5 Subject Access Points in Electronic Retrieval*

BIRGER HJØRLAND
University College of Borås

LYKKE KYLLESBECH NIELSEN
Royal School of Library and Information Science, Copenhagen

INTRODUCTION

This is the first *ARIST* chapter devoted to subject access points (SAPs, also called search fields or document representations) in databases. The term is used here in a much wider sense than just headings. ROWLEY & FARROW use this narrower sense and find that "the concept of the access point belongs to manually searched indexes, and is arguably irrelevant to databases with search systems allowing keyword access" (p. 253). In our wider sense, SAPs are fundamental to any kind of document retrieval. This subject has earlier been scattered in many different chapters (especially those on document representation, which have not been reviewed in *ARIST* since 1974 by HARRIS). A systematic cumulating of findings related to each kind of subject access data has never been undertaken in *ARIST* or elsewhere, although texts such as that by LANCASTER cover much of the relevant findings. This review cannot cover all relevant studies but concentrates on the broader theoretical perspective.

SUBJECT AND ACCESS DATA

Much more research has been done on searching and retrieving documents and information[1] and on users than on access points. Retrieval, however, is essentially the use of access data, and people

*We thank Raya Fidel, rector emeritus Tor Henriksen, and other reviewers for valuable feedback during the writing of this article.

[1] In this article the word "information" is used synonymously with data, which we believe is the ordinary sense of this word. In other papers we apply a more Shannon-inspired meaning of information. However, it is outside the scope of this article to discuss this concept further.

Annual Review of Information Science and Technology (*ARIST*), Volume 35, 2001
Martha E. Williams, Editor
Published for the American Society for Information Science and Technology (ASIST)
By Information Today, Inc., Medford, NJ

cannot use something that is not there. Access points determine in a rather firm way the objective possibilities that are provided for the talented user (or for any formalized, algorithmic, or automatic procedure). Therefore, it is essential in information science (IS) to develop knowledge about what kinds of subject data exist as well as the strengths, weaknesses, and relative contributions of each kind. For example, what proportion of a given set of relevant documents is missed by using only one access point such as words from titles? How much can additional access points increase recall, and how do they affect precision? Only from such knowledge are we able to study how the users or algorithms utilize these possibilities, which form the subjective factors in retrieval. For example, if people (or algorithms) do not use references as access points (in citation databases) because they do not know about this possibility or they misjudge it, then an objective possibility in retrieval is not utilized.

Knowledge about SAPs is also crucial in relation to the design of information systems because it is related to the fundamental question of which possibilities should be provided. It is rather trivial to think that systems should provide as many retrieval possibilities as possible—to believe, for example, that databases providing access to searching abstracts are better than those that do not, other things being equal. Faster access to more information is an important demand from users, but this is primarily provided by better computer technology, especially storage technology, not by IS. The availability of many kinds of access points in databases demands much space, which is provided by developments in information technology (IT). There has therefore been an IT-driven growth in subject access data that is outlined below. This growth is mainly quantitative, while the qualitative ways in which the technological potential has been utilized is a central issue for research in IS.

Information science is concerned about how IT developments can best be used to represent and to retrieve documents and information. This is related more to qualitative characteristics of subject access points than to quantitative issues. IS should ask questions such as: Given certain constraints, what are the optimal ways to design a system? Theoretically we should have a comprehensive knowledge of the kinds of access data and their characteristics. Each existing retrieval system should then be seen as realizing more or fewer of these possibilities.

The term information retrieval (IR) was introduced by MOOERS in 1951. He also introduced the term "information retrieval languages" as the generic term for classification codes, keywords, free-text retrieval, and other search elements or SAPs. At the same time the empiricist, experimental approach to document retrieval (references, surrogates, or information) was founded as an important research tradition in IS.

This tradition is normally termed the information retrieval tradition in IS, and it has some distinctive characteristics that distinguish it from other research traditions within the field, such as the facet analytic tradition, the cognitive approach, and semiotic approaches. Analytically it is important to distinguish IR as a field of study from IR as a specific approach or research tradition because different traditions may provide useful contributions to this field. (The IR tradition may, like empiricism in general, have certain blind spots.)

The basic element in IR is the user's interaction with a database (or with electronic information environments such as the World Wide Web). The user has a query[2] that has to match, more or less exactly or directly,[3] some elements, which may be termed access points, search keys, retrieval keys, data elements, or document representations. There are many kinds of such access points, they have many different functions, and they have different informational values in different search situations. What are these subject access points?

Many texts in IS differentiate between subject access data and so-called descriptive data and other kinds of data such as call numbers. Metadata is the generic term for all such kinds of data. In major research libraries, librarians usually provide the descriptive data and subject specialists provide the subject data. Many people think that there is a clear and sharp functional division among subject data, descriptive data, and other kinds of metadata.[4] This was virtually true in the age of printed card catalogs, where the descriptive data allowed for searches for known items and subject data allowed for searches for known or unknown documents about a given subject. In the age of electronic retrieval, however, there is no clearcut functional division. All words in titles have become searchable, and titles are thus both descriptive elements and SAPs. Search profiles can include many kinds of data. Hypothetically, it may be relevant to limit a subject search according to the name of a publisher, a journal, or even a language code. Subject data are

[2] There have been attempts in IR to avoid queries, and systems that allow "navigating" seem to avoid this concept. We do not see this as a theoretical problem for our views on subjects; we do not discuss it here. We are also aware that advanced technologies, such as Latent Semantic Indexing (LSI), can retrieve relevant documents even when they do not share any words with the query. LSI uses statistically derived "concepts" to improve searching performance (see GORDON & DUMAIS). However, such "concepts" must be based on subject access points, so knowledge of these still is necessary.

[3] A direct match is obtained in systems based on Boolean logic. Such a match is between words (a lexical match), not between concepts (a semantic match). Implicit or latent semantic matches can be obtained by taking advantage of the implicit higher order structure in the association of terms with documents. Such structures represent important associative relationships that are not evident in individual documents (cf. BERRY ET AL.).

[4] Such a sharp dichotomy can be found in, for example, a Danish dictionary of information science, *Informationsordbogen*, published in 1996 by The Danish Standardization Organization (FRIIS-HANSEN ET AL.).

not strictly limited to specific kinds of data; under specific circumstances any kind of data may serve to identify documents about a subject (cf. HJØRLAND, 1997, pp. 11-37). But what is that "something" that subject data are meant to identify? What are subjects?

"Subject" is one of several related terms used in the literature. Terms that are sometimes considered synonyms and sometimes used with different meanings are shown below:

Subject (subject matter; subject-predicate)
Aboutness
Topic (topicality; topic/comment)
Theme (with "central theme" and the German "leitmotiv")[5]
Domain (cognitive domain, scientific domain)
Field (information field, field of knowledge, field of research)
Content
Information[6]
Other (including related terms such as "discipline" and "concept")

These concepts are considered very difficult both in IS and in linguistics, and when used in other fields such as semiotics, psychology, and cognitive sciences. One proposal for differentiation of some of these terms is given by BERNIER (p. 192). In his opinion, subject indexes are different from, and can be contrasted with, indexes to concepts, topics, and words. Subjects are what authors are working and reporting on. Presentations can be organized into topics and use words and concepts. A document can have the subject of Chromatography. Papers using Chromatography as a research method or discussing it in a subsection do not have Chromatography as subjects. Indexers can easily drift into indexing concepts and words rather than subjects, but this is not good indexing. Bernier does not, however, differentiate authors' subjects from those of the information seeker. A user may want a document about a subject that is different from the one intended by its author. From the point of view of information systems, the subject of a document is related to the questions that the document can answer for the user. Such a distinction between a content-oriented and a request-oriented approach is emphasized by SOERGEL (1985). A request-oriented approach implies that subject analysis should thus predict the questions that the document will help to answer. Based on such analyses, HJØRLAND (1997) proposes that subjects are the epistemological or informative potentials of documents, and he sees the job of the indexer as that of predicting the most important future applications of

[5] Theme is opposed to rheme: what an author tells about a theme.
[6] "Information analysis" is, for example, used for subject analysis in the INSPEC database.

the document. This view corresponds to the functional theory about sources in history, which states that what counts as an information source is always relative to the question that it is supposed to answer.

In linguistics, the corresponding concept is mostly known as "topic" (which is contrasted by the notion of "comment," i.e., what is said about a given topic). A concise encyclopedic article on this topic with further references is provided by VAN KUPPEVELT. A 1975 conference was devoted to Subject and Topic at the University of California, Santa Barbara (LI). In one of the papers, CHAFE treats a range of phenomena related to subject: topic, point of view, givenness, contrastiveness, and definiteness. In her text, NORD (1991) addresses subject matter from the point of view of translation theory. In psychology, subject/predicate has been treated by HORNBY.

In recent years the terms "topic" and "topicality" have been popular in IS. Many writers (e.g., BOYCE and WANG & SOERGEL) agree that topicality is only one of many factors influencing relevance, but they have not succeeded in defining this concept in a clear way. GREEN (1995) and GREEN & BEAN found that there is not one kind, but rather many kinds of relationships between texts and questions that are perceived as being "on the same topic." They have not, however, considered how concepts such as aboutness, theme, or subject relate to topic. These are different concepts that people use when searching for unknown documents, but we do not know much about how such concepts differ or overlap in ordinary use, nor have we any theory that provides a well-defined meaning for these concepts. According to JANES (p. 167): "Over the last several decades, a number of other words have been used to not only describe what goes on in people's heads when they make judgments about documents, but also to ask them to tell us about it. Our results might lead one to believe that these several concepts and terms overlap . . . . But it may go further than this. Perhaps what we have called 'topicality,' 'utility,' 'satisfaction,' 'pertinence,' and a variety of other names are in fact dimensions of a larger, multidimensional, dynamic concept . . . " This problem is still unsolved, although some hints are given. For example, WANG & SOERGEL suggest that "field" is (or should be used as) a broader term than "topic," and BOYCE (p. 109) suggests that the use of references or citation indexes is a recall-oriented technique in which each iteration brings in more and more documents of questionable topicality. This last suggestion points to a difference between a field defined as a network of citing papers and a topic defined as a conceptual or terminological structure. What kind of theory is needed to clarify these concepts further? Because they are concepts about structures in knowledge, epistemology is the most relevant discipline. Different theories in epistemology imply, however, different views of knowledge structures. Classical rationalism imagines a highly or-

dered universe of knowledge, in which every concept has its well-defined place in relation to all other concepts. The modern view is much more pragmatic—viz., that knowledge serves cognitive systems and that the structures of knowledge reflect the needs and behavior of activity systems and discourse communities. This view implicates that the concepts we are talking about (e.g., topic) are concepts we use about units or parts relating to (human) communication and that their definition must be grounded in sociocognitive theories.

Different kinds of SAPs describe the subject of a given document in different ways, such as more or less exhaustive, more or less general or specific, in a more-or-less open or closed way, and so on. Most importantly, they may describe the subject of a document from different interpretations of the relevance of the given document to future questions put to the database. Because any document can in principle answer an unlimited number of questions, subject analysis prioritizes the most important questions that the document is supposed to answer in the future. The most valuable SAPs are those that make it possible for the user to identify the most highly relevant documents, that is, make the highly relevant documents the most visible in the database at the expense of less-relevant documents.

Major Technology-Driven Stages in the Development of Subject Access Points (SAPs)

Manual indexing and classification in libraries. This first stage has deep roots in the history of libraries and comprises especially books and other physical units. A more formal research area was established about 1876 by Melvil Dewey and others. This stage concentrated mostly on the organization of specific physical collections of documents and enabling access either to known documents or to documents on specific subjects in these collections. Important developments in this stage were Charles A. Cutter's (1837-1903) rules for a dictionary catalog; Melvil Dewey's (1851-1931) Decimal Classification system, Henry E. Bliss's (1870-1955) Knowledge Organization, and principles developed by S. R. Ranganathan (1892-1972). This stage still influences some research traditions in library science. Classification research is built on theoretical traditions and assumptions other than the IR tradition. The most influential work in this tradition is Ranganathan's Colon Classification from 1933, and the most important kinds of SAPs in this stage are classification codes and subject headings. The main approach to subject access is a top-down division of "the universe of knowledge" according to some rational principles. A more empirical orientation was established by HULME (1911a) in the principle of bibliographical warrant or literary warrant, which states that a class or a subject heading must be

established only if there exists literature to be classified by that group. In this way subject retrieval was not only built on top-down analyses of the universe of knowledge but was also somewhat influenced by the existing literature in a bottom-up manner. SAPs in this stage are produced and controlled by librarians and information specialists (including subject specialists) and constrained by their subject knowledge. Another major constraint in this stage/tradition is that the principles were developed for subject access to physical units (e.g., books), not documentary units (e.g., journal articles). This implies a level of subject description and concepts that are often much broader than those needed by researchers in specific investigations. A third major constraint in this stage/tradition is that because the available space (e.g., on printed catalog cards) was very limited, the SAPs tended to contain scanty information. Nevertheless, this stage/tradition developed important principles that many researchers find useful in a fully electronic environment (see, e.g., POLLITT ET AL.). What de Grolier wrote in 1965 is still regarded by many as true.[7]

> We feared some years ago that classification was becoming useless, that the treatment of natural language texts by machines . . . would replace classification. Classification and the classificationists would become something like the dinosaurs, killed by the progress of evolution. This has proved to be a complete fallacy. When you examine the new literature you find that more and more classification . . . is considered as something quite essential in information retrieval . . . It is quite evident that hierarchies, generally speaking, are something which can not be avoided in an information retrieval language which is to be useful for the reader. (DE GROLIER, p. 11)

"Documentation" and scientific communication. "Documentation" is the name of a movement founded by Paul Otlet (1868-1944). The establishment of The International Institute of Bibliography in Brussels in 1895 (from 1937 called Fédération Internationale de Documentation (FID)) and of the Universal Decimal Classification (UDC) system in 1905 with the aim of universal bibliographical control, was a major achievement in this movement. The documentalists often regarded themselves as more service-minded, more technology-oriented, and more advanced than librarians. Where traditional librarians often had an orientation toward the humanities, the documentalists were mostly affiliated with science, technology, and business. They indexed single articles in journals and books and played a central role in establishing

[7] SALTON is an example of an explicit disagreement with this view.

international abstracting journals.[8] They were less interested in collection development and more concerned with providing better access to knowledge that is independent of specific collections. They were less interested in keeping books for their own sake or for broad cultural purposes and highly interested in establishing services that could stimulate the application of knowledge to specific purposes. The foundations of user studies (BERNAL) and bibliometrics (e.g., BRADFORD) are also part of this stage/tradition, which is primarily characterized by a more specific subject approach, a deeper level of indexing, and a more scientific attitude toward goals and problems.

