the same process.

### **3.1. DETERMINING THE SUBJECT MATTER**

The task of determining the subject matter of documents has traditionally been approached as a problem related to a specific kind of indexing language.

Therefore the teaching of indexing has traditionally been concerned with the last step in the subject indexing process, subject analysis, at the expense of the first step, document analysis. In Arlene Taylor's (1999) recent textbook on the organization of information she notices that, "many people who are currently working to apply index terms from a controlled vocabulary have only had instructions in assigning terminology from a particular list, and not instruction in the process of determining 'aboutness'" (Taylor 1999, 132). Other researchers who have taught and worked extensively with these issues and problems have reached similar conclusions. Bates (Bates 1986), for example, has observed 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, 360). Cooper (Cooper 1978) has noticed that there have been some investigations into the problems of the process and that these have led to a few insights. However, he concludes 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]" (Cooper 1978, 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” (Cooper 1978, 107).

Earlier in this dissertation it was stated that the problems of determining the subject matter of documents are only very little understood. The above citations support this supposition. However, one person has contributed significantly to the understanding of the problems and principles for determining the subject matter of documents. In a thirty-year-old essay Patrick Wilson discusses the concept of the subject in library and information science, and portrays different approaches to determining the subject matter of documents. Due to the importance of this essay, his discussions and conclusions of it are reviewed next.

### **3.1.1. Wilson’s Argument**

Patrick Wilson gave the most thorough discussion of the complexity of determining the subject matter in 1968. In his book, *Two Kinds of Power*, he devoted a full chapter, “Subjects and the Sense of Position,” to discussing the determination of the subject of documents. Patrick Wilson was educated in library science (MLS 1953) and philosophy (BA 1949 and Ph.D. 1960) and served on the faculty of the University of California at Berkeley, School of Library and

Information Studies throughout his career. Because of his background in philosophy he used a philosophical framework when discussing the problems related to determining the subject matter of documents.

Wilson starts his discussion by asking whether there is something to the notion of *the* subject of a document.<sup>6</sup> He notices that most systems for organizing documents require that documents be filed and stored in a single location. If each document is stored and filed in a single location, then each document must have a single subject, under which it can be filed and stored. Hence, a document must have a subject.

If we ask a person who is writing a book or paper what she is writing about, she will normally be able to tell us. Likewise if we ask a person reading a book or paper what the piece she is reading is about, she will normally be able to tell us. In these cases the question “what are you writing about?” can be replaced with the question “what is the subject on which you are writing?” The assumption is that it is possible to say what a document is about and that this is the same as saying what the subject of the document is. The implication, Wilson argues, is that “there will be just one perfectly precise description of what [a document] is about” (Wilson 1968, 71). There can of course be various formulations of the description but these will be synonymous or nearly so. This assumption further entails that if there are two “non-synonymous and non-equivalent” (Wilson 1968, 72)



descriptions of what a document is about, then it is really two documents. And, if there is not one single precise description of what the document is about then it is not about one thing.

Wilson argues that to understand what the document as a whole is about the reader needs to understand what the sentences which make up the document are about. To ask what a document is about would therefore equal asking what the individual sentences in the document are about. One way of determining the subject of a document is therefore to ask someone to list all the things each sentence in a document is about, that is everything named, mentioned, or referred to in the document. A problem would appear, of course, if two people produced two different lists for the same document. Would it be possible to say that one list was better or more correct than the other was? Wilson (1968, 75-76) argues,

We might say with confidence that omission of mention of an item corresponding to a grammatical subject was an error, that inconsistency in the treatment of analogous items was a mistake, that failure to write down words in the proper grammatical form to fit into the blank space in the frame sentence "The sentence is about \_\_\_\_\_" was an error, and that certain items on the list were definitely not anywhere named, mentioned, or referred to in the passage in question . . . But, other than by pointing out mistakes of this last sort, could we have any reason for saying that one list was *too long*?

---

<sup>6</sup> Wilson uses "writing" to denote what here is called a "document."

The point is that, besides obvious errors, it is extremely difficult to say that one list is more correct than another is. One could imagine a list containing everything the sentences in a document could possibly be said to mention, name, or refer to. However, it is unlikely that anyone will be able to say that the list exhausts all the possible statements of what the document is about.

Wilson refers in a footnote to philosophy of language literature that have discuss the problem of what a statement can be said to be about. One of the philosophers Wilson mentions is Goodman (Goodman 1961) who has pointed out just how difficult it is to determine exactly what a sentence is about. He defined the difference between absolute aboutness and relative aboutness by discussing the problem of defining a “general rule for deciding whether a given statement is or is not about a given thing” (Goodman 1961, 1). Goodman (1961, 2) gives as an example the sentence:

“Maine has many lakes”

This sentence is obviously about Maine and about lakes. Likewise the sentence:

“Portland has many lakes”

could be taken to be about Maine, since Portland is in Maine. So therefore, since Maine is a part of New England, the sentence:

“New England has many lakes”

is also about Maine.

We now have two general rules for determining what a sentence is about:

1. A sentence is about what it directly denotes. For instance the sentence “Maine has many lakes” is taken to be about Maine, since Maine is mentioned in the sentence.
2. A sentence is about anything that is contained in (whether as part of, member of, or member of member of, etc.) what the sentence denotes, names, or mentions. For instance the sentence “New England has many lakes” is taken to be about Maine, since Maine is a part of New England.

If we follow the second rule, then the sentence:

“Florida is democratic”

is about the United States, since Florida is part of the United States. But according to the second rule it is also about Maine, since Maine is part of the United States.

This means that ultimately every sentence can more or less be defined to be about even remotely implied things. Goodman defines two concepts of about, namely “*absolutely* about” (Goodman 1961, 3) and “*relatively* about” (Goodman 1961, 13) to be able to speak more precisely of what a sentence is about.

Absolutely aboutness limits what a sentence is about to what it directly denotes, names, or mentions. A sentence that mentions a class of things does not—in an absolutely about definition—imply that it mentions every member of that class. For instance the sentence:

“New England has many lakes”

is therefore *not* about Maine, and conversely a sentence that mentions Maine is not about New England.

Furthermore, when adhering to the “absolutely about” definition of aboutness, shifts in psychological emphasis do not influence what a statement is about. Goodman (1961, 7) gives as an example the sentence:

“Crows are black”

This sentence is absolutely about crows and things that are black. Ambiguous sentences, like this, are only about what they directly denote, name, or mention.

As another example, Goodman (1961, 8) gives the sentence:

“Paris is growing”

which is absolutely about Paris, “but we cannot tell whether it is about the capital of France or a certain town in Maine” (Goodman 1961, 8). This will depend on the context of the sentence.

*Relative* aboutness, on the other hand, is the aboutness of sentences that are not absolutely about, for instance, Maine, but nevertheless in some sense have something to do with Maine. For instance the sentence:

“Portland has many lakes”

is not *absolutely* about Maine since the sentence does not mention Maine, but the sentence is about Maine *relative* to the sentence:

“Portland is part of Maine”

Since Portland is a part of Maine, the sentence “Portland has many lakes” is relatively about Maine.

By referring to the discussion in the philosophy of language literature Wilson argues that it is impossible to say exactly what a sentence is about. The common answer might lie somewhere between absolutely about and relatively about, but it is impossible to define.

Furthermore, Wilson argues, it is impossible to say how one goes from knowledge of what the individual sentences are about to knowledge of what the document as a whole is about. That is, even if it was possible to produce a long list of what the sentences in a document named, mentioned, or referred to, how could one then convert this list into a single statement of what the document as a whole is about?

Wilson finds that if the persons were asked not only to list the things named, mentioned, or referred to, but also to list “concepts employed in addition to those employed in the naming, mentioning and referring to things” (Wilson 1968, 78), this would produce more similar lists. Such lists would of course also be much longer than the first kind of lists.

### **3.1.2. Wilson’s Methods**

However, the problem of how to determine the subject matter of the document remains. How does a person go from understanding what the parts (sentences) of a document are about to understanding what the document as a whole is about? Wilson defines four ways a person could determine what a document as a whole is about. These could be thought of as the four basic methods for determining the subject matter of documents. Wilson describes the four methods on the basis of methods and ways of determining the subject, from the library and information science literature. Although Wilson’s description of these methods are thirty years old they still remain the basic methods for determining the subject matter of documents.

The first method or way to determine the subject is named the *purposive way*. This method is oriented towards the author’s intentions with the document.

The idea is that the author's intention with the document is the subject matter of the document. The indexer's task is to find out what "the writer is trying to describe, report, narrate, prove, show, question, explain" (Wilson 1968, 78). The indexer should look for clues in the document, such as passages where the author explicitly writes what the purpose of the document is, such as "I will show that....," "It shall be proved that....," etc. (Wilson 1968, 78).

The purposive method is based on the assumption that when a person writes a document she is attempting to do a certain thing, to describe, report, narrate, prove, show, question, or explain something, and that the author is able to state her aims. Of course an author's purpose or aim with a document is seldom a single thing; authors often aim at several things at the same time. Further, there may be things done only for secondary purposes. For example, suppose an author covers subject X to lay the groundwork for arguments on subject Y. In this context X is a secondary subject and Y the primary subject, and we could argue that Y is the subject of the document. However, doing this "requires an ability to see which of the things done or attempted in the writing are done only because necessary as means to an end, and which are done 'for their own sake'" (Wilson 1968, 80). This is even more difficult in cases where the author does not state her intentions. In such cases the indexer would have to "speculate and frequently it

can be no more than speculation” (Wilson 1968, 80-81) which of the aims are primary and which are only secondary.

One of the problems with the purposive method, Wilson argues, is that it is difficult to discover a person’s initial intentions from the result of the person’s activity. It is difficult to determine an author’s intentions with a document from reading the document. The intentions may well have changed during the course of writing the document, so that what was primary at the beginning of the writing has become secondary at the end of the text. The author may not even aim at anything definite; the aim may be indefinite. Wilson concludes that the method on the surface may seem objective and neutral but the method depends on the indexer’s interpretation of the author’s intentions (Wilson 1968, 81),

Serious attempts to discover the subjects of writings by appeal to purposes and aims cannot rest on simple authorial declaration, for those are as open to question as are any other actor’s explanations of his own behavior. If the subject of a writing can be determined only by discovery of the writer’s purposes and aims, we must as often be in doubt and disagreement about the subject of a writing as we are in doubt and disagreement about the explanation of writer’s behavior, which is, if we are serious and even moderately skeptical, very often indeed.



The second method is named the *figure-ground* way. It is based on the assumption that some aspects or things treated in the document stand out as more dominant than other parts of the document. In the list of things named, mentioned or referred to in the sentences of a document a few things stand out as more central than others, and of these the most central is the subject of the document. The idea here is that among the many things on the list it will always be possible to point out those that are most dominant in the document. This is of course not a neutral and objective judgment, as Wilson (1968, 82) argues,

[W]hat seems to us to stand out depends on us as well as on the writing, on what we are ready to notice, what catches our interest, what absorbs our attention. The writer has some control over our sense of what dominates what, and the better the writer, the more control he has; but he does not have total control.

In other words, the dominant aspect, the subject, or thing treated in a document varies from person to person and we “cannot expect that everyone’s attention will be dominated by the same things” (Wilson 1968, 83).

The third method is an attempt to establish an objective way to determine the subject of a document. Wilson does not name this method, besides describing it as an “objective” way. It could, however, more precisely be named the

*constantly-referred-to method*. The idea here is that where the second method, the figure ground way, was based on an impression of dominance, the constantly-referred-to method is its objective correlate. The constantly-referred-to method is based on the assumption that if something is often referred to in a document, if it appears frequently on the list of things named, mentioned or referred to in the document, then this is the document's subject.

The obvious problem with this assumption is that things with high frequency might only be background information for the more predominant subject of the writing. Therefore the frequency of particular words says little about the subject of the document, as Wilson (1968, 83) argues,

Mere numerical predominance of references cannot uniquely determine the subject; this obvious truth is reinforced by the reflection that one can always rewrite a text in such a way as to reduce the number of references to any item and increase the number of references to any other without materially altering the general sense of the writing or even, if one were skillful enough, changing the balance of impressions of dominance and subordination.

However, the method could be made more complex and yield better results if it was modified slightly. Instead of basing the method on mere frequency of words, the things named, mentioned, or referred to could be grouped together.

This way, also indirect references would be counted and grouped together with direct references. Of course, that would be to give up the objective aspect of the method, as Wilson (1968, 85) argues,

The method of identifying subjects by counting references is a hopeless one unless we allow ourselves to count indirect references, to group items referred to and consider a reference to a member of the group to be an indirect reference to the group. But if we do allow ourselves to do this, it is quickly apparent that no unique results can be expected; for there are many possible ways of grouping any set of items referred to.

This way of determining the overall subject of a document rests on our ability to discover or invent concepts that cover the groups of items referred to. One difficulty with this approach is that the groups we discover or invent will seem natural to us because they occur in familiar environments. However, others in still other environments may discover or invent different groups and hence determine a different subject.

The fourth, and last, method could be named the *appeal to unity* method. Wilson does not specifically name this method but speaks of it as “appeal to unity, or to rules of selection and rejection” (Wilson 1968, 86). The method determines the subject by that which brings together the document as a whole, as opposed to

the previous two methods which were based on the assumption that the subject is that which is most dominant among the many subjects, either by being most central or most frequent.

The idea of the appeal to unity method is that in the creation of a document an author has made some decisions as to what should be included and what should be rejected in the argument. The author mentions a number of things in her argument but the core subject of the document is what makes the document hang together. However, the unifying element may not necessarily be that which is most central, stands out, or constantly referred to. Again, the determination of the subject relies much more on the indexer's opinion than on the author or the text, as Wilson (1968, 88) argues,

[I]t is too evident that this effort may result in a piece of artistry on our part rather than on the part of the writer; discovering how the writing hangs together may be discovery of one out of several possible ways in which we can make it seem reasonable unified, coherent and complete, ways whose success or failure could be judged only on predominantly aesthetic grounds.

### **3.1.3. The Subject is Indeterminate**

Of the four methods none could be argued to be inappropriate for the task of determining the subject matter of a document. As Wilson argues, "each seems to

have a reasonable claim to being *a* way of picking out the subject of a writing” (Wilson 1968, 88). However, using the different methods would not necessarily lead to the same result. Even an indexer who uses the different methods on the same document might possibly arrive at different results; and if several indexers used all the methods, they would likely reach a notable number of different results. Except for the most obviously incorrect and inappropriate descriptions, it would be impossible to decide which among all the results would be the best description of the subject of the document. Wilson (1968, 89) therefore argues,

The notion of the subject of a writing is indeterminate, in the following respect: there may be cases in which it is impossible in principle to decide which of two different and equally precise descriptions is a description of the subject of a writing, or if the writing has two subjects rather than one.

Simply devising multiple subject entries for each document cannot solve the problem, for the question remains of how many such entries to make. The degree of generality or exactness that subject entries should have also does not solve the problem. For example, attempting to keep subject descriptions at a vague or broad level would inevitably lead to multiple appropriate vague descriptions of the document. And attempting to keep subject descriptions at an exact or specific level would likewise lead to multiple specific and exact descriptions of the

document. In either case there would be no way to tell when one had been complete. The idea that a document has just one precise and accurate description must be given up because, frankly, a document does not have *a* subject. As Wilson (1968, 90) argues,

[W]e cannot expect to find one absolutely precise description of one thing which is *the* description of *the* subject, all other things being mere approximations to that one description, or being descriptions of what is not the subject. The uniqueness implied in our constant talk of *the* subject is non-existent.

This also means that under a given subject entry in the catalog we cannot be sure to find all that a “library has on a given subject” (to use Cutter’s [1904] terms), because the methods used to determine the subject and the rules of assignment prescribe nothing definite. Therefore no confident prescription could be made that a certain subject entry has gathered all in a library on a given subject.

Wilson concludes his essay by stating that his discussion and statements probably will not bring anyone who has a substantial amount of experience in classification and indexing to doubt the quality of his or her work. He believes that most practitioners do not see the kind of problems and difficulty which he has just described and discussed. He argues that the reason for this is that the

indexing languages most practitioners use “are mostly too coarse to allow or require the making of fine distinctions. They find a location which satisfies them, and count this as a success” (Wilson 1968, 91). The real problem is of course that the users of such services will not understand the product and the circumstances under which it is produced. The main lesson learned from Wilson is that nothing definite can be expected of documents found under a certain subject entry.

#### **3.1.4. Critique of Wilson**

Wilson’s essay centers around the discussion of whether subject indexing is an art or a science. In an article on indexing in the *Encyclopedia of Library and Information Science*, Rothman (1974, 287) states that,

Indexing is neither strictly an art nor a science, but mixes characteristics of each. It is an art in the sense that it requires sensitivity, intuition, and taste. It is a science because it requires the creation and recognition of patterns and rules, conscientious adherence to them, and painstaking accuracy and precision. It is not purely an art because it discourages and often even penalizes originality, individualism, and departures from the norm. It is not a science because it is pragmatic and empirical; it neither develops nor obeys universally applicable laws, and does not yield universally applicable results.

The features that make indexing a science are, according to Rothman, the creation of rules, the adherence to these rules, and the attempt to be objective and neutral. However, according to Wilson's discussion it is not possible to create rules or adhere to rules which make indexing objective and neutral. The representation of a document's subject will always depend on the methods an indexer chooses to use, and, perhaps more importantly, on the indexer herself. In this sense, indexing is an art which cannot be done objectively and neutrally. Furthermore, if indexing cannot be done objectively and neutrally it cannot be taught as a scientific task.

In the four methods Wilson chooses to describe and criticize as ways to determine the subject of a document, he examines two assumptions: 1) that it is possible to find *the* subject of a document, and 2) that the subject is something which can be found in the document itself. However, he only rejects the first assumption, that it is not possible to determine *the* subject of a document but only *a* subject. With respect to the second assumption that the subject of a document is to be found in the document itself, he seems not to have considered that the subject of a document might well be established by looking elsewhere than in the document itself. He hints at it a few times by saying that the determination of the subject matter rests more on the indexer than on the author or the document, but he does not discuss why different indexers would interpret documents differently. However, he does broach that issue somewhat in the chapter following "Subjects



and the Sense of Position.” Here he discusses whether the indexer should use internal or external criteria in judging the relative importance of an item (Wilson 1968, 97-98).

Wilson seems to follow a tradition within textual analysis which typically has stipulated that there are two ideas of interpretation. Eco (1994, 50) explains that the meaning of a text could be determined by looking at,

(a) what its author intended to say or (b) what the text says independently of the intentions of its author.

In the first of the four ways or methods Wilson discusses, the purposive way, the subject matter of the document is found by determining the author’s intention with the document. In the three other methods he investigates ways of determining the subject independently of the author’s intentions. Eco (1994, 51) argues that to interpret a text independently of the author’s intention could be done in two different ways--by finding out,

(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.

In the three methods that Wilson describes by which the subject matter of a document could be established independently of the author's intention, he focuses solely on the document or text as a closed system. This approach suggests that the subject can be determined by finding in the document the meaning or subject matter of the document. In none of his four methods is the analysis of the document's subject matter based on the readers' or indexers' expectations and approaches to the document. Again, he mentions that different indexers might index the same document differently, but he does not follow this path and does not discuss it in detail.

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

Since Wilson assumes that the subject of document should be determined by looking at the document itself he leaves out one major method of indexing which has received some attention lately, namely what could be named the *user-oriented method*. This method is essentially a variant of the purposive method. In the

purposive method the indexer attempts to identify the author's intentions with the document. In the user-oriented method the indexer attempts to identify the users' potential information needs and indexes the document accordingly. In such an approach to indexing it is not the document itself that determines its subject matter but the users' needs.

Wilson (1968, 89) gives one example where it will difficult to decide what the subject of a document is,

A writing might be such that there were equally good grounds for saying it had as its subject the commercial adventures of some imperialist power, with some supporting detail on military adventures, and for saying that the subject was the military adventures, with supporting detail on commercial adventures, or for saying that each of these was an independent and equal subject of the writing, that it had not one but two subjects.

The task is therefore not to determine *the* subject or *a* subject of the document.

The task is to represent the subject matter of the document according to the users' needs and tasks. The users' view of the document's subject could be used as the determining factor for determining the subject of the document. Depending on the users' work domain and needs, the document in Wilson's example, could for a given user group be said to be about either commercial or military adventures.

The central issue in Wilson's essay and the brief discussion above of his essay centers on whether a subject of a document is something that a document *has* or something a document is *given*. In the first case, it would be possible to determine the subject matter by investigating the document itself. In the latter case the subject matter is determined by something outside the document itself.

The chapter on semiotics introduces a theory of interpretation which is based on the assumption that texts and documents are open for multiple interpretations. This theory is then used to analyze and explain the subject indexing process and thereby supplement Wilson's argument by giving a general theory of indexing based on the assumption that the subject of a document is indeterminate.

The remainder of this chapter reviews other literature on indexing and also literature on the concept of relevance. The review of indexing literature will argue that most research on indexing has been concerned with finding the rules of indexing or determining what indexers do when they index.

### **3.2. STUDIES ON INDEXING**

If Wilson's arguments and the above discussion are sound, there will be no universal rules of indexing and therefore no rules to discover. Furthermore, if

there were such rules and they were discovered, they would be of little use because each indexer would apply them differently. However, much research into indexing has been concerned with finding such rules, and it has been conducted with the belief that because these rules are mental (although perhaps unconscious) they can be uncovered by studying the mind.

### **3.2.1. Perceptual and Conceptual Indexing**

Following the tradition of cognitive psychology and cognitive science, Farrow (1994) in his article on indexing in the Encyclopedia of Library and Information Science offers a cognitive model of the indexing process. Farrow's goal is to produce a model of the human information processing which takes place during indexing. For this purpose Farrow (Farrow 1991; Farrow 1994; Farrow 1995) examined the subject indexing process on the basis of speed reading studies.

Farrow argues that the indexing process could be analyzed as a top-down process or as a bottom-up process. The top-down process uses information for indexing which is not contained in the document, but that is part of the indexer's world knowledge. The bottom-up approach only uses information that is

contained in the document for indexing. He names the top down approach *conceptual* indexing, and the bottom-up approach *perceptual* indexing.

Perceptual indexing takes the form of scanning the text for cues. The cues might be long words, words that are italicized or underlined, words in headings, or words otherwise emphasized. The focus of 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 picks out subject information from the text.

Conceptual indexing, on the other hand, relies on the knowledge of the indexer. Farrow argues that it is well known that both the indexer’s knowledge of the subject matter and the structure of the text itself influence the quality of indexing. He also adds three more factors, which affect the process: knowledge about the system the indexers use, the users of the systems, and the general world knowledge which the indexer posits.

Farrow (1991, 163) concludes his discussion by providing a model for indexing, abstracting and classification. The model is based on a comprehension model and is simple and highly structured. Farrow states that the model might be weak but emphasizes its principles,

Even though the model is in many ways sketchy and incomplete, some clear principles emerge: the importance of conceptual processing and the consequent need for indexers to be familiar with the subject matter of the text they are working with; the limited comprehension obtainable by scanning and the types of perceptual cues in scanning text and their use; and the nature and importance of expertise in the performance of a professional task.

It should be noted that Farrow fails to mention the importance of the three additional factors which he delineated regarding the knowledge on which the indexing process relies, namely system knowledge, user knowledge and world-knowledge. And without explaining the omission, he argues that his model for understanding the subject indexing process pinpoints the “causes of inadequate or inaccurate indexing” (Farrow 1991, 163).

The value of Farrow’s discussion is the categorization of indexing into a perceptual and a conceptual process. This categorization of indexing underlines the principal distinction in indexing between an approach that believes that the subject matter could be determined solely by looking at the document, perceptual indexing, and another approach where additional knowledge is needed, conceptual indexing.

However, Farrow’s cognitive model of indexing adds no further knowledge or instructions to the process. He simply says that indexing is a mental process,

which can be explained by using models of human information processing from cognitive psychology. But these arbitrary models of minds, memory, and cognition explain little about the indexing process. In his discussion of conceptual indexing he states that indexing in this approach depends on a number of factors, but he does not discuss the influence of these factors.

Farrow bases his studies on the work of the linguist Teun Van Dijk and the psychologist Walter Kintsch. Beghtol (1986) also uses van Dijk and Kintsch's work. She too distinguishes between a top-down approach and a bottom-up approach to indexing, and she too sets out to describe a cognitive model of indexing.

Beghtol starts her investigation from the assumption that documents have a relatively permanent aboutness and a variable number of meanings. In other words, although documents may mean different things to different people, they will at the same time give relatively the same information to the same people for the document's aboutness. Beghtol therefore sets out to create a model for "aboutness analysis."

Beghtol defines the top-down and the bottom-up approaches similarly to Farrow, but she argues that an indexer uses both approaches. Again, she does not dwell too much on explaining what kind of knowledge top-down--conceptual--indexing depends on. Her focus is, like Farrow's, to explain the cognitive



processes. Her analysis is based on the premise that language is a system of codes added on to an already existing system of knowledge and ideas. Beghtol (1986, 87) explains,

Like transformational and generative sentence grammarians, text linguists accept the Chomskyan distinction between the deep and the surface structures of language. The deep elements of language are thought to be non-linguistic conceptualisations and cognitive integrations that are universal to human thinking but that are mapped onto the surface structure of a particular language in varied non-universal ways using an apparently unlimited array of linguistic devices. In addition, a single language may map the same deep logical concept onto the available surface verbal elements in more than one way. For example, 'Mary threw the ball' and 'The ball was thrown by Mary' are generally considered to contain the same deep, but very different surface, structures.

Beghtol's argument is that the task of the indexer is to:

1. transform the document's surface structure into its deep structure,
2. transform the indexing language's surface structure into its deep structure,
3. join together the two deep structures,

4. transform the resulting deep structure back into the surface structure of the indexing language.

She finds that Chomsky and Van Dijk have explained the first two steps whereas the third and fourth steps remain to be explained and understood.

As with Farrow's discussion, is it difficult to see how one gets from the models of the mental processing to the indexing of documents. Beghtol states that some of her theory still remains to be investigated but she suggests a test<sup>7</sup> that could be used to verify her hypothesis. But more importantly, both Farrow and Beghtol base their indexing theory on a rationalistic theory of language in which it is assumed that language is something which is added onto an already-there-world of ideas and knowledge. In such an approach to understanding language it is assumed that words only point out the ideas they refer to. Both Farrow and Beghtol therefore argue that their theories of indexing give access to *the* subject of the document.

The definition of language on which Farrow and Beghtol base their studies has been criticized by a number of people, including Wittgenstein (1969; 1958),

---

<sup>7</sup> Beghtol suggests that two groups of indexers index the same set of documents. Both groups are told to first describe the subject matter of the documents in their own words and then translate their descriptions into the vocabulary of a classification system, such as DDC. However, the first group is told at the beginning that they will use a particular classification system whereas the second group is not told so before they reach that point in the test.

Beghtol predicts that the two groups will differ in their descriptions of the documents, because the first group will be influenced by the structure of the classification system.

Heidegger (Krell 1978) and Gadamer. These critics do not separate the meaning of words from the people or the community in which the words are used. They argue that language is not a tool for pointing at the world, but “the very constitution of the world” (Introna 1998, 5). In this sense words and their meanings cannot be separated; “there is no meaning and word; the word *is* the meaning” (Introna 1998, 5).

When meaning and words cannot be separated into two different kinds of phenomena, the meaning of words cannot be defined by whatever the words refer to. The meaning of words is the very use of them. Language is therefore not a tool used to speak with, but the very social and cultural context in which the language is situated. In other words “I do not speak with language, as a tool, but *from* language” (Introna 1998, 8). The community we belong to has a language. Language is not something which is added on to the praxis. The practice is the language.

Therefore the meaning of words and the correct use of language cannot be studied separately from the community in which the words and language are used. Even though words come from an individual and are perceived by an individual, language is not the product of these individual persons. Language belongs to the community in which it is used. It is the community that defines and determines the meaning of the words used. Words therefore do not have an objective and true

meaning. Neither is the meaning of words fluid and individual. Introna (1998, 8-9) gives the example that one needs to start with the community's "already there language," even if one wants to disagree with the community,

I can not 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.

Wittgenstein (1958) defined these discourse communities as 'forms of life' which form the shared understanding of practice and reality. The meaning and correct use of words and discourses within these 'forms of life' are determined and established through 'language games.'

In this view of language and meaning there is nothing before language. Language is not used to describe some objective ideas or meanings. Words get their meaning through their use. In such an approach to language and meaning it is argued that documents do not have a permanent subject and various meanings, as Beghtol argues. Subjects are assigned to documents according to the users' domain, tasks, and needs.

### **3.2.2. Conceptions of Indexing**

Albrechtsen (1992; 1993) has proposed a framework for discussing subject analysis and indexing in which she incorporates the domain in which documents are used and lets the domain determine the subject matter of the documents.

Albrechtsen argues that there are three general conceptions or viewpoints of subject analysis and indexing. She argues that these conceptions relate to the type of information which constitutes the subject and to which method of indexing is used. The relations among these factors can be expressed in the table found in figure 3-1 that combines her 1992 and 1993 tables (Albrechtsen 1992, 141; 1993, 220)<sup>8</sup>.

---

<sup>8</sup> In a personal communication Hanne Albrechtsen explains that the content-oriented conception uses some explicit information and that the requirement-oriented conception uses some implicit information, which is the reason why the boxes in the figure do not align.

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. 3-1. Conceptions of Subject Analysis and Indexing

Her analysis incorporates three conceptions of subject analysis and indexing, the simplistic conception, the content-oriented conception and the requirement-oriented conception.

1. The simplistic conception views subjects as objective entities. This means that the subject information regarded as explicitly present is the text and that the indexing process simply needs to extract it. The indexing could therefore be executed automatically by means of statistical indexing methods.
2. The content-oriented conception is more complex since it involves an interpretation of the document to determine the subject. The subject is

therefore regarded as something that goes beyond the lexical and grammatical surface structure which the simple conception is based on. Within a content-oriented conception, analysis of documents involves an identification of subjects that are implicitly present in the document, and which can only be perceived by a human indexer with a high level of knowledge of the subject matter.

3. The requirement-oriented conception is simply a further development of the content-oriented conception. The requirement-oriented conception focuses on the needs of the users of the system that the subject analysis is part of, and the indexer's task is to determine the pragmatic information or knowledge contained in the document. The pragmatic information is regarded as part of the implicit information, but whereas the content-oriented conception searched for the complete description of the document's subject matter, the request-oriented conception only attempts to determine the part of the implicit information, which the users will need or use. Pragmatic information is useful information. The indexer should ask not "what is *the* subject of this document" but "how should I make this document, or this particular part of it, visible to potential users? What terms should I use to convey its knowledge to those

interested?” (Albrechtsen 1993, 222). Hence, a document is analyzed in order to predict its potentials for serving particular groups of users.

Albrechtsen argues that the requirement-oriented conception should be used in conjunction with a domain analysis. This method also involves a high degree of subjectivity and responsibility in choosing those aspects of the documents which should be indexed, and requires in turn a deep understanding of the users and their area of work or study.

From these two sets of studies it is concluded that the document analysis process may be based either on the text itself or on the world-knowledge of the indexer. Albrechtsen further argues that the analysis could be based on concepts that are explicit in the text, on concepts implicit in the text, or on the needs of future users. Albrechtsen has in a very simple but effective way drawn the lines 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.

In Albrechtsen’s requirement-oriented conception the users’ information needs are considered as those which should determine the document’s subject. This breaks with Wilson’s methods, in which the users are not present. His methods are all based on the assumption that the goal of indexing is to determine



the subject of the document. The same assumption is also present in Farrow and Beghtol's approach, as well as in Albrechtsen's Simplistic Conception and her Content-Oriented Conception. In her Requirement-Oriented Conception, however, what is determined as the subject is not *the* subject of the document, but the users' potential use of the document.

### **3.2.3. Information Needs and Indexing**

Albrechtsen's aim was not to discuss a particular user group's information needs and thereby exemplify her arguments but only to sketch the three conceptions. Weinberg (1988) has discussed the needs of scholars and researchers for information and why traditional indexing fails to satisfy their needs. 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" (Weinberg 1988, 3). Traditional indexes do not provide this kind of information, and indexers are not trained to derive this kind of information from the documents. She argues that scholars and researchers are interested in particular aspects of subjects and not the entire subjects themselves. Most indexes give access only to the subject and not to particular aspects or viewpoints of the

document. Weinberg concludes that researchers and scholars have to rely on extensive reading and prodigious memory

Weinberg points out an essential weakness of traditional indexes and subject analysis, namely that the aim is to determine the subject of the document instead of trying to determine the potential usage of the document.

Vickery (1968) discusses analysis of information, by which he means "deriving from a document a set of words that serves as a condensed representation of it" (Vickery 1968, 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" (Vickery 1968, 359-60). Vickery holds that any analysis must have the potential users of the documents in mind and base the analysis on the needs of these users.

