PERGAMON

Information Processing and Management 37 (2001) 231–254

INFORMATION PROCESSING & MANAGEMENT

www.elsevier.com/locate/infoproman

The nature of indexing: how humans and machines analyze messages and texts for retrieval. Part I: Research, and the nature of human indexing

James D. Anderson *, José Pérez-Carballo

School of Communication, Information, and Library Studies, Rutgers The State University of New Jersey,
4 Huntington Street, New Brunswick, NJ 08901-1071, USA

Received 4 February 2000; accepted 2 May 2000

Abstract

Does human intellectual indexing have a continuing role to play in the face of increasingly sophisticated automatic indexing techniques? In this two-part essay, a computer scientist and long-time TREC participant (Pérez-Carballo) and a practitioner and teacher of human cataloging and indexing (Anderson) pursue this question by reviewing the opinions and research of leading experts on both sides of this divide. We conclude that human analysis should be used on a much more selective basis, and we offer suggestions on how these two types of indexing might be allocated to best advantage. Part one of the essay critiques the comparative research, then explores the nature of human analysis of messages or texts and efforts to formulate rules to make human practice more rigorous and predictable. We find that research comparing human vs automatic approaches has done little to change strongly held beliefs, in large part because many associated variables have not been isolated or controlled.

Part II focuses on current methods in automatic indexing, its gradual adoption by major indexing and abstracting services, and ways for allocating human and machine approaches. Overall, we conclude that both approaches to indexing have been found to be effective by researchers and searchers, each with particular advantages and disadvantages. However automatic indexing has the over-arching advantage of decreasing cost, as human indexing becomes ever more expensive. © 2001 Elsevier Science Ltd. All rights reserved.

Keywords: Human indexing; Automatic indexing; Comparative research; Cognitive view; Social construction

* Corresponding author. Tel.: +1-732-932-7501; fax: +1-732-932-6916.
E-mail address: jda@scils.rutgers.edu (J.D. Anderson).

PII: S0306-4573(00)00026-1

1. Introduction

A recent article in *USA Today Tech Reviews* (Weise, 2000) proclaims "for once in the information revolution, the humans are pulling ahead. ... The computer-automated indexes that powered a majority of the Web's search engines gave ground to Web directories – listings that depended instead on the power of thousands of human minds to harness the limitless information of the Net. ... In December [1999], the top five search sites on the Net – Yahoo!, AOL, MSN, Netscape and Lycos – were all based mainly on human-generated directories rather than computer-created indexes, according to figures from market trackers Nielsen/NetRatings".

This is simply one of the latest of a long line of assertions touting the superiority of retrieval based on human indexing vs machine indexing – or vice versa! – going back to the beginning of automatic indexing. In this essay, we explore this question by reviewing the opinions and research of leading experts on both sides of this divide.

The entire essay focuses on the analysis of messages or texts, leaving aside other important aspects of indexing, such as size of documentary units (e.g., paragraphs, pages, complete articles, books and analogous options in other media), exhaustivity (number of terms assigned per documentary unit), indexable matter (titles, abstracts, full text, etc.), specificity (broad vs narrow terms for particular concepts), syntax (rules or patterns for term combination in index headings or search statements), vocabulary management (free vs controlled terms and links to related terms), browsable displays vs non-displayed indexes for machine matching, surrogates for texts, and the arrangement of displayed indexes. [1]

2. Analysis and indexing methods

In order to retrieve messages, texts, and documents via an IR database, they must be described and indexed ('indicated'). Description requires some kind of analysis. Two basic approaches are used for the analysis of messages, texts, and documents [2]: by human examination and by machine algorithm. Humans examine documents and texts in order to consider messages that texts represent, plus features of texts and the documents in which texts are recorded. Computers identify and compare components of texts – the symbols that comprise texts – sometimes consulting lexical, thesaural, discourse or other contextual data to expand and characterize sets of textual components; sometimes applying syntactic or pattern indexing algorithms to identify larger units of text; and sometimes calculating attributes for text components and documents based on available data.

[1] This essay is part of a larger project to address all aspects of indexing and the design of information retrieval databases. Issues not covered here are addressed in that larger project, to be published in book form by Scarecrow Press.

[2] 'Message' is used for the ideas, thoughts, emotions, or knowledge that a creator intends to convey to other people. 'Text' is used in the semiotic sense for the organized set of symbols chosen to represent the message. Thus, text is not limited to human language messages or language writing systems, but includes all symbolic methods for representing messages, as in music, the visual arts, dance, mathematics, computer programming, and chemistry. Document is used for the combination of text placed on or in a medium for transmission to recipients.

These two approaches are often called human indexing and automatic indexing. Our definition of 'indexing' simply means pointing to or indicating the content, meaning, purpose, and features of messages, texts, and documents.

Both approaches are widely used. Research comparing retrieval based on human vs machine indexing tends to show that the two approaches produce different results, but that users find them, on balance, more or less equally effective. Similar evidence comes from observing the behavior of expert searchers. When they have access to indexing based on both approaches, they generally use both types of indexing, preferring human indexing for some types of searches and automatic machine indexing for others. Personal preferences also play a role. Some users prefer one type of indexing or the other most or all of the time. (Research is discussed in more detail in the following section.)

Increasingly, IR databases are designed to provide more than one indexing approach in hopes of maximizing the effective retrieval of useful messages, texts, and documents. By offering multiple approaches, it may be possible to take advantage of the strengths and features of different approaches and also to respond to the needs and preferences of users in a variety of situations.

This essay focuses almost exclusively on the indexing of language text, as opposed to messages expressed in other types of text, such as pictorial images, music, the fine arts, or mathematical or chemical formulae. Research and experimentation on automatic indexing of language text has been under way for several decades, and there has been much progress and growing use of automatic indexing of language texts for retrieval. But the automatic indexing of image text has barely begun. Whereas automatic indexing of language and mathematical/chemical text is now routine and common, automatic indexing of image text is usually experimental. IR databases that provide access to most other types of text, especially image texts, rely for the most part on human indexing of messages and their texts and documents.

Recent summaries and assessments of the indexing process by been provided by Hjørland (1997), Fidel, Hahn, Rasmussen and Smith (1994), Lunin and Fidel (1994), and Weinberg (1988, 1998).

3. Research comparing automatic and human indexing

Research comparing the relative strengths and weaknesses of these two basic approaches to indexing has failed to convince die-hard opponents of the merits of either approach, and in fact, the clarity of research results has been disappointing. Sparck Jones (1981a,b) provides an excellent over-view and assessment of the first 20 years of serious comparative research, including the landmark Cranfield 1 and 2 experiments, conducted under the direction of Cleverdon in the UK.

Most of the earlier research was limited to relatively small collections with the evaluation of search results based on judgments by persons other than real users with real information needs or desires. Since then, there have been increased efforts to put users in the center of information retrieval research, with the recognition that user variables are as important, perhaps more important, than any variation in indexing methods for determining the effectiveness of IR systems and IR databases. The study of searchers by Saracevic, Kantor, Chamis and Trivison (1988) is an example. Their subjects were functioning intermediary searchers dealing with real queries submitted by real clients, who provided relevance judgments tied to their actual information need perceptions.

In all IR research, it has been very difficult to isolate particular variables or differences in order to assess their specific impact on overall performance. The two major components of any information retrieval situation are the user on the one hand and the IR system, including IR databases, on the other hand. Here the focus is on the IR system component. Quite apart from human vs machine indexing of messages, texts, or documents, other key variables, each of which deserves separate and focused attention, tend to co-occur in varying degrees with human or automatic indexing. Nevertheless, unless these other variables are accounted for, in addition to the type of indexing (machine vs human), it cannot be clear which variables have the major impact on the results of comparisons. These other key variables for IR research include the following.