Information storage and retrieval by computers. This stage has been developing mainly since 1950 and can be seen as a technological modernization of documentation (American Documentation Institute (ADI), founded in 1937, changed its name in 1968 to American Society for Information Science (ASIS), then ASIS in 2000 added "Technology" (ASIS&T)). The establishment of computer-based abstract services, such as Chemical Abstracts and MEDLINE, in the 1960s was important during this stage. The development of descriptor-based and free-text retrieval (mainly based on titles and abstracts), Boolean logic, field-specific subject access, as well as the measurements of recall and precision and other innovations were extremely important in document retrieval. Information retrieval (IR) as a research tradition started with the Cranfield experiments in the 1950s, and today's Text REtrieval Conference (TREC) full-text experiments continue this tradition (see NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY).

This third stage improved information services and research efforts in IS in an important way. Computer technology made it possible to use many kinds of SAPs, both the traditional kinds produced by information specialists and the use of words from the documents themselves (e.g., titles and abstracts). It removed the monopoly of librarians and information specialists over subject access and established a direct competition between SAPs produced by different agencies.

An underlying premise in this stage has often been that the length of the searchable record itself was the most important parameter in retrieval (LANCASTER, pp. 6-8). SAPs were often seen merely as "semantic condensations" of the texts represented (implying that the ultimate goal was full-text representation and nothing more). Research was dominated by quantitative methodologies, and not much research on qualitative differences (semantics or meanings) among different kinds of SAPs was established. The premise was empiricist, first and foremost, in its attempt to measure the efficiency of subject retrieval

[8] The history of the abstract journal goes back, however, to 1665 (cf. MANZER).

points empirically (e.g., by measuring recall and precision). It was also empiricist in its avoidance of "metaphysical"-based classifications and in its favoring of "atomist" SAPs, such as the Uniterm system devised by Mortimer Taube in 1951 and similar systems that depended on specific words from the documents themselves.

One associated tendency in this stage was the attempt to formalize and to automate retrieval and to eliminate human interpretation and subject analysis. We must distinguish between the economic pressure to automate practical systems on one side and the scientific evaluation of the performance of various aspects of human-based and mechanized retrieval systems on the other side. It is legitimate and highly desirable to reduce costs and improve efficiency in information systems. Basic research, however, should illuminate basic strengths and drawbacks in different approaches and not be blinded by the pressure to use automated or cheap solutions. Because of such tendencies, important approaches related to interpretation were neglected, and the research did not yield as satisfactory a body of knowledge as desired.

Citation-based retrieval (1963-). Eugene Garfield's introduction of the Science Citation Index in 1963 marks the fourth important stage in the development of SAPs. The possibility of retrieving documents according to the citations they receive represents a real innovation in IR, and this technique is able to supplement all forms of term-based retrieval in very important and qualitative new ways. This innovation has also prompted research on motives to cite other documents, on sociological patterns in citing, on the relative role of terms and references as SAPs, and on the semantic relations between citing and cited papers.

In this way, citation-based retrieval has changed our understanding not only of subject relatedness but also of the concept of subject matter and of the fundamental aim of IR itself. Because it may be relevant to cite papers that have no words in common with the citing papers (or no simple semantic relation such as narrower terms, broader terms, and synonyms), naive conceptions of subject relatedness or subject matter can no longer persist. Semantic relations may be implicit or latent. Semantic relations in science are determined by theoretical advances, which may change the verbal description of the research phenomena completely; this is why statistical patterns in vocabulary may sometimes be a less efficient measure of subject relatedness than patterns in citations.

Citation behavior is extremely important because the goal of IR is to provide the references that are useful in solving a specific problem. A scientific article is a documentation of how a specific research problem is solved. The problem is formulated in the article, and the problem has

determined the kind of information needed[9] by the author to solve the problem. Based on need, information was sought and selected, and the documents actually used were finally cited in the article. Each of the thousands of articles produced weekly is a kind of case study in IR. Every article not only poses a definite IR problem, but the list of references provided by the author is also the key to how that particular person has solved the problem. Thus, it is possible to check theories of IR against how they match the actual documents cited. According to the traditional view in philosophy of science, science should be able to predict future events. In other words, theories and models of IR should be able to predict citations that will appear in particular papers. Most research on relevance and on IR seems to have overlooked this fact. From what we do know, it seems extremely unlikely that an algorithm would be able to select references from electronic databases and end up with exactly the same references that appear in a given article. From this point of view, theories of IR seem naive and unrealistic (and the goal of prediction seems to be wrong). A more detailed study of citation behavior can illuminate the real problem of IR, which is that cited documents are not simply a set of documents sharing a fixed set of attributes that are not represented in the nonselected items. Documents that are similar from the point of view of retrieval algorithms need not be co-cited, whereas documents that are not similar are often co-cited. Ordinary retrieval algorithms and citation practices seem simply to reflect different theories about subject relatedness.

Because authors may cite other papers in order to flatter or to impress, the prediction of which references a given author will finally select for a given paper cannot be used as a valid criterion in IR. The criteria for IR should not be based on social or psychological motives but on epistemological principles for the advancement of public knowledge. In this way, our insight from citation indexes has profoundly changed not only the methods of IR but also the concept of subject relatedness itself and the basic aim of retrieving information. We can no longer regard the prediction of individual use as the ideal criterion for IR, nor can we regard IR as a value-free technique. Instead, we have

[9] Information need is an important concept in IS. People may have many needs with complicated interrelations. A more precise need arises when a specific decision is made to write a paper. From that point and until the paper is printed, the author seeks information, selects information, and decides what to cite in the paper. The references in the paper represent only one stage in the development of the author's information need. However, they are the most tangible, public, and available expression of how the author has seen and resolved his or her needs. People who are used to reading and interpreting papers can evaluate authors' conceptual horizons, compare them with others, and study their development and how they are influenced. In this way scholars may have methods to determine information needs other than behavioral methods.

to face the fact that the goals of IR are deeply rooted in epistemological norms for what should be regarded as good science and good citation behavior.

Full text, hypertext, Internet, and digital libraries. Full-text retrieval marks the fifth and final step in the development of SAPs. Until this point, space limits were a major constraint in the development of subject access systems because length of the record in itself is an important parameter in retrieval. At this stage, every single word and every possible combination of words in full-text documents are potential SAPs, as is every conceivable kind of value-added information provided by authors, readers, or intermediaries. Given full-text representations, the first important theoretical problem that arises is whether any kind of value-added information is necessary. Can the extra information provided by abstracting and indexing, at least in principle, increase recall and/or precision? If not, then we seem to have reached the end of the line in that no further contributions from research or practice in IS are needed. The answer to this question is closely linked to theoretical views on the concept of subject. POULSEN sees a subject as something that is expressed in the literature (in a transparent and self-evident way?). By defining subjects in this way it is impossible even to pose the problem of whether a given text always represents the optimal representation of itself. By defining subjects as informative or epistemological potentials, HJØRLAND (1992; 1997) established the possibility that documents may be implicit or even wrong about their own subject matter; hence, information professionals are still needed. To take an extreme example, a document about Jews written by a Nazi author should not only be indexed as being about Jews, but it is also important to make the Nazi view visible in the subject analysis (e.g., to index it as Nazi propaganda about Jews). Subjects are not objectively "given" but are influenced by broader views, which are important for the information seeker to know and should therefore be part of the subject analysis. Whether this is also practical, economic, and realistic is another question that must be explored by evaluating specific subject access systems.

Toward a Taxonomy of Subject Access Points

Figure 1 outlines some important criteria for the classification of SAPs. In general, access points should be regarded as a system wherein each element contributes to the overall performance of the retrieval system. For example, in research libraries, it would be a waste of resources to provide subject access to articles in the library catalog if this access is redundant with the subject information that can be found in, for example, CD-ROM databases in the same library.

Access Points Classified by Provider or Agent
Author-generated (e.g., document titles, abstracts, and keywords)
Value-added, including those provided by publisher or editor (e.g., journal name, publisher name, and cover information); indexer/abstractor/information specialist (e.g., classification codes, descriptors, identifiers, and abstracts); reviewers, readers, and other writers (e.g., reviews with links on Internet, best-seller statistics, citations, and citation indexing)

Access Points Classified by Kind
Verbal vs. nonverbal (nonverbal is sometimes called symbolic)
Long forms vs. short forms (e.g., abstracts vs. single keywords or classification codes)
Controlled vs. uncontrolled forms (or closed vs. open systems)
Derived vs. assigned forms (e.g., titles vs. identifiers)
Forms based on checklist or facet analysis vs. forms based on free analysis
Explicit vs. implicit (e.g., descriptors vs. references, journal names, or publishers. Implicit SAPs are mostly made for purposes other than IR. Titles are explicit SAPs when the authors intend them to be used for IR)
Content-oriented (or descriptive) vs. question-oriented (or evaluative)
Precoordinated vs. postcoordinated indexing forms
Syntactic indexing forms vs. forms without syntax (syntactic devices are, e.g., roles and links; they are also applied in the PRECIS indexing system)
Manually produced vs. computer-generated (computer-generated access points are sometimes produced by retroconversions in databases)

Figure 1. Some taxonomic criteria for subject access points

It is evident that a comprehensive description of all potential kinds of access points generated by the authors of documents implies a comprehensive typology of kinds of documents and a description of the structure (architecture or composition) of each kind of document listing all types of SAPs. Because document structures develop in response to different demands, they are also influenced by epistemological positions or paradigms. Figure 2 shows the potential SAPs in a typical scientific article.

Norms (of scientific method and philosophy of science external to the article)	Elements Contained in the Article	Value-Added Information (Subject access points, access, and evaluation information)
Observation and description	Bibliographical identification	Bibliographical description
Problem statement	(Journal name, volume, pages)	Relationship to other editions
Hypothesis	Title	Biographical information
Experiment	Author(s)	Institutional information
Theory building (According to the basic view formulated in HJØRLAND (1997), there exist different epistemological views (and each implies different standards or ideals regarding the structure of documents. Thus a typical empiricist article reflects the development of the empiricist research tradition.)	Corporate affiliation and address	Indexer abstracts
	Author abstract	Indexer descriptors and identifiers
	Author keywords	Classification codes
	Introduction	Language codes
	Apparatus and materials, method, results, discussion	Document type codes
	Conclusion	Editorial comments
	Acknowledgements	Links to citing papers, reviews, and criticism
	References	Information about availability of document
		Evaluations
		Target group information
		"Key word plus" and "research fronts"
		Other kinds of links and semantic networks

Figure 2. Structure and elements in a typical scientific article

In monographs, additional subject access points could be based on their composition—e.g., books/volumes, parts, chapters, sections, subsections, and bibliography and index. Internet documents form a third kind. The Internet search engine AltaVista provides the SAPs shown in Figure 3.

Searchable by Search Engine AltaVista

(Search codes in brackets)

- Words or phrases contained in the URL (Uniform Resource Locator) of the document [url:]
- Title [title:]
- Links (URL to other documents to which there is a reference) [link:]
- Word from the clickable text to a link [anchor:]
- Words in filenames of pictures contained in documents [image:]
- Words and phrases in full text (except image tags, links and URLs) [text:]
- Java Applets [applet:]

(Also searchable are domain names, host names, and "similar URLs")

Figure 3. Subject access points in Internet (HTML) documents (Based on ALTAVISTA: Advanced Search Cheat Sheet)

Other kinds of documents, such as newspapers, popular magazines, patents, pictures, and sound recordings, present different structures and different kinds of potential access points and retrieval problems.

The information to be derived from a document depends on the information contained in that document. Some documents have, for example, author-generated titles, abstracts, and keywords while others do not; the need to add such elements is more evident, but not necessarily redundant, in the last case. A taxonomy of derived SAPs thus clearly must be based on a taxonomy of documents and document structures. Some research in this area has been done in such fields as composition studies (e.g., BAZERMAN, 1988) and genre analysis (e.g., MALMKJÆR). In this still new and relatively unexplored field, we lack a taxonomy of document types, their composition and elements, and consequently the relative contributions of such elements in IR. We know more about scientific research articles than about all other kinds of documents,

including scholarly monographs. Thus, unless otherwise stated, this review considers only primary scientific articles.