Vickery also argues that the content analysis of a document entails four steps<sup>9</sup>:

---

<sup>9</sup> It could be argued that the last step in Vickery's four-step procedure actually is a sub-procedure of the third step. The fourth step covers the procedure of determining the exact syntax of the subject entry in the particular indexing language.

- 1) selecting words or phrases that could represent the subject content of a document,
- 2) transforming the words into standard form and standard terminology,
- 3) translating the words into a code,
- 4) choosing certain of these words, terms or codes as access points.

The first and last steps are obligatory and will take place in any subject indexing process. The second and third steps will take place in some but not all systems. Vickery further divides the first step into two sub-steps, namely 1) scanning the text to select a number of words that represent the subject content of the text, and 2) choosing some of these words to represent the document according to the needs of the users of the system. In practice these steps are performed concurrently, since indexers choose only words that are relevant to the purpose of the system.

Vickery concludes his discussion of the first step by stating that this is "the key operation in subject analysis, yet is the least discussed and the least reducible to rule" (Vickery 1968, 360). He further notes that there are two types of indexers, namely those who fully understand or attempt to fully understand the subject content of 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 which the authors have emphasized, such as

title section headings, conclusions and summary. Vickery notes that very little are known about how it is possible for indexers to achieve such an understanding.

Vickery's text is in many ways a good example of discussions of the subject indexing process. He stresses the importance of the first step in the analysis process and states how difficult it is to do this analysis, but does not shed much light on the process. Vickery only states that the analysis of documents must depend on the needs of the users, but he ultimately says little of how these needs should be uncovered.

Cooper (1978) has developed an indexing method, Gedanken Indexing--thought experiment indexing--which 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 or not to assign a given index term (Cooper 1978, 114):

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).

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

The method requires that a list of suitable index terms is actually available. Cooper does not specifically discuss how one derives the subject matter of a document, but assumes that choosing one index term over another could represent the appropriate subject matter of a document. The indexer therefore does not have to “wrestle directly with philosophical question of what it means for a document to be 'about' something” (Cooper 1978, 112). The method further assumes that the user population is fixed and well known by the indexer. The indexer should know the user population well enough, in fact, to be able to determine how many will find the index term useful.

The indexing method might be useful, but it probably will be very time consuming even in a *very* small library, say a library with only a small user population around 100 or less. For indexing in large-scale databases where the user population will likely be large and not well known, the indexing method does not seem feasible.

In a brief article, Ward (1996) has described the precise skills a good indexer needs. He has done this from his own practice as coordinator of the indexing service at a large engineering consulting company. By this he hopes to

provide an answer to the crucial question of whether the human indexer has a future. Ward provides the answer by enumerating and discussing skills required by a good indexer. These skills include:

- Prior knowledge of the field. The indexer needs knowledge of the subject field, the company the indexers are working for, and knowledge of the users' general and specific information needs.
- A sense of judging what to index as well as at what level.
- The ability to read the implicit meaning of a text. This includes being able to read the documents as if the indexers were the users, as Ward says, "We read the text as if we were the end-users" (Ward 1996, 218). When Wards and his coworkers index they try to reflect the "corporate mind" (Ward 1996, 218). They can only do this by being a part of the company and through effective commitment to the company.
- Good knowledge of languages, both native and common foreign languages.
- The ability to omit redundant information.
- Being able to evaluate the document that is determining how valuable the document would be for the particular users.

Ward provides a number of examples to illustrate his points, all of which are taken from his own practice as an indexer.

Ward finishes his article with an argument against the possibility of automating the indexing process. He concludes by stating that it does not seem at all likely that the human indexer will be replaced by automatic indexing techniques. He states that automatic indexing is only suitable for texts with a simple structure, and asks whether "a degree of vision and common sense . . . [will] not produce a better and more rich result" (Ward 1996, 227).

Ward's article is of interest because it summarizes the high level of judgment an indexer makes. Much of the knowledge involved in the process requires that the indexer is able to judge which information is relevant or useful for a particular user group, but this knowledge cannot be prescribed precisely. Following Ward's argument, indexing can only be learned by doing it, not by formal training. This idea follows Dreyfus and Dreyfus' (1986) argument, which was described in chapter two above. They found that learning a particular skill develops over five steps, and that only novices are able to describe exactly what they do when performing a task. Experts, on the other hand, are not able to explain what they do. Ward's list illustrates that an expert indexer such as Ward is able to explain what kind of knowledge he uses when indexing, but he is not able to explain exactly what he does.

Weinberg, Vickery, Cooper, and Ward's articles show that other knowledge than that which can be derived from the document itself is needed in the indexing process. They all argue that especially some information about the users is needed. However, they are not able to explain exactly what knowledge is needed or how this information is found. In this respect, whereas Farrow and Beghtol attempted to explain the mental processes involved in indexing, it does not seem possible to explain the involvement of contextual information. Nevertheless the contextual information seems rather important in the determination of a document's subject matter.

None of Wilson's four methods were based on contextual information. They all assumed that the subject matter could be determined from the document itself. However, it seems plausible to argue that the subject matter of a document cannot be determined by investigating the document alone. Some information about the users', or potential users', information needs is required to determine the subject matter of a document. The problem is that it is difficult, perhaps even impossible, to prescribe how to do this.



#### **3.2.4. Empirical Investigations**

A few attempts have been made to investigate empirically how indexers index. Most of these studies fall in the category of inter-indexer consistency studies, however. Very few empirical studies have been concerned with the indexing process itself, and none have been concerned with determining the subject matter of a document based on the users' need for information. The studies are based on the assumption that *the* subject of the document can be determined. None of the studies discuss particular methods--like Wilson's four methods--for analyzing the documents. What these studies strongly suggest is that the subject indexing process cannot be studied empirically or at least only very little can be inferred from such studies.

In the first of two important studies, Chu and O'Brien (1993) state that most studies of indexing are in fact about indexes. This focus has meant that the first step of the subject indexing process--in which the subject matter of a document is determined--has been neglected "to an extent that insufficient emphasis has been given a close examination of how text is analyzed" (Chu and O'Brien 1993, 439).

In the first and only empirical study devoted to the first step in the indexing process, Chu and O'Brien try to answer two basic questions 1) how is a text analyzed to determine its subject and 2) with what ease is this done. They expect

that their findings will provide a better understanding of the subject indexing process, and especially of the first step.

The authors asked a total of 104 students from UCLA's and Loughborough University of Technology's departments of library and information studies to read three short popular articles and analyze the documents to determine the subject matter. The participants were asked to state the subject of each article in the form of a sentence, and then to evaluate the ease or difficulty experienced in doing so. One of the articles selected represented the sciences, one represented the social sciences, and one represented the humanities. The authors determined the subjects of the three articles. These subjects were considered the correct analysis of the documents. The authors also determined a primary and secondary subject for each article. The participants' answers were later compared to these definitive statements and a participant's answer was considered right or wrong based on the participant's ability to determine the same subject as the authors.

The participants were further asked what they would consider the primary and secondary subject of the article and were then asked to evaluate the relative importance of the subjects. They were further asked how easy it had been to determine the subject and whether the layout (defined as the bibliographical apparatus, e.g. title, abstract, first and last paragraph, keywords, illustrations, etc.,

and the physical presentation) of an article had helped the participants determine the subject.

The authors concluded that it is possible for novice indexers to determine the main subject of an article, although they performed better for the science and social science articles. Chu and O'Brien ascribe the poor performance of the indexers in determining the subject of the humanities article to the subjective character of the text and "the little available bibliographic apparatus . . . making it difficult to find a prominent topic " (Chu and O'Brien 1993, 452). Later they modify this conclusion by stating that "no generalisations should be drawn" (Chu and O'Brien 1993, 453) because the articles used in the study all "were short, of popular nature and required very little specialised knowledge" (Chu and O'Brien 1993, 453). But they added that there seems to be "a serious problem when participants are required to isolate primary and secondary topics" (Chu and O'Brien 1993, 453). However, how troublesome it is to determine respectively the primary and the secondary subject is very unclear. At one point Chu and O'Brien (1993, 444) state that,

In a case where the primary topic was quite clear . . . there was more variability in the identification of the secondary topic.

A few pages later they seem to conclude the opposite (Chu and O'Brien 1993, 447),

In this case where the primary topic was not quite clear, there was variability in the identification of a secondary topic.

Nonetheless, they conclude, “inevitably, a secondary topic is difficult to isolate, as its identification depends on what has been chosen as primary” (Chu and O'Brien 1993, 453). Whether it is more difficult to determine the secondary subject when the primary subject is clear or when it is unclear is not clear.

Chu and O'Brien's second research question dealt with the ease of performing the analysis of the articles. It was found that the indexers had no problems determining the subject matter of the science and the social science articles but they “experienced some difficulty in determining the general subject of” (Chu and O'Brien 1993, 452) the humanities article. It was concluded that three aspects make indexing easier, namely:

1. when bibliographic apparatus is available,
2. when the content is informative,
3. when the text is clear.

However, of these three aspects, they found that the most important is the bibliographic apparatus since even when the text is complex the "bibliographical apparatus is a major factor in determining the general subject matter of a text" (Chu and O'Brien 1993, 453). Of the twenty different kinds of bibliographic apparatus mentioned by the participants, the most cited are title, subtitle, abstract, paragraph headings, initial paragraphs and body of text.

There are a number of problems with this study. First of all the participants were asked to determine the subject of relatively short texts, i.e. 1-2 pages long. It is not possible to generalize about longer articles and books on the basis of such short articles. In most cases it must be assumed that longer texts could be interpreted to have more subjects, depending on who reads the text. If a text has more subjects it will be more difficult to determine the subject of the document. It should also be noted that Chu and O'Brien used novice library and information science students and used popular articles which required very little (or no) specialized knowledge. Conclusions about the behavior of how novice indexers index popular literature says nothing about the behavior of how professional indexers index scientific or otherwise specialized documents.

Chu and O'Brien defined no user group, no working setting, no context, and no purpose and did not use an indexing language, all of which could have helped

the indexers analyze the articles. Their methodological setup is so far from any real indexing situation that the findings do not tell anything about how indexers index. Chu and O'Brien told the participants to determine the subject of the articles and then evaluated the participants' performance against their own analysis of the articles. But they do not state why their analysis is better and more correct than any the participants' analyses.

In conclusion very little about how indexers index can be inferred from this study. Furthermore, Chu and O'Brien's findings regarding the bibliographic apparatus add nothing to what is already incorporated into the ISO standard and the DDC guideline.

A second empirical study of indexing was done by Bertrand and Cellier (1995). They intended to "make a complete analysis of the indexing process" (Bertrand and Cellier 1995, 460) and to "measure the role of knowledge previously acquired by the indexers" (Bertrand and Cellier 1995, 460). Their study therefore "complements that of Chu and O'Brien" (Bertrand and Cellier 1995, 460).

Bertrand and Cellier used 25 participants. The participants were grouped into four groups according to 1) their expertise in indexing (professional or novices), 2) their expertise in the use of the RAMEAU<sup>10</sup> indexing language (experts or novices), and 3) knowledge about the subject area (experts in

psychology, experts in economics, or no subject knowledge). The four groups were:

- I) 5 professional indexers, experts in RAMEAU, and experts in economics.
- II) 5 professional indexers, experts in RAMEAU, and experts in psychology.
- III) 5 professional indexers, novices in RAMEAU, and experts in psychology.
- IV) 10 beginner indexers, novices in RAMEAU, and no subject knowledge.

The 25 participants were asked to index eight books from either psychology or from economics. The books were indexed using the RAMEAU indexing language. The participants were asked to use a two-step procedure in indexing, that is, they were asked to first determine the subject of the books and then to translate the subject into RAMEAU. After each of the two steps the participants wrote down the words they have found. Each participant created 8 lists of words after the first step (called concepts) and 8 lists of words after the second step (called indexing terms). Furthermore, the authors made 32 audiovisual recordings of participants indexing books.

---

<sup>10</sup> Répertoire d'Autorité Matière Encyclopédique et Alphabétique Unifié.

Bertrand and Cellier divided their findings into four areas: 1) performance analysis, 2) analysis of the indexing process, 3) kinds of expertise and objectives, and 4) indexing strategies. The findings of each of these areas are summarized in the following.

The performance of the participants indexing was measured by calculating the average number of indexing terms each participant gave each document. However, no difference between the four groups was found. The performance was further measured by comparing the overlap between the concepts (words after the first step) and the indexing terms (words after the second step). By this the authors found the “intra-indexer concepts-indexing terms consistency average rate” (Bertrand and Cellier 1995, 463). It was found that there were no difference between expert indexers (groups I, II, and III) and novice indexers (group IV), but experts RAMEAU users (groups I and II) were more consistent than novice users of RAMEAU (groups III and IV). The authors therefore conclude that knowledge of the indexing language affects the selection of indexing terms.

Bertrand and Cellier analyzed the indexing process by identifying the actions taken by the participants during the indexing process. For this purpose they made a total of 32 audiovisual recordings. It was found that the participants explored several parts of the books to determine the subject, especially title, tables of contents, and abstract. It was further found that the indexing language was



explored to determine possible indexing terms and not only to determine whether a particular concept was present or not.

Under the heading “kinds of expertise and objectives” the authors describe the behavior of the participants in different situations. The attempt is to describe some general procedures indexers take. However, the authors only confirm their finding that the participants frequently used the RAMEAU indexing language to extract, not only indexing terms, but also concepts. The participants further used the indexing language for hints and ideas when they reached a dead-end, and they searched the indexing language to see whether an idea could be expressed in the language. The expert users of RAMEAU (groups I and II) frequently checked the availability of a term in the indexing language before they decided a concept<sup>11</sup>. Lastly it was observed that the expert indexers used more time (between six and 11.5 minutes) indexing each book than the novice indexers (3 minutes). The reason is the fact that the novice indexers only explored the title, the abstract, and the table of contents of the books.

The last area of findings was the indexing strategies used by the participants. These strategies were found by analyzing the lists of concepts and indexing terms and in “the identification of specific objectives” (Bertrand and Cellier 1995, 468). What the latter means is not clear. The participants used three

---

<sup>11</sup> Although this explains the high degree of consistency between concepts and indexing terms for these groups, Bertrand and Cellier do not make this observation.

different strategies according to their level of knowledge of the RAMEAU and to their knowledge of indexing in general. These strategies were:

- 1) “Indexing orientated by the knowledge of potential indexing terms”  
(Bertrand and Cellier 1995, 468). The expert users of RAMEAU went directly to the indexing language to extract indexing terms. They knew what they were looking for and therefore went to the indexing language before writing down concepts.
- 2) “Indexing orientated to the users’ needs” (Bertrand and Cellier 1995, 469). Expert indexers who were not familiar with the RAMEAU language gave a ‘finer’ description of the books. The authors explain this by stating that the indexers wish to serve the needs of the users. However, no users were involved in the study, and it is therefore unclear how the authors reach such a conclusion.
- 3) “Indexing helped by the documentary language” (Bertrand and Cellier 1995, 469). Novice indexers used the indexing language to guide them through the indexing process. Instead of spending much time on analyzing the books they only looked at the title, the table of contents and the abstract. They then went to the indexing language to search for general terms and used these as starting points for the indexing process.

Bertrand and Cellier's study confirms a number of observations from second chapter in this dissertation. The idea of breaking down the indexing process seems reasonable for the purpose of studying the indexing process empirically. However, as it was argued in chapter two, the stages between the different steps are of very different nature, and a comparison of words uttered at different stages reveals little about the nature of the indexing process. Furthermore, in chapter two it was argued that novice indexers and expert indexers perform the indexing task quite differently. It was argued that novices use rules with no regard for the particular situation. Experts, on the other hand, do not break down the task into several steps and do not think of the situation as solving problems or making decisions, but simply as indexing. As Bertrand and Cellier show, expert users of the RAMEAU language actually index quite differently from novice indexers and professional indexers who were not familiar with the RAMEAU language. However, a comparison of the performance of experts and novices reveals little about the nature of the indexing process.

In conclusion, although Bertrand and Cellier's study confirms some of the assumptions made in the previous chapter, little new knowledge about indexing is gained. Their intention to "make a complete analysis of the indexing process" (Bertrand and Cellier 1995, 460) is not fulfilled. They merely confirm what was

already known about indexing. Their attempt to “measure the role of knowledge previously acquired by the indexers” (Bertrand and Cellier 1995, 460) is limited to stating that experts in a particular indexing language use their knowledge about the indexing language when they analyze a document.

To study the indexing process empirically seems almost impossible. As shown here neither Chu & O’Brien nor Bertrand & Cellier gain new knowledge about indexing from their studies. They merely confirm what is already known. In both studies an artificial controlled test situation was set up to determine what indexers do when they index. Chu and O’Brien use students as test indexers and not professional indexers. They thereby study something other than how indexers index, namely how non-professional indexers determine the subject matter of a document. The underlying assumption must be that novice indexers and professional indexers determine the subject matter of a document in the same way. It was shown in chapter 2 that Dreyfus and Dreyfus (1986) argued that novices and experts perform the same task very differently. Conclusions reached about how novice indexers index cannot therefore be generalized to how indexers index in general. Bertrand and Cellier, on the other hand, do use professional indexers in their test but the indexers’ working environments are not taken into account. The assumption must be that professional indexers determine the subject of a document in a certain way and that this way is independent of the environment in

which the indexing is done. However, if indexing is done conceptually (Farrow and Beghtol's terms), rests on the indexer as much on the document (Wilson's terms), and somehow involves the users' need for information (following Albrechtsen, Weinberg, Vickery, Cooper, and Ward), the entire situation surrounding the indexing must somehow influence the indexing process. But if it is believed that humans follow some unique mental rules when they index, then of course one can study indexing in artificial controlled test situations and say something general about how indexers index.

The basic problem in studying indexing empirically is that there are so many variables in studying indexing that they will be almost impossible to control. At a minimum, empirical studies should study professional indexers in their normal working environments, using their normal indexing language, indexing for their normal user group, working under their normal conditions. Only then will the indexing process be studied as close as possible to reality. And only then will all the possible variables come into play. However, imposing these conditions also limits the potential results of a study of this kind. It will be extremely difficult to say exactly which of the variables influenced the indexing process. Therefore only a few generalizations can be made of a study under such conditions

Nonetheless, studies of inter-indexer inconsistency are limited by the same constraints but have made a common generalization. It is often stated that "it is

well known that indexers differ significantly in their judgment as to which terms reflect most adequately the content of a given document” (David et. al 1995, 49). A statement like this is often encountered in the literature. However, these inter-indexer studies either study how non-professional indexers (mostly students) index documents (c.f. e.g. Lilley 1954; Bates 1977a; 1977b; Leonard 1977, 31-33; Furnas et. al. 1982; Markey 1984; Lancaster 1998, 71-76) or compare professional indexers working in different environments, often with different indexing languages (c.f. e.g. Moeller 1981; Chan 1989; Iivonen 1990; 1995; Iivonen & Kivimäki 1998). Again, the assumption must be that novice indexers index documents in the same way as professional indexers, and that studying the behavior of novice indexers will say something about the behavior of professional indexers. When comparing the work of professional indexers the assumption must be that their task is to determine *the* subject of the document.

Although none of the authors of these empirical studies state this explicitly, they are all based on Wilson’s methods of indexing. In other words, they all assume that the subject matter of a document can be determined from the document itself. None of these empirical studies incorporate contextual information of any sort.

### **3.2.5. Criticism of Mentalism in Indexing**

The quest for the rules of indexing underpins most research in indexing. In one way or another researchers have believed that their task is to discover the rules of indexing. It is believed that there are some rules that govern how indexers index and that the object of research in indexing is to discover these rules. Frohmann (1990, 82) explains the assumption underpinning most research on indexing,

It seems obvious enough that the indexer performs an intellectual operation of some kind in representing the text by means of an indexing phrase. But exactly what kind of operation? We believe that if we understood its rules, the benefits would be obvious. Not only would our understanding of a fundamental operation of information science advance by a quantum leap, but we may also gain the added bonus of solutions to persistent and embarrassing problems. A clear distinction, for example, between the essential core of the indexing operation . . . and its merely peripheral fringe . . . , might yield a method for reducing the notorious inconsistency of indexers.

The assumption is that, since indexers perform some kind of intellectual operation, then research on indexing must have this intellectual operation as the object of study. The final goal of research in indexing is to explain the intellectual

operation and be able to represent it in such a way that two people will be able to perform the intellectual operation alike. A further assumption is that operation is guided by some mental rules. Frohmann (1990, 84) elaborates,

[T]he most fundamental intellectual operation of indexing is, in principle, explicable by internally realized and tacitly known rules that generate an indexing phrase from a given text. It seems that there must be some rules guiding the mental activities of indexers, for otherwise it becomes impossible to explain *how* they are able to utter or to write down an indexing phrase for the text. The problem is to discover the precise form of these rules.

The discovery of the rules of indexing is explicit in Farrow and Beghtol's studies. However, the above assumptions are present in most other studies. As it has been shown in this review, most studies of indexing in fact investigate how indexers index. This tradition follows Wilson's discussion of the four methods for determining the subject matter of a document. However, Wilson's points that there is no one correct way to determine the subject matter of a document and that there is no one precise description of a document's subject matter seem to be missed. Research on indexing is generally based on the assumption that there are some mental rules that need to be uncovered and that when these rules are found



they will provide guidance to prescribe the best way to determine the subject matter of documents.

The major goal in indexing research is to explain the mental operation in indexing. In the first chapter of this dissertation it was pointed out that more research into the cognitive process of indexing is needed. Shaw and Fouchereaux point out that one of the areas that needs more research is “the cognitive processes involved in indexing and classification” (Shaw and Fouchereaux 1993, 25). Milstead likewise points out 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’” (Milstead 1994, 578).

Frohmann sets out “to evaluate the implications for indexing” of this kind of research “for the proper theoretical orientation of information science” (Frohmann 1990, 81). He claims that this kind of research is based on the assumption that “thoughts are mental processes occurring in minds. Because thoughts arise before language and language is used to express thoughts, albeit often inadequately, thoughts are only contingently related to language” (Frohmann 1990, 82). In other words, language is used to express ideas and denote physical entities that exist independently of humans. The word ‘cat’ refers to the physical living creature we call cat, the word ‘love’ refers to the idea of love, etc.

The individual words we use to describe ideas and physical entities are only labels for the concepts we have in our minds. These mental concepts are connected by some unknown mental rules and the aim of research is to discover these rules and make them accessible. Or in other words, “the project consists in bringing the unconscious into conscious” (Frohmann 1990, 84). For indexing this means that the project is to uncover the mental rules indexers follow when they index.

Frohmann brings forth Wittgenstein’s remarks on following a rule to expose the illusions that are hidden in the mentalist approach to indexing theory and research. In short, Wittgenstein argues that to follow a rule is to understand the social practice in which it is embedded. It is not possible to follow a rule if one does not understand its social context. Rules are used to guide, direct, instruct, prescribe, influence, and determine our actions. To use a rule is therefore “a matter of course” (Frohmann 1990, 87). One does not need to think, consider, and reflect. It is just done. We only need the rule itself to follow it. Furthermore, only the rule itself is needed to determine whether a rule has been followed in particular situation. Rules are therefore also used to justify our actions. However, rules themselves do not determine a person’s actions. The rule itself does not activate some mental process that makes sure that the person acts according to the rule. The person acts according to the rule because she is a member of a social

practice in which it is decided that one does so and so when one is given a certain rule. The judgment of whether a rule is followed and should be understood is decided among the members of the social practice. The criterion of following a rule is therefore public. Following a rule therefore “cannot consist in the possession of internal representations generating their own applications in the recesses of the mind because the mental arena excludes public criteria” (Frohmann 1990, 92).

One does not follow the rules of indexing “unless the point, purposes, intentions, aims, ends and strategies of” (Frohmann 1990, 92) the rules are the same as those of indexing in the social context in which the indexing is done.

Therefore, according to Wittgenstein, to understand a rule is to understand the social practice in which the rule is used. The two things--rules and practice--cannot be separated. Furthermore, there is no such thing as the rules of indexing, there are different sets of rules of indexing for each social practice. Studies of indexing therefore need to start with social practice in order to understand how indexers in that particular setting index. It will thereafter be possible to say something about how indexers in that social practice index, but such findings will not be general and cannot say anything about how indexers in a different social practice index or should index.

However, Frohmann notices that effective information retrieval systems are built on the flawed mentalist foundation, and he therefore asks himself what the danger of the mentalist research approach is. The danger, he says, is that it has the “power to deflect attention from theoretical problems central to the development of effective information retrieval systems” (Frohmann 1990, 94). In other words, if research in indexing keeps following the mentalist approach, it will be possible to enhance effectiveness of information retrieval systems, but it will not be possible to reach theoretically sound conclusions about the nature of indexing.

Frohmann suggests that the mentalist approach conceals "fruitful directions of inquiry in at least the following five areas" (Frohmann 1990, 94).

1. Mentalism focuses on *discovering* the rules of the indexer, where 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, which 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, which occur in the mind, mentalism conceals the crucial social context of rules.

Frohmann criticizes research approaches to indexing that attempt to discover the underlying rules that govern the subject indexing process. He makes a strong case that such an attempt is bound to fail, if one accepts Wittgenstein's remarks on following a rule.

Blair (1990) also rejects the mentalist approach and uses Wittgenstein in his study on the nature of information retrieval. He finds that empirical studies of information retrieval are "prohibitively difficult and costly" (Blair 1990, 121). Blair (1990, 121) therefore argues that instead of large empirical studies we should determine,

[W]hat the essential nature of information retrieval is, and then within this context try to define and solve puzzles which in their solution (and attempted solution) will advance our understand of how information retrieval systems work and lead to new puzzles to solve.

Since Blair finds that the fundamental question for all information retrieval is to decide how to represent documents for retrieval, this is where he starts. What is needed, he says, is a theoretical foundation for understanding the representation of

documents. Previous studies in indexing theory have “not gone deep enough” (Blair 1990, 122). He wants to redirect research and claims that the process of representing documents for retrieval is “fundamentally a linguistic process” (Blair 1990, 122) and therefore basically about how language is used. Whether or not a researcher is explicit about the linguistic foundation she bases her information retrieval or indexing research on, she will in fact base it in a linguistic tradition, because “any theory of indexing or document representation presupposes a theory of language and meaning” (Blair 1990, 122).

To find the proper linguistic basis for indexing and information retrieval theory, Blair examines a few major theories of linguistics. The first of these is semiotics. The second is Wittgenstein’s pragmatic philosophy of language.

The central point in Blair’s search for a sound theory is whether signification is taken to be prior to communication or not. If one believes that signification is prior to communication, it is assumed that language is only used to label some ideas which are present before communication. In other words, the words merely denote some ideas and physical entities that exist objectively and independently of the speaker. In such a theory there is a close connection between reality and the structure of language. Reality must somehow influence the structure of language. Whether a statement is true therefore depends on whether the statement’s claim correlates with reality.

On the other hand, if one believes that signification takes place through communication, then the words do not necessarily correlate with reality. The two are seen as independent of each other. The words therefore get their meaning through their use, and the decision of whether a word is used correctly is decided jointly in the particular social context or practice in which it is used.

Blair finds that signification takes place through communication and he therefore finds that semiotics does not deliver a proper theory of meaning and language to be used for a theory of representing documents. Semiotics, he says, argues that an expression “cannot stand by itself. It must be (correctly or incorrectly) correlated with a content or meaning, otherwise it ceases to be an expression and remains a cipher” (Blair 1990, 126). For a person to understand a given expression the person must already know the meaning of the individual words of the expression. In other words, the content or meaning of the individual words must be given to the person prior to communication, her job is then to translate the words into concepts. Blair therefore argues that “the cornerstone of semiotics is . . . the relation between expressions/signifiers and contents/signifieds” (Blair 1990, 131).<sup>12</sup> Because “the content/signified is an idea

---

<sup>12</sup> As it will be shown in the next chapter on semiotics Blair’s claims about semiotics are only valid for the European structuralistic oriented semiology as it has been promoted by Saussure (1966), Barthes (1968), Hjelmslev (1969), and others. Blair’s claims are not valid for Peirce’s semiotics.

(feeling, sense, fantasy, etc.)” (Blair 1990, 131) semiotics is a mentalistic theory of meaning.

Blair (1990, 132) finds that there are two points which make mentalistic theories of meaning problematic:

1. Because ideas are private it will not be possible to verify whether a person has understood an expression correctly. It is not possible for her to get access to the original idea in the mind of the speaker and therefore not possible for her to verify whether she understands the speaker’s expression.
2. The very nature of the ideas is problematic. These ideas are often explained as mental pictures of the meaning of words. But even if we had some clear mental image/picture/idea/concept of what is said when we hear an expression, we cannot easily attach the meaning of the words to the mental picture because the picture will require some interpretation to be understood.

The consequence of mentalistic theories of meaning and language is that there is no clear definition of the content of an expression. The content is a mental idea and its status is therefore reduced to an implicit assumption. The problem, Blair



argues, is that mentalistic oriented theories of language and meaning have asked the wrong question. They have been too occupied with the question of what a given expression means or signifies, because that question assumes “that there does exist a ‘something else’ that an expression means/signifies” (Blair 1990, 135). However, this is the natural question to ask in a tradition that believes that signification is prior to communication. If it is believed that words point out some object or idea in reality, then it is natural to ask what these ideas or objects are. Blair argues that instead of “asking ‘what does an expression mean/signify’ we shall now ask ‘how is an expression used?’” (Blair 1990, 136).

According to Strawson (1970) the meaning of a word is determined by how the word is used. Rules for the use of words are established in the social context in which the words are used. In other words, the meaning of words is closely tied to the social context in which the words are used. The meaning of a word cannot be determined independent of the context of the word. Although this may imply a relativistic approach it does not claim that words can be used however one feels like. Such a claim would entail that any word could mean what the speaker believes it means. A particular use of a word and thereby a particular meaning of the word has to be accepted by the social context in which it is used. Wittgenstein (1958, par. 43) argued a similar view,

For a large class of cases--though not for all--in which we employ the word "meaning" it can be defined thus: the meaning of a word is its use in the language.

To understand the meaning of words is thus to be able to use the words correctly. But to use a word correctly is not to know its inherent meaning or what the word refers to. To use a word correctly is to understand the practice in which the word is used. In the same manner as rules and practice (as it was explained in the review of Frohmann's paper), words and practice cannot be separated. The meaning of a word is therefore closely connected to the particular practice in which it used. The same word might not be used in the same way in different practices.

To explain this understanding of the relation between the meaning of words and the use of words Wittgenstein (1958, par. 11) compared the use of words to the use of tools,

Think of the tools in a tool-box: there is a hammer, pliers, a saw, screw-driver, a rule, a glue-pot, glue, nails and screws. --The function of words is as diverse as the functions of these objects. (And in both cases there are similarities).

Although there are plenty of ways to use a hammer, there is usually a limited number of correct ways. These have been established through experience and conventions. This holds for words as well. In principle, any word could be used as one feels like, but since words are used to convey meaning to other people, there are only a few correct uses of a particular word. Blair (1990, 139) gives another example,

[T]he primary use of a screwdriver would be to turn screws of a certain size and type, and an accepted secondary use of the screwdriver would be as a lever to pry open a can of paint. A creative use of the screwdriver might be as a wedge under a door to hold it open.

In the same manner, a word can be used in a range of different manners. Some are more central or typical than others. The point is that the meaning of the word depends as much (or even more) on the people using the word as on the word itself. This puts the concern for the signification of an expression in a better light, as Blair argues, “to ask a carpenter what one of his tools ‘means’ or ‘signifies’ would be a strange question, indeed” (Blair 1990, 139). Instead of a definition of the tool we would ask him for a demonstration of the tool. In the same way, instead of asking what the meaning of a word is, we should ask how the word is used.

Blair argues therefore that we should not ask what a particular document or subject description is about or what it means. This question “will lead us to some kind of mentalistic or referential explanation, which may be briefly satisfying, but ultimately of not much use” (Blair 1990, 157). To understand what the subject matter of a document is we need to understand how the document will be used. To understand how the document will be used requires that we understand the social practice in which it will be used. This in turn requires that we understand what is done in that social practice. In other words, it is not possible to represent the subject matter of a document if the indexer does not have an understanding of the practice for which she indexes.

Blair’s arguments take Frohmann’s argumentation a step further in that they outline a direction for future research within a pragmatic theory of indexing. Frohmann limits himself to merely pointing out the limitations of the mentalist approach. However, taken together they form a strong argument for a redirection of research in indexing, that is, if one accepts the assumptions made in the later Wittgenstein’s philosophy of language.

This limitation is exactly the weak point in their studies. They merely discuss indexing at a general and philosophical level. They never in any way demonstrate that their theories will improve the quality of indexing or the efficiency of indexing systems. They are, however, aware of this. It is not their

aim to actually put forward new and better ways to index. They merely want to lay the groundwork for a reorientation in studies of indexing.