- Size of documentary units: Human indexing tend to focus on larger documentary units, such as complete periodical articles, complete chapters in collections, or even complete monographs. With the wide-spread availability of full-text documents within IR databases, automatic indexing now routinely retrieves individual paragraphs of texts, rather than complete documents. But of course, humans *could* in principle analyze and index at the paragraph level of documentary unit, so this variable is *not* directly tied to indexing method.
- Extent of indexable matter: Tied to the availability of full-text documents are differences in the extent of indexable matter. Automatic indexing is now routinely based on the complete text, whereas much human indexing may be limited to an abstract or other summarization of the complete text. This difference is tied closely also to exhaustivity, because the lower exhaustivity of typical human indexing can be accommodated with briefer indexable matter. But again, humans *could* analyze complete texts and *could* index at greater levels of exhaustivity, so these variables are *not* the same as indexing method.
- Exhaustivity: Automatic indexing tends to be exhaustive, considering most, if not all words in indexable matter as potential indicators of content. On the other hand, human indexing tends to be selective, indexing only topics or aspects that appear to be of most importance for *summarizing* the content, meaning, or purpose of a message. But, as just noted, humans *could* analyze and index at higher levels of exhaustivity, so exhaustivity is *not* directly tied to indexing method.
- Specificity: Automatic indexing tends to use very specific terminology (and therefore a very large and varied vocabulary), because it uses, or at least begins with, the actual language of the text. Human indexing tends to use more generic terminology (and a much smaller vocabulary over all) in an attempt to summarize topics and to avoid too much scatter of closely related topics. However, humans *could* use larger vocabularies, so that high specificity is not necessarily an attribute unique to automatic indexing.
- Browsable displayed indexes: Browsable displayed indexes with multi-term context-providing headings are certainly possible with automatic indexing, but they are not as common as with human indexing, and the types of term combination tend to be more limited for most types of automatic indexing. Tied to this variable is human searching vs machine matching. When humans search (browse, inspect) displayed indexes, they generally make judgments about possible relevance or usefulness during the searching process. When machine matching is used, users submit search terms to an IR system, which performs some kind of matching and then presents a list of retrieved items for users to evaluate. These different methods may effect

perceptions of performance, but they are *not* tied to either human or automatic methods of indexing.

- Searching syntax, display syntax: Although the increasingly sophisticated methods for selecting, combining, manipulating, and weighting terms for machine matching can also be used with human indexing, most often these techniques are used with automatic indexing. On the other hand, the syntactic possibilities for the combination of terms to create context-providing headings tend to be richer for human-assigned indexing than for automatic indexing. Such headings are meant to facilitate human searching and browsing (as opposed to machine matching). For human-assigned terms, a wide range of syntactic patterns, ranging from traditional subject headings to modern contextual string-indexing procedures, are available, whereas the presentation of automatically selected terms is usually limited to KWIC, KWOC, KWAC (key-word in, out of, or along-side context) or permuted formats.

  Thus very different types of syntax are typically used with these different approaches to indexing, but nevertheless, differences in syntax are *not* the same as differences in indexing methods.
- Vocabulary management: This variable is closely related to specificity. Although there is no necessary connection between type of indexing on the one hand and vocabulary control or management on the other, nevertheless, the provision of cross references linking synonymous or equivalent terms, pointing to related terms, and distinguishing among ambiguous homographs tends to accompany human indexing more commonly than automatic indexing. This type of vocabulary management is increasingly common in automatic experimental systems and more advanced publicly available systems.
- Surrogation: Closely related to several of these key variables is the amount, nature, and style of information provided to the user about documentary units. For browsable displayed indexes, this will be connected to the amount and style of information provided in index headings, but also in subsequent documentary unit records that are linked to index headings. For machine matching systems, this variable relates to the size and style of the documentary unit records provided to the user for evaluation, ranging from very brief (such as titles only) to very lengthy (citations, abstracts, outlines, index terms, etc.) Newer methods of using visual displays (such as icons, graphs or network nodes) to characterize retrieved or relevant sets of messages have been more closely tied to automatic indexing techniques, but there is no inherent reason why they could not also be used with human indexing in the context of electronic IR database displays.

Because variables such as these have typically not been separately analyzed, it has been difficult, if not impossible, to determine whether the results of particular IR systems are due to automatic vs human indexing, or to different documentary units, different levels of indexable matter and exhaustivity, different types of search options provided (such as browsable displayed indexes vs machine matching), different levels of vocabulary specificity, different types or levels of vocabulary management, different types of surrogation, or to combinations and interactions among these features.

In 1978, Cooper commented on this problem in indexing research: "Reflecting the importance of the problem, the indexing process has been investigated extensively and a few insights have been achieved. [However] ... of the ... research that *has* been addressed to the central issue of finding the normative criteria that ought to govern human and automatic indexing, most has been burdened by the almost insurmountable methodological obstacles involved in making comparative evaluations of retrieval systems as wholes" (p. 107). When IR systems are considered only 'as wholes', then it is difficult, if not impossible, to suggest exactly which aspect of the system is contributing or detracting from desirable results.

Conflation of distinct variables continues to be a problem in IR research. In 1994 in an important anthology assessing the status of human and machine indexing by leaders in the indexing and IR research community (Fidel et al., 1994), Rasmussen characterizes the differences between automatic indexing and human indexing as that between the "relative effectiveness of controlled vocabulary vs free text" (Rasmussen 1994, p. 241). With the advent of full-text IR databases, this comparison has progressed to 'full-text searching' vs 'controlled vocabulary indexing' (p. 245). In each of these examples, two different variables have been conflated. It is possible to present controlled vocabulary terms for searching based on either automatic or human analysis, so the first of these comparisons should appropriately focus on the presence or absence of vocabulary management, separating that attribute from automatic vs human indexing. Similarly full-text searching has to do with exhaustivity and indexable matter, so in a genuine comparison between human vs machine indexing, or between free-text terms vs controlled vocabulary, these attributes (level of exhaustivity, extent of indexable matter) should be as similar as possible. All the papers in this anthology are valuable and useful, but they also illustrate the continuing difficulty of isolating the many different aspects of IR database design for assessing the impact of each variable.

If research into the merits of automatic vs human indexing has been inconclusive, the actual experience of IR database producers and users is persuasive. The fact that IR databases that rely solely on automatic indexing have been economically successful means that the users who are paying for them (either in actual financial outlay or in time spent using them or both) find them sufficiently effective to justify the cost. In some situations, no other options are available.

Fidel (1991) has verified preferences and usage patterns of professional searchers. When they have a choice between automatic indexing and human indexing as the basis for a search, they often opt for automatic indexing, depending on a whole array of other considerations, which Fidel explores. Again, however, choosing automatic indexing means also choosing, in most cases, a greatly expanded level of exhaustivity, much larger indexable matter, much smaller documentary unit, a higher level of specificity, a much larger indexing vocabulary, and little or no vocabulary management. It also provides access to different types of indexing syntax and searching options, which can be much more flexible in certain situations. On the other hand, choosing automatic indexing usually limits a user to electronic searches, as opposed to browsable displays. Thus, when a searcher chooses automatic indexing, it is not clear which features are the most influential. These are *not* simple choices limited to automatic vs human indexing.