In our view, the essential quality of SAPs is their ability to express that aspect of a given document that would be most useful in answering the questions put to the specific database from which the SAPs' performance is to be evaluated. Poor titles, bad indexing, and in general poor SAPs are those that express unimportant (or perhaps even false) information about a given document. All questions concerning the choice of formal aspects of retrieval language (e.g., standardization, pre- vs. postcoordination, length of representation) are subordinate.

If a need for value-added information is to be justified in future systems, it must be done by arguments about the ability of information specialists to interpret documents in relation to other documents and to the specific user group they are serving. Meaning, semantics, and epistemology become the most important theoretical perspectives that can be generalized from specific domains.

RESEARCH ON SPECIFIC SUBJECT ACCESS POINTS

Document Titles

A title is the name of a document given by the author and influenced by existing norms at the given time. According to BERNARD, there exists an entire discipline within literary history called titrology, which confines itself to the study of titles. For nearly 30 years it has generated an impressive number of publications (mostly in French). One survey of titrology is given by GENETTE, who defines the functions of titles in the following way: "The first function, the only mandatory one in literary practice and institution, is the function of designation or identification. It is the only one to be mandatory, but impossible to separate from the others, since under the semantic pressure of the environment, even a simple opus number can be invested with meaning. The second one is the descriptive function: thematic, rhematic,[10] mixed, or ambiguous . . . [the last] is the function called seductive" (GENETTE, pp. 718-719). Whereas most books and journal articles have titles, other kinds of documents (e.g., pictures, and nonprinted documents such as letters) may lack them. Names may characterize what they name, and their use in retrieval is based on this assumption, which, however, is not always true. The most common measure of title informativity has been the number of "substantive" words that it includes (e.g., by counting all words except trivial words, such as articles from a stop list). Because

[10] A rhematic title indicates the kind of document considered rather than what the document is about—e.g., the terms "novel," "letter," "dissertation" are examples of rhematic titles.

titles can express many different things, this method gives a very rough measure and can be misleading.

According to NORD (1995) titles can be intended to achieve six communicative functions, four of which (referentiality, expressivity, appellativity, and phatic function) can be universally assigned to all texts and text types. The other two (metatextuality and distinctive function) can be observed as specific functions of particular text types; the distinctive function is typical of names or labels, and the metatextual function is found in metatexts such as text commentaries, reviews, abstracts, and summaries. Therefore, titles are not just texts but typical texts presenting a complex hierarchy of communicative functions. In spite of their complex functionality, titles present simple syntactic–semantic structures. Nord found only four macrostructural types (simple titles, title–subtitle combinations, duplex titles with "or," and title series), six syntactic forms (nominal titles, verbal titles, sentence titles, adverbial titles, attributive titles, and interjection titles), and a limited number of microstructural patterns such as "NP & NP"= nominal phrase + connective + nominal phrase (as in *John Jakes: Heaven and Hell*). Therefore, title elements have to be polyfunctional if the title is to achieve its intended functions, which is also typical of other communicative signs.

The design or form of a title varies over time, culture, subject matter, and document type. BERNARD analyzed a representative sample of French monographs from 880 to 1991 and found that titles in the nineteenth and twentieth centuries are distinctly shorter than those of the seventeenth and eighteenth centuries, whereas titles from 880 to 1673 are as short as recent ones. Books republished in modern times often bear titles that are abbreviations of their original title. In modern terms, Renaissance titles served as both title, subtitle, signature, and fourth cover page. The development of carefully structured titles and subtitles legitimizes the use of the title without the subtitle. Another development is homonymic works. Sometimes there is an intentional repeat of a title, with, for example, parody or location within a tradition as the objective. In general, books from the Middle Ages and Renaissance did not, however, take the precaution of attaching to their works a unique label, which we consider so important today.

Titles are intended to indicate what the document is about (its subject). Authors usually choose a name that draws potential readers, indicating the document's content at a glimpse and thus contributing to its initial selection or rejection. We have little knowledge of how titles are actually used or should be interpreted in selection processes. Among the few studies on this subject are those by ATKINSON, BAZERMAN (1985), and NAHL-JAKOBOVITS & JAKOBOVITS. Studies such as the one by SARACEVIC on the comparative effects of titles, abstracts, and

full texts on the relevance judgment of documents are pertinent. He found that of 207 answers judged relevant from full text, 131 were judged so from titles and 160 from abstracts.[11] He also found that it seems to be easier for users to recognize nonrelevant documents than to recognize relevant documents from the title.

A title normally constitutes the most concise statement of a document's content. It is often used as a surrogate for the document in bibliographies, databases, indexes, tables of contents, current-awareness services, and reference lists, and it is heavily used in IR. However, because the title is a name, it is the author's decision as to how informative it will be, and what kind of information is given priority. The great importance of informative titles is almost unanimously emphasized in the literature by many writers, journal editors, and authors of guidance books for scientific and professional authors (YITZHAKI, 1996).

When we are evaluating titles as SAPs, we have to consider the kind of skills, motives, and norms that may influence the author's choice of title and hence its subsequent possibilities and limitations in IR. For example, an author may want a title that "sounds good," perhaps poetic. Metaphorical language is one of the most common problems with titles in IR. A title such as "The Conflict between Egypt and Israel: A Nightmare in Modern Politics" is a problem for the psychologist who is seeking information about nightmares by looking in Social Sciences Citation Index using titles for subject access. Another problem with title words is the lack of control of synonyms and homonyms. In a given time period of the Social Sciences Citation Index, "AIDS" is a useful access point for the illness, but when it is used in the total time span of the database, other meanings such as "teaching aids" may cause a very low precision rate.

In composition studies, CROSBY suggests a high correlation between the quality of a written composition and its title. The shuttlecock process of finding an appropriate title stimulates creativity, unity, revision, and significance. He classified 300 titles according to their apparent purpose in order to infer certain lessons for writers. The classification includes:

- Titles announcing the general subject, such as "The Age of Adolescence" and "The Collective Corporation";
- Titles indicating a specific topic, including "The Decline of Courtesy" and "Toward a New Morality";
- Titles indicating the controlling question; some titles

[11] The ability to evaluate relevance from bibliographical records seems to be much better in the study reported by SARACEVIC than in the study by WELWERT, reported in English in HJØRLAND (1988).

indicate the question that the writer is answering, and they go a long way to help the writer stay focused: e.g., "Is Culture Worthwhile?" and "How Can We Recover Our Joy?";

- Titles announcing the thesis, such as "This Thing Called Love is Pathological" and "The Rip-Off Age is the Clue to Nation's Ills"; and
- Titles that bid for attention. Some methods of attracting attention include alliteration, deliberate ambiguity, intriguing word coupling, allusions from serious and pop culture, and the twist (something unexpected).

The length of a title is also important for retrieval. The longer the title, the more words it contains and the greater should be the probability that it will be retrieved by a given query. This is not always the case, however. KELLER found that masters theses with 1 to 12 words in the title had a greater chance of being retrieved than did titles with 13 to 18 words, showing that factors other than number of words are at work.

The difference between titles in professional scientific journals and in popular science journals is not just a question of length but also of emphasis (see Figure 4). It should be remembered that the title is always a choice among possible alternatives. What is considered the core subject by the author is not necessarily the same as the searcher's core interest. A paper may be relevant for a searcher from a point of view different from the one expressed in the title (or expressed explicitly at all). Titles often express more general claims than are covered by the paper; they may be seductive or inflated, and a given subculture may stimulate a kind of marketing of a paper that resembles commercial thinking more than scientific precision.

The hard sciences tend to have longer titles than the softer and popular sciences. An analysis by BUXTON & MEADOWS (1977) and YITZHAKI (1992; 1996; 1997) demonstrated a trend toward longer (and more informative) titles, which occurred over a wide range of subject fields and was apparent before KWIC indexes and computer-based searching of title words became common. Although this trend preceded the introduction of these tools, the tools undoubtedly contributed greatly to the growing awareness of the importance of title informativity. In the humanities a somewhat similar trend seems to have occurred but in a weaker way and at a slower pace. (These studies do not discuss alternative hypotheses such as the need for longer titles because of increasing specialization in research, creating a need for more words to express a given piece of research.)

VOORBIJ studied the relative roles of title keywords and subject descriptors of monographs in the humanities and social sciences held

Articles for Professional Audiences	Articles for Popular Audiences
Insects as Selective Agents on Plant Vegetative Morphology: Egg Mimicry Reduces Egg Laying by Butterflies (K. Williams and L. Gilbert, *Science*, 1981)	Coevolution of a Butterfly and a Vine (L. Gilbert, *Scientific American*, 1982)
Female Sex Pheromone in the Skin and Circulation of a Garter Snake (W. Garstka and D. Crews, *Science*, 1981)	The Ecological Physiology of a Garter Snake (D. Crews and W. Garstka, *Scientific American*, 1982)
The Reproductive Behavior and the Nature of Sexual Selection in Scatophaga stercoraria L. (Diptera: Scatophagidae). IX. Spatial Distribution of Fertilization Rates and Evolution of Male Search Strategy within the Reproductive Area (G. Parker, *Evolution*, 1974)	Sex around the Cow-pats (G. Parker, *New Scientist*, 1979)

Figure 4. Comparison of professional and popular titles (Based on MYERS, p. 275)

by the online public access catalog (OPAC) of the National Library of the Netherlands. He found that 37% of the records were considerably enhanced by a subject descriptor and that 49% were slightly or considerably enhanced. In a second study he found that when subject librarians performed subject searching using title keywords and subject descriptors on the same topic, the relative recalls were 48% and 86%, respectively. Failure analysis revealed why so many records that were found by descriptors were not found by title words. First, the title of a publication does not always offer sufficient clues for retrieval. Second, and more important, is the wide diversity of expressing a topic in titles. Descriptors remove the burden of vocabulary control from the user. While the study clearly demonstrates the benefits of descriptors over title words, it does not consider the functions of those descriptors in relation to other kinds of subject access data that will probably soon be available from other sources (such as tables of contents and book descriptions as used, for example, by Amazon.com).

A study of COMPENDEX by BYRNE comparing titles and abstracts as subject access points found that titles retrieved 22% of citations, abstracts retrieved 61%, and titles and abstracts combined retrieved 75%. This study did not, however, report any percent for precision, but it indicates that titles alone perform very poorly compared with abstracts. COMPENDEX is dominated by articles, and we must expect that this problem is even greater with monographs. In another study, BARKER ET AL. examined chemical databases and found that summaries increased recall over titles by 68% but at the expense of a 23% drop in precision. Keywords increased recall by 35% with a 10% drop in precision.

HODGES tested the effectiveness of title keywords in retrieval and concluded that less than 50% of the relevant titles were retrieved by words in titles. Surprisingly, this study found that the social sciences had better retrieval from titles (48%) than the hard sciences (42%); arts and humanities retrieved 31%. This low rate of retrieval from titles was attributed to three sources: (1) titles themselves, (2) ignorance by the user and information specialist of the subject vocabulary in use, and (3) general language problems. Even the best efforts of users and specialists are not likely to improve this rate significantly. Hodges argues, however, that in many instances this recall is more than adequate for the user. Many students and faculty do not require the entire body of literature on a topic; they are just trying to determine the kinds and amount of material being written on a given topic, or they wish an introduction to a topic or an entry point into the literature. Also, because of their timeliness and economy, title-word indexes will, in her view, remain an important element of indexing.

When titles are used for retrieval, their words are merged with those from other titles in the same journal, other journals, other kinds of documents in the domain, and perhaps also words from titles in other domains. IR is always done in one or more specific collections, and the actual context determines the most rational search strategy. The principal disadvantages in having authors rather than professional indexers provide access points may be related to the fact that authors do not have the same overview of the total database (or total literature in the field). Hence, they may have difficulty in predicting the discriminative value of words and their combinations. Their selections can easily be either too specific or too general.

Because titles are different in their informational values, they have a different status in different databases. Some printed bibliographies (e.g., ERIC) use titles as document surrogates or document representations in the index (under each descriptor), while others (e.g., Psychological Abstracts), apply a value-added index phrase with a higher informational value. (This may of course reflect a decision that is not

grounded in a difference in the informativeness of titles in educational and psychological research.)

PERITZ examined the frequency of noninformative titles in library and information science (LIS) and in sociology. Noninformative articles totaled 21% in LIS and 15% in sociology. For both fields the study showed that the noninformative articles were concentrated in a few journals.

Conclusion. Investigations of titles as access points tend to emphasize quantitative aspects, such as length, number of "substantive" words, and differences between domains and over time. Studies of qualitative aspects of titles are scarce and are found mostly in disciplines outside IS (e.g., linguistics and composition studies). If we assume that different theoretical views or paradigms have different views on a given paper and on what in that paper is of interest, then such different views should be able to express different criteria for the informativity of given titles. For example, we might expect positivist-oriented information seekers to value titles that express the kind of statistical methods used in a paper, and hermeneutical-oriented seekers to value titles that express the interpretative attitudes of the author. This implies that title informativity cannot be measured by an objective standard, for example, by number of words. Nor is such informativity simply a subjective or cognitive value in an individual, psychological way. The epistemological view implies that the informativity of titles is something to be inferred theoretically by views formulated in epistemology.

Abstracts

According to ALTERMAN, text summarization is not a single phenomenon. There are many different kinds of summaries, such as abstracts, epitomes, overviews, abridgements, digests, and recapitulations. Alterman does not, however, describe the differences among them. We can add the following: annotations, briefs, cuts, extracts, part texts (e.g., half texts as opposed to full texts), précis, and Zentralblätter. However, in IS the two most common distinctions are indicative vs. informative abstracts—respectively, evaluative (or critical) vs. nonevaluative abstracts.