### **3.2.6. Summary**

The majority of research on indexing has been occupied with one basic question, namely “what do indexers do when they index?” This review of literature on indexing gave examples of such studies. As a prelude to the review Wilson’s thirty-year-old essay on the determination of subjects was reviewed. It was noted that Wilson’s four methods or ways to determine the subject of a document all presume that the subject matter in one way or another should be found in the document itself. In other words, in all of Wilson’s methods, textual structure is viewed as something that should “be analyzed in itself and for the sake of itself” (Eco 1994, 44). The “pragmatic aspect of reading” (Eco 1994, 44) is left out in Wilson’s methods. At the end of the review Frohmann’s and Blair’s studies were mentioned as examples of scholars who base their understanding of text, documents, and representation on a pragmatic tradition.

To answer the question of what indexers do when they index a number of research paths have been followed. One important path has been cognitive studies in which mental models of indexing have been established. Another path has

been empirical studies of the behavior of indexers. Albrechtsen pointed out that one important aspect of determining the subject matter of document might be the users' information needs. There have in fact been a number of studies from this perspective. They are, however, in line with the mainstream tradition of attempting to describe what indexers do when they index.

Frohmann and Blair both use Wittgenstein's pragmatic philosophy of language to argue for a redirection in the study of indexing. They argue that all previous research on indexing has been based on a mentalistic conception of language and meaning. They conclude that research on indexing has therefore asked the wrong questions, namely what the subject of a document is and how this subject is determined. Much previous research on indexing has therefore assumed that documents have subjects. Blair and Frohmann, however, point out that documents are given subjects according to their use. They thereby lay the grounds for further studies that will build a pragmatic theory of indexing.

The present dissertation will build on the foundation laid by Blair, Frohmann, and others. In the next chapter Peirce's semiotics will be introduced as a framework for understanding the subject indexing process.

### **3.3. RELEVANCE STUDIES**

One of the central goals of information retrieval (IR) theory is to measure the performance of a given IR system or IR technique. At the heart of this goal lies the concept of relevance, an idea widely discussed in the IR literature and with many variant definitions. Accordingly, there is a huge corpus of literature on this topic. Here, however, only a fraction will be reviewed. This selection is hopefully representative of the range of definitions of relevance in the field. The purpose of this small review is to show how closely related the concept of relevance is to the concept of subject, and to indicate that these two concepts are studied in isolation.

The term relevance is used as a way to speak generally of how well a particular IR system or IR technique retrieves documents that are ‘relevant’ to a user. Particular measures usually employ more specific terms such as precision, recall, probability, and utility when discussing the measuring process. And almost all such measures test relevance by beginning with specific information requests or information needs. In this section the literature of relevance studies will not be examined in order to portray such measures, but rather to look for suggestions in them about how researchers in this area define the concept of subject. The underlying assumption here is that researchers on the concept of relevance must operate (explicitly or implicitly) with a concept of a document’s subject in order

to measure how well a system or technique retrieve documents on a particular subject. And, if they operate with a particular concept of subject, that concept might have bearings on the conception of the subject indexing process.

Saracevic (1970) has in a clever way clarified the confusing terminology in literature on relevance. On the basis of earlier definitions of relevance, he developed a chart (see figure 3-2) in which he joined various definitions (Saracevic 1970, 120).

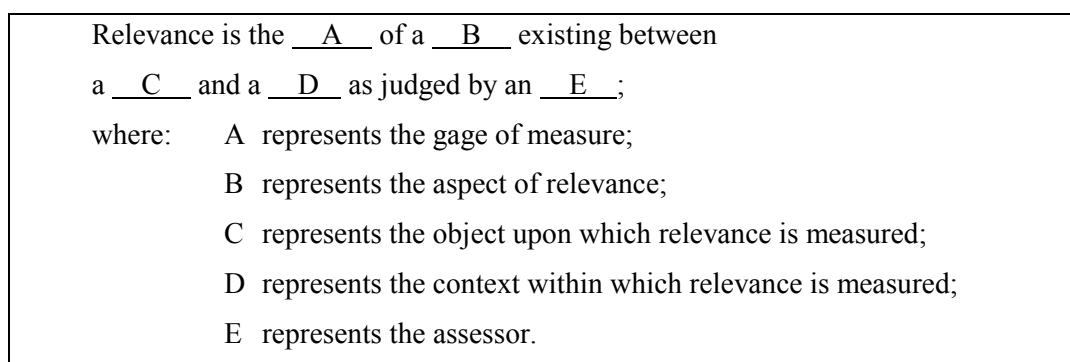


Fig. 3-2. Variables in Relevance Judgments

With the appropriate words inserted, a number of statements could be formulated, this is illustrated in figure 3-3. (The figure is a slight modification of Saracevic's [1970, 121] figure).



<b>Relevance is the</b>	<b>of</b>	<b>between a</b>
a. measure	a. utility	a. document
b. degree	b. matching	b. document representation
c. extent	c. informativeness	c. reference
d. judgment	d. satisfaction	d. textual form
e. estimate	e. appropriateness	e. information provided
f. appraisal	f. usefulness	f. fact
g. relation	g. correspondence	g. article
h. quantity	h. importance	
i. dimension	i. connection	
	j. fit	
	k. similarity	
	l. applicability	
	m. closeness	
	<b>and a</b>	<b>as judged by a</b>
	a. question	a. requester
	b. question representation	b. intermediary
	c. research stage	c. expert
	d. information need	d. user
	e. information used	e. person
	f. point of view	f. judge
	g. request	g. information specialist
		h. librarian
		i. delegate

Fig. 3-3. Definitions of Relevance

By combining the terms from the five categories, different connotations of a definition of relevance may be given. At the same time, the basic problem remains the same: a judgment of relevance is a relative measurement. For example two people with what appears to be the same information need will not have the same perception of the relevance of a document. Or, again, the same person will judge the same document's relevance differently over time. And still

again, one document's relevance might depend on the retrieval of another document, or on the absence of another document. Saracevic's definition of relevance clearly shows the many variables the notion of relevance has and hence how problematic the concept is. This conception of relevance should make clear that it is not possible to speak of the relevance of a document, in the same way it is not possible to speak of the subject of a document. Saracevic does not include any discussions of the concept of subject. However, his definition of relevance is close to a definition of the determination of the subject matter of a document.

Cooper (1971) distinguished between two uses of relevance, namely relevance as usefulness for the user and relevance as a logical relationship between the topic and the document. With respect to the first, 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. But while the statement seems viable, it is nevertheless intractable. 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. Cooper argues, however, that since the utility of documents is relative to different users it would not be possible to measure such a concept of relevance. Given this reality, he fell back on his second use of the concept, that relevance should be defined as the

relation between topics of the stated need and the retrieved documents. In short, the aim must become the measurement of relevance as the relationship between the stored information and the user's information need as it is stated in linguistic terms. Cooper's definition of *logical relevance* is hence not a definition of what the users perceive as relevant, but merely a definition of the goodness of match between the topics of the request and the retrieved documents. Here, however, Cooper faces the same problem as others who have deferred to this solution. His definition of relevance requires a clear concept of subject. But, like others, he has not supplied this concept.

Swanson (1977) has defined two frameworks for understanding relevance. Within the first framework relevance is understood as something which the requester or user creates or constructs from whatever new knowledge the user derives from the document. In this sense, relevance is not something that is present in the document and cannot be measured or judged. The relevance judgment is solely based on the requester's use of 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 request which the user enters into the system. A person other than the user can make the judgment of relevance in this sense. But it is important to notice that even though

a document and a stated request have the same topic, the document might not necessarily fulfill the user's underlying need for information.

Swanson does not discuss how one determines the subject of a document, nor what a subject of a document is. His purpose was simply to distinguish between two frames of reference--one that is based on the user's information need and one that is based on the subject of the document.

Swanson further critiqued three central and highly cited IR effectiveness tests, Cranfield II, Lancaster's MEDLARS tests and Salton's SMART-MEDLARS comparison. He found that there were serious problems with the design of each. Among these problems was that their relevance judgments were delegated to someone other than the user.

Swanson's distinction between the two frameworks of relevance is significant, but he does not clearly argue how persons other than users can judge relevance in the second framework. To do so he would need a clearly defined concept of subject, which he has not supplied.

Buckland (1983) has distinguished three uses of the term *relevance*. He defines these as:

- I. *responsiveness*, which is the measurement of the system's ability to retrieve correct data on the basis of the attributes used as the basis for retrieval,
- II. *pertinence*, which is a narrower use of responsiveness, namely when the attribute used for retrieval is the subject matter,
- III. *beneficiality*, which is 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 (responsiveness and pertinence), where it is the ability of the system to respond to an inquirer's formulated inquiry.

Buckland argues that measurements of relevance must omit users' subjective and individual information needs, that is relevance defined as beneficiality. The reason is that it will not be possible to measure the effectiveness of an IR system objectively on the basis of how users utilize documents retrieved from the system. On the other hand, a measurement of an information retrieval system based on *pertinence*, where 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. This aspect is missing in Buckland's discussion. He assumes that it is possible for a third person to compare the subject of a request with the subject of a document. However, he does not discuss how the subject matter of a request and document is determined. Such a discussion would probably have shown that the distinction between pertinence and beneficiality is not as sharp as Buckland makes it. The determination of the request's and the document's subject matter requires that these are seen in relation to some specific situation, and some specific user group.

Green (1995) takes a position directly opposite than Cooper, Swanson, Buckland, and others. She identifies the same two approaches to relevance by calling them strong and weak relevance. Strong relevance is the relationship between the users' need for information and the documents, whereas weak relevance is the relationship between the topic of the users' request and the topic of the documents. Green states that the ideal system must consider relevance as "the property of a text's being potentially helpful to a user in the resolution of a need" (Green 1995, 647). The heart of her argument is that the IR system should provide *useful* information to the user. Given this assumption, the measurement of the system should rely on the system's ability to provide the user with useful information. It is not enough to simply measure the system's ability to retrieve topically relevant information.

Others have made similar observations. Barry (1994) for instance argues that there are factors other than the topical appropriateness which influence users' evaluation of the relevance of the retrieved documents. Barry therefore argues that since "users approach information retrieval systems in hopes of finding information that has some *meaning* for them" (Barry 1994, 151 [italics added]), retrieval mechanisms based primarily on topical matching may fail to address the needs that users bring to the systems. And Park (Park 1994) argues that topical relevance ignores the individual's particular context and needs.

The upshot of Green's, Barry's, Park's and others' user-oriented definition of relevance is that measurements of relevance need to take the users' individual information needs and work situations into account and base the evaluation of the IR technique or IR system on these. Although none of them explicitly discuss the concept of subject, they all make clear that the subject matter of requests and documents are not easily determined.

### **3.3.1. Summary**

The central discussion of the concept of relevance seems to be whether someone other than the person with the information need which generated the request can judge the relevance of a particular document. If that is possible,

there is no need to involve the users in the evaluation, but if the judgment of relevance is closely tied to the individual user, users need to be part of the evaluation. Although none of the works cited here make the comparison between studies of indexing and studies of the concept relevance it seems obvious that the two areas of studies have common interests.

These two areas of study, indexing and relevance, have the same main problem, namely whether the subject matter of documents (and also requests in relevance studies) can be determined objectively and neutrally. In both areas of studies, concern has been raised that this cannot be done. However, in the case of relevance, very little research has been done to show how IR systems and techniques can be measured if the measurement is based on the users' judgments. As it has been suggested by Ellis (1996a; 1996b, 189-191) one fruitful path of future research might be to shift from quantitative methods to qualitative methods of study.

Since both areas are concerned with the same basic question, there is reason to believe that closer collaboration between the two areas of research might lead to fruitful insights. A judgment of the relevance of a document and a determination of a document's subject matter are two sides of the same coin.



### **3.4. SUMMARY AND CONCLUSION**

A number of studies, all of which relate to knowledge representation, have been reviewed here. The review was divided into three sections. The first dealt mainly with Wilson's discussion of ways to determine the subject matter of documents. The second--and largest--section reviewed literature on determination of the subject matter of documents. And the third section reviewed a small number of works on the concept of relevance to determine the relationship to studies of indexing.

One common assumption in the literature reviewed in the first section is that research in indexing should be concerned with *how* indexers index. Further, all authors appear to be driven by the need to find the rules which an indexer follows in representing the subject matter of documents properly. Another common assumption shared by the authors is that the subject matter of a document can be determined based on the document alone. A few authors have challenged these assumptions. Albrechtsen, for instance, outlines three approaches to indexing, one of which is based on the users' explicit needs for information. Frohmann and Blair also challenge these common assumptions. They argue that indexing and subject representation need to be explicitly based on a theory of language and meaning. They both use Wittgenstein's pragmatic philosophy of language to discuss the nature of indexing.

In the literature reviewed no one has investigated the importance of the interpretive nature of the subject indexing process. Some, like Farrow and Albrechtsen, argue in favor of an interpretive nature of the first step, but not of the whole process. Others have discussed the relation between knowledge representation and information need, but only as a process of matching these, not as interpreting needs and documents. Still others have discussed the relation between indexing and the indexing language.

To map the different approaches to studies on indexing and subject representation, a number of conceptions of indexing and subject representation can be enumerated. In the following, five different approaches are listed. These are based on the present review and condense the views expressed in the literature reviewed here.

#### **3.4.1. Five Conceptions of Indexing**

There is a trend in indexing literature to discuss the advantages and disadvantages of different approaches to indexing. Such discussions have often lead to dual conceptions of indexing. Soergel (1985) discussed the difference between entity and request oriented indexing, Fidel (1994) the difference between a document-oriented approach and a user-oriented approach and, as shown here,

Albrechtsen (1992; 1993) the difference between content and requirement oriented indexing. Others have proposed similar dual approaches to 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 indexing process reveals that it is possible to enumerate at least five different conceptions. These conceptions are syntheses of the conceptions which have been discussed earlier in this dissertation, but they are furthermore argued to be founded in different epistemological positions.

Hjørland (1997) has argued that there are four basic epistemological positions, which are of interest to knowledge representation. He defines these as:

- i) 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 (Hjørland 1997, 59-61).
- ii) 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 (Hjørland 1997, 69-72).
- iii) Historicism, the view that the principles for obtaining knowledge develop 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 (Hjørland 1997, 73-75).

- iv) Pragmatism, the view that knowledge primary develops from praxis, that the future use determines the context for knowledge, and that knowledge is context dependent (Hjørland 1997, 75-76).

In the following the relation between different conceptions of indexing and different epistemological positions will be tied together. The five basic conceptions of indexing are as follows:

- a) The *simplistic conception* of indexing. This conception is similar to both Wilson's constantly-referred-to method and Albrechtsen's "simplistic conception" and focuses solely on automatic extraction and statistical manipulation of words.

This conception of indexing is linked to empiricism, since the assumption in the simplistic conception is that the subject matter of documents can be derived on the basis of the word occurrence in the

document. The idea is that the sum of the words in the document constitutes the subject matter.

- b) The *document oriented* conception focuses on the information that is present in the document. This relates to Farrow's definition of perceptual indexing and to Wilson's figure-ground method and purposive method. In this conception the indexer investigates specific parts of the document, for instance the title, the headings, the beginning of paragraphs. The indexer then picks out the information of the documents which is used for indexing. The conception only uses information found in the document, but a human indexer determines the importance of the information.

The document-oriented conception is closely related to the rationalistic position. The idea behind a document-oriented conception is that it is possible by pure reasoning to objectively determine the subject matter of documents. Such an approach to indexing will define some fundamental principles by which it is possible to determine the subject matter objectively and eternally.

- c) The *content oriented* conception attempts to describe the content of the document as fully as possible. This is an objectivistic conception, which, in

the extreme, would claim that there is only one correct analysis of a given document. An indexer with this conception will strive to cover all the topics of the documents and represent the full meaning of the document.

As with the historicist epistemological position, the content oriented approach assumes that it is possible by carefully investigating the different interpretations of the document to determine the subject of the document. This conception will determine the subject matter on the basis 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 which determine the subject matter of the documents.

- d) The *user oriented* conception focuses on users. The focus of indexing is directed towards the users' general knowledge or the users' work or research domain. The indexer pays special attention to the knowledge of the users. Users at a public library require a different kind of indexing than users at a university library, even for the same document. Public libraries express subjects more generally than university libraries. If the indexer works in a organization that serves users with a particular interest or work domain, the indexer should pay special attention to this. For instance, if the indexers work in a library which only serves biologists, then the indexer should index

documents from a biologist's point of view. The same holds for more specific domains.

The user-oriented conception is clearly based on a pragmatic position in that the conception first and foremost focuses on the future use of the document. The potential user group shapes the analysis of the document. In such an approach the subject matter of documents can not be determined objectively and one time for all. It will change as the members of the user group change and as the interests and tasks of the user group change.

- e) In the *requirement oriented* conception, the indexers are aware of the users' individual information needs and work tasks. In the user-oriented conception the indexer had some general knowledge of the group of users who use the library, in the requirement-oriented conception the indexers have knowledge about individual users' information needs. A requirement-oriented conception will only be useful in smaller organizations, say a consulting firm with 15 consultants, in which it is possible for the librarian to know the needs and tasks of each of them. This is not possible for a large research library with hundreds or thousands of users.

As with the user-oriented conception, the requirement oriented conception is also based in the pragmatic position. Whereas the user

oriented conception focuses on a specific user group, the requirement-oriented conception is narrower in focus, since it bases the document analysis on specific information needs. But common for both conceptions is that they base the analysis on the future use of the documents. Indexing done in the requirement-oriented conception will, like the user-oriented conception, change over time.



	<b>Simplistic</b>	<b>Document-oriented</b>	<b>Content-Oriented</b>	<b>User-oriented</b>	<b>Requirement oriented</b>
<b>Type of information</b>	Explicit	Explicit	Implicit	Implicit	Implicit
<b>Determining factor</b>	Words in the document	Composition of the document	Analysis of the document's content	Based on the users' work domain	Users' explicit need for information
<b>Objectivity</b>	Neutral	◀			▶ Value added
<b>Time to perform</b>	Quick	◀			▶ Time consuming
<b>Consistency</b>	High	◀			▶ Low
<b>Cost, short term</b>	Cheap	◀			▶ Expensive
<b>Cost, long term</b>	Expensive	◀			▶ Cheap
<b>Environment</b>	Any	Any	Any	University/special/public	Small organizations
<b>Indexing method</b>	Extraction	Extraction	Assignment	Assignment	Assignment
<b>Epistemology</b>	Empiricism	Rationalism	Historicism	Pragmatism	Pragmatism

Fig. 3-4. Aspects of the Five Conceptions of Indexing

These conceptions should not be seen in isolation. There probably is not any indexing unit which solely base its indexing on just one of the above conceptions. The conceptions are related to each other, as on a continuum, as shown in figure 3-5. The idea of the continuum is that the indexing practice becomes more and more dependent on the individual users and their needs as one

moves from left to right. Conceptions that are at the right side of the continuum will use aspects of the conceptions on the left side. This means that the content oriented conception involves methods and aspects which are also used in the document oriented conception, while the user oriented conception involves methods and aspects from the content oriented conception. In other words a conception will always involve methods and aspects from conceptions which are placed at a position to the left of that conception.

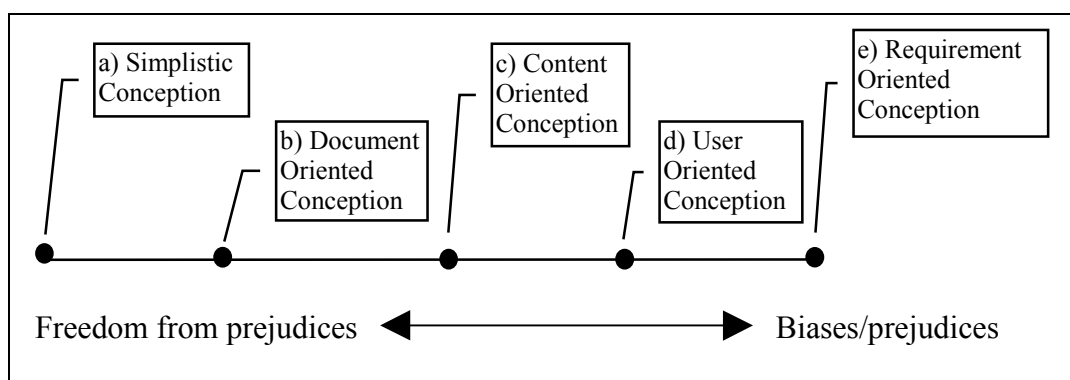


Fig. 3-5. Five Conceptions of Indexing

An indexing unit will most likely choose a combination of two of the five different conceptions. The 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. The university and public libraries might

choose to combine the user-oriented conception with a content oriented conception, again to ensure a lasting value of the indexing.

The upshot is that it is impossible to claim that an indexing unit only uses one conception or approach to indexing. It is most likely that combinations of conceptions are used. It is furthermore impossible to argue that one conception of indexing works better than others - it all depends on the particular situation of the indexing unit. Moreover, as it was shown here, each conception is closely tied to different epistemological positions which place the conception in different traditions of thinking.

#### **3.4.2. Information Needs, Relevance, and Indexing**

The essence of the discussion of the concept of relevance ultimately turns on the level of information need that is in operation, where the level of need forms the basis for evaluating the IR system or IR technique being used. Ultimately, the evaluation should also influence how the documents are represented. However, none of the reviewed authors deals specifically with the subject indexing process and the concept of subject is only touched upon lightly. Furthermore, the authors tend to focus solely on the utility of the document for the user and not on the relationship between the representation of a document and its retrieval.

Nearly every researcher who is concerned with the relevance measurement problem seems to agree that relevance is some relationship between an information need and information contained in documents. They merely disagree about the level of the information need which the information contained in documents should be held against.

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

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

As an 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 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. But if one of the latter stages is 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 as *topical relevance*.

In Taylor's framework this means that the information need moves from a need to solve a particular problem, to an information need for some particular documents. The first two stages in Taylor's categorization of needs are needs for knowledge about a particular problem and relevance judgments are here based on the individual's specific problem and previous knowledge. When the need moves to the latter stages, it is transformed from a need to solve a particular problem to a need for some particular literature on a specific subject. It is assumed that the user in the latter stages expresses his/her needs in stating a demand for literature *about* a specific subject.

When testing an IR system's ability to retrieve relevant documents it makes a difference which of the two kinds of needs that is used. Suppose a researcher is writing a paper about a group of nurses who traveled around Texas in the 1840s to visit poor farm villages to care for the ill and brought a number of books that they sold to wealthier people. This researcher now wants some more information about these nurses. This is the writer's information need. She brings this need to the IR system and reformulates it into a number of demands for information. These demands take the form of a number of sentences such as: "I want literature about farm villages in Texas in the 1840s," "I want literature about nurses' work

conditions in the 1840s,” “I want literature on the reading habits of wealthy people in Texas in the 1840s,” etc. 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 meet each of the writer’s demand for literature, although it does not solve the problem that generated the information need. Was the IR system successful? The answer to this depends 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 to retrieve *topical* relevant information (topical relevance).

Researchers mainly focus their concerns with the problem of relevance on one side of the problem, namely defining the state of the information need that the documents should be related to. They assume that the representations of the documents are given and present no problem or variable. The assumption is that only the user side of the problem is relative, and that the subject of the document can be objectively defined.

A subject representation of a document has two primary functions: 1) it should represent the subject content of a document, and 2) it should direct the users to documents which might solve their information need. The challenge is to find a balance between the two functions.

If one solely focused on the representation aspect and more or less ignored users, one might risk developing subject representations which are of no or little use for the user. An indexer who does not pay much attention to users encounters problems. The indexer might choose to represent subjects of documents which are of no interest to the user, or might use another vocabulary than the users, or might represent the subject too broadly or too narrowly for the user. On the other hand, if the indexer pays too much attention to the users of the system, the indexer might index documents in such a way that the subject representation of the documents only serves current users and current information needs.

In relevance studies the balance between the two functions is very important, and evaluation of relevance which only test 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 need 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. On the other hand is it very difficult, and rather odd, to discuss relevance without taking into account the users.

To advance research from simplistic or document oriented conceptions of indexing towards a user or requirement oriented conception, a solid theoretical

foundation for understanding the pragmatic aspects of indexing is needed. In the next chapter such a foundation is provided, and in chapter five the implication of such a new orientation in studies of representation is given.



## **4. Semiotics**

Chapter 2 argued that the subject indexing process takes a number of steps. It further suggested that these steps could be viewed as interpretations. Chapter 3, having reviewed the literature of indexing research, argued that representing documents or knowledge could not be understood independently of the documents' social contexts. For the purpose of explaining the interpretive nature of the subject indexing process, and having the pragmatic philosophy of language aspect in mind, this chapter introduces Peirce's semiotics.

Peirce's semiotics is concerned with meaning, production of meaning, and representation of meaning. The core problems in the subject indexing process have to do with meaning, production of meaning, and representation of meaning. Peirce's semiotics can therefore be used to understand the subject indexing process.

Some scholars within the LIS field have begun to find Peirce's semiotics and philosophy useful. Brier (1996) argued that semiotics together with second order

cybernetics and Wittgenstein's pragmatic philosophy of language could form the theoretical foundation for the field. Karamüftüoğlu (1996) has used semiotics to analyze the information retrieval process, and Warner (1990) has noticed that there is a conceptual overlap between semiotics and LIS, which has not yet been investigated thoroughly. In a NSF funded research project, Pearson and Slamecka (Pearson 1980; Pearson and Slamecka 1977) used Peirce's semiotics to form the foundation of a pragmatic approach to programming and understanding information systems.

Perhaps the best known discussion of semiotics in LIS is Blair's (1990) analysis of language and representation problems in information retrieval. In his book *Language and Representation in Information Retrieval*, Blair argued that theories of indexing and retrieval have to include explicit theories of language and meaning in their foundation. Blair especially used Wittgenstein's pragmatic philosophy of language for understanding information retrieval.

The major part of Blair's book is an analysis of the importance of language in indexing and representation. Blair argues that Wittgenstein's philosophy of language has significant bearings on the understanding of indexing and representation of documents. However, Blair rejects semiotics as a possible foundation for understanding indexing and information retrieval. He argues that "semiotics begin from the perspective that certain words/expressions exist and

that they need explanation” (Blair 1990, 145). This may be true of Saussure’s *semiology*, but not of Peirce’s *semiotics*. Semiotics, in Peirce’s understanding, can be defined as the study of meaning, what meaning is, how and where meaning comes into existence and how meaning is transformed and combined. Semiotics does not focus on what a specific phenomenon means, but rather on why meaning exists. Instead of semiotics Blair argues that the later Wittgenstein’s theories are useful as a foundation for understanding how to represent documents for retrieval. If Blair had been aware of the difference between Peirce’s semiotics and Saussure’s semiology, he would have seen that Peirce’s and the later Wittgenstein’s philosophies are very much alike.

Hoopes (1991) has discussed the usefulness of semiotics and has noticed that humanist scholars have described language and thought as processes which greatly resemble Peirce’s semiotic conceptions. Hoopes finds that these humanistic scholars’ orientation toward language and meaning could conveniently be grouped under a broad rubric called the interpretive revolution. Hoopes (1991, 3) further argues that this group of scholars could benefit greatly from Peirce’s philosophy,

The spokespersons for these methodologies attribute an extraordinary degree of either freedom or transitoriness to human thought, language and interpretation—with little explanation of how such a universe is possible or

of how reasonable honest and thoughtful people could have wrongly overestimated the constraints within which human beings live. These methodological radicals would do well to consider Peirce's semiotic realism, according to which interpretation proceeds under the weight of many natural, logical limits.

This chapter will introduce the framework Hoopes suggests for understanding and describing precisely what interpretation is. After an initial definition of Peirce's semiotics and a brief biographical introduction to Peirce, the chapter will introduce and explain three aspects of Peirce's semiotics: the sign itself, the concept of unlimited semiosis, and the ten categories of signs. The latter is by far the most complex of the three aspects and will require a short introduction to Peirce's phenomenology and fairly long and detailed explanations.

This introduction to semiotics will provide a framework for understanding the many kinds of interpretations and hence provide a tool for further investigations into the interpretive nature of the subject indexing process.

#### **4.1. DEFINITION OF SEMIOTICS**

Semiotics is generally defined as the *study of signs*. Two traditions of the study of signs can be identified, a European and an American. The European tradition is based on the work of the French linguist Ferdinand de Saussure (1857-

1913). This school is usually named *semiology*. The American tradition is based on the work of the American philosopher Charles Sanders Peirce (1839-1914) and is called *semiotics* (or semeiotic, as Peirce preferred to spell it). Both traditions are concerned with the meaning of signs, but unlike Saussure (1966) who was primarily concerned with semantics, Peirce was also concerned with how signs are attributed meaning. Although there have been attempts to define a unified theory of semiotics, most notably by Eco (1984; 1990), the two traditions are distinct, even though they might complement each other. Saussure's theory is a theory of how to derive meaning from words. Peirce's theory, on the other hand, is about how signs in general, and not only words, are attributed meaning.

Johansen (1985, 225-26) has discussed the distinction between the two traditions.

As a contra distinction to the concept of sign of continental structuralism (Saussure, Hjelmslev), defining the sign as an immanent solidarity between two formal entities (an element of expression and one of content), Peirce conceives the sign as an element in a signifying process.

In short, Saussure operated with a dual concept of the sign. He suggested that words are not merely names that represent *things*, but are expressions that *stand for* some content. By this he separated words and their content. Saussure argued

against the notion that words have an inherent quality, as earlier linguistics had suggested. Instead he argued that the connection between a word and its content is *arbitrary*. His theory is centered on how to *derive* meaning from words.

#### **4.2. C. S. PEIRCE: A BIOGRAPHICAL SKETCH**

Charles Sanders Peirce was born in 1839 in Cambridge, Massachusetts, and died in 1914 in Milford, Pennsylvania. He is known as a philosopher, scientist, and mathematician, but perhaps most importantly as one of the founders of the philosophical tradition generally called ‘pragmatics.’ Peirce himself named the tradition ‘pragmaticism’ (Peirce 1958), because he wanted to distinguish his own logical formulation of pragmatism from the psychological version of pragmatism found in the works of William James (Wiener 1958).

Peirce received a degree in chemistry in 1863 from Harvard University, where his father, Benjamin Peirce was professor of mathematics and astronomy. In 1861 he began work for the US Coast and Geodetic Survey, and he remained in its service until 1891.

Peirce was instructor of logic at Johns Hopkins University, the first graduate school in the US, from 1879-1884. After five years of teaching, to no more than a dozen students in any of his courses, he was not re-appointed, and he never

resumed a teaching position at any college or university. However, he gave lectures on philosophy and logic at various universities throughout his lifetime (Wiener 1958).

The only book Peirce published in his lifetime, *Photometric Researches*, first appeared in 1878. He wrote this book during an assistantship at Harvard Observatory.

Peirce devoted his retirement to philosophical work. From 1900 until his death in 1914 he lived with his second wife, Juliette Tourtalai, in poverty and ill health, and on the charity of relatives and friends, such as William James.

Although it is commonly said that Peirce published little--perhaps based on the fact that he only published one book in his lifetime--he actually wrote more than 10,000 pages which are scattered among various journals, periodicals, scientific annuals, dictionaries, and encyclopedias. He contributed to many sciences--mathematics, chemistry, astronomy, and psychology--as well as philosophy, where he is primarily known for his contributions to logic and metaphysics (Hoopes 1991). Besides these writings, he wrote much more which was never published. After his death, his wife sold his manuscripts to Harvard University for \$500; the manuscripts still remain in the Houghton Library at Harvard. It was not until 1951 that his work became accessible through the first of six volumes of "Peirce's Collected Papers."

Since then a number of other collections have been published. The newest of these enterprises is pursued by Mary Keeler from the Center for Advanced Research and Technology in the Arts and Humanities of the University of Washington, and Christian Kloesel, Indiana University and Purdue University, who will make Peirce's manuscripts accessible on CD-ROMs (Keeler and Kloesel 1997).

#### **4.3. THE SIGN**

Central to Peirce's semiotics is the concept of *sign*. 'Sign' is a common word and therefore has a colloquial meaning, but various scholars use it in more technical senses. In the following, Peirce's definition of the concept will be introduced and it will thereafter only be used to denote Peirce's definition of the word.

Peirce (1955) defined a sign as a relation among *three* entities, the sign itself, the referent of the sign, and the meaning which is derived from the sign. Peirce's concern was how meaning is derived from a sign and transformed into another sign. He operated with a three-sided, or a *triadic* concept of sign, which he (Peirce 1955, 99) defined as,



A sign, or *representamen*, is something that stands to somebody for something in some respect or capacity. It addresses somebody, that is, creates in the mind of that person an equivalent sign, or perhaps a more developed sign. That sign which it creates I call the *interpretant* of the first sign. The sign stands for something, its *object*. It stands for that object, not in all respects, but in reference to a sort of idea.

Peirce distinguished between the physical entity, for example *words*, the *idea* that these words refer to, and the *meaning* one derives from words. He refers to the possibility that people will derive different meanings from the same words.