The bottom line is clear, however: automatic indexing works! And it appears to work just as well as human indexing, just differently. Automatic indexing is also considerably faster and cheaper than indexing based on human intellectual analysis. Automatic indexing can be applied to enormous collections of messages (such as the world-wide web) where the volume of texts and constant change, both in individual texts and in the composition of the collection as a whole, makes human indexing impractical, if not impossible.

The challenge for IR database designers is to determine, for particular clientele, particular types of messages, texts and documents, in particular subject areas and for particular purposes, how expensive human analysis and fast, cheap machine analysis can best be deployed to maximize effective retrieval results at the lowest overall cost.

4. Human analysis for indexing

Ironically, much more is known about automatic machine methods of analysis for indexing than about human methods, because machine methods must be rigorously described in detail for the computer to carry them out. Human indexing has been performed for millennia, at least since the invention of methods for recording messages on long-lasting media, but understanding exactly how it is done is limited to the rather vague guidelines that IR database producers provide to their staffs and the distinctly general explanations that experts suggest in textbooks and training materials, as well as very preliminary results of research into the workings of the human mind and brain.

Brain scientists, neuroscientists, and cognitive psychologists are making progress in understanding how humans think and perform mental tasks. Recent advances were well described for the lay person in a series of articles in *The New York Times* (Hilts, 1995), but the steps that might take place in the mind of an indexer are still only suggested hypotheses. A few scholars have begun to address the specific act of indexing, as a kind of problem solving (David et al., 1995). Their "research program is an attempt to apply theories and methods from cognitive psychology to the study of indexing behavior" (p. 49). Members of the David et al. research team have reported related research in Bertrand-Gastaldy, Lanteigne, Giroux and David (1995), Bertrand and Cellier (1995) and Bertrand, Cellier and Giroux (1996). Comparing subject-matter experts vs non-experts in the 1996 study, they found that indexers "not familiar with the content based their judgments on surface-level features of the information. ... Identifying important concepts could be due to perceptual processing based on specific cues, as well as conceptual processing based on prior knowledge of the documentary language and the domain to be indexed" (p. 419). The 'documentary language' is the indexing language or controlled vocabulary used. Earlier psychological research related to human indexing is summarized by Farrow in "A cognitive process model of document indexing" (1991). Farrow notes that "the comprehension of text for indexing differs from normal fluent reading" in the following ways: time constraints; rapid text scanning for perceptual cues "to aid gist comprehension"; task-oriented rather than learning-oriented comprehension; immediate production of some text representation (abstract, index heading or terms, classification category or notation); and the repetition of text processing "by experienced indexers working with a restricted range of text types". He explores the interplay of perceptual (cues from text) and conceptual (prior knowledge) processing and the "allocation of mental resources to text processing" (p. 149).

The general consensus among indexers and theoreticians is that human indexers perceive (read, view, examine, listen to) a text, interpret the message encoded in the text as they understand it (influenced by previous experience and current personal knowledge, including their interpretations of any instructions given them), and then describe their version of the message, plus any important text or document features, in accordance to rules and patterns for the type of index they are working on. Not much more detail than that is provided by experts in indexing. Here are examples of explanations provided by leading experts in human indexing.

4.1. Nancy Mulvany

In *Indexing Books* (1994), Mulvany says:

> "I do not believe that indexing can be taught. ... [T]he ability to objectively and accurately analyze text and to produce a conceptual map that directs readers to specific portions of the text involves a way of thinking that can only be guided and encouraged, not taught. ... Indexing cannot be reduced to a set of steps that can be followed" (p. vii–viii).
>
> "...[T]he indexer's ability to thoroughly digest the intentions of the author and anticipate the needs of the readers, thereby producing a knowledge structure that is sensible and usable, involves the application of abilities and skills that are inherent in some individuals and not in others" (p. 39).
>
> "An indexer with a clear idea of the scope of the book itself and a general understanding of the subject matter and the audience will be in a position to distinguish between relevant and peripheral information.
>
> "Distinguishing between relevant and peripheral information involves judgment. Careful exercise of such judgment is what sets a true index apart from a computer generated list of words" (p. 45).

Later, in Part II on automatic indexing, we shall see that modern indexing algorithms go well beyond simply generating lists of words, and that indeed, judgments are made based on a wide range of criteria, including those encoded in knowledge bases reflecting the significance of subject area and cultural understanding of their creators. Nevertheless, effective human indexing relies on a very sophisticated use of human intelligence. Machines are very far from simulating the work of a human indexer. Part of what a human indexer does is to interpret the text (understand the message). Human indexers do this in the context of their cultures and their personal experiences, including their prejudices, as well as taking into consideration user needs and desires. Consequently, an index based on human indexing may not travel well between cultures. A freedom fighter in one culture may be a terrorist in another.

But the machine also has a culture: the culture imposed by its programmers. For example, a knowledge-base that would associate certain strings of language text with the concept of 'terrorism' would use the understanding of that concept in the context of the culture of the programmers. A simpler index, 'a computer generated list of words', that would use only a simple manipulation and accounting of the symbols found in the text, would be much closer to an objective index that could be used across cultures, and even across languages.

4.2. Lois Mai Chan

Cataloging is the application of indexing procedures to a particular collection of documents. Classification is indexing that results in conceptual groupings of topics, rather than alphabetic arrays of headings. Chan has written widely on cataloging and classification. Here is what she says about subject analysis in her popular introductory textbook, *Cataloging and Classification: An Introduction* (1994):

> “No matter what the subject access system within which a subject cataloger is working, subject analysis of a particular work or document involves basically three steps: (1) determining the overall subject content of the item being cataloged, (2) identifying multiple subjects and/or subject aspects and interrelationships, and (3) representing both in the language of the subject headings list at hand.
>
> “The most reliable and certain way to determine the subject content is to read or examine the work in detail” (p. 166).

4.3. Robert Fugmann

Writing on “recognizing and selecting the essence of a text”, German indexing theorist Fugmann (1993) says:

> “Essence recognition is a most fundamental and cognitive process in science. The kind of subjectivity which is inherent in this process does not detract from its fundamentality. To the contrary, all progress in cognition has been achieved through subjectivity. At some time, a genius saw or hypothetically assumed lawful relations which up to then had been hidden to everybody” (p. 74).

4.4. Dagobert Soergel

Soergel, whose book *Organizing information: principles of data base and retrieval systems* (1985) won the ‘Book of the Year’ award from the American Society for Information Science, emphasizes the importance of “request-oriented indexing”. This means not just indexing according to the message of a text, but according to what users are looking for. He portrays indexers as scouts who are sent out on behalf of users to look for answers to particular questions. Of course, it is economically unfeasible to have an indexer for every information seeker, looking through masses of documents for answers to a single query, so queries must be aggregated or batched, and indexers should look for answers to all of these anticipated requests as they examine documents (p. 50–56). This is why a fairly detailed subject scope statement for IR databases is so important – a statement of the kinds of questions that users will want (and therefore should be able) to ask of an IR database. For indexers the subject scope statement can serve as a kind of questionnaire that needs to be answered for each document that is indexed.

4.5. F. W. Lancaster

Widely recognized as an authority on indexing, abstracting, and vocabulary management, Lancaster echoes Soergel’s ideas when he writes:

> "Effective subject indexing involves deciding not only what a document is about but also why it is likely to be of interest to a particular group of users ... The same publication could be indexed rather differently in different information centers and should be indexed differently if the groups of users are interested in the item for different reasons" (1991, p. 8).