In the philosophy of science there is an important argument—viz., that one's observations are not independent of one's theoretical assumptions (cf. CHALMERS, chaps. 1 and 2). This principle is also valid concerning the observation/reading of documents and the interpretation of their essential or core information (or rather their informational potentials) and thus the summarization of them. As a consequence, even nonevaluative abstracts cannot just be regarded as objective de-

scriptions of a document but are influenced by norms, interests, and epistemological positions.

Today most scientific journals publish authors' abstracts for all their articles. These abstracts may be used directly in bibliographical databases, or they may be edited, revised, or replaced by an abstract written by a professional abstractor, who usually then signs it. We call such value-added abstracts "indexer abstracts."

LANCASTER believes that the length of a given search field is the most important factor in information retrieval:

> For retrieval purposes, the longer the abstract the better. At least, the longer the abstract the more access points it provides, and the more access points the greater the potential for high recall in retrieval. At the same time, it must be recognized that precision is likely to deteriorate: the longer the abstract, the more "minor" aspects of the document that will be brought in and the greater the potential for false associations. (LANCASTER, p. 21)

> Because the brief abstract provides more access points than title or selective indexing, the item it represents will be more retrievable. Likewise, the exhaustive indexing may make this item more retrievable than it would be in a search on the brief abstract but less retrievable than it would be in a search on the expanded abstracts . . . . The longer the record, the greater the chance that spurious relationships will occur. Spurious relationships, of course, cause lower precision. (LANCASTER, pp. 227ff.)

From our point of view, however, this quantitative measure—that is, the length of the field—is less interesting than how well it will satisfy the needs of users in given situations. Because some subject analyses are simply better than others, the strategy of unlimited aliasing, which implies that as many different subject descriptions as possible be put into the document representations, is not a correct theory or strategy. This can be disproved both theoretically and empirically (cf. BROOKS). Therefore, we need a theory about what should be expressed in different SAPs (viewed as a system) and what is the abstract's role in this system. The ability to see what is important and to express it in a way that maximizes its visibility to the user must be the only factor that matters.

LANCASTER writes further:

> At the present time, authors and publishers have little incentive for "embroidering" abstracts to make the underlying

> work seem more attractive than it really is. Price . . . has argued that this could become a danger in a completely electronic environment . . . . Publishers would want to promote use because they would probably be paid on this basis. Authors would want to promote use if this factor became, as it might, a criterion used in promotion and tenure decisions. The term "spoofing" has been used to refer to the embroidering of Web pages to increase their retrievability . . . . (LANCASTER, p. 116)

This quotation is the key to understanding the role of value-added information provided by information specialists. Their perspective is different from those of authors and publishers. Ideally they read on behalf of the user (or on behalf of science or some collective goals and values). Perhaps the commercial or self-promoting embroidering of abstracts is rare in the printed world, but a more "scientistic" "embroidering" of the whole text including name dropping, for example, may be the rule rather than the exception (and some embroidering may be unconscious and subtle). Abstractors can—at least ideally—have an overview of the system in which the single document is going to be organized. They have an implicit knowledge of the visibility and retrievability of different documents in the database, and they can improve the visibility of those aspects of a given document that will be most useful. Most importantly, because all documents are based on implicit assumptions, information professionals can make a difference in explicating such epistemological assumptions. Two specific examples of how this can make an important difference are given by HERRELL and by WINDSOR.

The work of abstractors can be guided by thesauri, classification systems, checklists, and facet analysis (FIDEL, 1986). In this way their specific subject analysis can be somewhat formalized. The most important factor is not the degree of formalization but the fact that the abstractor write on behalf of the users and from the perspective of a more-or-less specific collection or database with more-or-less well-defined functions in the information environment.

Conclusion. Abstracts are important in IR as access points and as indicators of the relevance of documents during a search. Abstracts increase recall and precision much better than titles and keywords. Their efficiency depends not only on their length but also on their content. With titles they share the problem of providing users with a relevant description of the document being represented. Such a description is in principle not value free or neutral but always biased in one direction or another. In information systems, abstracts should ideally be written on behalf of the user and from the perspective and goals

implicit in the specific system. This is why many information systems have their own abstractors and do not rely on author-created abstracts.

References/Citations

Searches that use the references in documents as SAPs, directly or via citation indexes, are called chain searches.[12] They represent a qualitatively different method from term searching. How should we evaluate the relative strengths and weaknesses of term searching vis-a-vis citation or chain searching?

Chain searching is often quite valuable (e.g., see WELWERT, which is (reported in English in HJØRLAND (1988)). A search for the subject "reading comprehension vs. listening comprehension" resulted in 79 relevant references using database searching, 47 using manual sources, and 82 using chain searching in the references that were located. The last 82 references could, of course, not have been found without the previous bibliographic search, but this example indicates the significance of chain searching. It may also indicate the high degree of uncertainty of bibliographic searching in that so many references were not found by a thorough search of databases and printed bibliographies.

Chain searching vs. bibliographic searching can be further illustrated by field studies (PAO, 1993) and controlled studies (PAO & WORTHEN) in terms of literature references vs. terms as search criteria. These studies, which were performed in medicine and which built on a pool of common references in MEDLINE (a database that maintains a high level of indexing quality) and SCISEARCH (a science citation index), cannot be regarded as definitive, but they do indicate the following:

- The level of overlap is low (4-5%) when terms and references are used for searching.
- Given a high quality of indexing, term searching seems to be more efficient than reference searching (term searching in MEDLINE gave a mean recall of 77% and a precision level of 56%; reference searching in SCISEARCH gave a recall of 33% and a precision level of 60%).
- Compared with term searching alone, reference searching increased recall by a mean of 24%. Moreover, the overlap between the two search strategies had high precision.

[12] An advanced way to do chain searching is by using the Web of Science produced by Institute for Scientific Information.

Unfortunately, these studies lack a closer analysis of the nature of the terms and references that result in few or no results. These kinds of studies are typically quantitative rather than qualitative. If recall can be increased by 24% by including reference searching, it would be relevant to analyze what kinds of concepts typically should have been included in the bibliographic records, but were not. Might these kinds of experiences lead to new instructions for the indexers so that indexing practices could be improved? HARTER ET AL. also found that the subject similarity among pairs of cited and citing documents typically is very small, indicating that term searching and chain searching are complementary methods.

GREEN (2000) compared chain searching with the use of standard bibliographic tools in the humanities and found that less than 5% of the references were found by both types of searches. Precision of retrieval based on bibliographical references from "seed documents" appears to be high. Whereas bibliographical tools generally observe a well-defined boundary of coverage relative to subject, date, format, and language, the relevant literature may not respect the same boundaries, especially in the humanities. This is one reason chain searching is so important. Green also found (p. 224) that although most of the sample documents were covered in the bibliographic tools being used, only 10% were assigned index terms that matched the user's need in terms of both breadth and depth. She says, "Suffice it to say that there are no trivial or easy solutions to the overwhelming problem of assigning subject descriptors to documents that will consistently enable users to locate all, but only, the literature relevant to their needs" (p. 225).

The efficiency of bibliographic searching is, of course, determined by how much of the relevant literature has been recorded, analyzed by subject, and described in a way that allows searchers to locate it via bibliographies, databases, and reference literature. The bibliographic approach is characterized by formal rules that determine what is included in a bibliography or database and how it is described (e.g., by using descriptors). The document description is largely an expression of the competence that is tied to the administration of a set of rules. The efficiency of the result depends in particular on whether the formal rules are able to ensure the design of a product that meets the users' needs. The strength of the formal approach is that little material is excluded because of value-based criteria. The weakness is that because they are formal, these systems do not give priority to materials according to relevance. They may, for example, include all books longer than 49 pages or exclude book reviews or not index parts of a document. A lack of resources or of adequate rules to carry out the formal program might lie behind the random inclusion of both the highly relevant and

the nonrelevant references. In real life, there is almost always a lack of resources, which means that highly relevant references are often absent. Such formal omissions should not be expected in references, which may, however, contain other kinds of omissions.

The efficiency of chain searching—assuming that one can identify relevant seed documents—is determined by how well the document identifies and cites relevant information in the reference list. The method presupposes that the scientific literature in the field is neither unrelated to other research in the field nor simply redundant. In other words, it assumes that researchers are extremely conscientious in their literature searching and their referencing to relevant sources and that the references are selected with a view to informing the reader of important literature. It also presupposes that the scientist does not cite on purely formal or presentational grounds, for example. Most importantly, it presupposes that authors are not biased in selecting information but give even consideration to papers that argue both for and against their own view. This last assumption seems to contradict the results of psychological research:

> As shown by a multitude of studies, such information-seeking processes often are not balanced: people prefer information that supports their favored or chosen decision alternative compared to information that opposes it. . . . the preference for supporting (consonant) compared to conflicting (dissonant) information occurs if people have decided voluntarily and with a certain degree of commitment for a particular alternative . . .We will refer to this preference for supporting information as *confirmation bias*. . . . Therefore, it can be concluded that individuals carry out biased information seeking while making decisions, and that this happens from the moment they commit themselves to a particular alternative. (SCHULZ-HARDT ET AL., p. 655)

In citation studies MACROBERTS & MACROBERTS (1988; 1989) have considered authors' motives for not citing relevant documents, just as they represent—together with SEGLEN (and GARFIELD himself)—some of the most qualified and dedicated critics of the misuse of citation indexes. Psychological factors are important in studying why authors quote other documents. As GARFIELD (p. 85), points out, there are many kinds of citation motivations:

- Paying homage to pioneers;
- Giving credit for related work (homage to peers);
- Identifying methodology, equipment, and so on;

- Providing background reading;
- Correcting one's own work;
- Correcting the work of others;
- Criticizing previous work;
- Substantiating claims;
- Alerting to forthcoming work;
- Providing leads to poorly disseminated, poorly indexed, or uncited work;
- Authenticating data and classes of facts—physical constants, and so on;
- Identifying original publications in which an idea or concept was discussed;
- Identifying original publications or other work describing an eponymic concept or term;
- Disclaiming work or ideas of others (negative claims); and
- Disputing priority claims of others (negative homage).

SEGLEN (p. 29) also lists a range of problems concerning selection of references:

- References are selected because of their usefulness for the author, which is something different from their quality;
- Only a small fraction of all used material is cited;
- General knowledge is not cited;
- Knowledge is often cited from secondary sources;
- Documents supporting an author's arguments are cited more often than other documents;
- Flattering (citing editors, potential referees, and other authorities);
- Showing off (citing hot new "in" articles);
- Reference copying (references provided by other authors);
- Conventions (in biochemistry, for example, methods are cited but not reagents);
- Self citations; and
- Citing colleagues (often reflecting informal transfer of information)

Such research says something about the usefulness of references vs. descriptors in information seeking. To the degree that the conventions can be generalized and described, they are of immediate relevance. For example, with the knowledge given above, we can state that citation

indexing should perform well on a search for biochemical methods but rather badly on a search for a reagent. There are many studies in this exciting area of citation behavior that directly or indirectly illuminate both the strengths and weaknesses of citations as SAPs, but space limitations prevent us from referring to more of them.

It should be clear that the evaluation of the possibilities of chain searching is connected with studies of cooperation and competition among scientists and subsequent citation behavior. Studies in the sociology of science and in epistemology are highly relevant. It is not difficult to see the importance of, for example, KUHN's well-known theory of scientific paradigms, which directly explains how groups of scientists develop different criteria for relevance and subsequent citation behavior.

Conclusion. A given subject access point (e.g., descriptors, references) cannot be expected to have a fixed information value regardless of conventions in the knowledge domain and the writing culture. This is a serious argument against positivistic approaches, which try to develop general algorithms and measures without regard for the contents and the context of the information. To the extent that the demands on "optimal citation behavior" are met, the reference list of every document represents a perfect, "selective" bibliography in the field or together with other articles is part of a network that represents a perfect bibliography. Inclusion in the bibliography that is formed by the reference list is characterized in particular by the fact that the bibliography expresses more limited disciplinary and paradigmatic priorities. The strength of chain searching is that, within a scholarly discipline, there is little risk of overlooking the most important documents. The weakness of this method is related to the fact that any given assumptions within a field can be reevaluated. The documents that become relevant after a reevaluation (e.g., a paradigm shift) typically will not be found by chain searching because references are selected according to paradigmatic norms. In addition, the motives of the scientists are not always pure; these authors may not inform the reader of important sources because they wish to reap the fruit of these at some later date.

Both bibliographic searching and chain searching depend on certain conditions that determine their efficiency. Neither method can a priori be said to be the more systematic, and to some degree they are prerequisites for each other. In areas where quality bibliographies exist, bibliographic searching will be strong. In areas where the scientific standard is very high, chain searching will be strong. In the end, scientific work might develop into an efficient bibliography and efficient bibliographies into products of scholarly quality. Under these conditions the subject bibliography will represent the best map of the research area, a

sort of empirical map of the structure of a field, while the article or book and its reference list will be the most accurate answer to a well-defined question, a sort of microsounding in its structure. In other words, bibliographies are more metascientifically oriented. The two products will be able to use each other in this process.

Full Text

There are today several prominent research projects and different research strategies concerned with the retrieval of full-text documents (or parts of these). Among the most important are the Text REtrieval Conference (TREC)[13] experiments, the Digital Libraries Initiative (DLI),[14] research on the Internet including studies on hypertext markup language (HTML), and programs devoted to the analysis of linguistic problems in natural-language processing (NLP). This chapter can present only a selective review of this research that focuses on our theoretical approach to SAPs. The reader should also consult other reviews, including the review of TREC by SPARCK JONES and various *ARIST* chapters on metadata, information retrieval, and full-text databases.