Peirce's concept of a sign is represented as a triangle, as shown in figure 4-1, which is based on a figure by Johansen (1993). The triangle is sometimes referred to as the Ogden Triangle, although it is evident that Ogden and Richards (1923) got their inspiration from Peirce (Fisch 1986, 344).

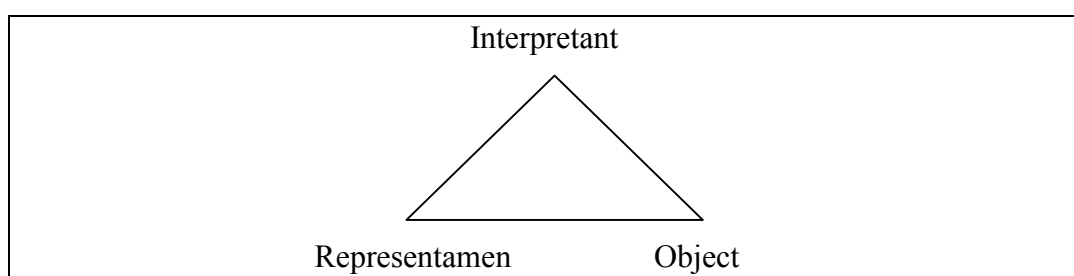


Fig. 4-1. The Semiotic Triangle

The representamen is that which represents the sign, often in the form of a physical entity or at least manifested in some form. The representamen is, in other words, the entity of the sign relation that is perceived and therefore often denoted the “sign.” Peirce contributes to this confusion, for instance, when he in the definition of the sign above writes, “the sign, or *representamen*, is...” (Peirce 1955, 99), he thereby suggest that the sign and the representamen are the same. However, later he states that the three entities, which he calls members, of the sign relation constitute the sign as a whole, at one point, for instance, he writes that “the triadic relation is *genuine*, that is its three members are bound together” (Peirce 1955, 100).

The representamen represents an object. However, it is not a one-to-one relationship between the representamen and an object, the object is not some identifiable entity that exist independent of the sign. Peirce (1955, 101) states about the object that,

The Objects—for a Sign may have any number of them—may each be a single known existing thing or thing believed formerly to have existed or expected to exist, or a collection of such things, or a known quality or relation or fact, which single Object may be a collection, or whole of parts, or it may have some mode of being, such as some act permitted whose being does not prevent its

negation from being equally permitted, or something of a general nature desired, required, or invariably found under certain general circumstances.

The sign can only represent the object and tell about it, it cannot furnish acquaintance with or recognition of the object. The object, therefore, is not some objective entity that exists and which can be known or realized through the sign. The object is “that with which . . . [the sign] presupposes an acquaintance in order to convey some further information concerning it” (Peirce 1955, 100). The object should be understood as the background knowledge that one needs to understand the sign, or the range of possible meaningful statements that could be made about the sign.

The connection between the representamen and its object is made by the interpretant, which is the third entity in the sign relation. The interpretant is *not* a person who interprets the sign, but rather the sign that is produced from the representamen. In other words, when the representamen is perceived as a sign, a new and more developed sign is created on the basis of the representamen. In other words, the person who interprets the sign makes a connection between what she sees (which is the representamen) and her background knowledge (which is the object) and thereby creates an understanding or meaning of the sign (which is

the interpretant). This process is called *semiosis*. Semiosis is the act interpreting signs.

The connection of the representamen and the object to create the interpretant as a process of semiosis is emphasized in the Y-leg model of the sign in figure 4-2. This figure is based on a figure by Larsen (1993).

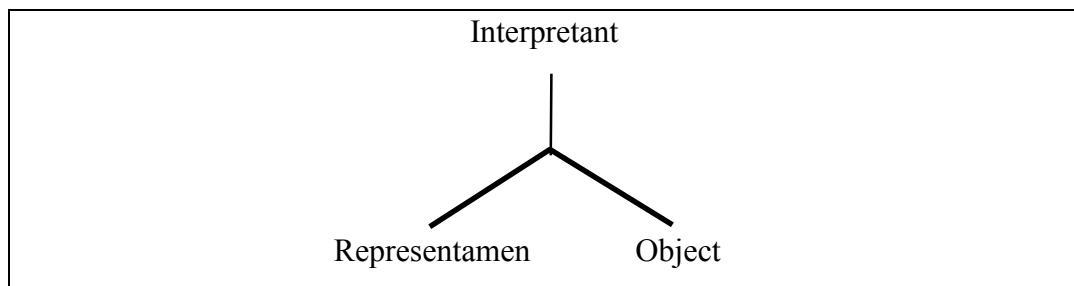


Fig. 4-2. The Y-leg Model

The bold line from *representamen* to *object* stresses the connection between the primary sign (the representamen) and its referent (the object). The connection between these two entities is the meaning of the representamen. The meaning is represented as the interpretant. This idea is stressed in the Y-leg model, but is less clear in the semiotic triangle. Throughout this dissertation both models will be used, however.

Peirce's concept of sign can hence be characterized as *realistic*, in the sense that Peirce argued that there is a material world independent of our perceptions. At the same time it could also be characterized as *idealistic*, in the sense that Peirce argued that our perception of the material world is the only knowledge we can gain about it.

These two conceptions, realism and idealism, are sometimes seen as opposites. However, it is important to make clear that Peirce's worldview combined realism (the view that objects of sense impressions exist independently of the observer) and idealism (the view that reality is essentially a property of the mind rather than a phenomenon itself). He did so by claiming that we always observe reality (which exists before and independently of any encounter with it) from a 'semiotic distance.' Reality really exists out there independently of humans, but we will only learn about it through our interpretations of it.

Hookway (1985, 37) discusses Peirce's arguments for defining his view as realistic.

The motivation for calling the view realist is that it permits Peirce to reject a nominalist view of universals or 'generals.' We must free ourselves from the nominalist prejudice that the only things that are real are particulars.

Hookway concludes that Peirce's arguments that what is real is what is knowable would generally ally him with views that are considered anti-realistic. Wiener, on the other hand, argues that Peirce rejects the subjectivism of idealism. Wiener (1958, xi) argues that Peirce condemned the subjectivism of psychological approaches.

The relativizing of all knowledge and reduction of all meaning to sensations or ideas leaves us with the particular things of experience and uniformities or laws as final fixtures of reality. This relegates the universal traits of existence and general principles of logic, science, and ethics to the passing figments of the mind.

Ultimately, Peirce distinguished between the subjective meaning of our beliefs and the objective properties of what our beliefs imply about things and events. The latter are independent of our thoughts and labels. This distinguishes Peirce from William James' definition of pragmatics. As Mounce (1997) recently showed, there are in fact two different traditions of pragmatics.

Peirce's own epistemological standpoint is variously referred to as objective idealism (Hookway 1985, 143), evolutionary objective idealism (Hookway 1985, 264), medieval realism (Eco 1979, 193), common sense realism (Wiener 1958, viii), and, perhaps most appropriately, semiotic realism (Hoopes 1991, 9).

#### **4.4. UNLIMITED SEMIOSIS**

A key element in Peirce's theory of semiotics is the notion of 'unlimited semiosis'. This notion could be seen as the connecting of sign or the process that one sign produces another sign. This is called semiosis. Unlimited semiosis is based on the fundamental idea of semiosis; that a sign (b) is generated on the basis of another sign (a). When a new sign (c) is generated on the basis of the second sign (b), still another semiosis process occurs. Because new signs will always generate still more signs, this process can continue indefinitely. The process is, therefore, unlimited, hence the term, unlimited semiosis.

The unlimited semiosis process is represented in figure 4-3 (based on a model by Johansen [1993, 80]). The interpretant of the first sign in unlimited semiosis changes to become the representamen in the second sign. There is a relation between these, but the object in each case remains independent of both the representamen and the interpretant. The object will change throughout the process. Each object relation in the unlimited semiosis process will be unique to that sign relation. The single objects in the unlimited semiosis process are independent of each other. The latter is crucial for Peirce's theory of semiotics.

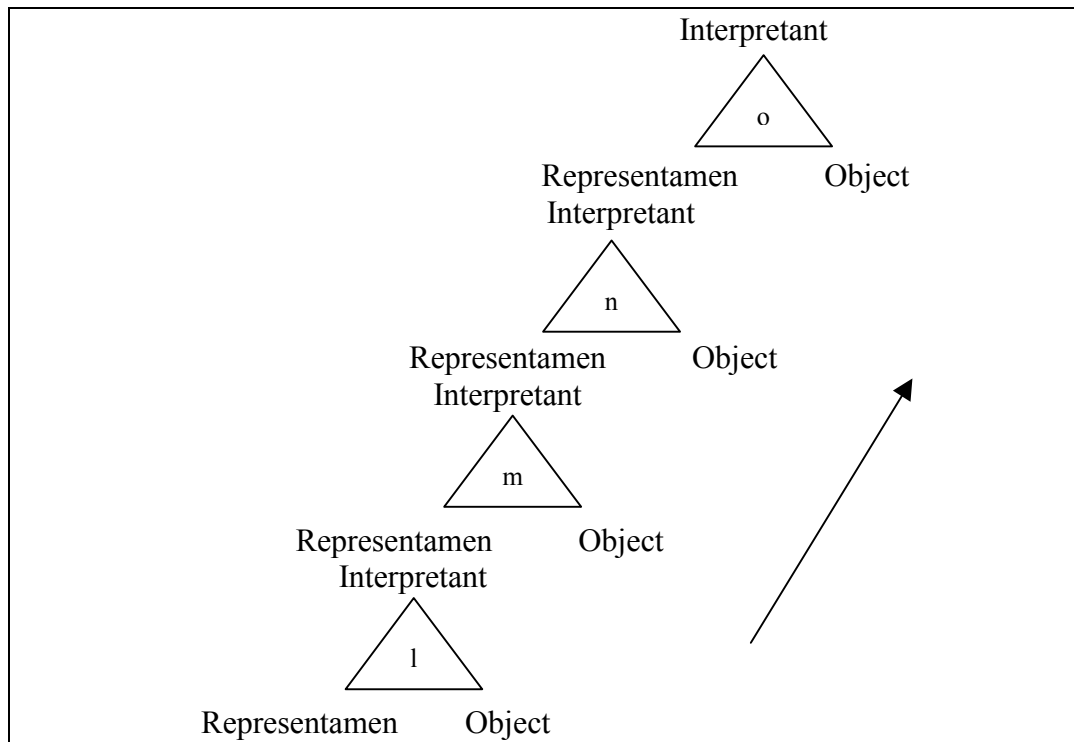


Fig. 4-3. Unlimited Semiosis

In figure 4-3, the triangles in the figure will continue to generate new triangles. It should be understood that there have been triangles before them, and there will be triangles after them. What this means is that an understanding of something is always based on an understanding of something else; and it will always generate still another understanding.



By way of illustrating this, suppose that Jones and Smith are having a conversation when a fat cat walks by. Jones looks at the cat. The act of looking at the cat is the first sign relation. At this point, the cat is the representamen. A cat stands for a range of ideas and meanings related to “catness.” In Jones’ mind, looking at the cat creates an interpretant (i.e. his interpretation). This interpretant is created on the basis of the representamen, the cat, and limited by its object, “catness.” Jones then utters “isn't that a fat cat?” This creates a second sign relation. The sentence “isn't that a fat cat?” is the new representamen, and the object is the range of ideas and meanings associated with this sentence, which are of course large. These ideas and meanings have to do with “fatness,” “catness,” cats in general, the particular cat itself, etc. The interpretant of the second sign relation is Smith's understanding of the sentence. There are, of course, several other interpretants created by the same sentence. For instance, Jones will also create an interpretant, on the basis of his own sentence. In order to keep this illustration simple, however, only one sign relation will be followed here. If Smith then utters, “well it isn't as fat as Mrs. Swan's dog,” a third sign relation is created. Here the sentence “well it isn't as fat as Mrs. Swan's dog” is now the new representamen. The object now becomes the range of ideas and meaning related to this sentence. These are numerous, including “dogness,” “fatness,” Mrs. Swan,

cats and dogs in general, etc. The new interpretant is the meaning, which Jones now generates from the sentence.

Two important ideas are illustrated here. First, the different sign relations have different objects, each of which is solely dependent on the person for whom the interpretant is created. There are no necessary relations between the different objects throughout the conversation. Second, each representamen is based in turn on an interpretant, which again is based on a representamen.

Actually, it could be argued that there are even more sign relations in the above sequence. The change from interpretant to representamen could also be seen as a sign relation. This, of course, all takes place within the mind of the person for whom the changing of the interpretant to a representamen occurs and is, therefore, not accessible for inspection. This is worth noting, however, because it illustrates the complexity of the process.

#### **4.5. PHENOMENOLOGY**

Peirce categorized signs into a number of categories to illustrate their different kinds. One set of sign categories commonly associated with his work consisted of icon, index, and symbol. This approach to categorization grouped signs on the basis of their relation to their referent/object. In this respect, an icon

sign is based on resemblance (like the sign on a bathroom door), an index sign points to what the sign refers to (like smoke to a fire), and a symbol sign refers to a convention (like language).

The categorization into *icon*, *index*, and *symbol* is a simple representation of Peirce's full categorization of signs. To reach this categorization, Peirce defined three modes of each entity (interpretant, representamen and object) of the sign. The three modes of each entity are based on Peirce's phenomenology, in which he argued for a division of the world into three modes of phenomena—or three modes of being. Before Peirce's categorization of signs is more fully explored, however, his phenomenology must be introduced. Only after that will his triadic division of the sign be presented.

#### **4.5.1. Three Modes of Being**

Peirce argues that everything which exists in the world, including feelings, ideas and thoughts, belong to one of three fundamental modes of being. These are the mode of being of positive qualitative possibility, the mode of being of actual fact, and the mode of being of law (or conventions). Peirce named these respectively firstness, secondness, and thirdness. Generally they can be thought of as:

- Firstness: Monadic, properties, qualities.
- Secondness: Dyadic, relations between two phenomena, facts.
- Thirdness: Triadic, mediated relations between three phenomena, thoughts.

*Firstness* is the mode of being that consists of the category of qualities of phenomena, such as red, bitter and hard. These are monadic properties. Members in this category are mainly seen as qualitative possibilities. The understanding of these qualities is reasonably alike for all humans. That is, there is general agreement on what is red, bitter, or hard. Peirce (1955, 80-81) argues that this is the mode of being of things which simply just are,

[T]here are certain qualities of feeling, such as the colour of magenta, the odor of attar, the sound of a railroad whistle, the taste of quinine, the quality of the emotion of contemplating a fine mathematical solution, the quality of feeling of love, etc. I do not mean the sense of actually experiencing these feelings, whether primarily or in any memory or imagination. That is something that involves these qualities as an element of it. But I mean the qualities themselves which, in themselves, are mere maybes, not necessarily realized.

Firstness is sheer thisness, or existence of things (Hoopes 1991, 10). This existence is neither dependent on its being in the mind of some person, whether in the form of sense or in thought, nor on its being in the form of some material thing possessing the quality (Peirce 1995, 85).

*Secondness* is the mode of being that tells something about other objects. This mode of being is represented by dyadic properties, the actual facts about the phenomena. These properties make it possible to recognize and identify the phenomena directly. Secondness can be illustrated by a situation where a person leans against a locked door. The person concludes that because the door cannot open somebody must have locked it. Peirce (1955, 88) explains the difference between experience and perception.

Some writers insist that all experience consists in sense-perception; and I think it is possibly true that every element of experience is in the first instance applied to an external object . . . We perceive objects brought before us; but that which we especially experience--the kind of thing which the word "experience" is more particularly applied--is an event. We cannot accurately be said to perceive events.

The concept of experience is, according to Peirce (1955, 88), broader than the concept of perception. If we hear a sound, say a whistling locomotive, we

perceive the whistle, the change in notes. We might experience the perception of the whistle differently. But the perception is the same, the perception is factual. The perception is the same for everyone but the experience differs. Experience, therefore, includes much, which is not perception.

Secondness is the relations between things (Hoopes 1991, 10), Peirce furthermore describes secondness as facts. It is the direct relation between things, for instance, between the whistling locomotive and the perception of the whistle.

*Thirdness* is the triadic relation between something first and something second. This relation reveals information about something third. This can generally be defined as *meaning*. Meaning is not inherent in signs, but something one *makes* from signs. Language is based on habits. Or more correctly, people think of habits as laws. Thirdness is the class of these habits, which, at the same time, are the vagueness in language and our ability to communicate. If a person utters “Ms Jones is terribly slow,” it is not clear if the person means physically or mentally slow. From the context, however, it will be clear what the person means. Peirce (1955, 91) argues that the meaning of words and our actions are closely related,

It is impossible to resolve everything in our thought into those two elements [of firstness and secondness]. We may say that the bulk of what is actually done consists of secondness -- or better, secondness is the predominant

character of what *has been* done. The immediate present, could we seize it, would have no character but its firstness. . . . In general, we may say that *meanings* are inexhaustible. We are too apt to think that what one *means* to do and the *meaning* of a word are quite unrelated meanings of the word “meaning,” or that they are only connected by both referring to some operation of the mind.

Peirce sometimes speaks of thirdness as the category of law (e.g. Peirce 1955, 90), by this he means that thirdness is a relation between two things, which is established by humans. The relation between the word “cat” and real living cats is a relation, which has been established by humans. This relation does not exist independently of social contexts, as the two first categories do. Thus thirdness can be said to be representational (Hoopes 1991, 10).

#### **4.6. CATEGORIES OF SIGNS**

Before a categorization of signs can be developed, the three aspects of the sign (the representamen, the object, and the interpretant) must be divided into three levels, called *trichotomies*. The trichotomies reflect the aforementioned phenomenology, viz. firstness, secondness, and thirdness.

In the following section the three trichotomies will first be presented and defined. Then the ten sign categories, which Peirce developed from the

trichotomies, will be presented and defined. Lastly the relations between the ten categories will be discussed to explain their differences and define the categories individually.

#### **4.6.1. The Trichotomies**

Each sign consists of *three* entities: representamen, object and interpretant, which all must be present to make a sign. The three entities of the sign have three elements, which reflect the three modes of being; firstness, secondness and thirdness. This trisection of the sign, and the further trisection of its components is essential to Peirce's semiotics. The trisection of the trisection is represented graphically in figure 4-4 (based on a figure by Christiansen [1988]).



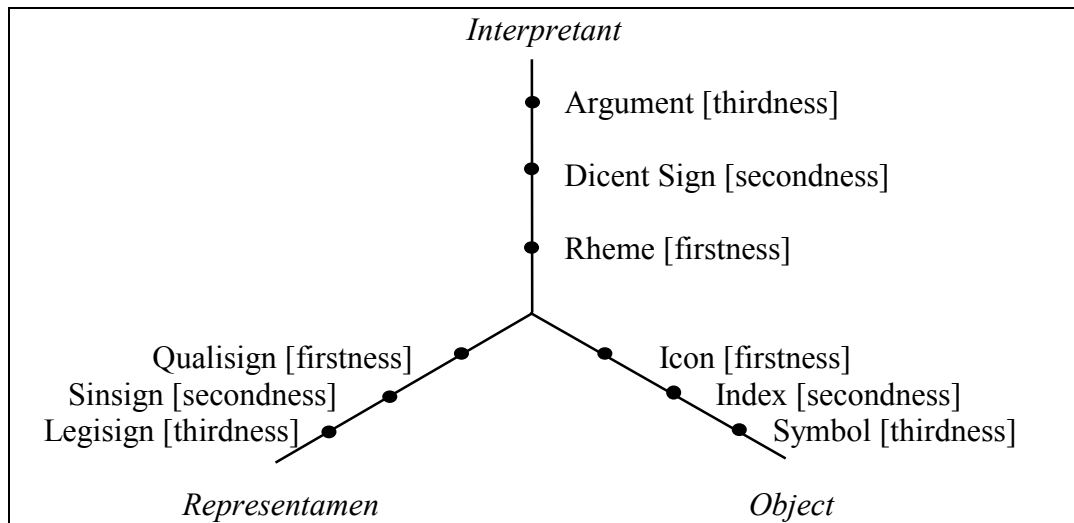


Fig. 4-4. Triadic Classification of Signs

The three inner categories--*rheme*, *icon*, and *qualisign*-- represent firstness. The middle categories--*dicent sign*, *index*, and *sinsign*--represent secondness. And the outer categories--*argument*, *symbol*, and *legisign*--represent thirdness. These categories will be introduced here.

The following definitions of the categories and division of categories are based on Peirce (1955, 101-104) and Merrell (1995, 93-96).

- The *representamen* is divided according to whether the sign itself is a mere quality (*Qualisign*), an actual existent (*Sinsign*), or a convention (*Legisign*).

1. A Qualisign is a quality, which is a sign. It is an appearance of quality before someone is conscious of it. It cannot act as a sign until it is embodied although its “embodiment has nothing to do with its character as a sign” (Peirce 1955, 101).
2. A Sinsign (the syllable *sin* is derived from singularity as in *single*, *simple*, etc.) is an actual existent thing, an individual object, an act, or an event. “It can only be so through its qualities; so that it involves a qualisign, or rather, several qualisigns. But these qualisigns are of a peculiar kind and only form a sign through being actually embodied” (Peirce 1955, 101). In other words the sinsign is thisness, or in philosophical terms, *haecceity*, which is an fundamental actuality of an entity. Therefore the sinsign is “a once-occurring token” (Merrell 1995, 93) in the sense it represents specific objects, acts or events.
3. A Legisign is a general type, law, habit, or convention, which is established by humans. Conventional signs are always legisigns. Qualisigns are qualities, which are regarded as signs. Signs are individual instances of particular objects, acts, or events. And legisign are only actual in the sense that they have been described. “Thus, the word ‘the’ will usually occur from fifteen to twenty-five times on a page. It is in all these occurrences one and the same word, the same legisign. Each

single instance of it is a replica. The replica is a sinsign. Thus, every legisign requires sinsigns. But these are not ordinary sinsigns, such as are perculiar occurrences that are regarded as significant. Nor would the replica be significant it were not for the law which renders it so “ (Peirce 1955, 102). In other words, the meaning of the different occurrences of the word ‘the’ is determined by the convention for using the word in the particular instances.

For instance the sign ‘A’ could be considered 1) black lines or the quality of black ink on paper (i.e., a qualisign), 2) a good example of the class of letter ‘A’, which would be an actual existent (i.e., a sinsign), or 3) an expression of satisfaction with a term paper, that is, a convention (i.e., a legisign).

- The *object* is divided according to the sign’s relation to the object it represents. The sign could either have some character in common with its object (icon), some existential relation to that object (index), or only have a representational relation to its object (symbol).
- 1. An Icon is a sign that refers its object merely by sharing some property with its object whether any such object actually exists. “It is true that unless there really is such an object, the Icon does not act as a sign; but this has nothing to do with its character as a sign” (Peirce 1955, 102). What

makes something an Icon is that there is some kind of likeness between the Icon and that which the Icon represents. An Icon can therefore be a quality (a Qualisign), a specific object, act or event (a Sinsign), or a convention (a Legisign).

2. An Index is a sign, which refers to its object by being affected by that object. Therefore the sign has “some actual or imagined relation to its object” (Merrell 1995, 93). An Index, therefore, cannot, be a qualisign, because qualities are independent of anything else. “In so far as the Index is affected by the object, it necessarily has some quality in common with the object, and it is in respect to these that it refers to the object. It does, therefore, involve a sort of an Icon, although an Icon of a peculiar kind; and it is not the mere resemblance of its object, even in these respects which makes it a sign, but it is the actual modification of it by the object” (Peirce 1955, 102). In sum, the Index refers to its object by a direct connection between the sign and what it represents.
3. A Symbol is a sign, which refers to its object through law, habit, or convention. This usually takes place through an association of general ideas by which the Symbol is interpreted as referring to its object. Since the Symbol is a general type or law it will always be a Legisign. It therefore produces an interpretation of the object, which does not share

characters with the object (as the Icon) or is directly affected by the object (as the Index). This interpretation is called a replica. It is not only a general law or convention itself, but the object to which it refers is of general nature. “Now that which is general has its being in the instances which it will determine. There must, therefore, be existent instance of what the Symbol denotes, although we must here understand by ‘existent,’ existent in the possibly imaginary universe to which the Symbol refers” (Peirce 1955, 102-103). The Symbol will, of course, be affected by those instances and therefore involve an Index. But the replica of the symbol will only indirectly and through an association be affected by the object.

An example of an icon is a pictogram where the sign resembles the object.

An example of an index is a footprint, which points to a person. And an example of a symbol would be a sign based on the context in which it occurs, such as a street sign with the letter P, which means “parking allowed”.

- The *interpretant* represents the sign as a sign of possibility (rheme), a sign of fact (dicent sign), or a sign of reason (argument).
  1. A Rheme is a sign with qualitative possibilities, that is, understood as representing a certain kind of possible object. It is a sign “of quality or feeling” (Merrell 1995, 93). “Any Rheme, perhaps, will afford some

information; but it is not interpreted as doing so” (Peirce 1955, 103). The meaning of a Rheme is easily understood, and the understanding yielded is almost alike for all humans. It requires almost no knowledge of context to interpret a Rheme.

2. A Dicent Sign is a sign of actual existence or “of a statement of presumed fact” (Merrell 1995, 93). The Dicent sign, therefore, cannot be an Icon, because the Icon provides no ground for an interpretation of its referent. “A Dicent Sign necessarily involves, as part of it, a Rheme, to describe the fact which it is interpreted as indicating. But this is a peculiar kind of Rheme; and while it is essential to the [Dicent Sign], it by no means constitutes it” (Peirce 1955, 103). The Dicent Sign is more complex than the Rheme, which means that it requires more knowledge to determine the meaning of the Dicent Sign than to interpret a Rheme.
3. An Argument is a sign of reason or law and is understood to represent its object in its character as sign. This is opposed to a Rheme, which is a sign that represents its object in its characters merely, and a Dicent Sign which is a sign that is understood to represent its object in respect to actual existence. The truth of the Argument is a mental act of the person who seeks the truth in the Argument. But the focus should not be on the interpreter’s mental operation. Instead it should be on the interpreter’s

actions. Only this will determine the truth of the Argument. The Argument should, therefore, be “contemplated as a sign capable of being asserted or denied. The sign itself retains its full meaning whether it can be actually asserted or not. The peculiarity of it, therefore, lies in its mode of meaning; and to say this is to say that its peculiarity lies in its relation to the interpretant” (Peirce 1955, 104).

Examples of these three are: rhemes are nouns (e.g. ‘house,’ ‘car’), sinsigns are propositions (e.g. ‘the house is green,’ ‘the car is fast’), and legisigns are arguments, i.e. meaningful links of propositions (e.g. ‘Jones has a green house and a fast car. Smith on the other hand does not like to drive and therefore prefer to bike...’).

The above examples are all rather weak since any sign is defined as a combination of the elements of the sign, such that each sign consists of one element from each trichotomy.

By combining the different categories of the sign elements Peirce defined ten categories of signs. Although a total of  $3^3$ , or 27, different categories could be defined<sup>13</sup>, Peirce only enumerated ten, since some possible signs are logically

---

<sup>13</sup> Many other numbers have been considered, Seboek (1994) for instance, expanded the three basic signs--icon, index, and symbol--into six signs. Marty (1982) enumerated twenty six, and Weiss and Burks (1945) sixty six. Since each sign possess its own triads, Peirce argues that a total of  $3^{10}$ , or 59,049, signs could be enumerated (Merrell 1997).

excluded. A qualisign will for instance always be a rhematic icon (because a mere quality cannot be a convention). A symbol will always be a legisign (a symbol is a representation of its object based on context, and a legisign is a sign based on convention). An argument will always be a symbolic legisign (since an argument always is thirdness to the interpretant and requires a high degree of interpretation).

#### **4.6.2. The Individual Categories**

Each of the ten categories of signs is loosely defined. Though they are not clearly distinguished, they can be viewed as ten points on a continuum from the mere sense of a feeling to a complex statement. These ten categories of signs are presented next.

The following description of the ten categories is based on Peirce (1955, 115-118) and Merrell (1997, 193-195).

First: A *Qualisign* “is any quality in so far as it is a sign” (Peirce 1955, 115).

It is “a feeling, a sensation, for example, the sense of ‘blueness’ upon one’s being subjected to a blue object” (Merrell 1997, 193).



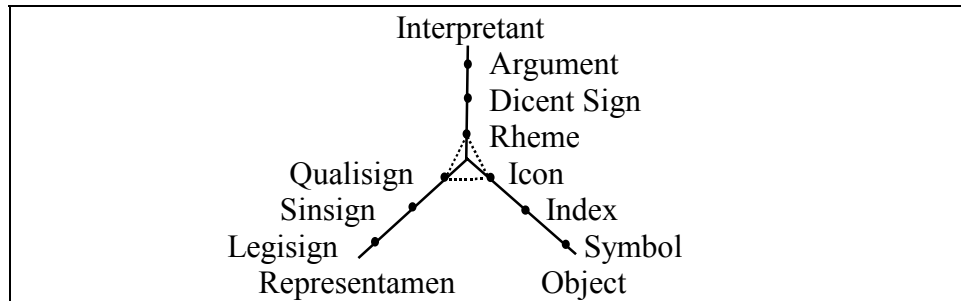


Fig. 4-5. Qualisign

A qualisign is firstness, an Icon, in relation to the object, but this relation should be understood as firstness of the ‘semiotic *object*,’ “so labeled in order to distinguish it from the ‘real object’ *an sich*” (Merrell 1997, 193). A Qualisign can, therefore, only denote an object by virtue of some similarity with the object.

A qualisign will always be a Rheme because a quality is a mere possibility and can only be interpreted as a sign of essence. The qualisign is firstness of the interpretant, which means that the meaning derived from the sign will be more or less alike for all humans. The Qualisign is firstness of the representamen, a rheme, which means that all humans will perceive and experience the sign alike.

The Qualisign is pure firstness, in the sense that it is firstness in all aspects and therefore as close as possible to being ‘the same’ as its object.

Second: An *Iconic Sinsign* is any “object of experience in so far as some quality of it makes it determine the idea of an object” (Peirce 1955, 115). The Iconic Sinsign is “something not yet clearly distinguished from something else. For instance, a diagram apart from that to which it is possibly related; a shape, however vague it may be at this stage of semiosis” (Merrell 1997, 194).

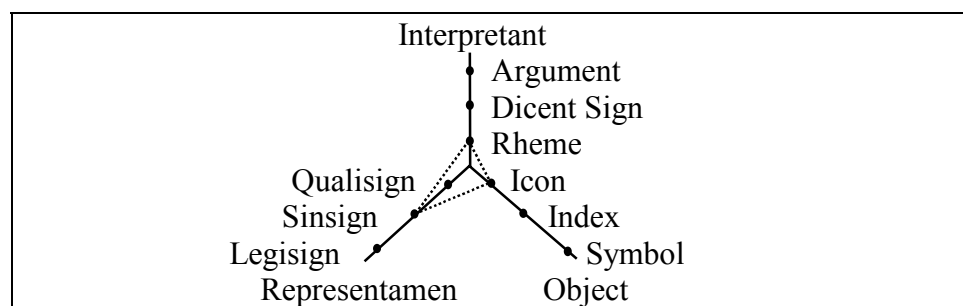


Fig. 4-6. Iconic Sinsign

The Iconic Sinsign is firstness of the object, an Icon, and thus a sign of likeness of whatever it represents. Such a sign can only be interpreted as

a sign of essence, a Rheme. It is therefore, as the qualisign, firstness of the interpretant.

The difference between the Qualisign and the Iconic Sinsign is that the Iconic Sinsign's relation to the representamen changes from firstness (qualisign) to secondness (sinsign), in other words from a quality to an actual existence.

Third:     A *Rhematic Indexical Sinsign* is any “object of direct experience so far as it directs attention to an object by which its presence is caused” (Peirce 1955, 115). The Rhematic Indexical Sinsign involves a kind of an Iconic Sinsign, but they are quite different since the Rhematic Indexical Sinsign brings the interpreter's attention to its object. Where the Iconic Sinsign resembles its object (since it is an Icon), the Rhematic Indexical Sinsign is an Index and therefore points to its object.

A Rhematic Indexical Sinsign is a Rheme (firstness of Interpretant) and therefore requires little knowledge of the interpreter, and it is a Sinsign (secondness of Representamen) and therefore is an actual existent.