4.6. Robert Fairthorne

Writing earlier, the British information scientist Fairthorne (1971) also deals with the thorny issues of aboutness and purpose:

> "What discourse speaks of, – that is, what it mentions by name or description – , are amongst its extensional properties. What discourse speaks on, – that is, what it is about – , is amongst its intensional properties. Thus, its topic cannot be determined solely from what it mentions. For this, one must take into account extra-textual considerations, such as who is using it for what purpose, what purpose the author intended it to be used for, and for whom or for what the librarian, or other manager of messages, acquired it. ... [T]opics are not the properties of text marks as such, but of discourse. ... [T]o create or assign topics to a text we must consider it in the wider context of what kind of person uses it for what, what other texts are used, and in what ways do these texts depend on each other" (p. 361, 362).

4.7. Brian O'Connor

In his philosophical *Explorations in indexing and abstracting: pointing, virtue, and power*, O'Connor (1996) defines *subject* as "a relationship between each individual and the squiggles that constitute the document. If the subject were a single, self-evident entity then subject representation would be only a slight challenge. ... The circumstances of the patron and the nature of the squiggles combine to generate a unique, user-dependent meaning for each engagement with each document" (p. 51). O'Connor's 'squiggles' and Fairthorne's 'text marks' are the symbols used to create a text that represents the message of the creator of the text. Later, O'Connor addresses the concept of 'aboutness', which is usually central to the human indexer's analysis and subsequent description of a message: "Aboutness is the behavioral reaction of a person to a document. Each patron may have a different experience with the same document" (p. 147). It is clear that indexers, as well as patrons, each have different experiences with messages and texts.

4.8. Hans Wellisch

Veteran indexer and scholar of indexing, Wellisch writes in the first edition of his encyclopedic *Indexing from A to Z* (1991) that "the mental activities resulting in the formulation of index entries cannot be observed and can therefore not be objectively described, measured, or reduced to fixed rules similar to those that govern the purely technical aspects of indexing such as filing or capitalization of words" (p. 175). A little later he explains that "The problem is that the topics or subjects dealt with in a document (or the 'aboutness' of that document) and their relevance for its intended or prospective users is in many if not most cases vague and difficult or even impossible to pin down exactly, because it is almost always a matter of subjective opinion" (1991, p. 178).

Wellisch expands on these ideas in the second edition of this book (1996):

> "Beginning indexers often ask whether there is a theory of indexing. If by this is meant a coherent system of propositions explaining the mental activities involved in transforming a text into its index, the only honest answer is that we do not have such a thing. ... All we know is that indexing is a highly complex intellectual process involving the use of language in a specific and somewhat artificial way, and that it is also to a considerable extent a matter of intuition, the workings of which cannot be reduced to fixed rules. ... In this respect, indexing is similar to other mental operations such as the recognition of faces and voices: we know that we can do it, but cannot describe in so many words how we do it, nor can we reduce it to a set of rules" (p. 218–219).

4.9. Patrick Wilson

One of the most detailed analyses of the challenges of human indexing appears in the chapter 'Subjects and the sense of position' in Wilson's classic treatise *Two kinds of power: an essay on bibliographic control* (1968). He writes: "It is difficult enough in any field of human behavior to discover a man's purposes by examining the results of his activity; and the difficulties must be much greater than ordinary in the case of those most complex products of human effort, writings" (p. 81). He concludes: "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" (p. 89).

Finally:

> "Any actual physical object is, as the old philosophers would have said, 'determinate in every respect'; whether we can decide on its actual shape and size and weight and color, it must have some definite shape and size and so on, at any moment. There are no doubt limits to the precision of measurement and description possible to us, but there must be some descriptions which are the exactly correct descriptions of its various characteristics, even if we cannot, because of physical limitations, tell which ones those are. Things are what they are; our descriptions may be vague and imprecise and indefinite, but there can be no vagueness or indefiniteness about the things themselves. Now we have an inclination to say that what is true of things must be true of writings 'about' things; a writing must have a definite subject, and there must be some description of the subject that is absolutely precise and accurate, all other descriptions being imprecise or inaccurate. It is this inclination which must, I think, be resisted; of course we can always formulate descriptions which are obviously and definitely *not* descriptions of what a writing is about, but we cannot expect to find one absolutely precise description of one thing which is *the* description of *the* subject, all others 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" (p. 89–90).

4.10. Arlene Taylor

Taylor (1999) has analyzed Patrick Wilson's commentary on concept analysis and has named the approaches to analysis, or types of analysis, that he identified (p. 138–139):

> "Purposive method. One tries to determine what the author's aim or purpose is. If the creator of the information package gives a statement of purpose, then we can presume to know what the work is 'about'. ..."

> "Figure-ground method. Using this method, one tries to determine a central figure that stands out from the background of the rest of the information package. However, what stands out depends on the observer of the package as well as on its creator. What catches one's interest is not necessarily the same from person to person, and may not even be the same for the same person a few weeks later"

> "Objective method. One tries to be objective by counting references to various items to determine which one vastly outnumbers the others. Unfortunately, an item constantly referred to might be a background item (e.g., Germany in a work about World War II). ..."

> "Appealing to unity or to rules of selection and rejection. When using this method one tries to determine what holds the work together, what cohesiveness there is, and what has been said (selection) and not said (rejection). Again, the observer of the information package has to be objective and also has to know quite a lot about the subject in order to know what was rejected."

4.11. Birger Hjørland

We close this section with quotes from one of the more recent analyses of the nature and purpose of human indexing, and more broadly, the whole field of information science: *Information seeking and subject representation: an activity-theoretic approach to information science* by Hjørland (1997). Hjørland urges the study and practice of indexing (or more broadly, the facilitation of information seeking through subject representation), within the context of an activity-theoretic focusing on the working domains of users as the context and impetus for their information seeking. This might encompass, as well, various social or cultural domains for users seeking informative or uplifting or entertaining messages relating to such life concerns as occupational or career options, spiritual life, or entertainment. Hjørland writes:

> "... [K]nowledge is organized in learned institutions, in professionals [i.e. professions?], in journals, in libraries, and so on. Knowledge is produced as a part of human activities and tied to the division of labor in society. From the point of view of activity theory, this is the primary organization of knowledge. The organization of knowledge in IS [information science] is secondary or derived, and so is, in certain ways, the cognitive organization of knowledge in individual minds. The organization of knowledge is determined by evolution

> of different kinds of functional forms and principles which vary according to specific needs, contents, and conditions" (p. 45).

Comparing activity theory or the domain analytic approach to other current paradigms in information science (the information object paradigm, the cognitive paradigm, the behavioral paradigm, and the communication paradigm), Hjørland writes:

> "The domain analytic paradigm is a theoretical approach to information science (IS) which states that the best way to understand information in IS is to study the knowledge domains as discourse communities, which are parts of the society's division of labor. Knowledge organization and structure, cooperation patterns, language and communication forms, information systems, and relevance criteria are reflections of the objects of the work of these communities and of their role in society. The individual person's psychology, knowledge, information needs, and subjective relevance criteria should be seen in this perspective" (p. 106).

In this context:

> "What constitutes a subject according to activity theory is not independent of purpose, viewpoint, or theoretical influences. What constitutes a subject for one discipline or theory need not constitute a subject for another (p. 84). ... Another consequence of this view is that different theoretical backgrounds, paradigms, world views or metaviews – which can be either disciplinary, interdisciplinary, or cross-disciplinary views – are central to subject analysis" (p. 85).