One of the key components of the success of the World Wide Web is HTML, which has been formalized according to the rules defined by the INTERNATIONAL ORGANIZATION FOR STANDARDIZATION (1986), which defines Standard Generalized Markup Language (SGML). Research and development on this issue has a strong bearing on SAPs and is illuminating for people interested in the underlying intellectual structure of texts rather than just the physical display of that text on paper or screen. BRYAN is an influential source, showing how markup languages operate with document type definitions (DTDs) as well as document analysis and information modeling. Bryan also describes the structures of different kinds of documents such as letters, textbooks, and scientific articles and provides explicit coding of all elements in each type of document, showing why they are important tools for improved subject access based on specified text elements.

---

[13] The First Text REtrieval Conference (TREC-1) was held in Gaithersburg, MD, November 4-6, 1992. The eighth Text REtrieval Conference (TREC-8) was held in Gaithersburg, MD, November 17-19, 1999. See also http://trec.nist.gov/.

[14] The Digital Libraries Initiative (DLI) Phase One (1994-1998) comprised six projects at six research universities under the joint initiative of the National Science Foundation (NSF), the Department of Defense Advanced Research Projects Agency (DARPA), and the National Aeronautics and Space Administration (NASA). The DLI's goal is to advance the means to collect, store, and organize information in digital forms and make it available for searching, retrieval, and processing via communication networks in user-friendly ways. The following sites contain conference information, DLI publications, DLI workshop series, and related projects and resources to the DLI. URL: http://dli.grainger.uiuc.edu/national.htm and http://www.dli2.nsf.gov/.

In IS, ELLIS ET AL. explore the retrieval effectiveness of creating hypertext links in full-text documents, while BATES (1998) discusses human and domain factors in indexing for digital libraries and the Internet. MALET ET AL. describe how medical document retrieval on the Internet can be enhanced utilizing medical core metadata, such as the National Library of Medicine's Medical Subject Headings (MeSH) vocabulary and MEDLINE-type content descriptions. TURNER & BRACKBILL found that the use of the keywords-attribute in a META tag substantially improves accessibility. They suggest that HTML document authors should consider using keywords attribute META tags and that more search engines should index the META tag to improve resource discovery.

One example of relevant research on SAPs using the natural-language processing approach (NLP) is the article by PIRKOLA & JÄRVELIN, who studied the effect of anaphor[15] and ellipsis[16] resolution on proximity searching in a newspaper article database. Their findings indicate a recall increase of 38.2% in sentence searches and 28.8% in paragraph searches when proper-name ellipses were resolved. The increase in recall was 17.6% in sentence searches and 10.8% in paragraph searches when proper-name anaphora were resolved. This result suggests that some simple and computationally justifiable resolution method might be developed for proper-name phrases to support keyword-based full-text IR. PEREZ-CARBALLO & STRZALKOWSKI describe "stream-architecture," a method they designed to combine evidence from different document representations by also applying NLP. DR-LINK, described by LIDDY & MYAENG and by MANNING & NAPIER INFORMATION SERVICES, is an advanced approach to NLP, in which it is possible to search for causes and consequences of events, for example.

With this brief introduction to current research, we now look at a few important studies. TENOPIR & RO report some experiments in full-text retrieval. In a study of *Harvard Business Review* online, they found that full-text searching retrieved 7.4 times more documents than did abstracts, 5.7 times more than controlled vocabulary, and 3.4 times more than the bibliographical union (abstracts, controlled vocabulary, and titles). They define relative recall as the proportion of relevant documents a searcher would retrieve if searching only with that one method.

[15]An anaphor is the repetition of a word or phrase in successive clauses as a literary device—e.g., "for them he worked, for them he went hungry, for them he was tempted to steal." (LEXICON PUBLICATIONS)

[16] An ellipsis is a construction that omits one or more words that must be understood for the grammatical completeness of a sentence, for example, in "it's a book I would d . . . d well like to read." The dots indicate that a word, words, or part of a word has been omitted, in this case "d . . . d" for damned." (LEXICON PUBLICATIONS)

On average almost three-fourths of all relevant documents could be retrieved by full-text searching without any value-added fields. Controlled vocabulary contributed on average 28%, abstracts 19.3%, and the bibliographical union 44.9%. These results indicate the value of full-text searching in this database. However, Tenopir and Ro also suggest the importance of value-added fields because in some queries certain documents would not be retrieved without them, and, as hypothesized, full-text searches have a lower precision ratio than do abstracts or controlled-vocabulary searches.

Whereas most studies (e.g., SARACEVIC; SIEVERT ET AL.; TENOPIR, 1985a) compare the overall performance of full-text retrieval with value-added fields, some studies try to illuminate the parts of a full-text document that contribute to its retrievability. VOOS & DAGAEV studied the placement of citations to four highly cited articles in the citing papers. Dividing the articles into four parts—introduction, methodology, discussion, and conclusion—they found that "on the average, the source articles, when highly cited, seem to occur more in the introduction than anywhere else in the article" (p. 20). They conclude that the value of a citation to a researcher depends not only on the number of times it is referenced but also on its placement in the citing article. In the same way we may assume that future full-text retrieval systems may consider the relative information value of terms from different parts of documents.

BISHOP describes DeLIver, a web-based testbed that is a part of the Digital Libraries Initiative at the University of Illinois. DeLIver contains the full text of recent articles from more than 50 science and engineering journals and has the capacity, through Standard Generalized Markup Language (SGML) and enhanced search features, to support retrieval of newly foregrounded document components. Information in individual parts can be disaggregated from the surrounding textual package and retrieved for use in a way not possible with traditional bibliographic retrieval systems. One can search for terms in particular components of documents (e.g., "spectrum" in a figure caption) to enhance the precision of the search. Users can either execute a search "anywhere in article" or limit the search to title, abstract, table text, figure caption, cited references, and more. (The body of the article itself is not distinguished according to introduction, methods, and conclusion). In DeLIver, one can also view certain components before retrieving the full text of the article (including full texts of documents referred to in references if they are included in DeLIver).

A central theme in Bishop's article is a discussion of the need to replace the traditional linear structures in documents with a free combination of "info-bricks." The traditional structure of documents is seen as an artifact of both the technology of printing and beliefs about the

scientific method that prevailed in the seventeenth century. This raises the question of whether the unit to be retrieved in IR should be seen as a document or another kind of unit, such as an info-brick.

This question is important in the search process, and much valuable research has been done on passage or paragraph retrieval (PR). Studies such as those by AL-HAWAMDEH ET AL., AL-HAWAMDEH & WILLETT, and LALMAS & RUTHVEN can provide knowledge on the function of parts of texts as SAPs. Studies in PR divide documents in segments based on different principles. A motivated segment can be determined by the content (semantics) or explicit structure of the document (including SGML). An unmotivated segment (or "window") can be determined by number of words (e.g., 25 or 1,000 words). Strangely, experiments by CALLAN and others suggest that motivated segmentation of a given text does not always perform as well as windows. It is, however, too early to draw firm conclusions on this.

If searchers need only a part of a document, they will usually need the whole document as the reference (HJØRLAND, 2000). From our point of view, the most interesting point is not PR as an aim in itself but how the retrieval of whole documents can be improved by using SAPs in full text (in general or by using searches limited to parts of full texts, as for example, the use of conclusions to enhance precision[17]).

Because different kinds of texts have different structures with different consequences for retrieval, we first need a typology of documents. Newspaper articles, for example, usually are organized in a pyramid structure, with the most important information in the heading, then less important information in the first paragraph, and so on. This is done in order to keep the attention of the reader as long as possible. Information retrieval from full-text newspaper databases should take advantage of this structure, whereas IR in scientific articles could vary the retrieval strategy depending, for example, on whether methodological issues or conclusions are of most interest.

DIODATO offers a study on how title words appear in parts of research papers. Given the assumption that title words reflect article content, they propose some interesting ways in which more relevant search terms from the text itself could be identified. Despite a general similarity among the disciplines, they found some important differences. First, the absence of a significant change over time in the number of title words per article in history and philosophy indicates that IR systems would expect changes to occur more slowly in the vocabulary of these two fields than in the other fields. Second, the better matching

[17] In Boolean searches recall cannot be improved by PR. This is not the case if other retrieval methods such as vector-space models are used.

in history and philosophy than in chemistry of title words with first-paragraph words emphasizes that IR systems should be aware that history and philosophy articles often begin with long introductory paragraphs, while chemistry articles assign some of the important introductory material to abstracts. Extraction of terms from a chemical abstract may well be comparable to extraction of terms from the first paragraph of a history or philosophy article. Third, the better match in history, philosophy, and economics than in chemistry and mathematics between title words and last-paragraph words suggests the tendency of the former group of journals to use last paragraphs for recapitulation. The latter group often terminates articles when the final result has been demonstrated or the final theorem proved. An IR system that extracts data from only the last section of a chemistry or mathematics journal would get an incomplete picture. Fourth, the better match in chemistry, mathematics, and economics than in history and philosophy between citing and cited titles is partly due to many non-English language titles cited by the latter group. The use of the bibliography of an article for clues to its content would find this a more effective strategy in chemistry, mathematics, and economics than in history and philosophy.

BLAIR & MARON reported on the problems of language in full-text IR in the STAIRS experiment. How can one identify, for example, all documents about a certain train accident? The searcher will think of some obvious terms, and there is a good chance that these will retrieve some relevant documents. However, the searcher may not realize that many other relevant documents will not contain the terms "train" and "accident" or obvious synonyms. Blair and Maron write that this occurs because natural language can be used to discuss a subject using an unpredictably varied and creative combination of words and phrases. The size of such problems is illustrated by the results that recall was on average no better than 20% with a 79% mean precision level. According to the authors, these results were achieved in an environment that was unusually favorable for effective retrieval.

What Blair and Maron do not say—but what is implied in their example–is that relevant documents can describe events leading to the accident, which is not terminologically linked to documents about the accident itself. Then retrieval is not just a matter of the creative expressiveness of natural language, but it is also a matter of real knowledge of what is searched (e.g., the accident). In the process of retrieval, searchers must learn about the object about which they are seeking information, and this subject knowledge must then be fed into the retrieval process to expand the search criteria (iterative searching). For example, an accident can be caused by a failure in a certain kind of signal; thus, the name of the manufacturer of the signal could be a relevant search

term. No linguistic theory can provide such knowledge. Searchers thus face the problem of predicting three interacting levels of problems:

- What is in reality (e.g., causes of train accidents)? This is substantive knowledge. At the most fundamental and general level this is ontological knowledge.
- What is known and described so that it can be retrieved and trusted (e.g., engineering studies of train accidents and newspaper reports on train accidents)? These are problems related to theory of knowledge, science studies, and theory of information sources.
- How is recorded knowledge described (e.g., engineering terminology, legal language, and ordinary language)? These require familiarity with document composition and discourse communication and thus particularly relate to terminological, linguistic, and library and information science knowledge.

Such knowledge is *not* the same as subject knowledge as ordinarily taught at universities, although people with subject knowledge often have implicit knowledge about methodological problems, publication patterns, and terminology. Normally, however, they are not experts in such issues. Theories of information seeking and retrieval should provide more explicit knowledge of such questions. Information scientists studying bibliometric patterns, terminological problems (e.g., thesauri), and the like have some advantages in relation to ordinary subject specialists in this respect (which is in accordance with the view expressed by BATES (1999)).

So far these problems have not been seriously addressed theoretically in IS, but mostly by common-sense approaches to ontology, epistemology, and text theory. Controlled systems for information selection and vocabularies normally reduce the searcher's load of predicting such knowledge. Retrieval of documents, for example, on train accidents, is very different in a dedicated journal or database about accident research and prevention than in a merged journal or database. The cognitive and social organization of knowledge in disciplines and literatures facilitates the retrieval of information by reducing the semantic distances between documents and searchers (and in the variance among the documents). A well-designed thesaurus could provide information about, for example, the manufacturer of signals.

Conclusion. Full-text databases form the ultimate challenge to information professionals and to information science. We have put forward empirical and theoretical evidence demonstrating that full-text databases without value-added information are not performing with 100%

effectiveness and that value can be added successfully. We have also tried to show that further investigation into the typology and architecture of the texts themselves has potential for the further advancement of full-text retrieval systems.

Descriptors, Identifiers, Classification Codes, and Other Kinds of Access Data

Classification and indexing are big areas in library and information science with a lot of literature that cannot be reviewed here. We limit ourselves to a few principal aspects related to the overall perspective of this review.

When indexers assign keywords to a record, they are influenced by the title, the abstract, and other access points already given. (Often, for example, the subject headings given by the Library of Congress and printed on the colophon in books affect the way books are classified and indexed in other libraries.) This fact presents a problem in interpreting the relative role of such access points. In other words, the value-added services provided by classifiers, indexers, and abstractors are not always independent interpretations of a document's subjects. If they were (or to the degree that they are independent), their relative importance in retrieval could be determined in relation to those provided by the document itself (i.e., by the author). Certainly, empirical evidence tells us that descriptors and other indexer-assigned keywords do improve retrieval considerably (e.g., PAO, 1994). However, the nature of this improvement is not described well today, although FUGMANN (1993; 1994), among others, has contributed much to the theoretical clarification.