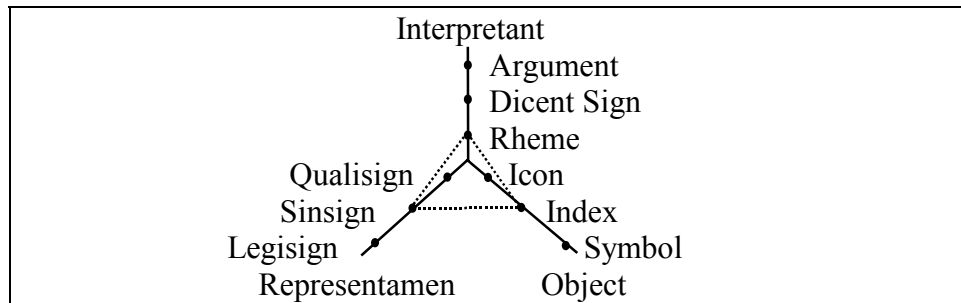


Fig. 4-7. Rhematic Indexical Sinsign

A Rhematic Indexical Sinsign can be characterized as “a recognition or sudden jolt of surprise regarding the existence of the sign (a spontaneous cry, or attention abruptly drawn toward an unexpected object, act, or event)” (Merrell 1997, 194).

Fourth: A *Dicent Sinsign* is “any object of direct experience, in so far as it is a sign, and, as such, affords information concerning its object” (Peirce 1955, 115).



Fifth: An *Iconic Legisign* is “any general law or type of sign, insofar as it manifests some likeness with something other than itself. It further requires that each of its replicas incorporate a definite quality which renders it fit to evoke in the mind the idea of its respective semiotic *object* (e.g. a diagram coupled with and related to that of which it is a likeness, a form and the idea is a form of)” (Merrell 1997, 194).

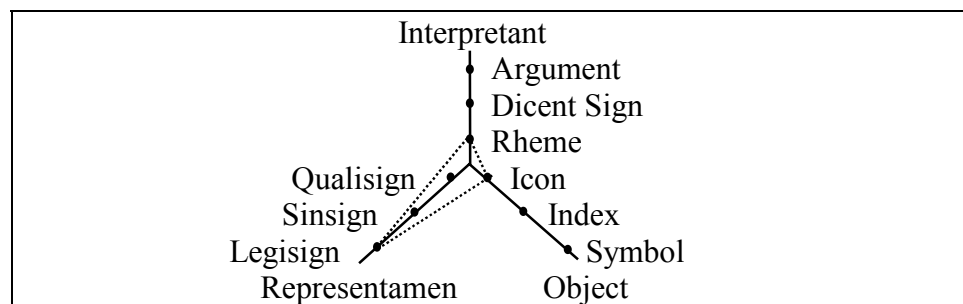


Fig. 4-9. Iconic Legisign

Signs, which are Icons, will always be Rhemes. The Icon (firstness of the object), represents its object through resemblance, and it is therefore easy to interpret. An Icon will, therefore, always be a Rheme which is firstness of the interpretant. The Iconic Legisign represents through likeness and is therefore an Icon and a Rheme.

“Being a Legisign, its mode of being is that of governing single replicas, each of which will be an Iconic Sinsign of a peculiar kind” (Peirce 1955, 116). Legisigns are conventional signs. The meaning of the Iconic Legisign, thus, is a convention or law. However, the understanding of the Iconic Legisign will be more or less alike for all humans. It could for instance be a sign on a restroom door signifying gender.

Sixth:     *A Rhematic Indexical Legisign* is “any general type or law of sign, however established, which requires each instance of it to be really affected by its [semiotic] object” (Peirce 1955, 116). It does so by merely drawing attention to the object it indicates. The object, however, remains distinguished from the sign.

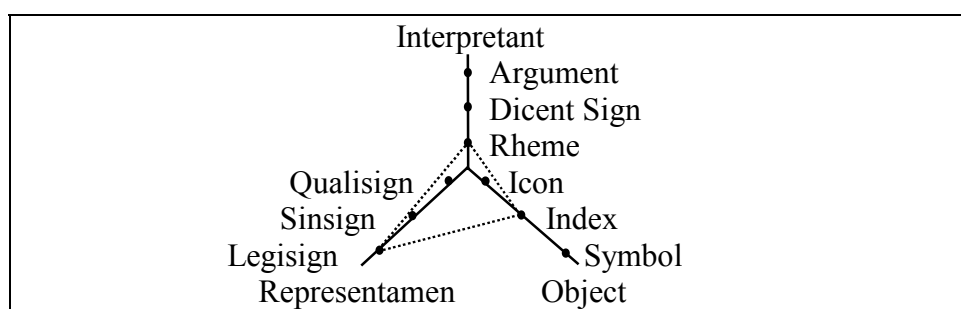


Fig. 4-10. Rhematic Indexical Legisign

The Rhematic Indexical Legisign is secondness to the object. It therefore represents by drawing attention to the object. It is, however, firstness to the interpretant, and therefore requires little or no special knowledge of the interpreter. It is a Legisign and is therefore a convention or a social or cultural construct. In other words, everyone within a specific social or cultural context will understand the Rhematic Indexical Legisign with ease and alike.

Examples of Rhematic Indexical Legisigns are a “demonstrative pronoun, the Mercedes Benz logo standing in for a car, an image of a cup standing in for coffee, and a photo of Michael Jordan standing in for Nike” (Merrell 1997, 194).

“Each replica of it will be a Rhematic Indexical Sinsign [sign category number 3] of a peculiar kind. The interpretant of a Rhematic Indexical Legisign represents it as an Iconic Legisign [sign category number 5]; and so it is, in a measure—but in a very small measure” (Peirce 1955, 116). When we interpret Rhematic Indexical Legisign, we do it with ease and are therefore not conscious about the complexity of the sign.



Seventh: A *Dicent Indexical Legisign* is “any general type or law, however established, which requires each instance of it to be really affected by its object in such a manner as to furnish definite information concerning that object” (Peirce 1955, 116).

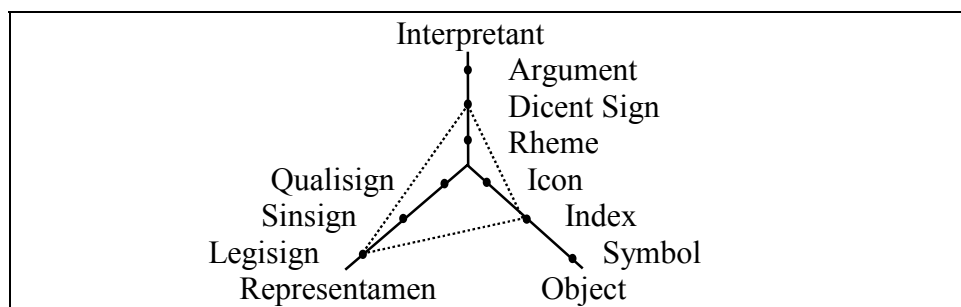


Fig. 4-11. Dicent Indexical Legisign

The Dicent Indexical Legisign will represent its object by pointing to it (a Symbol). The Dicent Indexical Legisign is based on conventions (a Legisign) and it is a sign of actual existence (a Dicent Sign).

Examples of Dicent Indexical Legisigns are “a street cry such as “Watch out!,” commonplace expressions such as “Hi!,” “You doin’ all right?,” “Bless you,” or a hand extended to open a door for someone, or to shake another hand as a salutation” (Merrell 1997, 195).

The Dicent Indexical Legisign involves an Iconic Legisign (sign category number 5) to signify the information it yields, and a Rhematic Indexical Legisign (sign category number 3) to denote the subject of that information. Interpretations of the Dicent Indexical Legisign will be a Dicent Sinsign (sign category number 4) of a peculiar kind.

Eight:      A *Rhematic Symbol* is a “sign connected with its object by an association of general ideas. By this its replica calls up an image in the mind, which image, owing to certain habits or dispositions of that mind tends to produce a general concept, and the replica is interpreted as a sign of an object that is an instance of that concept” (Peirce 1955, 116).

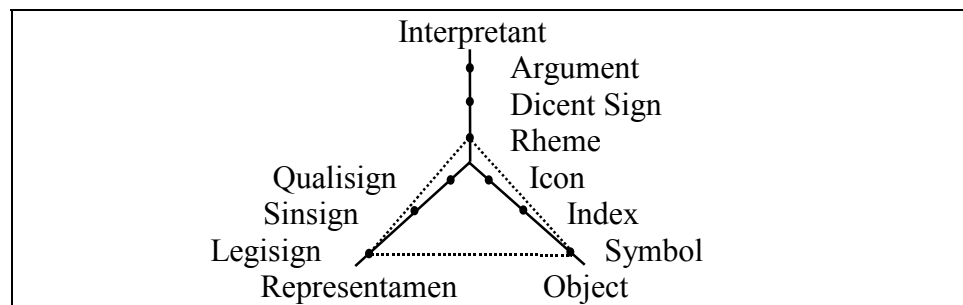


Fig. 4-12. Rhematic Symbol

The Rhematic Symbol is based on conventions, a legisign, and it therefore requires little knowledge of its context, and will yield the same information for all humans, since it is a Rheme. It represents its object through an interpretation or association—it is a Symbol.

Given the foregoing, the Rhematic Symbol is a general term. Like any Symbol, the Rhematic Symbol is necessarily a general type, and is thus a legisign. The replica is a Rhematic Indexical Sinsign (sign category number 3) since the interpreter of the Rhematic Symbol acts upon a symbol already in their minds which gives rise to a general concept. Peirce (1955, 116-117) gives an example of this,

A replica of the word “camel” is likewise a Rhematic Indexical Sinsign, being really affected, through the knowledge of camels, common to the speaker and auditor, by the real camel it denotes, even if this is not known to the auditor; and it is through such real connection that the word “camel” calls up the idea of a camel. The same thing is true for the word “phoenix.” For although no phoenix really exists, real descriptions of the phoenix are well known to the speaker and the auditor; and thus the word it really affected by the object denoted.

The interpreter of the Rhematic Symbol sometimes interprets it as a Rhematic Indexical Legisign (sign category number 3), sometimes as an

Iconic Legisign (sign category number 5). To a certain degree the Rhematic Symbol is like these, however it is much more complex. Examples of Rhematic Symbols “are a common noun, an adjective, adverb or verb” (Merrell 1997, 195).

Ninth: A *Dicent Symbol*, “or ordinary proposition, is a sign connected with its object by an association of general ideas, and acting like a Rhematic Symbol, except that its intended interpretant represents the Dicent Symbol as being, in respect to what it signifies, really affected by its object, so that the existence or law it calls to mind must be actually connected with the indicated object” (Peirce 1955, 117).

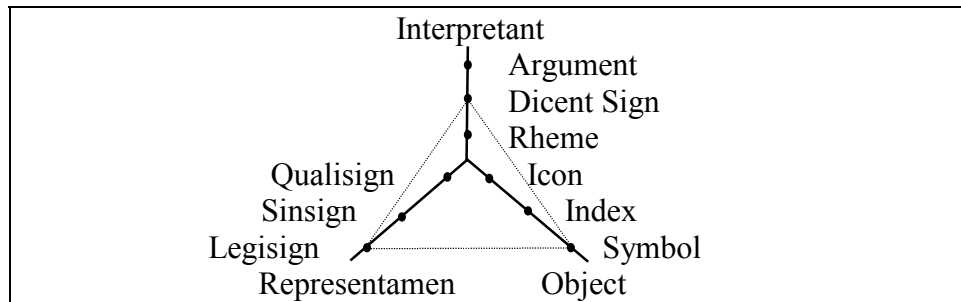


Fig. 4-13. Dicent Symbol

The difference between the Rhematic Symbol (sign category number 8) and the Dicent Symbol is that the sign's relation to its Interpretant has moved from firstness to secondness. This means that the sign is more complex, that it requires greater knowledge of the interpreter. Now the interpreter needs knowledge of the context in which the sign is used to understand the sign.

The interpretation of the Dicent Symbol is a certain kind of Dicent Sinsign (sign category number 4). Just like the Dicent Sinsign, the Dicent Symbol gives information about actual facts. The paradigm example of a Dicent Symbol is a proposition or sentence.

Tenth: An *Argument* is a sign “whose interpretant represents its object as being an ulterior sign through a law, namely, the law that the passage from all such premises to such conclusions tends to the truth” (Peirce 1955, 117-118).

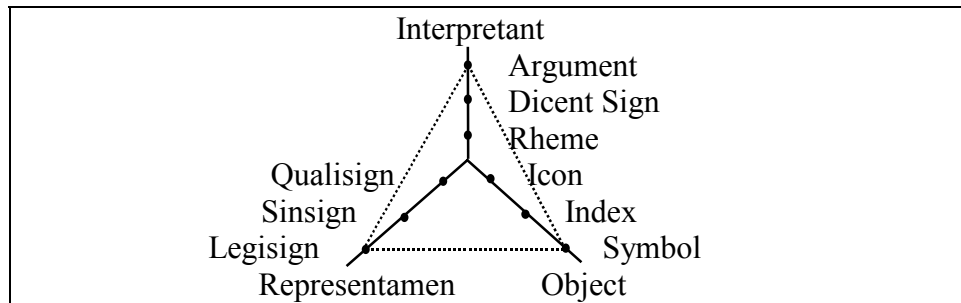


Fig. 4-14. Argument

The Argument is the most complex sign. It is a sign of pure thirdness, that is, it is thirdness in respect to all three aspects of the sign. It is a Legisign, and therefore based on conventions. It is a Symbol and therefore represents through associations. And it is an Argument (thirdness to the Interpretant) and therefore requires that the interpreter has knowledge about the context of the sign. The Argument is, as its name suggests, an argument or text.

These ten categories of sign are regarded as basic and provide a framework for discussing different kinds of interpretation, in the sense that different kinds of signs require different kinds of interpretation. The following section will elaborate the relations between the ten categories of signs.

#### **4.6.3. Relations Between the Ten Categories of Signs**

To show the relationship between the categories of signs, Peirce arranged the categories in ten boxes. The boxes, shown in figure 4-15 (based on a model by Peirce [1955, 118]), are arranged according to their relationship. This relationship is based upon the number of aspects the categories share. Each neighboring pair of boxes share two aspects, e.g. the categories in boxes I and II share the aspects “rhematic” and “iconic,” boxes IV and VII share the aspects “dicent” and “indexical,” boxes VIII and X share the aspects “symbol” and “legisign.” Exceptions to this pattern are boxes separated by a thick line, i.e. boxes II and VI, VI and IX, III and VII, which only share one aspect. The categories in boxes separated from each other by another box also only share one aspect. Boxes that are more than two boxes away from each other, e.g. boxes I and IV, share no aspects.

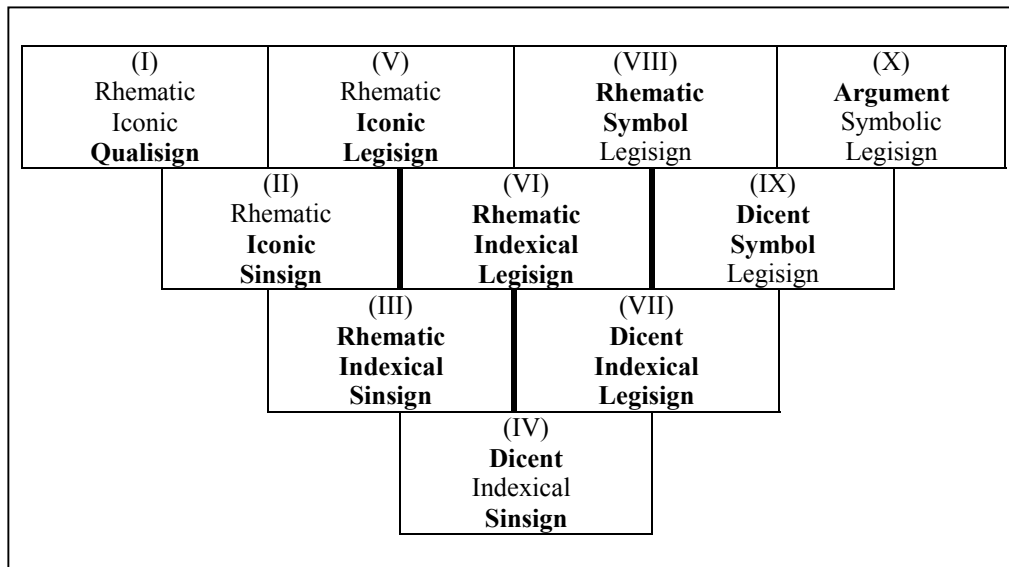


Fig. 4-15. Ten Categories of Signs

One principle in this enumeration of signs, in figure 4-15, is that more complex signs contain less complex signs. The least complex sign, the qualisign, is placed at the upper left-hand corner in figure 4-15. If we move diagonally downward to the dicent sinsign at the bottom of the figure we meet more complex signs. If we again start at the top of the figure with the iconic legisign and move diagonally downward to the dicent indexical legisign, then we also meet more complex signs. In other words, as we move from category I toward category X we meet more complex signs. The more complex signs owe their existence to the



preceding signs and contain them so to speak. This means that complex signs, or more developed signs, also have the features of the preceding signs. Merrell (1997, 205) explains this principle,

Observe the containing-contained relations, the subtle overlapping, the embedment of the less complex signs into the more complex signs as if they were a variation of that simile of the onion--albeit an irregular and ‘nonlinear’ onion in this case--being peeled, shell by shell, and with occasional criss-crossing, until the core remains.

The foregoing point is illustrated in figure 4-16 (a modification of a figure by Merrell [1997, 205]), where it is seen that less complex signs are embedded in more complex signs. Sign category III, for instance, is embedded in sign category VI. This means that what is true for signs in category III is also true for signs in category VI. But signs in category VI are more complex than signs in category III.

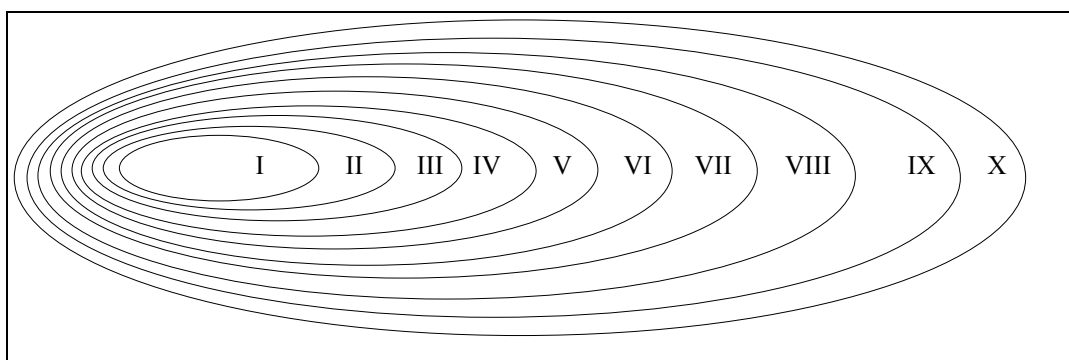


Fig. 4-16. Signs Embedded in Signs

It might further be useful to classify the categories of signs according to their respective qualities. These qualities are derived from the sign's relations to interpretant, representamen, and object. By arranging the categories of signs according to the particular trichotomy that the aspect of the sign belongs to, a classification of the categories according to qualities occur. In other words, a classification of the categories according to whether the signs belonging to the category have a firstness, secondness or thirdness relation to respectively the interpretant, representamen and object. Such a classification will furnish an explanation of the basic differences between the categories of signs.

First (see figure 4-17) the signs are arranged in three classes according to their relations to the representamen.

Firstness;	Qualisign:	I
Secondness;	Sinsign:	II, III, and IV
Thirdness;	Legisign:	V, VI, VII, VIII, IX, and X

Fig. 4-17. The Signs Relations to the Representamen

By far the largest category of signs have a thirdness relation to the representamen; they are legisigns. Legisigns are signs which represent through conventions. This means that by far the most signs represent through conventions. Only one category of signs belongs to the category of qualisigns, which makes sense, since qualisigns are mere qualities, hence there can only be one category of these.

This categorization follows the complexity of the signs, such that less complex signs belong to first category--qualisigns--and more complex signs belong to the last category--legisigns.

Secondly (figure 4-18) the signs are arranged in three categories according to their relations to the object.

Firstness;	Icon:	I, II, and V
Secondness;	Index:	III, IV, VII, and VI
Thirdness;	Symbol:	VIII, IX, and X

Fig. 4-18. The Signs Relations to the Object

Icons, indexes, and symbols are sometimes considered the basic sign categories. Peirce often refers to these categories of signs as the only kinds of

signs. The reason is that these categories of sign could be regarded as the basic families of signs, such that signs that are chiefly *iconic* are grouped together in the upper left corner of figure 4-5; signs I, II and V. Chiefly *indexical* signs are grouped together in the lower part of figure 4-5, these are sign categories III, IV, VI and VII. Lastly, signs that are chiefly *symbolic* are grouped together in the upper right corner of figure 4-5; sign categories VIII, IX, and X.

Thirdly, see figure 4-19, the signs are categorized according to their relation to interpretant.

Firstness;	Rheme:	I, II, III, V, VI, VIII
Secondness;	Dicent:	IV, VII, IX
Thirdness;	Argument:	X

Fig. 4-19. The Signs Relations to the Interpretant

Since arguments are defined as text, only one category of signs can include arguments. A large group of signs--rhematic signs--are transformed into an interpretant quite easy, and the interpretant very much resembles the sign. In other words, the kind of interpretation is that of immediate recognition. The middle category includes symbols, indexes, legisign, and sinsign.

#### **4.6.4. Kinds of Signs**

After this introduction to Peirce's categorization of signs it should be clear that the everyday use of the concept of sign is rather limited in scope. There are in fact many different kinds of signs, a difference that can be ascribed to the way signs are attributed meaning, and to the way they are interpreted. It should also be clear that there are different kinds of interpretation. Interpretations are sometimes a mere translation of a sign into an action, and at other times an interpretation requires an involved understanding of the social context in which the sign is used.

#### **4.7. SUMMARY AND CONCLUSION**

The semiotics and philosophy of C. S. Peirce have been introduced in this chapter. Special emphasis has been put on Peirce's notion of unlimited semiosis and his extensive categorization of signs.

More specifically, Peirce's definition of the sign was presented. He separates the sign into three elements, the sign itself, the object the sign refers to, and the interpretation of the sign. This concept of the sign was not limited to understanding language but could be used in the analysis of any phenomena. The concept of sign is further an extension of Peirce's ontology where he argued for a

mixture of idealism and realism, resulting in what has been called semiotic realism. It was further noted that this conception of the sign is distinct from the structuralistic definition of the sign argued by, for instance, Saussure.

The concept of unlimited semiosis was presented. This concept refers to the notion that any understanding is based on prior understanding, and that these interpretations of understandings inevitably will continue infinitely.

Finally, Peirce's categorization of signs into ten categories stresses that a sign refers to the various aspects of the sign (representamen, object, and interpretant) in different ways. This is an important distinction to make in order to understand what kind of interpretation is needed to understand a given sign. Peirce's categorization of signs is an elaboration of his phenomenology. Each of the ten categories of sign was presented in detail in order to show the clear distinction between the signs.

The conceptions of the sign, unlimited semiosis and the above categorization will be used in the next chapter to analyze the subject indexing process.

## **5. Semiotic Analysis of the Subject Indexing Process**

Chapter 2 focused on the major problems and variations of the subject indexing process. In order better to analyze the subject indexing process, that process was divided into four elements and three steps. It was further argued that only little is known about the individual elements and steps and in particular about the subject indexing process' initial steps. At the end, the argument was put forth that subject indexing could best be understood as an interpretive process.

Research in indexing, which was reviewed in chapter 3, was found to primarily focus on finding the rules of indexing, i.e., finding out what indexers do when they index. Much research assumes that the subject matter of a document can be determined easily and is for all practical purposes unambiguous. However, in chapter 3 it was also argued that language, meaning, and documents cannot be seen in isolation, but rather are closely related to and dependent on the social contexts in which actions take place. All this goes to suggest that a strong approach to understanding the nature of indexing must take into account the social

context of the process. Here, such an approach is found to reside in viewing indexing as an interpretive process. More specifically, an interpretive approach can be implemented by recourse to semiotics as found in the writings of Charles Sanders Peirce. In the light of the latter, chapter 4 presented a semiotic framework for understanding and talking about interpretation in a more precise manner than everyday language allows. Here, the subject indexing process, described as a series of elements and steps in chapter 2, will be discussed and analyzed within the semiotic framework presented in chapter 4.

Section 5.1 will briefly portray the subject indexing process in terms of Peirce's ideas of unlimited semiosis. The two chief elements of this process, steps and elements, will then be further elaborated. Section 5.2 will explain in detail the nature of the steps that one takes in the subject indexing process seen as acts of interpretation in Peirce's unlimited semiotic process. Illustrating the steps in terms of an actual subject indexing case will enhance this explanation. Section 5.3 will thereafter show how Peirce's categories of signs can provide insight into the nature of interpretation that has to take place at each step in the process. Finally, section 5.4 will summarize and draw conclusions regarding the results of these applications.



### 5.1. THE SUBJECT INDEXING PROCESS AS UNLIMITED SEMIOSIS

The subject indexing process consists of four elements (document, subject, subject description and subject entry) and three steps (document analysis, subject description and subject analysis). These elements and steps are interconnected in such a way as to be explainable in terms of Peirce's ideas of unlimited semiosis and signs. This section will show briefly how unlimited semiosis fits the case. It will be followed by more detailed explanations of how the individual steps of the subject indexing process are to be viewed (section 5.2) and how Peirce's categories of signs lend insight into the nature of the interpretation that occurs (section 5.3).

The subject indexing process can be expressed in terms of Peirce's idea of unlimited semiosis. The way unlimited semiosis applies here is that each *element* of the subject indexing process is to be regarded as a sign, with each *step* functioning as an act of interpretation linking the signs in a sequential process that is potentially unlimited depending on how many steps or acts of interpretation are added.

Here, the process begins with an initial sign, the document. The indexer initially makes an act of interpretation (the first step) in order initially to determine what the first sign, the document, is about. The product of this act is a new (or second) sign, the subject. A new act of interpretation (the second step) is

then made in order to convert what the indexer has come up with as a subject to something more manageable and concise for the purposes of indexing. The product of this act is still another new (this time, a third) sign, the subject description. Finally, still another act of interpretation (the third step) is made in order to fit the subject description into a given subject indexing system's vocabulary. This act in turn develops still another new (the fourth) sign, the subject entry. One could extend this process further, of course. For example, the user will come to the index and view the subject entry (a sign) and in an act of interpretation view it as a statement of aboutness for the document, though in this case, the aboutness will likely be related in some fashion to the reason for which the user is searching out information in the first place. The user's conclusions about what the subject entry means will constitute still another sign. And so on.

Unlimited semiosis was described in chapter 4 when Peirce's work was described. In that instance, however, it was not combined with the subject indexing process. Here it is, and in order to portray that combination visually, the entire process is presented in Figure 5.1.

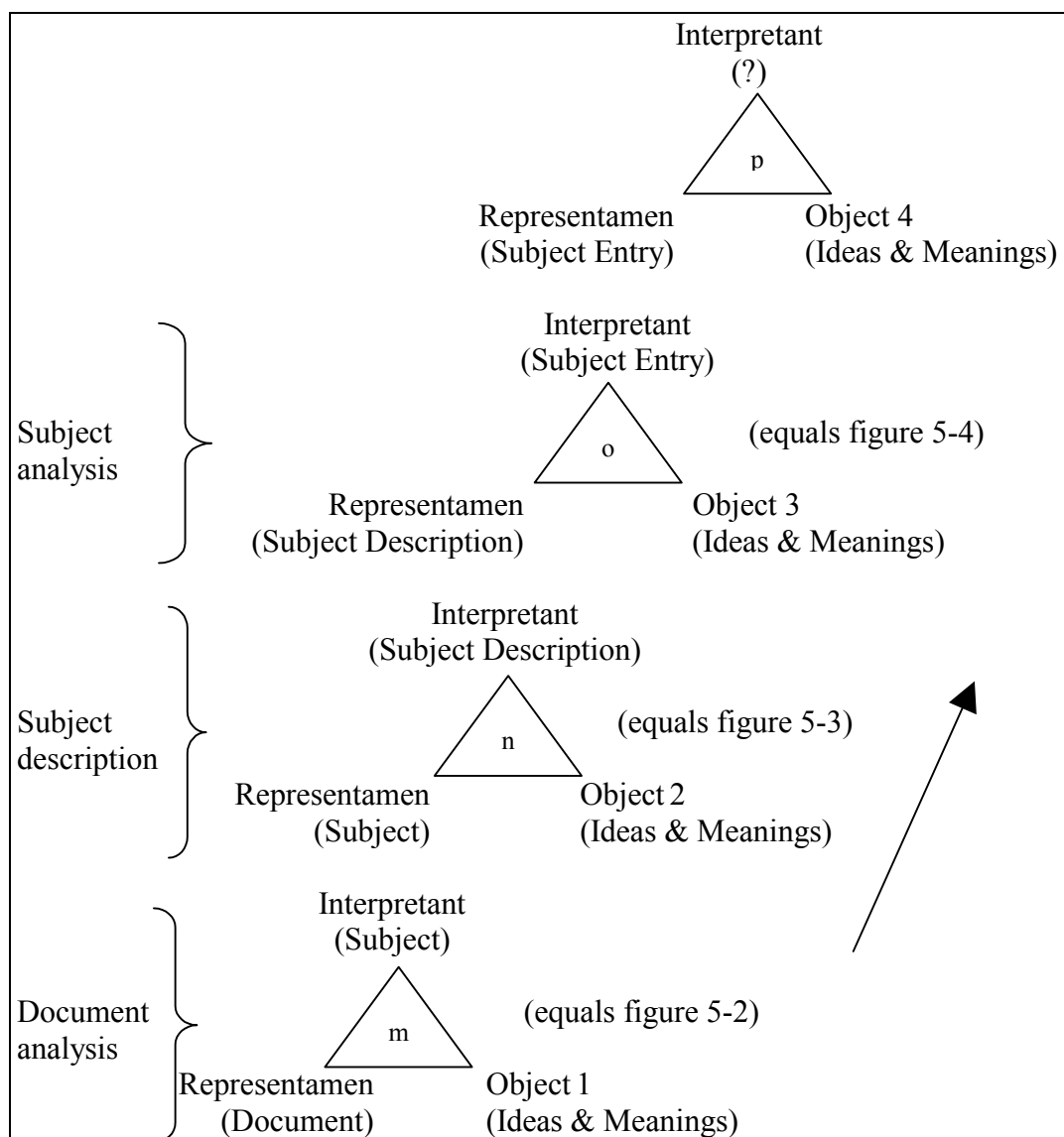


Fig. 5-1. The Semiotic Model of Indexing

In the diagram, it should first be noted that the triangles in figure 5-1 are called *m*, *n*, *o*, and *p* (rather than, say, *a*, *b*, *c*, and *d*) to emphasize the fact that in reality subject indexing is part of a much larger process of interpretation. Before an indexer begins the subject indexing process, the document will have been created in some sort of a discourse community, perhaps a scientific discourse community. Its very creation is the result of many acts of interpretation on the part of the document's author and on the parts of those to whom the author refers. Once completed and published and after the subject indexing process has made the document accessible, the document will be retrieved and used by a number of information users, some within that discourse community, and others outside it. The activities of those information users in consulting a catalog and focusing on the subject entry terms that represent the document in an indexing system (among other documents) and subsequent uses of the document as a whole or in part are likewise acts of interpretation. In short, the process of unlimited semiosis, confined here chiefly to the subject indexing process, started before the subject indexing process began and will continue after the subject indexing process is itself completed. The diagram simply represents an intercepted portion of the larger process.

The second thing to be noted of the diagram is the layout of its triangles. Each triangle is a sign that constitutes an element in the process of unlimited

semiosis. It should be remembered, as it was described in Section 4.1. in the preceding chapter, that the sign is defined as a relation between three entities: the representamen, the interpretant, and the object. This relation constitutes the sign and is in itself a process of semiosis, or interpretation. In other words, each element is a sign and the interpretation of the sign is a process of semiosis. As such, each triangle shows a process of semiosis with the beginning sign that is interpreted (called the representamen in Peirce's formal representation of the process) in its lower left corner, the newly created sign from the act of interpretation (called the interpretant in Peirce's formal representation of the process) at its apex, and the range of ideas and meanings associated with the representamen (called the object in Peirce's formal representation of the process) in its lower right corner. (This follows the description of the components of a sign as found in Section 4.3 in the preceding chapter.)

Given the foregoing layout, triangle *m* represents the first step of the subject indexing process (document analysis) in which the indexers' act of interpretation proceeds from an analysis of the document to the production of a new sign, what is called here the subject of the document. In this step the document is the representamen, the sign which is interpreted. The product of the act of interpretation is a new sign, the interpretant. It is called the subject of the document and its range of ideas and meanings is labeled object 1.

The subject, which is the second sign in the subject indexing process and which in actuality was produced as the result of step 1, is the content of triangle *n*. This triangle represents step 2, the second act of interpretation in the unlimited semiotic process. It begins with the subject as the representamen, produces a third sign (the subject description) as its interpretant, with the range of ideas and meanings of the sign denoted as object 2.

The subject description, which is the third sign in the subject indexing process and which in actuality was produced as the result of step 2 is the content of triangle *o*. This triangle represents step 3, the third act of interpretation in the unlimited semiotic process. It begins with the subject description as the representamen, produces a fourth sign (the subject entry) as its interpretant, with the range of ideas and meanings of the new sign denoted as object 3.