One conclusion that must be drawn from this survey of expert commentary on human intellectual analysis of messages is this: the one thing we definitely do know about human indexers is that they rarely agree on what is important in a message, or what to call it. Research on human indexing shows that human indexers share the enormous variability that characterizes all human use of language. Saracevic (1991), synthesizing "major findings from several decades of research on the magnitude of individual differences in information retrieval (IR) tasks", summarizes this variability in this way: "the degree of agreement (expressed by a variety of measures) in human decisions related to organizing, representing, searching and retrieving of information is relatively low and the range of performance relatively high. The agreement hardly reaches about one fourth of cases involved (and often it is lower), and the range of performance routinely varies 10-fold or more. However, the notion of 'low' or 'high' here may be inappropriate. The observed ranges may be all that is expected for these tasks, i.e., they may be 'normal"' (p. 85). 'Range of performance' refers to such measures as time taken to perform given tasks, error rates, number of documents retrieved in order to achieve a pre-determined level of recall, or recall and precision ratios (which measure the percentage of relevant documents retrieved and the ratio of relevant to irrelevant documents retrieved, respectively).

There is a large literature on indexer (and searcher) inconsistency. Overviews and summaries have been provided by Leonard (1977) and Markey (1984). Iivonen (1994) has analyzed the nature of indexer inconsistencies and distinguished differences in conceptual analysis from differences in naming. She found the latter differences – what to call topics – to be more prevalent than disagreements on key concepts.

Saracevic and Kantor (1988a,b) have investigated variation among searchers, verifying that searchers behave much the same way as indexers. After all, searchers describe (that is, they analyze and index) information needs or desires and hope that their indexing will match that of indexers of potential answers or responses. Lourdes Collantes (1995) focused on variation in naming among potential cataloging users.

5. Cognition vs social construction in human analysis for indexing

The prevailing view among indexers and indexing experts is that human indexing of messages and texts is largely a cognitive process. Bernd Frohmann has vigorously protested what he considers to be excessive preoccupation with the mental or cognitive processes of indexing. Yet almost everyone who has studied the indexing process has described it as a cognitive process governed or influenced by the workings of the individual minds of indexers. Emphasis has been on the essential characteristics of human mental operations. Proponents of this approach call it the cognitive view or approach to information retrieval. Frohmann considers this to be useless 'mentalism'.

In his 1990 article, 'Rules of indexing: a critique of mentalism in information retrieval theory', Frohmann quotes a number of scholars on the nature of human indexing, which he describes as "the implicit or explicit representation of a document by an indexing phrase". He summarizes current understanding of the indexing process as one that "continues to be lamented as an intellectual operation both fundamental to indexing yet so far resistant to analysis" (p. 82). Here are some examples of the passages that Frohmann quotes, not unlike those already quoted in the previous section.

5.1. A. C. Foskett

> "Scanning a text to decide what it is about is the key operation in indexing, yet it is the least discussed and the least reducible to rule" (1982).

5.2. Clare Beghtol

Basing her analysis on the work of van Dijk, Beghtol writes:

> "The ability to restate the semantic aboutness of a discourse ... originates in an automatic reductive cognitive process of summarisation that allows a reader to construct during reading a notion of the text topic and to store it in hierarchically-arranged memory structures for later recollection" (Beghtol 1986, p. 90).

5.3. James D. Anderson

Anderson is greatly honored to be classified, along with his contemporary Clare Beghtol, with leaders of an earlier generation of information scientists such as Foskett, Farradane, and Artandi. Frohmann says, "James D. Anderson (1985) is in the vanguard of library science's appropriation of mentalism by boldly representing the mind as a library, complete with a technical services department performing indexing operations of which, at least until we have read Anderson, we are completely unaware" (p. 84–85). Here is an excerpt from Frohmann's Anderson quote:

> "[T]he mind of a human indexer ... receives the symbols via normal perception processes, matches them against those stored in the mind, determines what concepts are represented and which are important, then chooses symbols to represent these concepts in the index" (p. 295).

Frohmann (1990) appeals to the philosopher Ludwig Wittgenstein to contest the 'mentalism' represented in current work on human indexing. Frohmann contends that indexing 'rules' are not (or should not be?) based on cognitive processes resident in the mind, but on socially constructed rules apprehended by indexers. So he argues that the focus must shift "indexing theory away from rule *discovery* and toward rule *construction*". He continues:

> "By Wittgenstein's lights, indexing rules governing the derivation of indexing phrases from texts are properly seen as instruments of particular social practices. Theory in indexing is therefore confronted with the challenge, not of discovering rules followed unconsciously, but of constructing, consistent with stated purposes, explicit, well-formulated, and strict rules which may be used to yield indexing phrases from texts. The problem of indexer inconsistency, for example, is not solved by first discovering and then bringing order to the motley of tacitly known rules unconsciously followed by indexers, but by replacing prevailing vague rules, for example, those providing no more guidance than 'express the subject of this text in a concise statement', which indexers perforce interpret variously, with rules sufficiently precise to serve as justifications, as standards of correctness, and as instruments of indexer training" (p. 94).

However, it is not at all clear how Wittgenstein's views might be applied to create more precise rules for human indexing. Frohmann does not suggest how this should be done, only that it ought to be done. If indexing rules are vague and difficult to formalize when they reflect an individual's cognitive processes, how can it be argued that they will become more precise when they reflect the 'stated purposes' of a user, a group, a culture, or society in general? Would the vague rule 'express the subject of this text in a concise statement' become somehow less vague if it reads something like 'determine and express the subject of this text according to the purposes of the intended user'? Indeed, it could perhaps be argued, according to Wittgenstein, that cognitive processes are themselves a reflection and a construction of the individual's culture. Persons, including indexers, understand a text based on what they are, who they are, when they are, where they are.

This controversy regarding the proper basis for indexing research and theory has attracted little attention in the primary literature of message and text analysis, indexing and cataloging, whether by humans or machines. It is much closer to similar and sometimes fierce debates in some of the newer post-modern disciplines. A prominent example is queer theory, "an ensemble of strategies of reading and interpreting texts (whether literary or social) that has emerged in the last decade and has been profoundly influenced by poststructural theory ... an eclectic and diffuse ensemble of practices influenced by the contestatory realms of psychoanalysis, Marxism, cultural materialism, semiotics, social constructionism, structuralism, and feminism" (Bredbeck, 2000). A dominating feature of queer theory is the on-going argument between the essentialists, who see sexual orientation as something innate and rather constant across humankind (and indeed, beyond our species), vs the social constructionists, who see sexual orientation as very much a social creation of every culture. The essentialist position in queer theory is comparable to that of the cognitive approach in human indexing (mentalism to Frohmann), which seeks to understand the essential nature of the mental processes of the human mind as applied to message and text indexing. Frohmann is clearly on the side of the social constructionists. It is likely that both camps possess some truth – perceptions that will lead to more complete and accurate understanding of complex phenomena. It may indeed be the case that there is something 'essentially' innate and constant about being homosexual, heterosexual, or bisexual, but what that means and how it is played out is certainly a product of one's culture. Similarly it is hard to imagine that fundamental (essential) human cognitive processes do not play a large role in human indexing, but the application of these processes are just as surely influenced, even determined by social forces and contexts. For a good summary of this debate in queer theory, with references to relevant publications, see Hogan and Hudson (1998).