In the literature of information science it is often thought to be ideal if different indexers are mutually consistent. However, as COOPER demonstrated, indexing can be consistently bad, which is why consistency is not necessarily a good criterion of quality in indexing. One can even imagine that indexers who are careless or mechanical in their work are much more apt to use keywords very similar to words from the title, for example. If the title is misleading, the indexing will be misleading. However, indexers and different SAPs could appear consistent and in a way confirm each other in a wrong subject analysis (which again may make the users judge those bibliographical records as relevant on an erroneous basis).

If indexing does not add information to a record, it is unnecessary. However, the repetition of words from titles in indexes is not always redundant. It is only redundant if the repetition is based on mechanical, noninterpretative indexing. Titles often contain metaphorical expressions, so searchers should avoid using titles as access points. In those

cases repetition of words from nonmetaphorical words is often necessary, and the indexer has contributed value-added information by distinguishing titles that are useful from those that are not.

The primary contribution of indexers and abstractors is the determination of the subjects of the documents to be indexed (which may vary according to different user groups, so that the indexing should be tailored to the target group). The secondary contribution from indexers is the formulation of the subjects in one or more languages, which facilitates retrieval. There are important investigations of the relative role of controlled vs. uncontrolled vocabularies in indexing (ROWLEY) and of closed systems as classifications vs. open systems as kinds of keywords. One of the most ambitious modern projects for establishing controlled vocabularies is the Unified Medical Language System (UMLS) of the NATIONAL LIBRARY OF MEDICINE.

Conclusion. Earlier IS research has been dominated by the search for one perfect all-purpose IR language that would accommodate users who prefer different languages, such as UDC, PRECIS, Bliss, descriptor-based systems, and citation indexes. Today the trend is to view different IR languages as complementary elements in a system. In other words, it seems important to define the relative strengths and weaknesses of different kinds of IR languages and to match them to special needs in different kinds of documents, media, domains, and user groups. The search for an ideal IR language seems to be related to the old philosophical dream of building a perfect language (cf. ECO).[18]

CONCLUSION

Studies have convincingly demonstrated that searchers who use different SAPs produce different but more-or-less overlapping results. PAO (1994) found that duplicate documents retrieved by the use of any two search fields had much higher odds of being judged relevant than those retrieved by only one of the fields. She concludes that the underlying principle of low overlap is still not well understood and that more research is needed.

What she and others have showed convincingly is that the quality of the subjective relevance evaluation increases dramatically when there are more and different cues in the records. This is not surprising. The quality of the judgment increases when its basis improves.

When researchers are attacking a problem—say, how to cure a disease—they are led by different hypotheses and assumptions about what is relevant. In this process they are using parts of the scientific

[18] The same rationalistic dream seems to lie behind the search for one perfect search algorithm in mainstream IR.

literature that are judged relevant on those premises. However, because this is a dynamic process, the relevance criteria are likely to be changing during the process itself (cf. HJØRLAND, 1997, pp. 165-166). The most tangible expression for what researchers find relevant are the references they include in the final document. However, some relevant documents may not be cited because they seem too general. Also, some nonrelevant documents may be cited for various reasons. Most importantly, if there appears to be a change in the theoretical approach in the field, the researchers may change their previous relevance criteria and reevaluate what they first considered relevant. When seeking new documents based on a changed concept of the problem, users will interpret every cue, which may indicate which documents will be relevant from the changed position. For example, a cue might be that the relevant papers cite other papers that demonstrate a similar conceptualization of the problem or that use a terminology developed to discriminate this conceptualization from others (or that is published in places or by journals devoted to such a conceptualization).

One problem is whether documents are judged relevant or are discarded given ideal conditions for studying them. Another problem is whether they are judged relevant or are discarded on the basis of certain cues (such as author, title, abstracts, recommendations). Even careful studies of single documents are often subjective and uncertain (as we know, for example, from book reviews and hermeneutic studies). Judgment of the relevance of single documents based on a quick examination of a few search fields increases this subjectivity and uncertainty in relevance evaluation dramatically. A given record may contain relevant words in the title or in the descriptors, it may cite well-known studies among its references, it may be published in a leading journal in the field, and so forth. Given the high degree of uncertainty in relevance assessment, it is not surprising that a given person is more likely to regard a document as being relevant if more than one cue indicates relevance. This finding is obvious. Overlap as a retrieval strategy can therefore be used to increase precision in searches, as PAO (1994) concludes. This occurs, however, at the expense of recall.

Because subjective relevance assessments are necessarily based on the available information, IS must focus more on the study of the objective informativeness of different SAPs, that is, on the given possibilities that searchers have, regardless of how they evaluate them and whether or not they understand how to use them. As BATES (1987) suggests, behavioral studies to date have not explained much of the variation in online search success; that is why a hard look at the information itself, especially its structure and organization, is likely to prove more fruitful. Although valuable behavioral research has since been carried out (e.g., FIDEL, 1991a, 1991b, 1991c; SARACEVIC & KANTOR),

the study of texts and "information" is still underrepresented. The more we know about how authors use titles and terminology, how they compose their documents, how they cite other documents, and how they are affected by metatheoretical trends, as well as the more we know about the indexing and abstracting process, the more we know about objective search possibilities. From here we can go on to study how those possibilities are actually used (the subjective, behavioral, and computerized side of searching).

BIBLIOGRAPHY

AL-HAWAMDEH, SULIMAN; DE VERE, RACHEL; SMITH, GEOFF; WILLETT, PETER. 1991. Using Nearest-Neighbour Searching Techniques to Access Full-Text Documents. Online Review. 1991 June/August; 15(3/4): 173-191. ISSN: 0309-314X.

AL-HAWAMDEH, SULIMAN; WILLETT, PETER. 1989. Comparison of Index Term Weighting Schemes for the Ranking of Paragraphs in Full-text Documents. International Journal of Information and Library Research. 1989; 1(2): 116-130. ISSN: 0953-556X.

ALTAVISTA. 2000. Advanced Search Cheat Sheet. Available WWW: http://help.altavista.com/adv_search/syntax.

ALTERMAN, RICHARD. 1991. Understanding and Summarization. Artificial Intelligence Review. 1991; 5(4): 239-254. ISSN: 0269-2821; CODEN: AIRVE6. Also available as: Text Summarization. In: Shapiro, Stuart C., ed. Encyclopedia of Artificial Intelligence. New York, NY: Wiley; 1992. 1579-1587. ISBN: 0-471-50307-X; LC: 91-37272.

ATKINSON, ROSS. 1984. The Citation as Intertext: Toward a Theory of the Selection Process. Library Resources & Technical Services. 1984 April/June; 28(2): 109-119. ISSN: 0024-2527.

BARKER, FRANCES H.; VEAL, DOUGLAS C.; WYATT, BARRY K. 1972. Comparative Efficiency of Searching Titles, Abstracts, and Index Terms in a Free-text Data Base. Journal of Documentation. 1972 March; 28(1): 22-36. ISSN: 0022-0418.

BATES, MARCIA J. 1987. Information: The Last Variable. In: Chen, Ching-chih, ed. ASIS '87: Proceedings of the American Society for Information Science (ASIS) 50th Annual Meeting; 1987 October 4-8; Boston, MA. Medford, NJ: Learned Information, Inc. for ASIS; 1987. 6-10. ISSN: 0044-7070; ISBN: 0-938734-19-9; CODEN: PAISDQ.

BATES, MARCIA J. 1998. Indexing and Access for Digital Libraries and the Internet: Human, Database, and Domain Factors. Journal of the American Society for Information Science. 1998 November; 49(13): 1185-1205. ISSN: 0002-8231; CODEN: AISJB6.

BATES, MARCIA J. 1999. The Invisible Substrate of Information Science. Journal of the American Society for Information Science. 1999 October; 50(12): 1043-1050. ISSN: 0002-8231; CODEN: AISJB6.

BAZERMAN, CHARLES. 1985. Physicists Reading Physics: Schema-Laden Purposes and Purpose-Laden Schema. Written Commmunication. 1985 January; 2(1): 3-23. ISSN: 0741-0883.

BAZERMAN, CHARLES. 1988. Shaping Written Knowledge: The Genre and Activity of the Experimental Article in Science. Madison, WI: The University of Wisconsin Press; 1988. 356p. ISBN: 0-299-11690-5; ISBN: 0-299-11694-8 (pbk.).

BERNAL, JOHN DESMOND. 1948. Preliminary Analysis of Pilot Questionnaire on the Use of Scientific Literature. In: Royal Society [Great Britain]. The Royal Society Scientific Information Conference: Report and Papers Submitted; 1948 June 21-July 2; London, UK. London, UK: Royal Society; 1948. 589-637. OCLC: 1820040.

BERNARD, M. 1995. À juste titre: A Lexicometric Approach to the Study of Titles. Literary & Linguistic Computing. 1995; 10(2): 135-141. ISSN: 0268-1145; CODEN: LLCOEI.

BERNIER, CHARLES L. 1980. Subject Indexes. In: Kent, Allen; Lancour, Harold; Daily, J.E., eds. Encyclopedia of Library and Information Science: Volume 29. New York, NY: Marcel Dekker; 1980. 191-205. ISBN: 0-8247-2027-X.

BERRY, MICHAEL W.; DUMAIS, SUSAN T.; O'BRIEN, GANIN W. 1995. Using Linear Algebra for Intelligent Information-Retrieval. SIAM Review. 1995 December; 37(4): 537-595. ISSN: 0036-1445; CODEN: SIREAD.

BISHOP, ANN PETERSON. 1999. Document Structure and Digital Libraries: How Researchers Mobilize Information in Journal Articles. Information Processing & Management. 1999; 35(3): 255-279. ISSN: 0306-4573; CODEN: IPMADK.

BLAIR, DAVID C. 1990. Language and Representation in Information Retrieval. Amsterdam, The Netherlands: Elsevier Science Publishers; 1990. 335p. ISBN: 0-444-88437-8; LC: 89-29881.]

BLAIR, DAVID C. 1996. Stairs Redux: Thoughts on the Stairs Evaluation, 10 Years After. Journal of the American Society for Information Science. 1996 January; 47(1): 4-22. ISSN: 0002-8231.

BLAIR, DAVID C.; MARON, M.E. 1990. Full-Text Information-Retrieval: Further Analysis and Clarification. Information Processing & Management. 1990; 26(3): 437-447. ISSN: 0306-4573; CODEN: IPMADK.

BORKO, HAROLD S.; CHATMAN, S. 1963. Criteria for Acceptable Abstracts: A Survey of Abstracters' Instructions. American Documentation. 1963 April; 14(2): 149-160. ISSN: 0096-946X.

BOYCE, BERT. 1982. Beyond Topicality: A Two Stage View of Relevance and the Retrieval Process. Information Processing & Management. 1982; 18(3): 105-109. ISSN: 0306-4573; CODEN: IPMADK.

BRADFORD, SAMUEL CLEMENT. 1948. Documentation. London, UK: C. Lockwood; 1948. 156p. LC: 49-12638; OCLC 1347246.

BROOKS, TERRENCE ALAN. 1993. All the Right Descriptors: A Test of the Strategy of Unlimited Aliasing. Journal of the American Society for Information Science. 1993 April; 44(3): 137-147. ISSN: 0002-8231; CODEN: AISJB6.

BRYAN, MARTIN. 1997. Web SGML and HTML 4.0 Explained. Available WWW: http://www.sgml.u-net.com/book/home.htm. Also published as: SGML and HTML Explained. 2nd edition. Harlow, UK: Addison Wesley Longman; 1997. 234p. ISBN: 0-201-40394-3.

BUCKLEY, CHRISTOPHER; MITRA, MANDAR; WALZ, JANET; CARDIE, CLAIRE. 2000. Using Clustering and SuperConcepts within SMART: TREC 6. Information Processing & Management. 2000 January; 36(1): 109-131. ISSN: 0306-4573; CODEN: IPMADK.

BUXTON, ANDREW B.; MEADOWS, A. J. 1977. The Variation in the Information Content of Titles of Research Papers with Time and Discipline. Journal of Documentation. 1977 March; 33(1): 46-52. ISSN: 0022-0418.

BUXTON, ANDREW B.; MEADOWS, A.J. 1978. Categorization of the Information in Experimental Papers and Their Author Abstracts. Journal of Research Communication Studies. 1978 August; 1(2): 161-182. ISSN: 0378-5939.

BYRNE, J.R. 1975. Relative Effectiveness of Titles, Abstracts, and Subject Headings for Machine Retrieval from the COMPENDEX Services. Journal of the American Society for Information Science. 1975 July/August; 26(4): 223-229. ISSN: 0002-8231; CODEN: AISJB6.

CALLAN, JAMES P. 1994. Passage-Level Evidence in Document Retrieval. In: Croft, W. Bruce; van Rijsbergen, C. J. eds. SIGIR '94: Proceedings of the Association for Computing Machinery Special Interest Group on Information Retrieval (ACM SIGIR) 17th Annual International Conference on Research and Development in Information Retrieval; 1994 July 3-6; Dublin, Ireland. Berlin, Germany: Springer Verlag; 1994. 302-310. ISBN: 3-540-19889-X.

CHAFE, W.L. 1976. Givenness, Contrastiveness, Definiteness, Subject, Topic, and Point of View. In: Li, Charles N., ed. Symposium on Subject and Topic; 1975 March; Santa Barbara, CA. New York, NY: Academic Press; 1976. 25-55. ISBN: 0-12-447350-4; LC: 75-43861.

CHALMERS, ALAN F. 1999. What Is This Thing Called Science? 3rd edition. Buckingham, UK: Open University Press; 1999. 266p. ISBN: 0-335-20109-1.