Finally, the subject entry, which is the fourth sign in the subject indexing process and which in actuality was produced as the result of step 3 is the content of triangle *p* at the top of the diagram. This triangle represents an unnamed further potential step or act of interpretation in the unlimited semiotic process—perhaps that of an information user. It begins with the subject entry as the representamen, produces still another new sign (unnamed here) as its interpretant, with the range of ideas and meanings of the sign denoted as object 4.

The third thing to be said of the diagram is that it should be noted that the clear distinction between the elements and the steps of the subject indexing process, which was drawn in Section 2.2, here collapses. In Section 2.2 it was argued that an element consists of an object that is acted upon and a step is the action taken upon the object. This argument was put forth in order to take the elements of the indexing process into consideration. Earlier explanations had merely focussed on the steps and ignored the position of the elements in the process. However, in view of the foregoing explanation of Figure 5-1 it should be clear that no precise lines of demarcation exist between the elements and the steps. Rather, the elements and steps collapse into one single act of interpretation, semiosis. When the indexer acts upon an element, she is in fact already thrown into the step leading to the next element. For instance, when the indexer views and acts upon the document, that act is in fact the first step, the document analysis, of the subject indexing process. The indexer cannot view or act upon the document and then afterwards go into the first step. The elements and steps cannot be separated into two different kinds of phenomena. However, in order to reach a better understanding of the subject indexing process, the following discussion will continue to analyze elements and steps distinctively, but it should be clear that this distinction in reality cannot be maintained.

The final thing to be said of the diagram is that references to Figures 5-2, 5-3, and 5-4 are placed in the diagram for the purposes of correlating this diagram with the discussions that are to follow. Those forthcoming discussions are expansions of the foregoing description of this diagram. The nature of the various acts of interpretation in the continuous semiotic process will be presented in section 5.2; and the nature of the particular signs in the process will be presented in section 5.3.

## **5.2. THE STEPS OF THE SUBJECT INDEXING PROCESS**

In order to provide a greater degree of understanding of how Peirce's process of unlimited semiosis can be used as a basis for understanding how the subject indexing process works, the individual acts of interpretation of that process will be discussed in greater detail here. One aspect of this discussion will be the inclusion of an example of subject indexing, in this case, determining the subject, subject description, and subject headings for the work by Arlene Taylor entitled, *The Organization of Information*, published by Libraries Unlimited in 1999.



### 5.2.1. Step 1: Document Analysis

The first step in the subject indexing process is to analyze a document in order to determine its subject matter. In this step, the document, as a sign that is being interpreted, is the representamen, and the product of the step is a new sign. The sign consists of its beginning point, the representamen (the document), the interpretant, or subject, and the object, the range of ideas and meanings associated with the document in the process of producing the interpretant, the subject. Figure 5-2 illustrates this in the form of a diagram, but it should be noted that this diagram merely represents the lowermost triangle in Figure 5-1 extracted as a separate diagram.

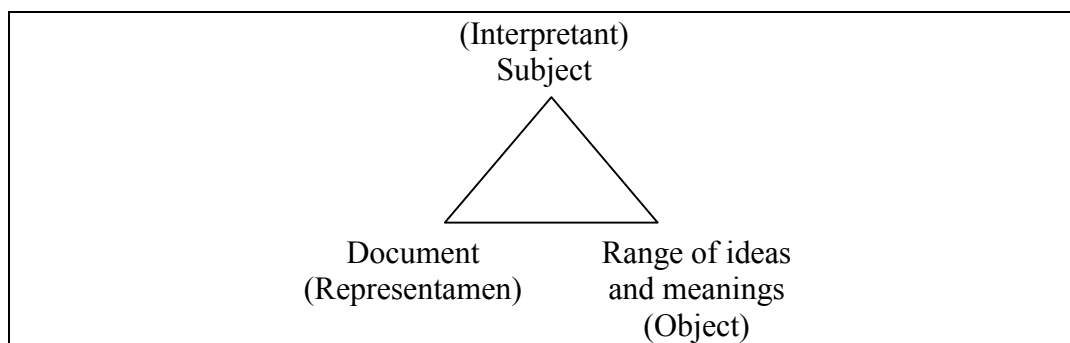


Fig. 5-2. Document Analysis

It would be nearly impossible, of course, for any single person or, in this case, any single indexer, to determine *all* of the ideas and meanings which might

be associated with any particular document, since there might always be potential ideas and meanings which different people at different times and places might find in the document. Furthermore, it would be well nigh impossible to predict precisely which of the many possible ideas and meanings that could be associated with the document would be specifically valuable to the users or that would have some sort of lasting value for the document. To recognize and accept this fundamental openness is of utmost importance. The indexer must realize from the start that she will never discover all the ideas and meanings that could be associated with the document and that, therefore, it is not possible to describe all these ideas and meanings.

It will serve to illustrate this process if it is applied to an actual document, in this case, *The Organization of Information*, a full-length book of nearly 300 pages authored by Arlene Taylor and published in 1999.

How might an indexer discover “the subject” of this book, for that is the first step in the subject indexing process?<sup>14</sup>

For this purpose the indexer would look at different places in the book—for example, the title, the tables of contents, the preface, etc. This would provide

---

<sup>14</sup> It should be noted that what is referred to as determining the subject in this semiotic process, is only taking an initial step in a more involved process. It should not be confused with the more common way of stating the process, which involves going all the way to the subject entry. In short, normally when an indexer or cataloger hears the statement, “discover the subject of the book” it does not get truncated to a single step. The subject, subject description, and subject entry are, in common parlance all the same thing. Here they are stages in a series of interpretive steps.

ideas about the topical content of the book. By doing this, a general impression of the document would begin to accumulate. By just looking at the title it might be supposed that the book is simply about knowledge organization. However, from reading the preface the indexer is informed that Taylor intends the book to precede Wynar's *Introduction to Cataloging and Classification*, a work on library cataloging and classification, the latest edition of which Taylor in reality helped to revise. This gives the indexer some idea about how the author herself viewed the content of the document. She intended it to provide information on matters of information organization that precede library cataloging and classification, the content of Wynar's work.

The titles of the individual chapters in the book at first glance suggest that the range of topics discussed in the book are much broader than simply library cataloging and classification. There are chapters on Organization in Human Endeavors, Retrieval Tools, Development of the Organization of Recorded Information in Western Civilization, Encoding Standards, Metadata, Verbal Subject Analysis, Classification, Arrangement and Display, and System Design. The book also deals with philosophical issues (Organization in Human Endeavors), historical issues (Development of the Organization of Recorded Information in Western Civilization) and technical issues (System Design).

At the same time, however, by looking at the titles of the sub-chapters the indexer will learn that the book not only covers a wide range of problems, standards, and issues in knowledge organization in general, but it also does so with a special orientation to library cataloging and classification. For example, the sources of some of the chapter discussions are clearly from library cataloging and classification sources and systems, rather than from some more general level of sources and systems. In short, while the book at first appears to be about knowledge organization in general, it also appears to treat that topic at least some of the time from the narrower standpoint of library cataloging and classification.

The sources that supported the foregoing ideas were found within the book. However, as the document analysis proceeds, external sources will also inevitably play a role. External sources could well include the indexer's general knowledge of library and information science, a possible knowledge of Arlene Taylor's other works, a knowledge of the users of the information system, and ultimately, interests that are based on the indexer's personal situation and experience. With respect to the latter, were I indexing this work for a given system, I would consider, for instance, whether the book could be used in courses I teach in knowledge organization, how the book supplements other standard works on cataloging and classification, and how Taylor talks about the subject indexing process.

By the foregoing process, the indexer ultimately collects what earlier was called the “range of ideas and meanings associated with a document”—the “object” of the sign in the document analysis step. The accumulated ideas and meanings are at this point, however, more like a collage of impressions of the book rather than some systematically organized statement about it. To arrive at a point in which the ideas are more formally organized and stated will require the second step in the ongoing process of unlimited semiosis as applied to the subject indexing process.

### **5.2.2. Step 2: Subject Description**

The second step (Figure 5-3), creating the subject description, begins with the subject which was reached in the first step. The representamen of the sign relation in the second step is now the subject matter of the documents that the indexer reached in the first step, rather than the document itself. And the interpretant of the sign relation, which is the product of the second step, is the subject description. The latter is more formalized and condensed than the subject matter that resulted from the first step.

To say that the subject description is more formalized and condensed than the subject is not a reflection about committing it to writing, although the indexer

may in fact write the subject description down at this point. It is rather a reflection of picking and choosing from among the range of ideas of meanings encountered in assembling the subject, or of combining elements of the collage assembled during step 1 in order to produce a sensible assertion or set of assertions about the document's subject.

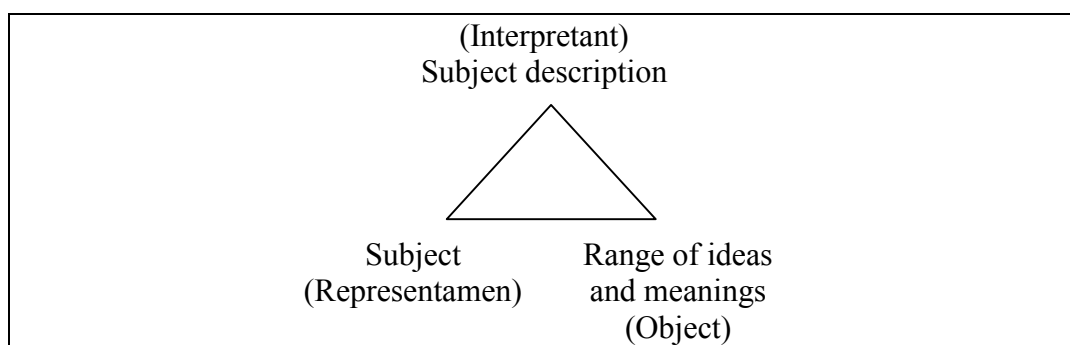


Fig. 5-3. Subject Description

At this point it is important to remember that the subject reached in the first step, the document analysis process, was primarily a mental matter. As such, it contained a great many associations and couplings the indexer found in the text and in other sources. By way of contrast, in the subject description process the indexer summarizes the information compiled in step one in a more or less formalized subject description. Such a description will not likely contain all the associations and couplings the indexer made during the first step but rather only

those that for various reasons the indexer concludes should eventually become statements of the document's subject matter within the system for which the indexer is working. The reasons why some might be used and not others will include things like limitations on how many indexing entries may be prepared per item, a sense that some of the ideas encountered in step one are better representatives of the document than others, and so on.

In order to provide a more realistic illustration of this process, it will be best at this point to return to the process of indexing the Taylor book.

The beginning point for this step, when applied to Taylor's book, is the subject collage that was accumulated as the product of the first step. A subject description of the subject collage accumulated for Arlene Taylor's book in step 1 might be something like the following:

*This book gives a broad introduction to the fundamentals of knowledge organization. It introduces and discusses the most important issues, concepts, and problems in knowledge organization and shows how the novice information scientist designs and implements information systems.*

This description focuses on the most obvious and broadest of the themes accumulated in the document analysis step. At the same time, it includes only a

part of the collage of ideas that were accumulated in the document analysis step. Another theme that was noticed at that stage was the fact that this book has a special relationship to the narrower knowledge organization practices known as library cataloging and classification. Were that also to be acknowledged in a subject description, it might appear something like the following:

*This book is somewhat of an introduction to the fundamentals of library cataloging and classification.*

This subject description is not nearly as accurate as the first, because the major theme of the book appears to be captured by the book. At the same time, this theme is also present.

Of course, still other ideas were represented in the accumulation of ideas and meanings found in the subject collage. But, many of those were simply sub-elements of one or the other of these two statements. And, if the goal is to find subject descriptions that cover or summarize or match the main scope of the content of the work, then the two written out here are about as close as one might get to the goal. Still, choosing these two over any others, or even between these two are the kinds of decisions the subject indexer must make, for that is the heart of subject description—creating more formalized statements that will be amenable



to conversion to subject entries, the final step in the process. It is important to note at this point, however, that this step is notably different than the first step. First of all, in this step, a sign with a narrower scope, a narrower “object” in other words, was examined in order to produce a new sign. The subject collage was its main object of focus, not the entire document and that the range of ideas and meanings associated with the subject was narrower than those ideas and meanings associated with the document itself. Second, the product—its interpretant, the subject description—was narrower still as an object. Third, this too was a process of interpretation, for just as one interprets to go from document to subject collage, so also one interprets anew to go from subject collage to subject description.

### **5.2.3. Step 3: Subject Analysis**

In the third and final step (figure 5-4), the subject analysis, the indexer moves from the subject description to a subject entry. The subject entry is itself based on an understanding of the subject description, and not directly on dealing with the initial subject estimation of the document. In this step, the representamen of the sign is the subject description already arrived at in step 2, and the product of the subject analysis is the subject entry, i.e. the interpretant of the sign. Again, the

object is a range of ideas and meanings, although this time that range is associated with and limited to the subject description.

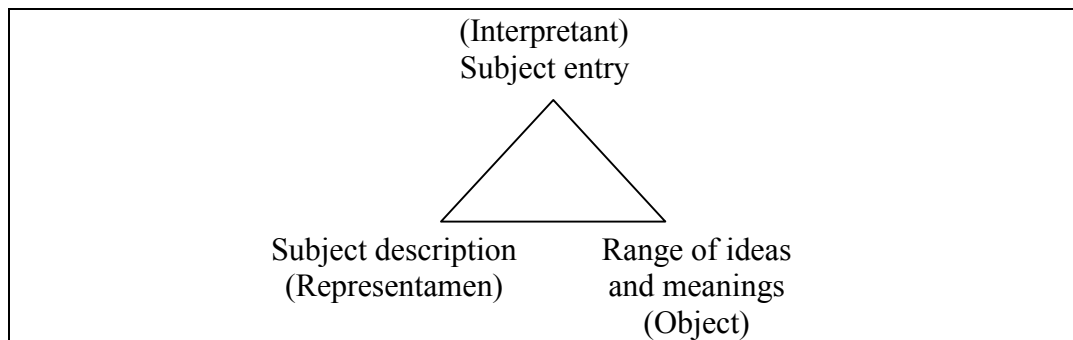


Fig. 5-4. Subject Analysis

As the indexer moves from the subject description to the subject entry she should convert the subject description directly into a subject entry according to the requirements of the system being used. The subject entry will result in expressing the subject description reached at the apex of the second step in terms of the particular indexing language the indexer works with. The third step may thus be characterized as being similar to translating a text from one language to another. In the latter, a translation is a mixture of the syntax, semantics, and pragmatics of the two languages involved. Likewise a similar thing occurs as the indexer proceeds from the subject description to the subject entry. These too

represent two languages: first in the indexers own language and second in the language of the indexing system.

It should be noted at this point that some have argued that the first step should be conducted without any focus on the indexing language. Langridge (1989, 7), for example, argues that because the first step “relates to the document and not to the system” the indexer should not be concerned about the system in this initial step. However, others suggest that indexers keep the particulars of the particular indexing system they are using in mind during all of the initial stages of the indexing process. It is the opinion here, however, that no matter how exemplary Langridge’s ideas may sound, in practice it will be almost impossible to do as he says in his warning, and in reality doing so may lead to unnecessary work. Avoiding the indexing system’s language will be impossible because the indexer will know which indexing language the subject indexing process will eventually be expressed in, and the indexer will, if not consciously, then unconsciously, think in the terms of the ontology of that system. By using the system over and over the indexer will have learned to represent documents according to this particular view.

Returning once again to Taylor’s book, step three takes place when the subject description is analyzed and translated into the indexing language. The

basis for this analysis is the written subject description reached in the subject description process.

The two themes of the subject description of Arlene Taylor's book that were produced in step two are converted into an indexing language. Here for the sake of the illustration, the indexing language to be used is that of the Library of Congress Subject Headings (LCSH).

The most obvious theme in the subject description is that the book is about organizing knowledge. However, the LCSH has no term that specifically matches that phrase. If the indexer confines herself only to the terms already extant in the LCSH and does not create a new term, then one is hard put to find a term that exactly matches the emphasis here. About the closest one might come is the term INFORMATION RETRIEVAL, but this is deceptive because all things being equal this term is ordinarily used to cover many subtopics that are not included in the Taylor work. That this is the case can be seen in the list of narrower terms associated with the term INFORMATION RETRIEVAL in the LCSH system. Thus, to choose it to represent the theme of knowledge organization would be to list this book in a system using LCSH among works that cover a much broader range of subtopics than Taylor's work covers.

Other terms are even less representative. One might consider, for example, INFORMATION STORAGE AND RETRIEVAL SYSTEMS, but this work is not

specifically about particular systems. Or, again, one might use INFORMATION SCIENCE, but this term is manifestly too broad for the theme of knowledge organization. This is the case because it is not limited to the process of knowledge organization but rather includes a wide range of subtopics that have to do with the field of information science.

The second theme that arose in determining the subject description is library cataloging and classification. Here too, the LCSH offers only modest help. There is no single term in LCSH for this entire theme. Rather, this theme is represented by two separate terms, CATALOGING, on the one hand, and CLASSIFICATION--BOOKS, on the other hand. To choose either of these terms would be to place Taylor's work among others that are fully about both topics when, in fact, Taylor's work is only partially about each topic.<sup>15</sup>

The important thing here, however, is not the particular terms that one has to contend with, but rather the fact that when the indexer moves from the subject description to subject entries, she is again engaging in an act of interpretation. This time, however, the sign being interpreted is neither the document nor even the subject as an interpretation of the document, but rather only the results encapsulated in the form of the subject description. And the result of the interpretation has a strong possibility of being even farther removed from the

---

<sup>15</sup> The Library of Congress Cataloging-in-Publication Data classifies the book under, 1. Information Science. 2. Information Science--United States.

reality of the document as a sign than even the more formalized statements that one arrives at in step two when a subject description is produced.

#### **2.2.4. Summary**

This brings to a close a discussion of the three steps in the subject indexing process. This discussion has elaborated on the individual acts of interpretation that takes place throughout the indexing process and as such furnished a better understanding of the portray of the indexing process as unlimited semiosis that was illustrated in Figure 5-1. Peirce offers still more insight into the process, however, when one views how his categories of signs may be applied to the elements that are linked by the steps. It was suggested in the foregoing discussion that the individual acts of interpretation that take place throughout the process are of different kinds. A categorization of the elements in the indexing process in terms of Peirce's categorization of signs will help understand the very nature of these different kinds of interpretation. And it is to those categories of signs that this discussion now turns.

### **5.3. PEIRCE'S CATEGORIES OF SIGNS AND THE FOUR ELEMENTS**

In order to elaborate on the differences between the elements in the subject indexing process, each element will here be categorized in terms of Peirce's categorization of signs, introduced previously in Chapter 4. The reason for doing this is because Peirce's categorization of signs highlights the potentially different kinds of interpretation that can be involved for any particular sign, and that these differences in interpretation are in reality related to the fact that signs are of different kinds. In other words, although each of the elements in the subject indexing process--document, subject, subject description, and subject entry--are signs, that each is a different kind of sign will make a considerable difference in how one approaches them and interprets them. However, by matching the elements of the subject indexing process to the most appropriate of Peirce's ten categories of signs, the kind of interpretation required at each step in the subject indexing process will become more evident.

Although the task of matching the elements with Peirce's ten categories of signs might seem like a straightforward task, it is actually a complex matter. This is the case because each of Peirce's ten categories of signs appears as a trichotomy with three modes of being involved, one for each of Peirce's ideas of firstness, secondness, and thirdness. Thus, the task here has been to determine which of the ten possible trichotomies of signs best fits each of the four elements in the subject

indexing process--document, subject, subject description, and subject entry--for each mode of meaning. What follows here is a discussion of what appears to be the best matches. A brief list of the matches is found in Figure 5-5.

<b>Sign</b>	<b>Best Fit among Peirce's Sign Categories</b>
Document	Sign Category X (Argument)
Subject	Sign Category IX (Dicent Symbol)
Subject Description	Sign Category VII (Dicent Indexical Legisign)

Fig. 5-5. Matches Between Categories of Signs and Elements

In each case, the mode of meaning in the sign will be described as to how it best fits the case and as to its effect on interpretation.

However, before going into the discussion of the best matches between categories of signs and elements in the subject indexing process, it should be remembered how Peirce developed these categories of signs.

It was described in Section 4.3 that a sign consists of three entities, the representamen, the interpretant, and the object. Each entity, however, can be divided into three different levels, called trichotomies. These trichotomies reflect Peirce's categorization of phenomena into three different modes of being, firstness, secondness, and thirdness. This categorization of modes of being was



described in Section 4.5. In turn the categories of signs were described in Section 4.6. It was shown how each entity of the sign could relate to the sign at three levels, such that the representamen could be a Qualisign, a Sinsign or a Legisign, the object could be an Icon, an Index, or a Symbol, and the interpretant could be a Rheme, a Dicent Sign, or an Argument. Each of these levels of the individual entities are then combined in order to define a category sign. However, Peirce did not define a total of 27 different signs that would have been possible if he had simply multiplied the number of possible combinations. Instead he went on to define only ten categories of signs, or kinds of signs. Therefore the matching of the elements of the subject indexing process and the categories, or kinds, of signs in the following are limited by this fact.

When the choice of category for each element in the subject indexing process is made in the following, each element will first be placed in the categorization of signs then afterwards, this choice will then be justified by looking at the trichotomies for each entity of the sign. In other words, after a category is chosen, then the relations between the sign and each entity in the sign is discussed. This involves discussing and justifying the choice of relation between the sign and the representamen as a qualisign, a sinsign, and a legisign; the object as an icon, an index, and a symbol; and the interpretant as a rheme, a dicent sign, and an argument. This order of argumentation is chosen because

otherwise it would be possible to determine a set of trichotomies, for instance a sinsign, an index and an argument, for which Peirce had defined no category of signs.

### **5.3.1. Peirce's Categories of Signs and the First Element—The Document**

The document is the initial element in the subject indexing process and is therefore the first element that is analyzed. A document represents a range of ideas and meanings, and it is the indexer's task to select from among these and capture them first in a more formalized subject description and ultimately in a subject entry. The document is a complex entity that involves many different statements. It is for that reason to be regarded as equal to Peirce's definition of the class of signs called Argument<sup>16</sup> (sign category number X, Figure 5-6).

---

<sup>16</sup> Peirce named the category 'Argument.' He did not use all three terms in the category--Argument, Symbolic, Legisign--to identify the category, although this may cause confusion, Peirce's use of names for the categories will be followed here.

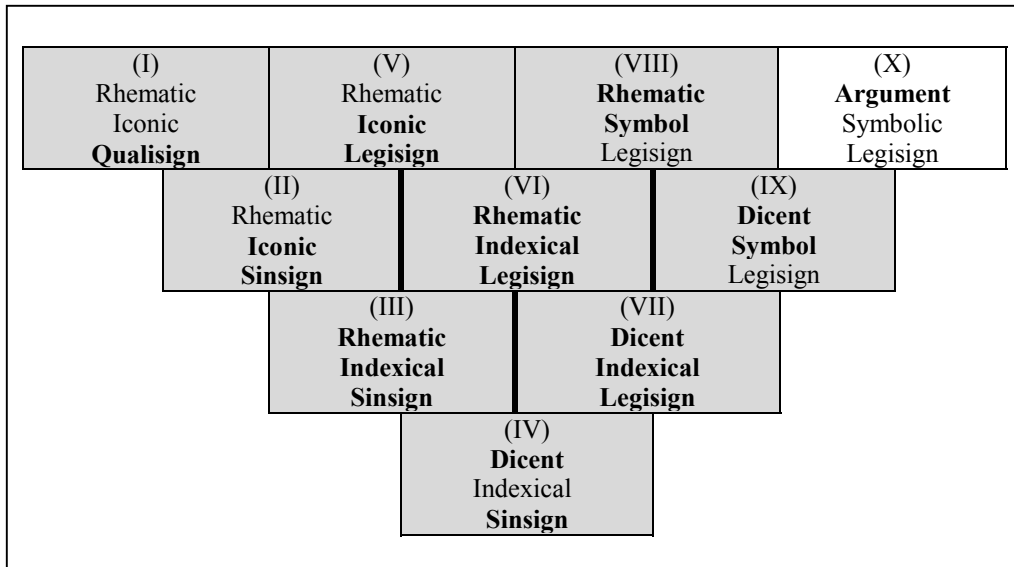


Fig. 5-6. Category X - Argument

The Argument is the most complex of the ten categories of signs in the sense that it embeds all the other kinds of signs in the other nine categories. It is the most developed of the ten categories of signs and its involves the greatest amount of interpretation and uncertainty in its result. The reason for this is that the interpretation of the sign is based on being part of a certain social and cultural context. Peirce (1955, 117-18) defined an Argument as a kind of law in the following way,

An Argument is a sign whose interpretant represents its object as being an ulterior sign through a law<sup>17</sup>, namely, the law that the passage from all such premises to such conclusions tends to the truth.

The interpretation of the Arguments therefore is therefore based on what Peirce called conventions. By this he meant the agreements of practice or customs based on a general consent within specific social and cultural contexts. The document is therefore seen as a product of certain conventions, and its meanings are to be understood in a social and cultural context. The determination of the document's aboutness cannot be separated from this social and cultural context, although there may be overlaps in determination of subject matter between different social and cultural contexts. However, this overlap is as much based on an overlap in conventions than determined by the particular document itself.

To justify this choice of category for the document, the document's relation to each of the entities of the sign will be elaborated in the following. A summary of the argumentation can be found in Figure 5-7<sup>18</sup> in which the choices of trichotomies for each of the three entities are underlined and connected with dotted lines.

---

<sup>17</sup> The word 'law' here means customs, habits, or practices.

<sup>18</sup> It could be confusing the term 'argument' also appears in this figure. However, it should be noted that 'argument' in this figure is distinct from 'Argument' in Figure 5-6. In Figure 5-6 the term 'Argument' denotes a category of signs and in this figure it denotes a relation to the interpretant.

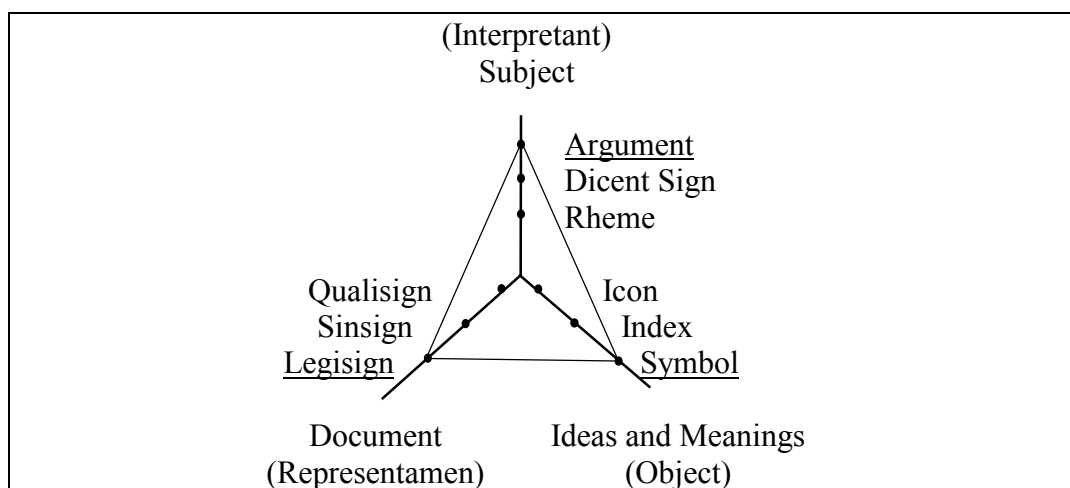


Fig. 5-7. The Document

The tenth category of signs, Arguments, is defined as those kinds of signs that are thirdness to all three entities of the sign. That is, the representamen is a legisign, the object is a symbol, and the interpretant is an argument, each of these representing thirdness. That category of signs best matches the document element in the subject indexing process. By investigating each of the three entities of the sign this match will become clear.

The *representamen*, or in this case the document itself, can either be a qualisign, a sinsign or a legisign and depending on what kind of phenomena the sign itself is. Qualisigns are mere qualities like colors and smell and sinsigns are specific objects like events or acts that only occur once. The legisign is a

convention that gives meaning through laws, or habits, which are “determined” culturally by humans. Accordingly, the document is a legisign, for it will be senseless to claim that the document is a mere quality or a once-occurring event.

The *object*, or in this case the range of ideas and meanings associated with the document, can either be an icon, an index, or a symbol. The icon represents through likeness in the sense that there is some kind of likeness between the sign and that which the sign represents. An index refers to its object by a direct connection between the sign and that which it represents. The relation between the document and the range of ideas and meanings associated with it, is not a relation of likeness or direct connection. It is therefore neither an icon nor an index. The symbol, on the other hand, refers to its object through a law, habit, or convention and is based on the social and cultural context of the document. The document therefore is a symbol. In practice this means that a document can have a different range of ideas and meanings, depending on social and cultural context.

The *interpretant*, or the understanding of the document’s subject matter, can either be a rheme, a dicent sign, or an argument. Rhemes are signs that are easily understood and understood equally alike for all humans, or at least for all human within a certain social and cultural context. It is understood as representing a certain kind of possible object. A rheme could therefore be compared to a single noun, like ‘cat.’ The word ‘cat’ will give nearly the same information to all who

speak English. A dicent sign is more complex than the rheme and as such requires more knowledge to interpret and yield different understandings for different people at different time. Dicent signs could be compared to sentences, like ‘the cat lies on the mat.’ Here there is greater variability in the interpretation of the sign. However, the argument is the most complex relation between sign and interpretant. It is a sign of reason or law and can only be understood in relation to the interpreter’s actions. It cannot be understood by itself and is as such highly dependent on the individual interpreter. The argument is thus to be compared to text, the interpretation of these varies highly dependent on the individual reader. The document element in the subject indexing process therefore is an argument.

In short, other sign categories could not be seriously considered because they 1) used rheme or dicent sign instead of argument, 2) qualisign or sinsign instead of legisign, or 3) icon or index instead of symbol, and as already shown, none of these alternatives are appropriate.

The categorization of the document in category X, Arguments, can be further elaborated and justified by referring back the previous example of the document analysis of Arlene Taylor’s book in Section 5.2.1. It was here seen how the indexer developed what was referred to as a collage of impressions of the book. This collage was developed by investigating different parts of the book and

consulting external resources like the indexer's knowledge about library and information science, knowledge about the users of the information system, etc. The book itself, the representamen, is a convention that is only understood through certain laws and habits related to the understanding of library and information science and the users of an information system, etc. The book refers to a range of ideas and meanings associated with the book, the object of the sign; however, this reference is made neither through resemblance or likeness nor through direct connection with the object. The book refers to its object through laws, habits, or social conventions and is therefore a symbol. In other words, Taylor's book belongs to sign category X, Arguments, because it will only be understood by referring to a certain set of laws, habits, or conventions. The collage of impressions developed by the indexer depends as much on her knowledge than on the book itself.

The interpretant of the document element of the subject indexing process is the subject of the document determined by the indexer. This is therefore the next element in the indexing process, and the discussion therefore turns to that next.



### 5.3.2. Peirce's Categories of Signs and the Second Element—The Subject

The subject is generally only present in the mind of the indexer as something like a collage of impressions and ideas taken from the document. The indexer has not at this point distinguished between what should be represented and what should not. The subject is whatever the indexer associates with the document. The range of such associations or such a collage of impressions could include, besides data from the document itself, such disparate things as other documents the indexer might think of, the fact that the indexer's brother could possibly make use of the document, or a discussion the indexer had with some friends on a previous day. In this respect, the subject is not solely something that has to do with the document and the text in the document. The subject has much to do with the person who reads, looks at, and evaluates the document as well as the document itself. In other words, at this point it makes less sense to say that a document *has* a particular subject than to say that a subject is something a document is *given*. In sum, the subject of a document could be almost anything. In most cases, however, there are only a limited number of accepted and useful interpretations of a document's subject matter, since the social practice in which the document shall be used determines the use, meaning, and subject matter of the document.

Social practice refers to the particular social context (including the domain, the users, the organization, etc.) of the indexing situation. This practice will limit the range of ideas and meanings the indexer will often associate with the document. This is the case especially to the extent that the subject is viewed as part of the subject indexing process. In this process, the indexer--the person who reads, looks at, or evaluates the document--must act as a professional indexer. That the indexer is aware of this serves in turn to limit the range of ideas and meanings to a smaller number than the entire total possible. One element of professionalism in indexing is an awareness of the particular setting in which the document should be used. The subject is categorized in Peirce's sign category number IX (Figure 5-8), which is the category of *dicent symbols*.<sup>19</sup>

---

<sup>19</sup> It should be noted that the category is referred to as a 'Dicent Symbol,' which is Peirce's denotation of the category. The category name 'dicent symbol' should not be confused with the *dicent sign*. The term *dicent sign* is used to denote a relation between the sign and the interpretant.