Another arena where the battle between the essentialists and the social constructionists rages unabated is in gender studies. Tavris (1998) does a good job of describing these two approaches, calling them "two antithetical trends in the current study of gender" (p. 127):

> "One, the oldest empirical tradition, takes an essentialist approach. Essentialists regard a gender-related attitude, trait or behavior as being something embedded in the person – internal, persistent, consistent across situations and time – and thus they tend to regard the sexes as 'opposites': men are aggressive, women pacifistic; men are rational, women emotional. ... For some feminist psychologists, men and women have inherently different ways of knowing, ways of speaking, ways of moral reasoning and the like. For neuroscientists, men's and women's brains operate differently. For sociobiologists, male promiscuity and female monogamy are opposite, hard-wired reproductive strategies. (When sociobiologists learned that the males of many species are nurturant and monogamous and the females of most species are promiscuous, they reconnoitered and decided that these reproductive strategies too are adaptive)".

As an aside, if we were to accept this essentialist approach to gender studies, then we must make sure that gender is considered as an essential factor in our cognitive study of human indexing! Carol Tavris continues:

> "In contrast, researchers who take a social constructionist approach vigorously dispute all forms of essentialism. Social constructionists hold that there is no 'essence' of masculinity and femininity, for these concepts and labels are endlessly changing, constructed from the

> eye of the observer and from the historical and economic conditions of our lives. 'Opposition', for example, is a social construction, not an empirical reality; it is a stereotype that blinds us to the greater evidence of gender similarity. … Constructionists regard gender as a performance, not an attribute. People don't *have* a gender, they *do* a gender, which is why their behavior changes so much depending on the situation''.

Returning to the arena of indexing, Frohmann is surely correct that much of indexing rests on rules informed by culture, and his efforts to get our field to focus on effective rules to guide our efforts at indexing are very important and to be encouraged. The cultural bias of classification schemes (a form of indexing language) has long been recognized, along with the prejudicial nature of many established subject headings in alphabetical indexing (such as the Library of Congress headings 'Pilgrim Fathers' and 'Hotel maids', changed to 'Pilgrims (Plymouth Colony)' and 'Hotel cleaning personnel' in 1976 and 1989, respectively). With respect to the human analysis process, however, it also seems clear that much of that process is also governed by the cognitive procedures of our minds. As in the contending sides in queer theory, gender studies, and many other of the human and social sciences, both sides, both approaches, contribute important aspects toward more complete and accurate understanding.

Here's one final quote (for the time being) from Frohmann (1990), emphasizing the social context of the rules that he advocates:

> ''… mentalism's focus on processes occurring in minds conceals the crucial social context of rules. Since we do not understand the rule we are constructing without understanding its social context, or the way it is embedded in the social world, its point, its purpose, the intentions and interests it serves, in short, the social role of its practice, indexing theory cannot avoid investigation into the historical, economic, political, and social context of the rules in its domain. Mentalism, on the other hand, either erases the social dimension altogether by conceiving rules as operating in disembodied, ahistorical, classless, genderless, and universal minds, or else acknowledges it only by expanding the set of rules of mental processing'' (p. 96).

We will return to these issues at the end of the next section on rules for human indexing.

6. Rules for human indexing

Just about everyone agrees that there is a two step process in human indexing: (1) the analysis of a text, resulting in the creation of some kind of notion, phrase or statement representing the meaning and/or features of a message and possibly also of its text and document; and (2) the translation of this notion, phrase or statement into the indexing language or format prescribed by the IR database producer or the design of the index.

Most of the rules regarding indexing, cataloging, and classification relate to the second step: the translation of the result of message-text-document analysis into terms and forms mandated by an indexing language or presentation format. These rules and procedures relate to term specificity, vocabulary management, syntax for strings and headings, surrogates and surrogate displays, and search interfaces and thus are outside the scope of this essay. Here the focus is on attempts to formulate rules, guidelines, or procedures, for the first step: the analysis of messages, texts and documents and the creation or production of the preliminary notion or statement of meaning, topic, importance or application. These attempts may be seen as at least initial efforts to respond, at least in part, to Frohmann's plea for the construction of effective rules that will contribute to better indexing.

Both the British Standards Institute (BSI) and the International Organization for Standardization (ISO) have issued standards (BS 6529: 1984, ISO 5963–1985) with recommendations on "methods for examining documents, determining their subjects, and selecting index terms" (International Organization for Standardization, 1985). These methods include the following list of questions an indexer should ask of a text (International Organization for Standardization, 1985, p. 2). The British standard contains the same list, in slightly different wording (British Standards Institute, 1984, p. 3):

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

These are offered as examples of general factors which are likely to apply in any subject field. Other questions may need to be formulated within a special discipline.

These guidelines are meant to suggest a general approach to analysis, but they certainly don't constitute anything like a rigorous procedure that would produce predictable results. They ask indexers to analyze "prominent" topics and features, reminding us of efforts by Wilson, Taylor and many others (described in previous section) to suggest ways in which 'prominent' or 'important' might be determined.

Commenting on the ISO guidelines, Hjørland notes, in line with his concern for domain analysis quoted previously, that:

> "Even though the ISO standard can in many ways be reasonable and useful, it can be noted that the prescribed guidelines for subject analysis are fairly document-centered. ... The standard does not offer any specific insights into how disciplines or user groups differ or explain the fact that they require particular domain-specific analyses. The document could, for instance, have mentioned that where social science and humanities disciplines are concerned, the indexers need to pay special attention to worldviews and theoretical orientation; that is, it could have emphasized the importance of the subject-object relation in these disciplines. The fact that this was not done shows the danger in publishing international standards in this area: paying attention to inconsequential external features in the analysis gives a false impression of general, objective criteria that in reality cannot be de-

> scribed in a standard, since subject analysis is a theoretical and intellectual process that is dependent on the concrete situation within the knowledge domain in question" (1997, p. 44–45).

As Hjørland points out, guidelines like the ISO/BSI lists consist of broad generic categories that can be applied to most if not all messages, texts and documents, with little or no attention to the special needs or interests associated with particular domains of human activity. Specialized domains deserve more specific lists of aspects to guide indexers. An example of such a specialized list is the category list used currently by the Modern Language Association of America for its international literary studies database:

1. MLA Categories for Analysis of National Literatures (Modern Language Association of America, 1997). Examples have been added for each category:
   - Specific literatures: e.g., English literature, American literature, Chicano literature, Puerto Rican literature.
   - Performance media: e.g., theatre, story-telling, recitation.
   - Languages (if different from language of national literature): e.g., English, Spanish, Swahili.
   - Periods: e.g., 20th century, 19th century.
   - Individuals (real): e.g., Thomas Hardy, Emily Dickinson, Abraham Lincoln, James Baldwin.
   - Anonymous works: e.g., Beowulf, Thousand and one nights, parts of the Bible.
   - Groups/movements: e.g., Avant Garde, Beat Generation, hippies, lesbian poets, African American writers, children.
   - Genres: e.g., poetry, drama, non-fiction novels.
   - Works: e.g., The wind in the willows, Alice in wonderland, Giovanni's room.
2. Further Description of Literary Topic:
   - Features: e.g., dialogue, poetic realism.
   - Literary techniques: e.g., visual metaphor, imagery, symbolism.
   - Themes/motifs/figures/characters: e.g., [treatment of] love, hate, war, Manifest Destiny, salvation, Huck Finn, Cinderella.
   - Influences (recipients): e.g., [influence on] Harlem Renaissance, Generation of 1898.
   - Sources: e.g., [influence of] Harlem Renaissance, Generation of 1898, The Bible.
   - Processes: e.g., characterization, translation.
3. Description of Document Author's Processes:
   - Types of scholarship: e.g., criticism.
   - Methodological approaches: e.g., sociological approach, psychological approach, Marxist approach.
   - Theories: e.g., Freudian theory, evolution (as theory).
   - Devices/tools: e.g., computers, concordances.
   - Disciplines: e.g., aesthetics, historiography.
   - Scholars: e.g., critics, folklorists (also particular individuals).
   - General/miscellaneous: A place for indexers to add anything that does not fit in the established categories!
   - Special types of documents: e.g., bibliography, film, slides, videotapes, multimedia.