CHEN, HSINCHUN; HOUSTON, ANDREA L.; SEWELL, ROBIN R.; SCHATZ, BRUCE R. 1998. Internet Browsing and Searching: User Evaluations of Category Map and Concept Space Techniques. Journal of the American Society for Information Science. 1998 May; 49(7): 582-603. ISSN: 0002-8231.

COLLISON, ROBERT LEWIS. 1971. Abstracts and Abstracting Services. Santa Barbara, CA: ABC-Clio; 1971. 122p. ISBN: 0-87436-078-1; LC: 78-149635.

COOPER, WILLIAM S. 1969. Is Interindexer Consistency a Hobgoblin? American Documentation. 1969 July; 20(3): 268-278. ISSN: 0096-946X; CODEN: AMDOA7.

CORMACK, GORDON V.; CLARKE, CHARLES L. A.; PALMER, CHRISTOPHER R.; TO, SAMUEL S. L. 2000. Passage-Based Query Refinement. Information Processing & Management. 2000 January; 36(1): 133-153. ISSN: 0306-4573; CODEN: IPMADK.

CROSBY, HARRY H. 1976. Titles, A Treatise On. College Composition and Communication. 1976 December; 27(4): 387-391. ISSN: 0010-096X; CODEN: CCCOAM.

DE GROLIER, ERIC. 1965. Current Trends in Theory and Practice of Classification. In: Atherton, Pauline, ed. Classification Research: Proceedings of the

2nd International Study Conference; 1965 September 14-18; Elsinore, Denmark. Copenhagen, Denmark: Munksgaard; 1965. 9-14. (FID Publication no. 370). OCLC: 9499510.

DIODATO, VIRGIL. 1982. The Occurrence of Title Words in Parts of Research Papers: Variations among Disciplines. Journal of Documentation. 1982 September; 38(3): 192-206. ISSN: 0022-0418.

DONG, XIAOYING; SU, LOUISE T. 1997. Search Engines on the World Wide Web and Information Retrieval from the Internet: A Review and Evaluation. Online & CD-ROM Review. 1997 April; 21(2): 67-82. ISSN: 1353-2642.

ECO, UMBERTO. 1995. The Search for the Perfect Language. [Translated from Italian: Ricerca della lingua perfetta nella cultura europea]. Oxford, UK: Blackwell; 1995. 385p. ISBN: 0-631-17465-6; LC: 94-29141.

ELLIS, DAVID; FURNER, JONATHAN; WILLETT, PETER. 1996. On the Creation of Hypertext Links in Full-Text Documents: Measurement of Retrieval Effectiveness. Journal of the American Society for Information Science. 1996 April; 47(4): 287-300. ISSN: 0002-8231; CODEN: AISJB6

FEDOSYUK, M.YU. 1978. Linguistic Criteria for Differentiating Informative and Indicative Abstracts. Automatic Documentation and Mathematical Linguistics. 1978; 12(3): 98-110. ISSN: 0005-1055.

FIDEL, RAYA. 1986. Writing Abstracts for Free-Text Searching. Journal of Documentation. 1986 March; 42(1): 11-21. ISSN: 0022-0418.

FIDEL, RAYA. 1991a. Searchers' Selection of Search Keys, 1: The Selection Routine. Journal of the American Society for Information Science. 1991 August; 42(7): 490-500. ISSN: 0002-8231; CODEN: AISJB6.

FIDEL, RAYA. 1991b. Searchers' Selection of Search Keys, 2: Controlled Vocabulary or Free-Text Searching. Journal of the American Society for Information Science. 1991 August; 42(7): 501-514. ISSN: 0002-8231; CODEN: AISJB6.

FIDEL, RAYA. 1991c. Searchers' Selection of Search Keys, 3: Searching Styles. Journal of the American Society for Information Science. 1991 August; 42(7): 515-527. ISSN: 0002-8231; CODEN: AISJB6.

FIDEL, RAYA; EFTHIMIADIS, EFTHIMIS NIKOLAOS. 1995. Terminological Knowledge Structure for Intermediary Expert-Systems. Information Processing & Management. 1995 January/February; 31(1): 15-27. ISSN: 0306-4573; CODEN: IPMADK.

FLYNN, PETER. 1997. W[h]ither the Web?: The Extension or Replacement of HTML. Journal of the American Society for Information Science. 1997 July; 48(7): 614–621. ISSN: 0002-8231; CODEN: AISJB6.

FRAENKEL, AVIEZRI S.; KLEIN, SHMUEL T. 1999. Information Retrieval from Annotated Texts. Journal of the American Society for Information Science. 1999 August; 50(10): 845-854. ISSN: 0002-8231; CODEN: AISJB6.

FRIIS-HANSEN, JENS B.; STEEN LARSEN, POUL; HØST, TORBEN; SPANG-HANSSEN, HENNING. 1996. Informationsordbogen: Ordbog for informationshåndtering, bog og bibliotek, 2. udg. [Dictionary of Information Terms]. Charlottenlund, Denmark: Dansk Standard (DS); 1996. 196p. (DS-Håndbog 109). ISSN: 0903-0484; ISBN: 87-7310-186-9.

FUGMANN, ROBERT. 1993. Subject Analysis and Indexing: Theoretical Foundation and Practical Advice. Frankfurt am Main, Germany: Index Verlag; 1993. 250p. (Textbooks for Knowledge Organization, vol. 1). ISBN: 3-88672-500-6.

FUGMANN, ROBERT. 1994. Representational Predictability: Key to the Resolution of Several Pending Issues in Indexing and Information Supply. In: Albrechtsen, Hanne; Ørnager, Susanne, eds. Knowledge Organization and Quality Management: Proceedings of the 3rd International Conference of the International Society for Knowledge Organization (ISKO); 1994 June 20-24; Copenhagen, Denmark. Frankfurt am Main, Germany: Index Verlag; 1994. 414-422. (Advances in Knowledge Organization, vol. 4). ISBN: 3-88672-023-3.

GARFIELD, EUGENE. 1965. Can Citation Indexing Be Automated? In: Stevens, Mary E.; Giuliano, Vincent E.; Heilprin, Laurence B., eds. Statistical Association Methods for Mechanized Documentation, Symposium Proceedings; 1964; Washington, DC. Washington, DC: National Bureau of Standards; 1965. 189-192. (National Bureau of Standards Miscellaneous Publication, no. 269). Also published in: Essays of an Information Scientist: 1962-1973. Volume 1. Philadelphia, PA: ISI Press; 1973. 84-90. Also available WWW: http://www.garfield.library.upenn.edu/essays/V1p084y1962-73.pdf.

GENETTE, GÉRARD. 1988. Structure and Functions of the Title in Literature. Critical Inquiry. 1988; 14(4): 692-720. ISSN: 0093-1896.

GERICK, THOMAS. 1999. Content-based Information Retrieval auf Basis Semantischer Abfragenetze. NFD Information-Wissenschaft und Praxis. 1999 July; 50(4): 205-209. ISSN: 1434-4653: CODEN: NADOAW.

GILLASPIE, DEBORAH L. 1992. Why Online Legal Retrieval Misses Conceptually Relevant Documents. In: Shaw, Debora, ed. ASIS '92: Proceedings of the American Society for Information Science (ASIS) 55th Annual Meeting: Volume 29; 1992 October 26-29; Pittsburgh, PA. Medford, NJ: Learned Information, Inc. for ASIS; 1992. 256-259. ISSN: 0044-7870; ISBN: 0-938734-69-5; CODEN: PAISDQ.

GILLASPIE, DEBORAH L. 1995. The Role of Linguistic Phenomena in Retrieval Performance. In: Kinney, Tom, ed. ASIS '95: Proceedings of the American Society for Information Science (ASIS) 58th Annual Meeting: Volume 32; 1995 October 9-12; Chicago, IL. Medford, NJ: Information Today, Inc. for ASIS; 1995. 90-96. ISSN: 0044-7870; ISBN: 1-57387-017-X; CODEN: PAISDQ.

GORDON, MICHAEL D.; DUMAIS, SUSAN T. 1998. Using Latent Semantic Indexing for Literature Based Discovery. Journal of the American Society for Information Science. 1998 June; 49(8): 674-685. ISSN: 0002-8231; CODEN: AISJB6

GREEN, REBECCA. 1995. Topical Relevance Relationships 1: Why Topical Matching Fails. Journal of the American Society for Information Science. 1995 October; 46(9): 646-653. ISSN: 0002-8231.

GREEN, REBECCA. 2000. Locating Sources in Humanities Scholarship: The Efficacy of Following Bibliographic References. Library Quarterly. 2000 April; 70(2): 201-229. ISSN: 0024-2519.

GREEN, REBECCA; BEAN, CAROL A. 1995. Topical Relevance Relationships 2: An Exploratory Study and Preliminary Typology. Journal of the American Society for Information Science. 1995 October; 46(9): 654-662. ISSN: 0002-8231.

HARRIS, JESSICA L. 1974. Document Description and Representation. In: Cuadra, Carlos A., ed. Annual Review of Information Science and Technology: Volume 9. Washington, DC: American Society for Information Science; 1974. 81-117. ISSN: 0066-4200; ISBN: 0-87715-209-8.

HARTER, STEPHEN P.; HERT, CAROL A. 1997. Evaluation of Information Retrieval Systems: Approaches, Issues, and Methods. In: Williams, Martha E., ed. Annual Review of Information Science and Technology: Volume 32. Medford, NJ: Information Today, Inc. for the American Society for Information Science; 1997. 3-94. ISSN: 0066-4200; ISBN: 1-57387-047-1.

HARTER, STEPHEN P.; NISONGER, THOMAS E.; WENG, AIWEI. 1993. Semantic Relationships between Cited and Citing Articles in Library and Information Science Journals. Journal of the American Society for Information Science. 1993 October; 44(9): 543-552. ISSN: 0002-8231; CODEN: AISJB6.

HERRELL, JAMES M. 1979. Abstract Thinking in APA Journals. American Psychologist. 1979 February; 34(2): 178-180. ISSN: 0003-066X.

HJØRLAND, BIRGER. 1988. Information Retrieval in Psychology. Behavioral and Social Sciences Librarian. 1988 December; 6(3/4): 39-64. ISSN: 0163-9269.

HJØRLAND, BIRGER. 1992. The Concept of "Subject" in Information Science. Journal of Documentation. 1992 June; 48(2): 172-200. ISSN: 0022-0418; CODEN: JDOCAS.

HJØRLAND, BIRGER. 1997. Information Seeking and Subject Representation: An Activity-Theoretical Approach to Information Science. Westport, CT: Greenwood Press; 1997. 213p. (New Directions in Information Management ; 34). ISBN: 0-313-29893-9.

HJØRLAND, BIRGER. 1998. Information Retrieval, Text Composition, and Semantics. Knowledge Organization. 1998; 25(1/2): 16-31. ISSN: 0943-7444.

HJØRLAND, BIRGER. 2000. Documents, Memory Institutions, and Information Science. Journal of Documentation. 2000; 56(1): 27-41. ISSN: 0022-0418.

HODGES, PAULINE R. 1983. Keyword in Title Indexes. Special Libraries. 1983 January; 74(1): 56-60. ISSN: 0038-6723.

HORNBY, PETER A. 1972. The Psychological Subject and Predicate. Cognitive Psychology. 1972 October; 3(4): 632-642. ISSN: 0010-0285.

HULME, E. WYNDHAM. 1911a. Principles of Book Classification: Introduction. Library Association Record. 1911; 13: 354-358. ISSN: 0024-2195.

HULME, E. WYNDHAM. 1911b. Principles of Book Classification: Chapter II – Principles of Division in Book Classification. Library Association Record. 1911; 13: 389-394. ISSN: 0024-2195.

HULME, E. WYNDHAM. 1911c. Principles of Book Classification: Chapter III – On the Definition of Class Headings, and the Natural Limit to the

Extension of Book Classification. Library Association Record. 1911; 13: 444-449. ISSN: 0024-2195.

IIVONEN, MIRJA; SONNENWALD, DIANE H. 1998. From Translation to Navigation of Different Discourses: A Model of Search Term Selection during the Pre-Online Stage of the Search Process. Journal of the American Society for Information Science. 1998 April; 49(4): 312-326. ISSN: 0002-8231; CODEN: AISJB6.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION (ISO). 1976. Documentation: Abstracts for Publications and Documentation. 1st edition. 6p. (ISO 214:1976). Available from: International Organization for Standardization, http://www.iso.ch/.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION (ISO). 1986. Information Processing—Text and Office Systems—Standard Generalized Markup Language (SGML). 1st edition. 155p. (ISO 8879:1986). Available from: International Organization for Standardization, http://www.iso.ch/.

JANES, JOSEPH W. 1994. Other People's Judgments: A Comparison of Users' and Others' Judgments of Document Relevance, Topicality, and Utility. Journal of the American Society for Information Science. 1994 April; 45(3): 160-171. ISSN: 0002-8231; CODEN: AISJB6.

KEEN, E. MICHAEL. 1992. Some Aspects of Proximity Searching in Text Retrieval-Systems. Journal of Information Science. 1992; 18(2): 89-98. ISSN: 0165-5515; CODEN: JISCDI.

KELLER, BARBARA. 1992. Subject Content through Title: A Masters Theses Matching Study at Indiana State University. Cataloging & Classification Quarterly. 1992; 15(3): 69-80. ISSN: 0163-9374.