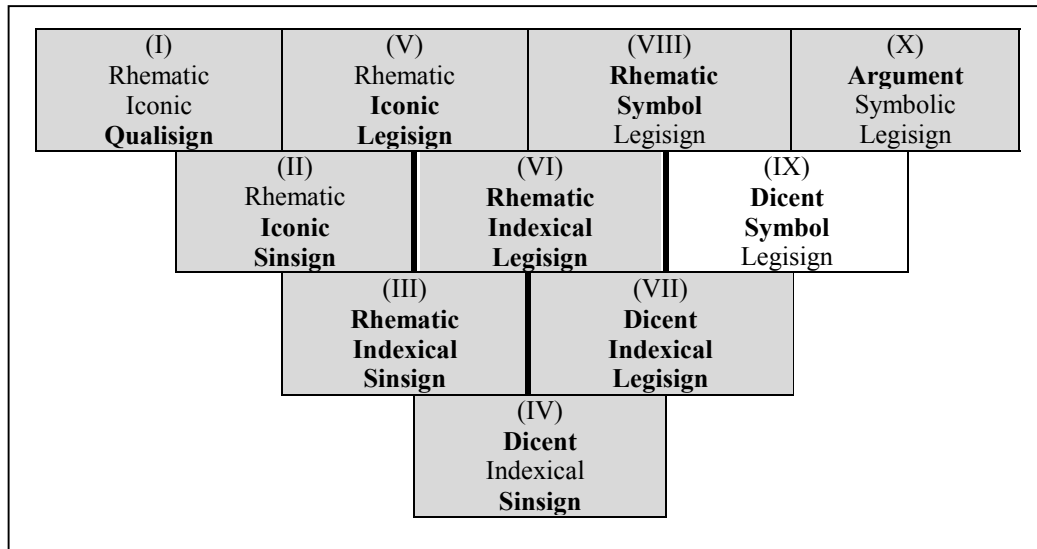


Fig. 5-8. Category IX – Dicent Symbol

A summary of the following argumentation for the choice of sign category IX for the subject can be found in Figure 5-9, the choices of trichotomies for each of the three entities are underlined and connected with dotted lines.

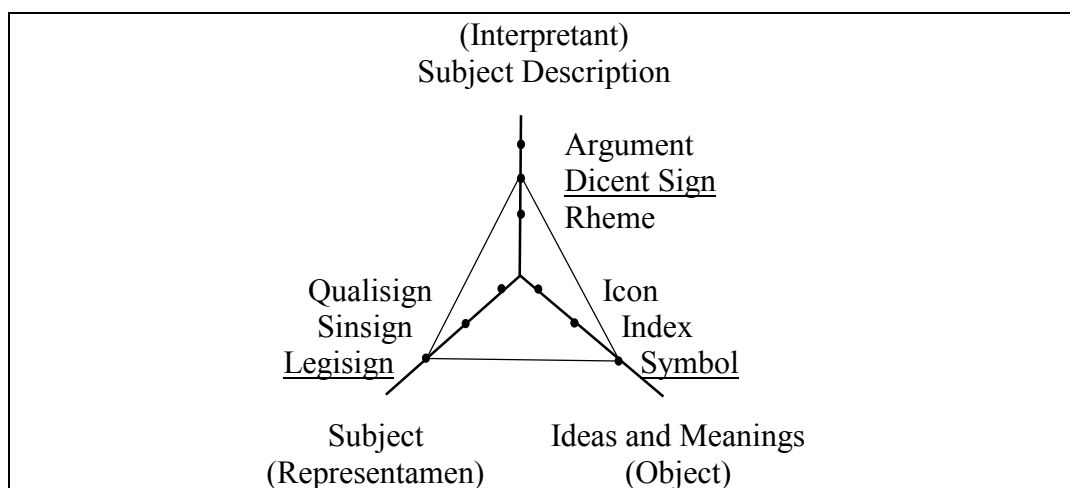


Fig. 5-9. The Subject

As it is seen in Figure 5-9 sign category IX, dicent symbol, is thirdness to the representamen (legisign) and the object (symbol), and secondness to the interpretant (dicent sign). This means that the dicent symbol is somewhat less complex than sign category X, argument, and it therefore involves a less complex kind of interpretation. This in turns means that it is less involved to interpret the subject than the document. The choice is justified next by investigating each of the three entities of the sign.

The *representamen*, or the subject, is a legisign. As already discussed in the previous section, where the document was categorized, legisigns are signs that represent through conventions. Even though the subject at this point exists chiefly

in a purely mental mode, the subject represents the ideas and meanings associated with it, through laws and conventions. It could be argued that the subject is a sinsign, since sinsigns are defined as actual existent things or events. However, one further implication of sinsigns is that they are embodiments of qualities of actual existences. Sinsigns, therefore, are signs of external, or physical, existence and the subject is therefore not a sinsign. As mentioned earlier, qualisigns are mere qualities, wherefore the subject cannot be a qualisign.

The *object*, or the range of ideas and meanings associated with the subject matter, is a symbol. If the subject were considered an index, this would mean that the subject should point out its object or that it would be directly connected with its object. But the subject represents its ideas and meanings not by pointing them out but through associations. Therefore the range of ideas and meanings are only represented through the subject by conventions and laws and are interpreted as referring to the range of ideas and meanings. The subject cannot be an icon, because icons represent through likeness, as mentioned in the previous section.

The *interpretant*, which in this case is the subject description, is a dicent sign. Peirce's definition of the dicent sign states that the dicent sign is sentence or proposition. As it was described in the previous section, an argument equals text and a rheme merely a noun. A subject description, being the interpretant of the subject is therefore best matched with the dicent sign. For the indexer, the subject

will only refer to a fraction of all the possible ideas and meanings, something that can be captured in a few sentences.

In short, other sign categories could not be seriously considered because they 1) used rheme or argument instead of dicent sign, 2) qualisign or sinsign instead of legisign, or 3) icon or index instead of symbol, and as already shown, none of these alternatives are appropriate.

This choice of category could further be illustrated by once again referring back to the indexing of Taylor's book. The subject of the Taylor's book is the collage of impressions and ideas that the indexer associates with the book after the document analysis. This was discussed in Section 5.2.1, and in Section 5.2.2 it was discussed how the indexer formulates a more condensed description of the book's subject matter based on that collage of impressions and ideas. The nature of the interpretation that is required for the subject is that which equals the ninth category of signs, dicent symbols.

The interpretant of the subject element of the subject indexing process is the subject description of the document. This is therefore the next element in the indexing process and the discussion turns to that below.

### **5.3.3. Peirce's Categories of Signs and the Third Element—The Subject Description**

The subject description is a formalization of the subject collage of impressions that resulted from the second step in the subject indexing process. It might be devised as a written statement of the subject or it may simply be a condensed implicit formulation of the subject matter. The essence of the formalization resides in the conscious effort to reduce the various elements of the subject collage to a sensible statement. Thus, in the second step, in which the movement is from the subject to the subject description, the indexer decides which of the many topics and other things associated with the document in the subject collage should be represented. Hence, the subject description will almost assuredly be considerably narrower in focus than the subject. Whereas the subject includes everything the indexer could possibly infer from and associate with the document, the subject description is limited to the information that the indexer concludes is worthwhile or important to represent in the indexing language. It is often recommended in this respect that the indexer produce a series of concise statements that describe the subject. Limiting the description this way will help the indexer to distinguish the different kinds of relations that the subject consists of and pick the most appropriate in the given situation.

The subject description, therefore, consists of a few sentences or statements that are made to represent more formally the content of the document. More specifically, it represents the document as viewed by the indexer, and reflects those topics of the document the indexer has chosen to emphasize. In terms of Peirce's sign categories, the subject description is a Dicent Indexical Legisign (sign category VII, Figure 5-10).

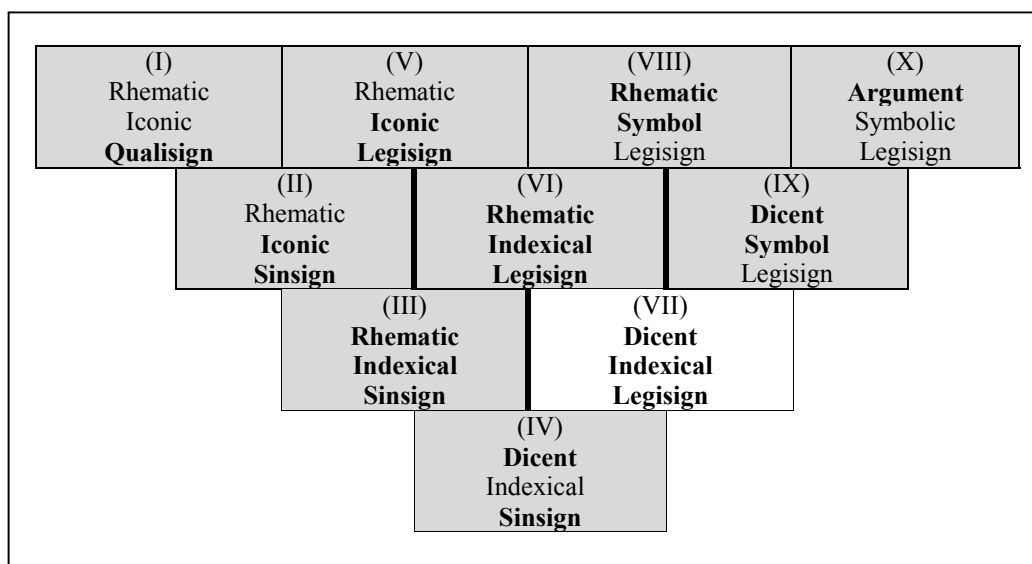


Fig. 5-10. Category VII– Dicent Indexical Legisign

A summary of the following argumentation for the choice of sign category VII for the subject description can be found in Figure 5-11, the choices of



trichotomies for each of the three entities are underlined and connected with dotted lines.

As it can be seen in Figure 5-11 the subject description is only thirdness in its relation to the representamen (legisign), whereas it is secondness to the object (index) and the interpretant (dicent sign). This means that the interpretation involved with the subject description is less complex than with the document and the subject. The production of the interpretant, the subject entry, of the subject description will therefore be less dependent on the individual indexer than the interpretation of the previous two elements.

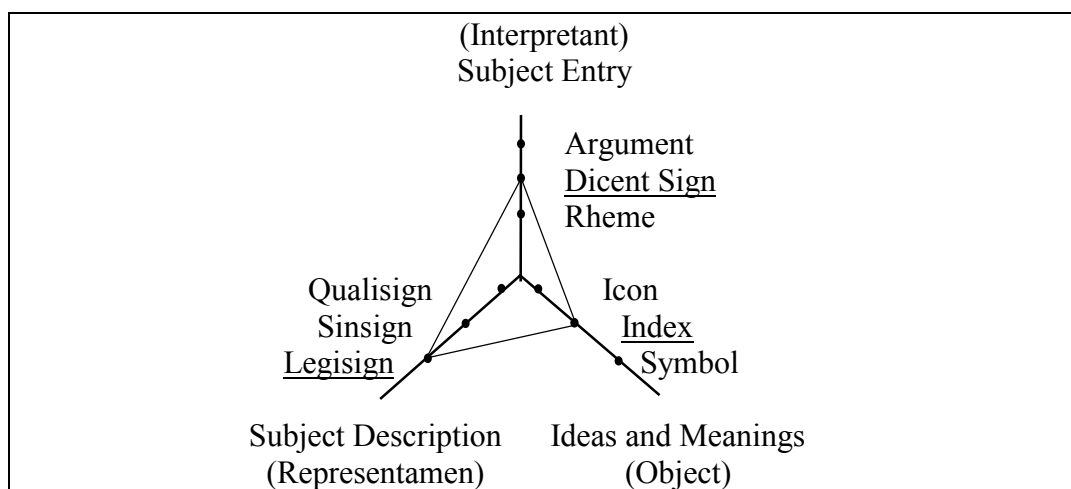


Fig. 5-11. The Subject Description

The reason for the choice of sign category VII, Dicent Indexical Legisign, for the subject description should become clear by going through the individual entities of the sign, next.

The *representamen*, which in this case is the subject description of the subject matter, is, like the previous two elements, a legisign. Legisigns represent through conventions, and the subject description must be regarded as such. The subject representation is not a sinsign because it is not a sign of actual existence and it is not a quality and therefore not a qualisign. Rather the subject description is a sign that represents through conventions.

The *object*, which in this case is the range of ideas and meanings associated with the subject description, is an index. The collage of impressions and ideas of the subject produced in the first step of the subject indexing process limits the range of ideas and meanings associated with the subject description. In other words, the subject collage that the indexer produced directly affects the interpretation of the subject description. Even if the indexer actually wrote down a few sentences that condensed the subject collage to a subject description, the range of ideas and meaning that the indexer associates with the subject collage directly affects the interpretation of these sentences of the subject description. The definition of an index includes that it points to its object and is affected by its object. The distinction between a symbol and an index is important here.

Symbols are interpreted as referring to their objects and thereby operate through laws and conventions. Indexes, on the other hand, are closely associated with their objects, since they are affected by them. Indexes point out their objects, whereas symbols are interpreted as merely referring to their objects. Even though it could be argued that the subject description under some circumstances might be regarded as a symbol--for example, when one indexer has analyzed the document for its subject matter and expressed this in a subject description and another indexer converts the subject description into the subject entry.<sup>20</sup> In such a case the subject description is not affected by its object and the subject description is therefore a symbol, and the subject description would belong to sign category IX, Dicent Symbol. However, it is here maintained that the subject description is an index. The subject description is not an icon, since icons are defined as likeness, and the subject and subject description are not alike, but different is the sense that the subject description is a condensed version of the subject collage.

The *interpretant* of the subject description, which is the subject entry, is a dicent sign. The three possible trichotomies for the interpretant are rheme, which is a sign that is easily understood and interpreted alike by most humans; the dicent sign, which is a more complex sign and involves more knowledge of the interpreter; and the argument, which is the most complex sign or a sign of reason

---

<sup>20</sup> A procedure similar to this was used for a number of years when the British National Bibliography used the PRECIS indexing language (Sørensen 1975a; 1975b).

or law. The three trichotomies can be exemplified by a noun (rheme), a sentence (dicent sign), and a text (argument). The subject description is a condensed version of the subject collage, which could be formalized in a few sentences. The best match for the subject description therefore is the dicent sign.

In short, other sign categories could not be seriously considered because they 1) used rheme or argument instead of dicent sign, 2) qualisign or sinsign instead of legisign, or 3) icon or symbol instead of index, and as already shown, none of these alternatives are appropriate.

The choice of sign category for the subject description could further be illustrated by the indexing of Taylor's book. It was shown in Section 5.2.2, about the second step, subject description, how the subject description could vary according to the focus or point of view the indexer choose to index the book from. Two examples of subject descriptions were given, one that focused on the broadest theme in the book, namely the introduction of the fundamentals of knowledge organization for the novice information scientist. The other description focused narrowly on the book as an introduction to library cataloging and classification. These two themes were converted to Library of Congress Subject Headings in Section 5.2.3, about the third step, the subject analysis. It was in Section 5.2.3 shown how the indexer interprets the subject description to create a subject entry. The indexer's interpretation of the subject description was

constrained by the object of the sign, the range of ideas and meanings associated with the subject description. Among these ideas and meaning were the indexer's initial subject collage of impressions of the document, but also knowledge about the users and knowledge about the indexing system itself. The indexer's final interpretation of the subject description was therefore based as much on her personal knowledge as on the subject description itself. In other words, the interpretation of the subject description for Taylor's book as being about, for example, INFORMATION RETRIEVAL or CLASSIFICATION—BOOKS are equally good and equally correct. However, whether one or the other interpretation is found best suited for the representation of Taylor's the book depends much more on the social and cultural context of the indexing than on the book itself.

#### **5.3.4. Peirce's Categories of Signs and the Fourth Element—The Subject Entry**

The fourth, and last, element of the subject process is the subject entry. The subject entry could be a verbal indexing term or it could be a notation in a classification system. The exact form of the subject entry will depend on the system for which the subject entry has been created. The categorization of the subject entry in Peirce's categories of signs depends on the exact situation in which the subject entry is interpreted and on whom interprets the subject entry.

The exact categorization of the subject entry in Peirce's categories of signs is therefore beyond the scope of this dissertation. The present dissertation is limited to the subject indexing process in which the subject entry is the end product. An analysis of the subject entry, therefore, belongs to a study of information seeking.

However, a brief discussion of the categorization of the subject entry will be valuable here to show the range of possible categorizations of the subject entry.

The interpretation of the subject entry will depend on various things, such as the interpreter's knowledge about the indexing system, about the users of the system, and the domain of the system and documents, etc. If the user of the indexing system is very familiar with it then the interpretation will be less involved and the interpretant will be a rheme, because the subject entry will be interpreted as easily as a noun and almost alike for all users in the same situation. The object will in this case be an index that simply points out the object of the subject entry or range of ideas or meanings associated with the subject entry. This situation would be the case, for instance, for professional librarians who have worked with a specific indexing system for a number of years. On the other hand, a novice user of an indexing system, one who does not know the indexing system very well, will involve a different kind of interpretation of the subject entry. In such a case the interpretation of the subject entry will be more involved. The interpretant will be a dicent sign, because the user had to interpret the subject

entry as statement or sentence of the content of a class or indexing term. The object, the range of ideas and meanings associated with the subject entry, will be a symbol, because the subject entry will be a convention and based on the social and cultural context of the subject entry.

The possible subject entries for Taylor's book might serve to better illustrate this point. A professional user of the LCSH system would know that subject headings such as INFORMATION RETRIEVAL and CLASSIFICATION—BOOKS cover many other topics than those that the headings imply. The professional user would therefore look under both of these and a number of other headings if she were to look for a book such as Taylor's. The novice user of the LCSH, on the other hand, would interpret the subject headings as statement of what could be found under these headings. The novice user would not off hand know that the LCSH lacks a heading that covers the focus of Taylor's book and would therefore, perhaps, not look under these headings for the book. The novice user might look for the topic under a broader heading such as INFORMATION SCIENCE, which she might interpret to cover books with a broad focus on knowledge organization and system design.

### **5.3.5. Summary**

The categorization of the elements in the subject indexing process highlights the different kinds of interpretation that is involved throughout the process and thereby explains the nature of the elements. The fact that each element in the process is a different sign, that requires a different kind of interpretation reveals that the subject indexing process is in fact a complex action of interpretation and choices on behalf of the indexer.

### **5.4. SUMMARY AND CONCLUSION**

The purpose of this chapter has been to apply Peirce's ideas regarding semiotics described in Chapter 4 to the subject indexing process. This has involved commenting on three important matters, showing how Peirce's idea of unlimited semiosis parallels and provides a general approach to understanding the subject indexing process, showing how the steps in the subject indexing process can be better understood in terms of Peirce's idea of unlimited semiosis, and lastly showing how Peirce's categories of signs help one to gain insight into the nature of the elements of the subject indexing process.

In section 5.1, the subject indexing process was described as an example of an unlimited semiosis process in which the individual elements and steps of the



subject indexing process are joined in a continuous sequence. When this is done, a new semiotic way to view the subject indexing process is produced.

In section 5.2, the steps in the subject indexing process as seen as parts of a process of unlimited semiosis, were examined in greater depth. Here too, an example of indexing was invoked in order to show how the process operates within a real indexing environment.

Finally, in section 5.3, the four elements of the subject indexing process were examined in greater depth by categorizing each element in terms of Peirce's categorization of signs.

It should furthermore be noted that the steps and elements of the subject indexing process cannot be separated as sharply as it was suggested in Chapter 2. When an element is interpreted as a sign the indexer is in fact already engaged in taking the next step in the subject indexing process. The discussion of the individual steps of the indexing process in Section 5.2 and the discussion of the categorization of the elements in Section 5.3, therefore, supplement each other. The focus in Section 5.2 was on explaining the choices that the indexer has to make during the indexing process and the focus in Section 5.3 was on nature of the elements, or signs, in the subject indexing process on which the indexer has to act.

One obvious conclusion that one may derive from applying Peirce's semiotics to the subject indexing process is to demonstrate how fundamentally interpretive and, therefore, variable, the entire process is. To portray the process in this way and to make a point of saying that it is a useful conclusion should not be seen as an attempt to somehow demean the process as something that will not yield itself to precision and exactness. Rather it is a way of showing how inexplicably profound and human the process is. Indeed, it is the profoundly human nature of the subject indexing process or task that makes it so impervious to being analyzed solely by quantitative empirical methods on the one hand, and so demanding of the need for qualitative and humanistic approaches to understanding it on the other hand.

## **6. Summary and Conclusions**

One of the central problems in the construction and use of systems for representation and retrieval of documents is how the subject matter of the documents is captured and represented. Although technological advances might provide means for fast access to documents, the problem of representing the subject matter of documents is first and foremost a problem of how people interpret and understand documents. The problem is how the interpretation of a document's subject matter can be represented to make it possible for others to gain access to the document. Indexing, therefore, is concerned with meaning and language -- and any theory of indexing explicitly or implicitly includes a theory of language and meaning. In other words, the study of indexing is closely related to the study of language and meaning and an understanding of indexing must begin with an understanding of how language works and how meaning can be expressed in language.

The subject indexing process is poorly described and documented in the LIS literature. The literature has had a number of calls for research on indexing; however, little into the nature of the process has actually been done. Furthermore, the existing manuals and guidelines designed to assist indexers in the subject indexing process are insufficient. No one has so far successfully described methods to be used in the indexing process. The purpose of this dissertation has been to address this lack by describing the subject indexing process in a new way and by analyzing its various components.

The understanding of the nature of the subject indexing process is more now critical than ever. Indexing of documents is being employed in settings far beyond the traditional library. Indexing is now being discussed in organizations that implement intranets and knowledge management. However, authors of books about knowledge management recognize that one of the most complicated tasks in the organization of knowledge is “to compile a set of meaningful terms by which your knowledge repository can be searched” (Davenport & Prusak 1998, 135). In fact, one of the techniques that Davenport and Prusak advocate for knowledge management is the human-generated thesaurus. In reality, however, the documents on an intranet will often be made accessible both through human-selected index terms and through automatically generated index terms. Although research in automatically generated indexing terms has a long history within

information retrieval it has received much attention lately by researchers focused on making information on the Internet accessible. Many of the automatic indexing techniques that were developed centuries ago are now being used commercially for the first time. Moreover the Internet community has not only applied known techniques but have also improved these and in many cases developed new techniques. Since many of these techniques are now being used commercially there is reason to expect that they will be improved during the next few years. However, no matter how sophisticated they become they will never be able to replace the human-selected index terms but in practice they supplement each other.

For the purpose of analyzing the nature of the subject indexing process in more detail, the process was portrayed in Chapter 2 as consisting of four elements and three steps. A general investigation into the process concluded that the process could be analyzed as a number of interpretations. The latter suggests in turn that the subject indexing process cannot be described in detail and performed and studied as a scientific task characterized by “objectivity” and “neutrality.” The subject indexing process is interlaced with philosophical problems regarding language and meaning. Research in indexing must, therefore, contend with such philosophical problems first before prescribing procedures for how indexing should be done. A “scientific” approach to indexing tends to focus on rules of

indexing, generalize particular indexers' problems and work, and exclude the indexers' environments from consideration. But more importantly, such research has implied that indexing can be studied through scientific methods. Indexing, however, is not governed by rules. There are no rules to discover and hence a nomological approach leads nowhere.

## **6.1. APPROACHES TO INDEXING**

Some authors have argued that there are different approaches to indexing. These authors typically argue that there are two conceptions of indexing. They argue that indexing could either take its basis in the document itself or in users' needs or potential needs for information. On one side it has been argued that the indexer should simply "stick to the text and the author's claim" (Lancaster 1998, 31). On the other side it has been argued 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, 222). However, these two basic conceptions of indexing are only points on a continuum and should be unfolded to five conceptions of indexing. The five conceptions of indexing summarize the various ways the indexing process has been characterized in the indexing literature:

- Simplistic indexing argues in favor of full text or automatic techniques.
- Document-oriented indexing bases the representation on specific parts of the document, for instance title, abstract, and headings.
- Content-oriented indexing attempts to describe the content fully and neutrally.
- User-oriented indexing bases the representation of the documents on knowledge about domain of work or study in which the documents will be used.
- Requirement-oriented indexing bases the representation on users' explicit statements for about what they need information about.

The upshot of dividing approaches to indexing into these five conceptions is show that indexing can be based on a number of different things. Neither of these can be said to better or more correct than others and neither is more realistic or unrealistic in practice than any other is. Any indexing agency needs to define the approach to indexing they which to take, such an approach would most likely combine a number of the conceptions suggested here.

The categorization of these five approaches furthermore stresses the difference between user-oriented indexing and requirement-oriented indexing. In

user-oriented indexing the focus is on the domain rather than on particular individual users, the focal point in the requirement-oriented approach to indexing. The user-oriented conception is closely related to the domain analytic approach devised by Hjørland and Albrechtsen (1995).

Relevance might be the most important and central concept in the LIS field. A small but important corpus of literature on this concept was reviewed. The purpose of the review was to search for conceptions of the subject indexing process and the concept of subject. Most researchers in relevance studies do not define the concept of subject and none of them discuss the subject indexing process per se. There is a tendency to investigate the utility of documents at the expense of investigations into the relationship between representation and retrieval of documents. Theorists and researchers of relevance and evaluation generally do not address problems regarding representation of the subject matter of the documents. A closer collaboration and understanding between these areas of study is needed.

Much literature on the subject indexing process is concerned with finding the rules of indexing. The major goal, which has guided research in the field, has been to find out what indexers do when they index. The goal of this kind of research has been to prescribe *how to* index. A few scholars in the field have challenged this goal, most notably Frohmann (1990) and Blair (1990), who argue



that it is not possible to find these rules of indexing and that even if they are found they will not be of much use. Frohmann shows that a rule has meaning only in a specific social context and Blair points out that subject representations cannot be understood independently of the social context in which they are produced and used.

## **6.2. INDEXING AS INTERPRETATION**

Indexing is not an activity whose “rules never let a doubt creep in, but stop up all the cracks” (Wittgenstein 1958, §84). It is an activity only partially determined by rules. Rules of indexing can never remove suspicion that indexing was not performed correctly or in the best possible way. It is, therefore, much more important to understand and be part of the social context in which the result of the subject indexing process will be used. Or, as Wittgenstein would have phrased it, to be a part of the same form of life and language game as the users. How the document should be represented is not something the indexer can decide in isolation. This decision should be made with the users of the representation in mind. And it should take into account the conclusion that Wilson (1968) appears to have reached, that a document does not *have* a subject. A more accurate statement would be that in reality a document is given one or more of a number of

possible subjects. Literary theorists have how readers derive meaning from text and they have recently begun to argue that the meaning of a text is what the readers find in it by virtue of their own systems of expectations, or as Stanley Fish (1980, 2-3) argues,

[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.

According to Fish's understanding a text does not have a meaning. Instead, the reader creates meaning as the text is read. A reader does not respond to the meaning of a text. The reader's response *is* the meaning of the text. According to this logic a document does not have a subject, but is given a subject by the reader. The indexer's task, therefore, is not to get the subject out of the document but to create the subject and to express this interpretation in the indexing language.

The indexer's interpretation of the document takes place within a given social practice; it is based on a particular understanding of the world, and is therefore limited. The indexer cannot give the document just any subject. The

interpretation is based on the language that constitutes the social practice. The social practice can be formed as a field of study, e.g. biology, economics, or sociology, or it can be formed as a group of people with a specific purpose, e.g. interdisciplinary studies, a company, or an organization. A social practice, therefore, is first and foremost a group of people forming a discourse community. The social practice will define--through language games--how words and documents can be used and the indexer should be a part of that social practice in order to understand the language and thereby choose the interpretation of the document that suits the community best. In other words, to have knowledge about the social practice includes knowledge about the domain, the organization in which the indexing takes place, the indexing language, the specific users, the potential use of the document to be indexed, and so forth. If the indexer possesses knowledge about the social practice then this knowledge will limit the number of possible interpretations of the document the indexer can choose among when indexing the document.

The indexer should represent the documents in agreement with the community the indexing serves. The indexer, therefore, needs to be part of that community. Because “speaking a language is part of an activity, or of a form of life” (Wittgenstein 1958, §23), the indexer needs to be part of that activity. Otherwise she will not speak the same language as the users. If so, she will

interpret the documents based on another form of life than that of the users. In such cases the indexing will at least be more difficult to use.

### **6.3. SEMIOTICS AND INDEXING**

If indexing is viewed as an interpretive process then a method for explaining more precisely what takes place during the process is needed. Although a number of philosophers of language have discussed signs and interpretation, Peirce's semiotics is the most in-depth analysis of the nature of the interpretation of different kinds of signs. He was a wide-ranging and important thinker whose work includes writings on logic, mathematics, metaphysics, cognition, language, and probably most importantly, the foundation of pragmatics. However, he viewed semiotics to be central to all his work and regarded his approach to other areas of inquiry to be consequences of his semiotics. In other words, semiotics should be regarded as a general theory of signs and phenomena.

Merrell (1995, 13) has shown that Peirce's philosophy, "represents a thoroughgoing attempt to undercut the bifurcating tendencies of modernism with a theory of the sign that includes, in addition to the abstract thought, the role of volition and feeling in the generation of meaning." Most studies and approaches to indexing have been caught in the bifurcation of wanting to study indexing by

using empirical statistical scientific methods on one side, and on the other having a process that has do with meaning, interpretation, and the understanding of texts. Peirce's semiotics is generally concerned with how different signs are understood and how signs are represented, it therefore includes a thorough explanation of how different kinds of signs are interpreted and is therefore of interest as a potential explanation of the foundation of indexing.

At the beginning of this dissertation it was argued that the subject indexing process consists of a number of interpretations. In order to understand better understand the exact nature of these interpretation a theory that provides such an explanation is needed. For this purpose Peirce's semiotics was suggested as a useful method and in Chapter 5 it was shown that it was indeed possible to provide an enhanced explanation of the subject indexing process by analyzing it with Peirce' semiotics.

Peirce's concept of unlimited semiosis states that any interpretation is based on previous interpretations and will generate new interpretations. When the subject indexing process is analyzed in terms of Peirce's idea of unlimited semiosis, the high degree of interpretation in the process is emphasized. This analysis results in the semiotic model of indexing, which is an extension of earlier models of the subject indexing process. The semiotic model of indexing shows how later elements of the subject indexing process depend on the interpretation of

earlier elements and how the referent in the process changes as the process takes place. These interpretations are of different kinds, however. The kinds of interpretation are described in Peirce's categorization of signs in which ten kinds of signs are described. Each kind of sign requires a different kind of interpretation.

Peirce defined ten categories of signs, in which he asserts that signs range from those that are qualities that all humans appears to experience or perceive alike, to signs that points out their objects, to signs that are based on agreement. The first kinds of signs are less complex than the latter kinds of signs in the sense that they will be interpreted more or less alike by all humans, whereas the latter kinds are more contingent upon the person who interprets the sign. The less complex a sign is, the less the interpretation of it depends on the person interpreting it. Conversely the more complex the sign is the more dependent the interpretation of it is on the person who interprets is.

To gain a fuller understanding of the nature of the subject indexing process each element in the process was analyzed according to Peirce's categorization of signs. This categorization demonstrates that the first step in the subject indexing process, the analysis of the document, involve a kind of interpretation that is highly contingent upon the social and cultural context of the indexer and the indexing process. Although this contingency has been discussed previously in the

literature this study demonstrates how inescapable the interpretation of the document is. Not only is the initial interpretation inescapable, the entire indexing process is made up of multiple inescapable interpretations. In this sense, the study of indexing is the study of documents and how documents are used. Furthermore, since the representation of documents is the first in a series of activities providing the user with requested documents, a clear understanding of the nature of the process is central to many activities and studies within library and information science. In other words, any study of information seeking, information retrieval, evaluation of information systems, and so on should take the fundamental and inescapable interpretive nature of the subject indexing process into account.

#### **6.4. LATER STUDY OF INDEXING**

The consequence of a semiotic approach to indexing is that indexing needs to be rethought. Indexing is usually thought of as the act of determining the subject matter of a document, but indexing should be thought of as the act of representing an interpretation of a document for future use. Therefore, a later study of the subject indexing process emerging from the present study would take a much different approach than earlier studies have taken. The aim of such a future study would be to determine how context plays a role in the indexing

process. In other words, it would be to find out what influences the indexer when she interprets a document and creates a subject entry. The focus for such a study would be to understand the relation among document, representation, and use. However, this study would be carried out as participation study where the researcher would participate in the indexing of documents in a given indexing setting for a period of time. This setting should preferably be a smaller indexing unit with only a limited group of users who work within a relatively identifiable domain of knowledge. Furthermore, in-depth interviews should be carried out with the indexers. These interviews should aim at understanding the practice of indexing in that setting; and they should attempt to understand the indexers' interpretive tasks.