The MLA list is based directly on a detailed analysis of the working habits and interests of literary scholars (Anderson, 1979).

Such specialized lists of categories are designed to increase the chances that indexers will not miss important aspects of topics or features, but they do not insure uniformity in the identification of important aspects. This variability appears to be due to those subjective, cognitive, 'mentalist' processes going on in our minds, and the fact that the mind of every individual is different.

Most guidelines for indexers, including those illustrated here, and the implied rules for analysis that they suggest, focus mostly on the content and features of messages, texts and documents, and less on potential uses by interested persons (this document would be good for...), and even less on the relevant characteristics of potential users and their information needs and information seeking behavior. In addition, they side-step entirely issues of quality, authority, accuracy, and appropriateness, other than some general categorization for intended audience by age level and possibly by occupation or level of expertise.

Soergel is a strong proponent of the need for qualitative judgments in request or user-oriented indexing – the indexing of messages *for* particular users, rather than just indexing the content and features of messages, texts, and documents. Thus, he suggests the use of index terms like 'read immediately' and 'danger to our business' which indicate the importance of messages for particular users or purposes and at specific times (it is urgent now!) (1985, p. 229).

Frohmann (1990) is seeking much more thorough and rigorous rules than those presented here – rules based on a careful analysis of social purpose. Here are questions that Frohmann asks us to consider in formulating such rules (p. 97–98):

> "What are the purposes of text retrieval in various social contexts and of various kinds of users? The problems involved in constructing rules for indexing languages incorporating the categories of the dominant social institutions of industry, research and development, commerce and finance, universities and the like, are not identical to those involved in designing text retrieval services for marginal groups outside the dominant institutions, such as the economically disadvantaged or the victims of racial, class or gender discrimination. ...
>
> "Some questions demand critical inquiry. Does text retrieval fulfil a need, or does it satisfy a want? Indexing rules will look quite different depending upon how this question is answered. Wants are explicitly recognised and admitted; they reflect the agents' goals, purposes, and intentions. Not all needs are known, and some fly in the face of wants. For example, not everyone knows what they need to prevent AIDS, and not everyone wants what they need."
>
> "Other questions demand political analyses. What and whose aims, goals, strategies, and intentions are fulfilled by text retrieval in the social world in which indexes, abstracts, online databases, catalogues, thesauri, bibliographies and the whole range of retrieval apparatus make their appearance? Is the retrieval of *truth* a desirable (or even feasible) retrieval practice, as Patrick Wilson (1978) has suggested? ... Is the spread of disinformation in the service of ruling elites the purpose of text retrieval? ..."

These are big questions, most of which our field has largely ignored.

7. Human indexing rules based on probabilistic analysis

Cooper (1978) is cited by Frohmann (1990) as one who merits special praise for his search for indexing rules: "Whether or not one agrees with Cooper's proposed indexing rules..., their theoretical base is nonetheless exemplary, because his explicit recognition of rules as instruments of training and as standards of evaluation locates them within a conception of rule-following as a practice, a custom, or a technique" (p. 94).

Cooper writes (1978):

> "It is widely acknowledged among information scientists that the problem of indexing ('cataloging', 'classification', etc.) is one of the fundamental problem areas, if indeed not *the* central theoretical problem, of document and reference retrieval. If correct normative rules of indexing could be formulated, the accessibility of man's entire store of recorded knowledge would be enhanced. However ... there is as yet no consensus among experts about the answers to even some of the most basic questions of what indexers ought to be told to do or of how an indexer's performance should be evaluated" (p. 107).

Using decision theory and utility theory, Cooper suggests a probability rule, which he reduces to this: "The assignment of a term to a document is justified if the average utility associated with that assignment is positive, and unjustified if it is negative" (p. 110). He then introduces a method of 'gedanken experimentation' (thought experimentation) by which an indexer may estimate average positive and negative utilities.

In the context of language texts, Cooper insists that "the gedanken indexer must read or at least scan the document to be indexed in order to do his job well, gain some idea of what it is about, and keep his user population clearly in mind" (p. 112). For a broad-based assessment of utility, the indexer must "take into account many aspects of utility often neglected in traditional subject cataloging, such as whether the document is written in a language and on a level which the users will be able to understand, whether it is up-to-date enough for his user population, even (if he can) whether the document is of 'high quality'" (p. 112).

After simplifying the recommended procedure as much as he can, Cooper asks the gedanken indexer to make two utility predictions: (1) the odds *against* satisfaction, to indicate the chances that a searcher using a particular term would *not* find a given document useful; and (2) the average predicted payoff – the amount a satisfied user would be willing to pay for the document.

In actual practice, this decision and utility theoretic approach is often simplified even further and shortened to: if a person using this term is likely to want this document, then use the term.

Cooper's rules for estimating the utility of documents for persons using particular search or index terms for topics and features can be combined with the lists of aspects to look for, with which this section on rules for human indexing began. When an indexer sees evidence of material related to a topic or features of known interest, then he or she should note it and assign an appropriate candidate term. The candidate term can then be analyzed by Cooper's decision-utility theoretic procedure as a way of gauging whether the treatment of the topic or the manifestation of the feature is really good enough or significant enough or clear enough or up-to-date enough (etc.!) to be tagged for retrieval.

Cooper's suggestions never caught on in the human indexing community. Adding numerical values to the procedures that expert indexers already follow (assessing needs and interests of potential users and comparing them to the characteristics of messages) did not seem to provide much added value. The fundamentals of just how user needs and desires might be assessed and message, text, or document characteristics be evaluated were not addressed. Cooper's ideas about collecting data about actual utility values for the purpose of training indexers and evaluating indexing performance had value, but they had no obvious advantage over other approaches to testing and evaluation of indexing effectiveness. The addition of numerical values doesn't change the basic fact that the judgments of human indexers are still very subjective, and their precise basis is still very difficult to specify. Despite praise by Frohmann, it does not appear that Cooper has laid down the rigorous rules that Frohmann was seeking.

8. Continued in Part II

Part II of this essay will address automatic indexing; the application of modern methods of indexing, both human and machine, by indexing and abstracting services; and possible ways for allocating both approaches for maximum benefit.

References

Anderson, J. D. (1979). Contextual indexing and faceted classification for databases in the humanities. *Information choices and policies: Proceedings of the 42nd Annual Meeting of the American Society for Information Science, volume 16; 1979 October 14–18. Minneapolis, MN* (pp. 194201). White Plains, NY: Knowledge Industry Publications.

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

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

Bertrand, A., & Cellier, J.-M. (1995). Psychological approach to indexing: effects of the operator's expertise upon indexing behaviour. *Journal of Information Science*, *21*(6), 459–472.

Bertrand, A., Cellier, J.-M., & Giroux, L. (1996). Expertise and strategies for the identification of the main ideas in document indexing. *Applied Cognitive Psychology*, *10*(5), 419–433.

Bertrand-Gastaldy, S., Lanteigne, D., Giroux, L., & David, C. (1995). Convergent theories: using a multidisciplinary approach to explain indexing results. In Forging new partnerships in information: Proceedings of the 58th Annual Meeting of the American Society for Information Science, volume 32; 1995 October 9–12; Chicago, IL. Medford, NJ: Information Today, (pp. 56–60).