The aim would not be to find some "objective facts" and generalize about the subject indexing process. Although more knowledge is needed about the process, the present study has shown how inexplicably profound and human the process is. This humanistic nature of the subject indexing process makes it highly impervious to being analyzed solely by quantitative empirical methods, on the one hand, and very demanding of the need for a humanistic approach to understanding it on the other hand.

The distinction between quantitative empirical research methods on the one side and humanities oriented approaches on the other side is a classic distinction



within the philosophy of science. Windelband is known for coining the distinction between nomothetic sciences and idiographic sciences. Nomothetic sciences are sciences that are based on experience and that attempts to uncover hidden laws that govern those experiences, the ultimate goal therefore is to explain what always will be the case in a certain situation. Idiographic sciences, on the other hand, are sciences that study particular phenomena in their real settings and for themselves, with no attempt of generalizing the findings. In short, nomothetic science search for the laws that govern the behavior of certain phenomena and idiographic sciences attempt to understand certain phenomena. Natural sciences are generally considered to be nomethetic sciences and concerned with *realia*, “the real things.” Human sciences are considered to be idiographic sciences, concerned with the human conditions expressed in language, history, art, and culture.

The aim of a later empirical study of indexing would be to gain an understanding of the kind of knowledge indexers use when interpreting documents and when creating subject entries for the documents. The approach taken will be a human oriented idiographic approach.

## **6.5. SUMMARY**

Studies of subject indexing are central to LIS, however, only little is actually known about the nature of the subject indexing process. This dissertation has shown that the subject indexing process is an interpretive process that involves a number of different interpretations. The nature of these interpretations has been described by analyzing them with Peirce's semiotics and categorizing the different kinds of interpretation in his ten categories of signs. The fact that each element in the process is a different sign, that requires a different kind of interpretation reveals that the subject indexing process is a complex action of interpretation and choices on behalf of the indexer. This insight further suggests that the indexing process cannot be studied as a process governed by laws or rules that can be uncovered. Indexing is a process of interpretation and should be studied by use of research methods that study texts and how texts are used.

## Bibliography

- Albrechtsen, Hanne. 1993. Subject Analysis and Indexing: From Automated Indexing to Domain Analysis. *The Indexer* 18, no. 4: 219-224.
- . 1992. PRESS: a Thesaurus-Based Information System for Software Reuse. In *Classification Research for Knowledge Representation and Organization*. New York: Elsevier.
- Albrechtsen, Hanne, & Jacob, Elin K. 1998. The dynamics of classification systems as boundary objects for cooperation in the electronic library. *Library Trends* 47, no. 2: 293-312.
- American National Standards. 1979. *American National Standards for Writing Abstracts*. New York: ANSI.
- Anderson, James D. 1985. Indexing Systems: Extensions of the Mind's Organizing Power. *Information and Behavior* 1: 287-323.

- Austin, Derek. 1974a. The Development of PRECIS: A Theoretical and Technical History. *Journal of Documentation* 30, no. 1: 47-102.
- . 1974b. *PRECIS a Manual of Concept Indexing and Subject Indexing*. London: The Council of the British National Bibliography, LTD.
- Barry, Carol L. 1994. User-Defined Relevance Criteria: An Exploratory Study. *Journal of the American Society for Information Science* 45, no. 3: 149-159.
- Barthes, R. 1968. *Elements of Semiology*. New York: Hill and Wang.
- Bates, Marcia. 1986. Subject Access in Online Catalogs: A Design Model. *Journal of the American Society for Information Science* 37, no. 6: 357-376.
- . 1977a. Factors Affecting Subject Catalog Search Success. *Journal of the American Society for Information Science* 28, no. 3: 161-169.
- . 1977b. System Meets Users: Problems in Matching Subject Search Terms. *Information Processing and Management* 13, no. 6: 367-375.
- Beghtol, Clare. 1986. Bibliographic Classification Theory and Text Linguistics: Aboutness Analysis, Intertextuality and the Cognitive Act of Classifying Documents. *Journal of Documentation* 42, no. 2: 84-113.

- Benediktsson, Daniel. 1989. Hermeneutics: Dimensions toward LIS Thinking. *Library and Information Science Research* 11: 201-234.
- Berman, Sanford. 1993. *Prejudices and Antipathies: A tract on the LC Subject Heads Concerning People*. Jefferson, NC: McFarland & Company.
- Bertrand, Annick, and Jean-Marie Cellier. 1995. Psychological Approach to Indexing: effects of the operator's expertise upon indexing behaviour. *Journal of Information Science* 212, no. 6: 459-472.
- Bertrand-Gastaldy, Suzanne, Diane Lanteigne, Luc Giroux, & Claire David. 1995. Convergent Theories: Using a Multidisciplinary Approach to Indexing Results. *Proceedings of the American Society for Information Science*. 32: 56-60.
- Blair, David. 1990. *Language and Representation in Information Retrieval*. New York: Elsevier Science Publisher.
- Bleicher, Joseph. 1980. *Contemporary Hermeneutics: Hermeneutics as Method, Philosophy and Critique*. London: Routledge.
- Bliss, Henry Evelyn. 1934. *The Organization of Knowledge in Libraries*. New York: H. W. Wilson.
- . 1929. *The Organization of Knowledge and the System of the Sciences*. New York: Henry Holt and Co.

- Bodoff, David. & Ajit Kambil. 1998. Partial Coordination. I. The Best of Pre-Coordination and Post-Coordination. *Journal of the American Society for Information Science*. 49, no. 14: 1254-1269.
- Bradley, Jana. 1993. Methodological Issues and Practices in Qualitative Research. *Library Quarterly*. 63, no. 4: 432-449.
- Bradley, Jana & Brett Sutton. 1993. Reframing the Paradigm Debate. *Library Quarterly*. 63, no. 4: 405-410.
- Brier, Søren. 1996. Cybersemiotics: A new Interdisciplinary Development Applied to the Problems of Knowledge Organization and Document Retrieval in Information Science. *Journal of Documentation* 52, no. 3: 296-344.
- Brookes, Bertram C. 1980. The Foundation of Information Science. Part I. Philosophical Aspects. *Journal of Information Science* 2: 125-133.
- Broadfield, A. 1946. *The philosophy of classification*. London: Grafton & Co.
- Buchanan, Brian. 1979. *Theory of library classification*. New York: K.G. Saur.
- Buckland, Michael. 1983. Relatedness, Relevance and Responsiveness in Retrieval Systems. *Information Processing and Management* 19, no. 3: 237-241.

- Budd, John M. 1995. An Epistemological Foundation for Library and Information Science. *Library Quarterly* 65, no. 3: 295-318.
- Chan, Lois Mai. 1994. *Cataloging and classification: An introduction*. New York: McGraw-Hill.
- . 1989. Inter-Indexer Consistency in Subject Cataloging. *Information Technology and Libraries*, Dec.: 349-358.
- Chan, Lois Mai, Phyllis A. Richmond, and Elaine Svenonius. 1985. *Theory of Subject Analysis : A Sourcebook*. Littleton: Libraries Unlimited.
- Christiansen, Peder Voetmann. 1988. *Charles S. Peirce: Musten og Mørtel til en Metafysik*. Roskilde, Denmark: IMFUFA.
- Chu, Clara M., and Ann O'Brien. 1993. Subject Analysis: the First Critical Stages in Indexing. *Journal of Information Science* 19: 439-454.
- Cleveland, Donald B., and Ana D. Cleveland. 1990. *Introduction to Indexing and Abstracting*. Englewood: Libraries Unlimited.
- Cole, Charles. 1997. Information as a Process: The Difference between Corroborating Evidence as 'Information' in Humanistic Domains. *Information Processing and Management* 33, no. 1: 55-67.

- . 1994. Operationalizing the Notion of Information as a Subjective Construct. *Journal of the American Society for Information Science* 45, no. 7: 465-476.
- Cooper, William S. 1978. Indexing Documents by Gedanken Experimentation. *Journal of the American Society for Information Science* 29: 107-119.
- . 1971. A Definition of Relevance for Information Retrieval. *Information Storage and Retrieval* 7, no. 1: 19-37.
- Cornelius, Ian. 1996. *Meaning and Method in Information Science*. Norwood: Ablex Publishing.
- Cutter, Charles A. 1904. *Rules for a Dictionary Catalog*. 4 ed. Washington: Government Printing Office.
- Davenport, T.H., & Prusak, L. 1998. *Working knowledge: How organizations manage what they know*. Boston: Harvard Business School Press.
- David, Claire, Luc Giroux, Suzanne Bertrand-Gastaldy, Diane Lanteigne, and Annick Bertrand. 1995. Indexing as Problem Solving: A Cognitive Approach to Consistency. *Proceedings of the American Society for Information Science*. 32: 49-55.
- Day, Ron. 1996. LIS, Method, and Postmodern Science. *Journal of Education for Library and Information Science* 37, no. 4: 317-324.



- Dick, Archie L. 1995. Library and Information Science as a Social Science: Neutral and Normative Conceptions. *Library Quarterly* 65, no. 2: 216-325.
- Dreyfus, Hubert I., and Stuart E. Dreyfus. 1986. *Mind over Machine : The Power of Human Intuition and Expertise in the Era of the Computer*. London: Basil.
- Dupré, John. 1993. *The disorder of things: Metaphysical foundations of the disunity of science*. Cambridge: Harvard University Press.
- Eco, Umberto. 1999. *Kant and the Playtapus*. London: Secker & Warburg.
- . 1994. *The Limits of Interpretation*. Bloomington: Indiana University Press.
- . 1990. Almen Semiotik og Sprogfilosofi. *Almen Semiotik* 1, no. 1: 12-23.
- . 1984. *Semiotics and the Philosophy of Language*. London: MacMillan.
- . 1979. *The Role of the Reader*. Bloomington: Indiana University Press.
- . 1976. *A Theory of Semiotics*. Bloomington: Indiana University Press.
- Ellis, David. 1996a. The Dilemma of Measurement in Information Retrieval Research. *Journal of the American Society for Information Science* 47, no. 1: 23-36.

- . 1996b. *Progress and Problems in Information Retrieval*. London: Library Association Publishing.
- Fairthorne, Robert A. 1985. Temporal Structure in Bibliographical Classification. In *Theory of Subject Analysis . A Sourcebook*. Edited by Lois Mai Chan, Phyllis A. Richmond, and Elaine Svenonius, 359-68. Littleton, Colorado: Libraries Unlimited.
- Reprinted from: Conceptual Basis of the Classification of Knowledge: Proceedings of the Ottawa Conference on the Conceptual Basis of the Classification of Knowledge, Ottawa 1971. Munich: Verlag Documentation, 1974.
- . 1969. Content Analysis, Specification, and Control. *ARIST* 4: 73-109.
- . 1968. *Functional Analysis of Information Retrieval*. New York: School of Library Science, State University of New York.
- Farradane, J. 1980. Knowledge, Information and Information Science. *Journal of Documentation* 2: 75-80.
- . 1979. The Nature of Information. *Journal of Information Science* 1: 13-17.
- Farrow, John F. 1995. All in the Mind: Concept Analysis in Indexing. *The Indexer* 19, no. 4: 243-247.

- . 1994. Indexing as a Cognitive Process. *Encyclopedia of Library and Information Science* 53, no. 16: 155-171.
- . 1991. A Cognitive Process Model of Document Indexing. *Journal of Documentation* 47, no. 2: 149-166.
- Fidel, Raya. 1994. User-Centered Indexing. *Journal of the American Society for Information Science* 45, no. 8: 572-576.
- . 1993. Qualitative Methods in Information Retrieval Research. *Library and Information Science Research* 15: 219-247.
- Fish, Stanley. 1980. *Is there a Text in This Class: The Authority of Interpretive Communities*. Cambridge, MA: Harvard University Press.
- Fiske, John. 1990. *Introduction to communication studies*. London: Routledge.
- Frohmann, Bernd. 1994a. Discourse Analysis as a Research Method in Library and Information Science. *Library and Information Science Research* 16, no. 2: 119-138.
- . 1994b. The Social Construction of Knowledge Organization: the case of Melvil Dewey. In *Knowledge organization and Quality Management: Proceedings of the Third International ISKO Conference*. Frankfurt/Main: INDEKS Verlag.

- . 1992. The Power of Images: A Discourse Analysis of the Cognitive Viewpoint. *Journal of Documentation* 48, no. 2: 365-86.
- . 1990. Rules of Indexing: A Critique of Mentalism in Information Retrieval Theory. *Journal of Documentation* 46, no. 2: 81-101.
- Furnas, G.W. et. al. 1982. Statistical Semantics: How can a Computer use what People Name Things to Guess what Things People mean when they name Things” *Proceedings of the Human-Computer Systems Conference*. New York, ACM.
- Gardner, Howard. 1985. *The Mind's New Science: A History of the Cognitive Revolution*. New York: Basic Books.
- Goodman, Nelson. 1961. About. *Mind* 70, no. 277: 1-24.
- Green, Rebecca. 1995. Topical Relevance Relationships. I. Why Topic Matching Fails. *Journal of the American Society for Information Science* 46, no. 9: 646-653.
- Gregory, Richard L. 1981. *Mind in Science: A History of Explanations in Psychology and Physics*. New York: Penguin Books.
- Grice, H. P. 1957. Meaning. *Philosophical Review* 66: 377-388.
- Hanson, Norrwood R. 1958. *Patterns of Discovery*. Cambridge: Cambridge University Press.

- Harris, Michael H. 1986a. The Dialectic of Defeat: Antinomies in Research in Library and Information Science. *Library Trends* 34, no. 3: 515-531.
- . 1986b. State, Class, and Cultural Reproduction: Toward a Theory of Library Service in the United States. *Advances in Librarianship* 14: 211-252.
- Harter, Stephen. 1996. Variations in Relevance Assessments and the Measurement of Retrieval Effectiveness. *Journal of the American Society for Information Science* 47, no. 1: 37-49.
- Hjelmslev, L. 1969. *Prolegomena to a Theory of Language*. Madison, Wis.: University of Wisconsin Press.
- Hjerpe, Roland. 1994. A Framework for the Description of Generalised Documents. In *Knowledge Organization and Quality Management: Proceedings of the Third International ISKO Conference*, Edited by Hanne Albrechtsen, and Susanne Ornager, 173-80. *Advances in Knowledge Organization*, no. 4. Frankfurt/Main: Indeks Verlag.
- Hjørland, Birger. 1998. The classification of psychology: A case study in the classification of a knowledge field. *Knowledge Organization* 25, no. 4: 162-201.
- . 1997. *Information Seeking and Subject Representation: An Activity-Theoretical Approach to Information Science*. Westport, Connecticut: Greenwood Press.

- . 1992. The Concept of “Subject” in Information Science. *Journal of Documentation* 48, no. 2: 172-200.
- Hjørland, Birger & Hanne Albrechtsen. Toward a New Horizon in Information Science: Domain Analysis. *Journal of the American Society for Information Science* 46, no. 6: 400-425.
- Hookway, Christopher. 1985. *Peirce*. London: Routledge & Kegan Paul.
- Hoopes, James. 1991. Introduction. In *Peirce on Signs : Writings on Semiotics by Charles Sanders Peirce*. Chapel Hill: University of North Carolina Press.
- Hutchins, W. J. 1978. The Concept of 'Aboutness' in Subject Indexing. *Aslib Proceedings* 30, no. 5: 172-181.
- Iivonen, Mirja. 1995. Consistency in the Selecting of Search Concepts and Search Terms. *Information Processing and Management*, 31, no. 2: 173-190.
- . 1990. Interindexer Consistency and the Indexing Environment. *International Forum on Information and Documentation*, 15, no. 2: 63-76.
- Iivonen, Mirja and Katja Kivimäki. 1998. Common Entities and Missing Properties: Similarities and Differences in the Indexing of Concepts. *Knowledge Organization*, 25, no. 3: 90-102.
- Ingwersen, Peter. 1992. *Information Retrieval Interaction*. London: Taylor Graham.

- Introna, Lucas D. 1998. Language and Social Autopoiesis. *Cybernetics and Human Knowing* 5, no. 3: 3-17.
- ISO. 1985. *Documentation -- Methods for Examining Documents, Determining their Subjects and Selecting Indexing Terms*. International Organization for Standardization.
- Jacob, Elin. K., & Shaw, Deborah. 1998. Sociocognitive perspectives on representation. In M.E. Williams (ed.), *Annual Review of Information Science and Technology*, 1998, vol. 33, pp. 131-185. Medford, NJ: Information Today for American Society for Information Science.
- Johansen, Jørgen Dines. 1993. *Dialogic Semiosis: An Essay on Signs and Meaning*. Bloomington: Indiana University Press.
- . 1985. Prolegomena to a Semiotic Theory of Text Interpretation. *Semiotica* 57, no. 3/4: 225-288.
- Jones, Kevin. 1983. How do we Index? A Report of some ASLIB Informatics Group Activities. *Journal of Documentation* 39, no. 1: 1-23.
- . 1976. Towards a Theory of Indexing. *Journal of Documentation* 32, no. 2: 118-125.
- Kaplan, Abraham. 1964. The Age of the Symbol: A Philosophy of Library Education. *Library Quarterly* 34: 295-304.

- Karamüftüoğlu, Murat. 1996. Semiotics of Documentary Information Retrieval Systems. In *CoLIS 2. Second International Conference on Conceptions of Library and Information Science: Integration in Perspective*. Copenhagen: The Royal School of Librarianship.
- Keeler, Mary, and Christian Kloesel. 1997. PORT: A Testbed Paradigm for On-Line Digital Archive Development. In *Digital Collections: implications for users, funders, developers and maintainers*, Edited by Candy Schwartz, and Mark Rorvig, 37-53. Proceedings of the ASIS Annual Meeting, no. 34. Medford, NJ: Information Today.
- Krell, David Farrell 1978. *Martin Heidegger. Basic Writings*. Ed. by David Farrell Krell. London , Routledge.
- Lakoff, George. 1987. *Women, fire, and dangerous things*. Chicago: University of Chicago Press.
- Lancaster, F. W. 1998. *Indexing and Abstracting in Theory and Practice*. Champaign: University of Illinois.
- Lancaster, F. W., Calvin Elliker, and Tschera Harkness Connell. 1989. Subject Analysis. *Annual Review of Information Science and Technology* 24: 35-84.
- Langridge, D. W. 1989. *Subject Analysis : Principles and Procedures*. London: Bowker-Saur.



- Larsen, Gorm. 1993. System og semiosis - eller forholdet mellem semiologi og semiotik. In *Tegn og Data. En semiotisk tilgang til humanistisk datalogi*. Edited by Jens F. Jensen, Inger Lytje, and Peter Øhrstrøm, 139-60. Aalborg, Denmark: Aalborg Universitetsforlag.
- Latour, Bruno. 1991. *We have never been modern*. New York: Harvester Wheatsheaf.
- Leonard, Lawrence, 1977. *Inter-Indexer Consistency Studies, 1954-1975: A Review of the Literature and Summary of Study Results*. Urbana, IL: University of Illinois. (Graduate School of Library Science: Occasional Papers).
- Lilley, Oliver L. 1954. Evaluation of the Subject Catalog. Criticisms and a Proposal. *American Documentation*, 5, no. 2: 41-60.
- Liston Jr., David M., and Murray L. Howder. 1977. Subject Analysis. *Annual Review of Information Science and Technology* 12: 81-118.
- Mai, Jens-Erik. 2000. Review of The Organization of Information / Arlene G. Taylor. *Journal of the American Society for Information Science*. 51(5): 491-492.
- . 1999. A Postmodern Theory of Knowledge Organization. In *Knowledge: Creation, Organization and Use*. Proceedings of the ASIS Annual Meeting, no. 36: 547-556.

- . 1999. Deconstructing the Indexing Process. *Advances in Librarianship*. Vol. 23: 269-298.
- . 1999. Review of Indexing and Abstracting in Theory and Practice / F.W. Lancaster. *Journal of the American Society for Information Science*. 50(8): 728-730.
- . 1998. Organization of Knowledge: An Interpretive Approach. In *Information Science at the Dawn of the Next Millennium*, Edited by Elaine G. Toms, D. Grant Campbell and Judy Dunn, 231-42. Proceedings of the 26th Annual Conference of the Canadian Association for Information Science. Toronto: Canadian Association for Information Science.
- . 1997. The Concept of Subject in a Semiotic Light. In *Digital Collections: implications for users, funders, developers and maintainers*, Edited by Candy Schwartz, and Mark Rorvig, 54-64. Proceedings of the ASIS Annual Meeting, no. 34. Medford, NJ: Information Today.
- . 1997. The Concept of Subject : On Problems in Indexing. In *Knowledge Organization for Information Retrieval, 6th International Study Conference on Classification Research*, Edited by I. C. McIlwaine, 60-67. FID, no. 716. The Hague: International Federation for Information and Documentation.
- . 1997. Semiotikken og dens anvendelsesmuligheder indenfor Biblioteks- og Informationsvidenskaben. *Svensk Biblioteksforskning*, no. 3: 43-62.

- Mann, Thomas. 1997. "Cataloging must change!" and Indexer Consistency Studies: Misreading the Evidence at our Peril. *Cataloging and Classification Quarterly* 23, no. 3/4: 3-45.
- Markey, Karen. 1984. Indexindexer Consistency Tests: A Literature Review and Report of a Test of Consistency in Indexing Visual Materials. *Library and Information Science Research*, 6, no. 2: 155-177.
- Maron, M. E. 1977. On Indexing, Retrieval and the Meaning of About. *Journal of the American Society for Information Science* 28, no. 1: 38-43.
- Martinich, A. P. 1984. *Communication and Reference*. New York: Walter de Gruyter.
- Marty, Robert. 1982. C.S. Peirce's Phaneroscopy and Semiotics. *Semiotica* 41, no. 1/4: 169-181.
- Merrell, Floyd. 1997. Do we really need Peirce's whole decalogue of signs? *Semiotica* 114, no. 3/4: 193-286.
- . 1995. *Semiosis in the Postmodern Age*. West Lafayette, Indiana: Purdue University Press.
- Mitchell, Joan S., Julianne Beall, Winton E. Matthews, Jr., and Gregory R. New, eds. 1996. *Dewey Decimal Classification and Relative Index. Devised by Melvil Dewey. Edition 21*. Albany: Forest Press.

- Miksa, Francis. 1998. *The DDC, the Universe of Knowledge, and the Post-Modern Library*.
- . 1992. The Concept of the Universe of Knowledge and the Purpose of LIS Classification. In: *Classification Research for Knowledge Representation and Organization*. Elsevier Science Publications.
- . 1985. Machlup's Categories of Knowledge as a Framework for viewing Library and Information Science History. *Journal of Library History* 20: 157-172.
- . 1983. *The Subject in the Dictionary Catalog From Cutter to the Present*. Chicago: American Library Association.
- Milstead, Jessica L. 1994. Needs for Research in Indexing. *Journal of the American Society for Information Science* 45, no. 8: 577-582.
- Moeller, Bente A. 1981. Subject Analysis in the Library: A Comparative Study. *International Classification*, 8, no. 1: 23-27.
- Moss, R. 1975. PRECIS (letter). *Journal of Documentation* 31, no. 2: 116-117.
- Mounce, M. O. 1997. *The Two Pragmatisms: From Peirce to Rorty*. London: Routledge.
- Natoli, Joseph P. 1982. Librarianship as a Human Science: Theory, Method and Application. *Library Research* 4: 163-174.

- Neill, Samuel. 1992. *Dilemmas in the Study of Information: Exploring the Boundaries of Information Science*. Westport: Greenwood Press.
- O'Connor, Brian C. 1996. *Explorations in Indexing and Abstracting: Pointing, Virtue, and Power*. Englewood: Libraries Unlimited.
- Ogden, C. K., and I. A. Richards. 1923. *The Meaning of Meaning*. London: Kegan Paul.
- Olson, Hope A. 1998. Mapping beyond Dewey's boundaries: Constructing classificatory space for marginalized knowledge domains. *Library Trends*, 47, no. 2: 233-254.
- Olson, Hope A. 1996. Dewey thinks therefore he is: The epistemic stance of Dewey and DDC. In *Knowledge organization and change: Proceedings of the Fourth International ISKO Conference*. Frankfurt/Main: INDEKS Verlag.
- Olson, Hope. 1995. Quantitative "versus" Qualitative Research: The Wrong Question. In *Connectedness: Information, Systems, People, Organizations*, Edited by Hope A. Olson, and Dennis B. Ward. Proceedings of the Annual Conference of the Canadian Association for Information Science, no. 23. Edmonton: University of Alberta, School of Library and Information Studies.

- Olson, Hope A., and Dennis B. Ward. 1997. Feminist locales in Dewey's landscape: Mapping a marginalized knowledge domain. In *Knowledge organization for information retrieval: Proceedings of the sixth International Study Conference on Classification Research*. The Hague: International Federation for Information and Documentation.
- Park, Taemin Kim. 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, no. 3: 135-141.
- Pearson, Charls. 1980. The Basic Concept of the Sign. *Proceedings of the 43<sup>rd</sup> ASIS Annual Meeting*. Vol. 17: 367-369.
- Pearson, Charls & Vladimir Slamecka. 1977. *Semiotic Foundation for Information Science. Final Project Report. National Science Foundation Grant GN-40952*. Atlanta, Georgia: School of Library and Information Science, Georgia Institute of Technology.
- Peirce, Charles Sanders. 1958. *Charles S. Peirce. Selected Writings (Values in a Universe of Chance)*. New York: Dover Publications.
- . 1955. *Philosophical Writings of Peirce*. New York: Dover Publications.
- Petersen, Toni. 1994. Introduction. In *Guide to Indexing and Cataloging with the Arts and Architecture Thesaurus*. New York: Oxford University Press.

- Polanyi, Michael. 1958. *Personal Knowledge: Towards a Post-Critical Philosophy*. Chicago: University of Chicago Press.
- Popper, K. R. 1972. *Objective Knowledge: An Evolutionary Approach*. Oxford: Clarendon Press.
- Putnam, Hilary. 1958. Formalization of the Concept of 'About'. *Philosophy of Science* 25, no. 2: 125-130.
- Qvortrup, Lars. 1993. The Controversy over the Concept of Information: An Overview and a Selected and Annotated Bibliography. *Cybernetics and Human Knowing* 1, no. 4: 3-24.
- Ranganathan, Shiyali Ramamrita. 1967. *Prolegomena to Library Classification*. Bombay: Asia Publishing House.
- Rees, A. M., and T. Saracevic. 1966. The Measurability of Relevance. *Proceedings of the American Documentation Institute* 3: 225-234.
- Richardson, Ernest Cushing. 1935. *Classification. Theoretical and Practical*. New York: H.W. Wilson Co.
- Robertson, Stephen. 1979. Indexing Theory and Effectiveness. *Drexel Library Quarterly*: 40-56.
- Rothman, John. 1974. Index, Indexer, Indexing. *Encyclopedia of Library and Information Science*.

- Rowley, Jennifer. 1994. The Controlled Versus Natural Indexing Languages Debate Revisited: a Perspective on Information Retrieval Practice and Research. *Journal of Information Science*. 20, no.2: 108-119.
- Rudd, D. 1983. Do we Really need World III? Information Science with or without Popper. *Journal of Information Science* 7: 99-105.
- Ryle, Gilbert. 1933. About. *Analysis* 1, no. 1: 10-12.
- Saracevic, Tefko. 1970. The Concept of Relevance in Information Science: A Historical Note. In *Introduction to Information Science*. New York: R.R. Bowker.
- Saussure, Ferdinand de. 1966. *Course in General Linguistics*. New York: McGraw-Hill.
- Schulte, Joachim. 1992. *Wittgenstein: An introduction*. Albany, NY: State University of New York Press.
- Schwartz, Candy, and Laura Malin Eisenmann. 1986. Subject Analysis. *Annual Review of Information Science and Technology* 21: 37-61.
- Seboek, Thomas A. 1994. *Signs: An Introduction to Semiotics*. Toronto: University of Toronto Press.



- Shaw, Debora, and Karen Fouchereaux. 1993. Research Needs in Information Science. *Bulletin of the American Society for Information Science* 19, no. 3: 25.
- Soergel, Dagobert. 1985. *Organizing Information: Principles of Data Base and Retrieval Systems*. San Diego: Academic Press.
- Sørensen, Jutta. 1975a. PRECIS – et emneindexeringssystem. 1. del. *Bogens Verden* 57, no. 4: 111-115.
- Sørensen, Jutta. 1975b. PRECIS – et emneindexeringssystem. 2. del. *Bogens Verden* 57, no. 5: 152-206.
- Strawson, P F. 1970. *Meaning and Truth*. Oxford: Oxford University Press.
- Sutton, Brett. 1998. Qualitative Research Methods in Library and Information Science. *Encyclopedia of Library and Information Science* 62: 263-284.
- . 1993. The Rationale for Qualitative Research: A Review of Principles and Theoretical Foundations. *Library Quarterly* 63, no. 4: 411-430.
- Svenonius, Elaine. 1997. Definitional Approaches in the Design of Classification and Thesauri and their Implications for Retrieval and Automatic Classification. In *Knowledge Organization for Information Retrieval*, Edited by I. C. McIlwaine, 12-16. FID, no. 716. The Hague, Netherlands: International Federation for Information and Documentation.

- Swanson, Don R. 1977. Information Retrieval as a Trial and Error Process. *Library Quarterly* 47, no. 2: 128-148.
- Swift, D. F. 1975. PRECIS (letter). *Journal of Documentation* 31, no. 2: 117-118.
- Taylor, Arlene G. 1999. *The Organization of Information*. Englewood, CO: Libraries Unlimited.
- . 1994. Books and Other Bibliographic Material. In *Guide to Indexing and Cataloging with the Arts and Architecture Thesaurus*. New York: Oxford University Press.
- Taylor, Robert S. 1968. Question-Negotiation and Information Seeking in Libraries. *College and Research Libraries* 29: 178-194.
- Travis, Irene L., and Raya Fidel. 1982. Subject Analysis. *Annual Review of Information Science and Technology* 17: 123-157.
- Vickery, Brian. 1997. Metatheory and Information Science. *Journal of Documentation* 53, no. 5: 457-476.
- . 1986. Knowledge Representation: A Brief Review. *Journal of Documentation* 42, no. 3: 145-159.
- . 1968. Analysis of Information. *Encyclopedia of Library and Information Science* 1: 355-384.

- Ward, Martin L. 1996. The Future of the Human Indexer. *Journal of Librarianship and Information Science* 28, no. 4: 217-225.
- Warner, Julian. 1990. Semiotics, Information Science Documents and Computers. *Journal of Documentation* 46, no. 1: 16-32.
- Weinberg, Bella Hass. 1988. Why Indexing Fails the Researcher. *The Indexer* 16, no. 1: 3-6.
- . 1987. Why Indexing Fails the Researcher. *Proceedings of the ASIS Annual Meeting*, no. 24. Medford, NJ: Information Today.
- Weiss, Paul, and A. W. Burks. 1945. Peirce's sixty-six Signs. *The Journal of Philosophy* 42: 383-388.
- Wersig, Gernot. 1993. Information Science: The study of a Postmodern Knowledge Usage. *Information Processing and Management* 29, no. 2: 229-239.
- Wiener, Philip P. 1958. Introduction. In *Charles S. Peirce: Selected Writings (Values in a Universe of Chance)*. New York: Dover Publications.
- Wilson, Patrick. 1968. *Two Kinds of Power: An Essay on Bibliographic Control*. Berkeley: University of California Press.
- Wittgenstein, Ludwig. 1969. *On Certainty*. New York: Harper and Row.

———. 1958. *Philosophical Investigations*. New York: Macmillan Publishing.

Wright, H. Curtis. 1978. Inquiry in Science and Librarianship. *Journal of Library History* 13, no. 3: 250-264.

———. 1976. The Immateriality of Information. *Journal of Library History* 11: 297-315.

Wynar, Bohdan S. 1992. *Introduction to Cataloging and Classification*. 8<sup>th</sup> ed. by Arlene G. Taylor. Englewood, CO: Libraries Unlimited.

## **Vita**

Jens-Erik Mai was born in Taulov, Denmark on February 8, 1968, the son of Kamma Mai and Milton Mai. After his parents divorced he lived with his father and stepmother, Birgit Irene Mai. After completing his high school level work at the Fredericia-Middelfart Handelsskole, Denmark, in 1988, he entered the Royal School of Library and Information Science, Aalborg, Denmark (formerly The Royal School of Librarianship). He received a bachelor degree in library and information studies in 1992 and a master's degree in library and information science in 1994. In September 1994 he entered the doctoral program in the Graduate School of Library and Information Science of the University of Texas at Austin. He has been a member of the faculty at the Royal School of Library and Information Science, Copenhagen, Denmark, since 1996 and has published a number of items in proceedings and journals in the area of library and information science.

Permanent address: Ingerslevsgade 130, 4. tv., 1705 Copenhagen, Denmark

The dissertation was typed by the author.