Bredbeck, G. W. (2000). Queer theory. In G. E. Haggerty (Ed.), *Gay histories and cultures: an encyclopedia* (pp. 728–729). New York: Garland.

British Standards Institution (1984). *British standard recommendations for examining documents, determining their subjects and selecting indexing terms*. London: British Standards Institution, 6 p. (BS 6529: 1984).

Chan, L. M. (1994). *Cataloging and classification, an introduction* (2nd ed.). New York: McGraw-Hill.

Collantes, L. Y. (1995). Degree of agreement in naming objects and concepts for information retrieval. *Journal of the American Society for Information Science*, *46*(2), 116–132.

Cooper, W. S. (1978). Indexing documents by gedanken experimentation. *Journal of the American Society for Information Science*, *29*(3), 107–119.

David, C., Giroux, L., Bertrand-Gastaldy, S., Lanteigne, D., & Bertrand, A. (1995). Indexing as problem solving: a cognitive approach to consistency. In Forging new partnerships in information: Proceedings of the 58th Annual Meeting of the American Society for Information Science, volume 32; 1995 October 9–12; Chicago, IL. Medford, NJ: Information Today, (pp. 49–55).

Drabenscott, K.M. See Markey, K. (1984).

Fairthorne, R. A. (1971). Temporal structure in bibliographical classification. In Chan L. M., Richmond P. A., E. Svenonius (Eds.), *Theory of subject analysis: a sourcebook* (pp. 359–366). Littleton, CO: Libraries Unlimited.

Farrow, J. F. (1991). A cognitive process model of document indexing. *Journal of Documentation*, *47*(2), 149–166.

Fidel, R. (1991). Searchers' selection of search keys: I. The selection routine; II. Controlled vocabulary or free-text searching; III. Searching styles. *Journal of the American Society for Information Science*, *42*(7), 490–527.

Fidel, R., Hahn, T. B., Rasmussen, E. M., & Smith, P. J. (1994). *Challenges in indexing electronic text and images*. Medford, NJ: Learned Information for the American Society for Information Science.

Foskett, A. C. (1982). The subject approach to information. 4th ed. London: Bingley; Hamden, CT: Linnet Books. Frohmann (1990) cites this book for a quote from Foskett. Similar statements occur in this book, but the statement quoted by Frohmann could not be found. A 5th ed. was published in 1996: London: Library Association Pub.

Frohmann, B. (1990). "Rules of indexing: a critique of mentalism in information retrieval theory.". *Journal of Documentation*, *46*(2), 81–101.

Fugmann, R. (1993). *Subject analysis and indexing: Theoretical foundation and practical advice*. Frankfurt/Main: Indeks Verlag.

Hilts, P. J. (1995). Brain's memory system comes into focus. *New York Times*. Section C: 1, 3; 1995 May 30. Behind the veil of thought: Fifth article of an occasional series about advances in brain research. "The first article in this series, about evidence for a new theory of consciousness, appeared on March 21. The second article, about the way the brain processes emotion, appeared on March 28. The third one, about the brain's "working memory", appeared on May 2, and the fourth, about music and the brain, appeared on May 16".

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

Hogan, S., & Hudson, L. (1998). Constructionism vs essentialism. In Hogan S., L. Hudson (Eds.), *Completely queer: The gay and lesbian encyclopedia* (pp. 149–152). New York: Henry Holt.

Iivonen, M. (1994). Consistency in the selection of search concepts and search terms. *Information Processing & Management*, *31*(2), 173–190.

International Organization for Standardization, (1985). *Documentation – methods for examining documents, determing their subjects, and selecting indexing terms*. Geneva: ISO (ISO 5963-1985 (E)).

Lancaster, F. W. (1991). *Indexing and abstracting in theory and practice*. Champaign, IL: University of Illinois, Graduate School of Library and Information Science (A 2nd ed. was published in 1998).

Leonard, L. E. (1977). *Inter-indexer consistency studies, 1954–1975: a review of the literature and summary of study results*. Champaign, IL: University of Illinois, Graduate School of Library Science. (University of Illinois. Graduate School of Library Science. Occasional papers; no. 131).

Lunin, L. F., & Fidel, R. (1994). Perspectives on indexing (special issue). *Journal of the American Society for Information Science*, *45*(6), 569–636.

Markey, K. (1984). Interindexer consistency tests: a literature review and report of a test of consistency in indexing visual materials. *Library and information science research*, *6*(2), 155–177.

Modern Language Association of America (1997). MLA bibliography worksheet for national literatures. New York: MLA.

Mulvany, N. C. (1994). *Indexing books*. Chicago: University of Chicago Press.

O'Connor, B. C. (1996). *Explorations in indexing and abstracting: pointing, virtue, and power*. Englewood, CO: Libraries Unlimited 182 p.

Rasmussen, E. M. (1994). Indexing and retrieval from full-text. Introduction. In Fidel R., Hahn T. B., Rasmussen E. M., P. J. Smith (Eds.), *Challenges in indexing electronic text and images* (pp. 241–245). Medford, NJ: Learned Information for the American Society for Information Science.

Saracevic, T., Kantor, P., Chamis, A. Y., & Trivison, D. (1988). A study of information seeking and retrieving. I. Background and methodology. *Journal of the American Society for Information Science*, *39*(3), 161–176.

Saracevic, T., & Kantor, P. (1988a). A study of information seeking and retrieving. II. Users, questions, and effectiveness. *Journal of the American Society for Information Science*, *39*(3), 177–196.

Saracevic, T., & Kantor, P. (1988b). A study of information seeking and retrieving. III. Searchers, searches, and overlap. *Journal of the American Society for Information Science*, *39*(3), 197–216.

Saracevic, T. (1991). Individual differences in organizing, searching and retrieving information. In *Systems understanding people: Proceedings of the 54th Annual Meeting of the American Society for Information Science*, volume 28; 1991 October 27–31; Washington, DC. Medford, NJ: Learned Information; 1991: (pp. 82–86).

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

Sparck Jones K. (1981a). The Cranfield tests. In Karen Sparck Jones (Ed.), for Cyril Cleverdon. *Information retrieval experiment*. London; Boston: Butterworths (pp. 256–284).

Sparck Jones K. (1981b). Retrieval system tests 1958–1979. In K. Sparck Jones (Ed.), for C. Cleverdon. *Information retrieval experiment*. London; Boston: Butterworths (pp. 213–255).

Tavris, C. (1998). The paradox of gender. *Scientific American*, *279*(4), 126–128.

Taylor, A. G. (1999). *The organization of information*. Englewood, CO: Libraries Unlimited.

Weinberg, B. H. (Ed.). (1988). *Indexing: the state of our knowledge and the state of our ignorance: Proceedings of the 20th Annual Meeting of the American Society of Indexers*; 1988 May 13; New York, NY. Medford, NJ: Learned Information; 1989. 134 p.

Weinberg, B. H. (1998). *Can you recommend a good book on indexing? Collected reviews on the organization of information*. Medford, NJ: Information Today.

Weise, E. (2000). Web changes direction to people skills. Neatly categorized information requires the human touch. *USA Today*. Jan. 24, 2000, p. 1D. Also available as: Search sites brush up on people skills, www.usatoday.com/life/cyber/tech/reivew/crg841.htm.

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

Wellisch, H. H. (1996). *Indexing from A to Z* (2nd ed.). New York: H. W. Wilson.

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

Wilson, P. (1978). Some fundamental concepts of information retrieval. *Drexel Library Quarterly*, *14*(2), 10–24.