KUHN, THOMAS S. 1962. The Structure of Scientific Revolutions. Chicago, IL: University of Chicago Press; 1962. 172p. LC: 62-19621.

LALMAS, MOUNIA; RUTHVEN, IAN. 1998. Representing and Retrieving Structured Documents Using the Dempster-Shafer Theory of Evidence: Modelling and Evaluation. Journal of Documentation. 1998 December; 54(5): 529-565. ISSN: 0022-0418; CODEN: JDOCAS.

LANCASTER, FREDERICK WILFRID. 1998. Indexing and Abstracting in Theory and Practice. 2nd edition. London, UK: Library Association Publishing; 1998. 412p. ISBN: 1-85604-268-5.

LANGRIDGE, DEREK W. 1989. Subject Analysis: Principles and Procedures. London, UK: Bowker-Saur; 1989. 146p. ISBN: 0-408-03031-3.

LARSON, RAY R. 1991. The Decline of Subject Searching: Long-Term Trends and Patterns of Index Use in an Online Catalog. Journal of the American Society for Information Science. 1991 April; 42(3): 197-215. ISSN: 0002-8231; CODEN: AISJB6.

LEXICON PUBLICATIONS. 1990. The New Lexicon Webster's Dictionary of the English Language. New York, NY: Lexicon Publications; 1990. 2000p. ISBN: 0-7172-04546-2.

LI, CHARLES N., ed. 1976. Subject and Topic: Symposium on Subject and Topic; 1975 March; Santa Barbara, CA. New York, NY: Academic Press; 1976. 594p. ISBN: 0-12-447350-4; LC: 75-43861.

LIDDY, ELIZABETH D.; MYAENG, SUNG H. 1993. Linguistic-Conceptual Approach to Document Detection. In: Harman, D.K., ed. The 1st Text REtrieval Conference (TREC-1); 1992 November 4–6; Gaithersburg, MD. Washington, DC: U.S. Department of Commerce, National Institute of Standards and Technology; 1993. 113-129. NTIS: PB93-191641.

LIU, MENGXIONG. 1993. The Complexities of Citation Practices: A Review of Citation Studies. Journal of Documentation. 1993 December; 49(4): 370-408. ISSN: 0022-0418.

MACROBERTS, MICHAEL H.; MACROBERTS, BARBARA R. 1988. Author Motivation for Not Citing Influences: A Methodological Note. Journal of the American Society for Information Science. 1988 November; 39(6): 432-433. ISSN: 0002-8231; CODEN: AISJB6.

MACROBERTS, MICHAEL H.; MACROBERTS, BARBARA R. 1989. Problems of Citation Analysis: A Critical Review. Journal of the American Society for Information Science. 1989 September; 40(5): 342-349. ISSN: 0002-8231; CODEN: AISJB6.

MALET, GARY; MUNOZ, FELIX; APPLEYARD, RICHARD; HERSH, WILLIAM R. 1999. A Model for Enhancing Internet Medical Document Retrieval with "Medical Core Metadata". Journal of the American Medical Informatics Association. 1999 March/April; 6(2): 163-172. ISSN: 1067-5027.

MALMKJÆR, KIRSTEN. 1995. Genre Analysis. In: Malmkjær, Kirsten, ed. The Linguistics Encyclopedia. London, UK: Routledge; 1995. 170-181. ISBN: 0-415-12566-9.

MANNING & NAPIER INFORMATION SERVICES. 2000. DR-LINK: Document Retrieval Using LINguistic Knowledge. Available WWW: http://www.textwise.com/dr-link.html.

MANZER, BRUCE M. 1977. The Abstract Journal, 1790-1920: Origin, Development and Diffusion. Metuchen, NJ: Scarecrow Press; 1977. 312p. ISBN: 0-8108-1047-6; LC: 77-24143.

MOOERS, CALVIN NORTHRUP. 1951. Zatocoding Applied to Mechanical Organization of Knowledge. American Documentation. 1951 January; 2(1): 20-32. ISSN: 0096-946X.

MOOERS, CALVIN NORTHRUP. 1972. Descriptors. In: Kent, Allen; Lancour, Harold; Daily, J.E., eds. Encyclopedia of Library and Information Science: Volume 7. New York, NY: Marcel Dekker; 1972. 31-45. ISBN: 0-8247-2107-1.

MYERS, GREG. 1990. Writing Biology: Texts in the Social Construction of Scientific Knowledge. Madison, WI: University of Wisconsin Press; 1990. 304p. ISBN: 0-299-12230-1; LC: 89-40263.

NAHL-JAKOBOVITS, DIANE; JAKOBOVITS, LEON A. 1987. Teaching the Analysis of Titles: Dependent and Independent Variables in Research Articles. Research Strategies. 1987 Fall; 5(4): 164-171. ISSN: 0734-3310.

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY (NIST). 2000. Proceedings of the 8th Text REtrieval Conference (TREC-8); 1999 November 17-19; Gaithersburg, MD. (NIST Special Publication 500-246). Available WWW: http://trec.nist.gov/pubs/trec8/t8_proceedings.html.

NATIONAL LIBRARY OF MEDICINE. 2000. Unified Medical Language System (UMLS). Available WWW: http://www.nlm.nih.gov/research/umls/umlsmain.html.

NORD, CHRISTIANE. 1991. Text Analysis in Translation: Theory, Methodology, and Didactic Application of a Model for Translation-Oriented Text Analysis. Amsterdam, The Netherlands: Rodopi; 1991. 250p. (Amsterdamer Publikationen zur Sprache und Literatur; 94; Translated from the German by Christiane Nord and Penelope Sparrow). ISBN: 9-05183-311-3.

NORD, CHRISTIANE. 1995. Text-Functions in Translation: Titles and Headings as a Case in Point. Target. 1995; 7(2): 261-284. CODEN: TARGEC; OCLC: 20768955.

OLSEN, KAJ A.; SOCHATS, KENNETH M.; WILLIAMS, JAMES G. 1998. Full Text Searching and Information Overload. International Information & Library Review. 1998 June; 30(2): 105-122. ISSN: 1057-2317.

PAO, MIRANDA LEE. 1993. Term and Citation Retrieval: A Field Study. Information Processing & Management. 1993 January/February; 29(1): 95-112. ISSN: 0306-4573; CODEN: IPMADK.

PAO, MIRANDA LEE. 1994. Relevance Odds of Retrieval Overlaps from Seven Search Fields. Information Processing & Management. 1994 May/June; 30(3): 305-314. ISSN: 0306-4573; CODEN: IPMADK.

PAO, MIRANDA LEE; WORTHEN, DENNIS B. 1989. Retrieval Effectiveness by Semantic and Pragmatic Relevance. Journal of the American Society for Information Science. 1989 July; 40(4): 226-235. ISSN: 0002-8231; CODEN: AISJB6.

PEREZ-CARBALLO, JOSE; STRZALKOWSKI, TOMEK. 2000. Natural Language Information Retrieval: Progress Report. Information Processing & Management. 2000 January; 36(1): 155-178. ISSN: 0306-4573; CODEN: IPMADK.

PERITZ, BLUMA C. 1984. On the Informativeness of Titles. International Classification. 1984; 11(2): 87-89. ISSN: 0340-0050.

PINTO, MARÍA; GÁLVEZ, CARMEN. 1999. Paradigms for Abstracting Systems. Journal of Information Science. 1999; 25(5): 365-380. ISSN: 0165-5515; CODEN: JIOSED.

PIRKOLA, ARI; JÄRVELIN, KALERVO. 1996. The Effect of Anaphor and Ellipsis Resolution on Proximity in a Text Database. Information Processing & Management. 1996 March; 32(2): 199-216. ISSN: 0306-4573; CODEN: IPMADK.

POLLITT, A. S.; TINKER, AMANDA J.; BRAEKEVELT, PATRICK A.J. 1998. Improving Access to Online Information Using Dynamic Faceted Classification. In: McKenna, Brian; Graham, Catherine; Kerr, J., eds. Online Information 98: Proceedings of the 22nd International Online Information Meeting; 1998 December 8-10; London, UK. Oxford, UK: Learned Information Europe; 1998. 17-21. ISBN: 1-900871-31-9.

POULSEN, CLAUS. 1994. Informationens skygge og foran: informationskvalitet, informationseksplosion og online kataloger. Roskilde, Denmark: Institut for Datalogi, Kommunikation og Uddannelsesforskning, Roskilde Universitetscenter; 1994. 198p. (Papirer om faglig formidling; 36). ISBN: 87-7349-264-7.

PRICE, DOUGLAS S. 1983. Possible Impact of Electronic Publishing on Abstracting and Indexing. Journal of the American Society for Information Science. 1983 July; 34(4): 288. ISSN: 0002-8231; CODEN: AISJB6.

RAO, ASHWIN; LU, ALLAN; MEIER, ED; AHMED, SALAHUDDIN; PLISKE, DANIEL. 2000. Query Processing in TREC-6. Information Processing & Management. 2000 January; 36(1): 179-186. ISSN: 0306-4573; CODEN: IPMADK.

ROWLEY, JENNIFER. 1994. The Controlled Versus Natural Indexing Languages Revisited: A Perspective on Information Retrieval Practice and Research. Journal of Information Science. 1994; 20(2): 108-119. ISSN: 0165-5515.

ROWLEY, JENNIFER; FARROW, JOHN. 2000. Organizing Knowledge: An Introduction to Managing Access to Information. 3rd edition. Aldershot, Hampshire, UK: Gower; 2000. 404p. ISBN: 0-566-08047-8.

SALTON, GERARD. 1996. A New Horizon for Information Science. Journal of the American Society for Information Science. 1996 April; 47(4): 333. (Letter to the Editor). ISSN: 0002-8231; CODEN: AISJB6.

SARACEVIC, TEFKO. 1969. Comparative Effects of Titles, Abstracts and Full Texts on Relevance Judgments 1. In: North, J.B., ed. Proceedings of the American Society for Information Science (ASIS) 32nd Annual Meeting: Volume 6; 1969 October 1-4; San Francisco, CA. Westport, CT: Greenwood Publishing; 1969. 293-299. OCLC: 8416080.

SARACEVIC, TEFKO; KANTOR, PAUL. 1988. A Study of Information Seeking and Retrieving. III: Searchers, Searches, and Overlap. Journal of the American Society for Information Science. 1988 May; 39(3): 197-216. ISSN: 0002-8231; CODEN: AISJB6.

SCHULZ-HARDT, STEFAN; FREY, DIETER; LÜTHGENS, CARSTEN; MOSCOVICI, SERGE. 2000. Biased Information Search in Group Decision Making. Journal of Personality and Social Psychology. 2000; 78(4): 655-669. ISSN: 0022-3514.

SEGLEN, PER O. 1996. Bruk av siteringer og tidsskrift-impaktfaktor til forskningsevaluering. [The Use of Citations and Journal Impact Factors for Evaluation of Research]. Biblioteksarbejde. 1996; 48: 27-34. ISBN: 87-88524-55-8.

SIEVERT, MARYELLEN; MCKININ, EMMA JEAN. 1989. Why Full-Text Misses Some Relevant Documents: An Analysis of Documents Not Retrieved by CCML or MEDIS. In: Katzer, Jeffrey; Newby, Gregory B., eds. ASIS '89: Proceedings of the American Society for Information Science (ASIS) 52nd Annual Meeting: Volume 26; 1989 October 30-November 2; Washington, DC. Medford, NJ: Learned Information, Inc. for ASIS; 1989. 34-39. ISSN: 0044-7870; ISBN: 0-938734-40-7; CODEN: PAISDQ.

SIEVERT, MARYELLEN; MCKININ, EMMA JEAN; SLOUGH, MARLENE. 1988. A Comparison of Indexing and Full-Text for the Retrieval of Clinical Medical Literature. In: Borgman, Christine L.; Pai, Edward Y.H., eds. ASIS '88: Information & Technology: Planning for the Next Fifty Years: Proceedings of the American Society for Information Science (ASIS) 51st Annual Meeting: Volume 25; 1988 October 23-27; Atlanta, GA. Medford, NJ: Learned Information, Inc. for ASIS; 1988. 143-146. ISSN: 0044-7870; ISBN: 0-938734-29-6; CODEN: PAISDQ.

SILLINCE, JOHN A. A. 1992. Argumentation-Based Indexing for Information Retrieval from Learned Articles. Journal of Documentation. 1992 December; 48(4): 387-405. ISSN: 0022-0418; CODEN: JDOCAS.

SOERGEL, DAGOBERT. 1985. Organizing Information: Principles of Data Base and Retrieval Systems. Orlando, FL: Academic Press; 1985. 450p. ISBN: 0-12-654260-0; LC: 83-15741.

SOERGEL, DAGOBERT. 1994. Indexing and Retrieval Performance: The Logical Evidence. Journal of the American Society for Information Science. 1994 September; 45(8): 589-599. ISSN: 0002-8231; CODEN: AISJB6.

SPANG-HANSSEN, HENNING. 1974. Kunnskapsorganisasjon, informasjonsgjenfinning, automatisering og språk. [Knowledge Organization, Information Retrieval, Automatization and Language]. In: Kunnskapsorganisasjon og informasjonsgjenfinning: Seminar arrangert 3.-7.desember 1973 i samarbeid mellom Norsk hovedkomité for klassifikasjon, Statens Bibliotekskole og Norsk Dokumentasjonsgruppe. [Knowledge Organization and Information Retrieval: A seminar held December 3rd-7th 1973: A Cooperation between the Norwegian Commitee for Classification, The Norwegian School of Librarianship, and the Norwegian Documentation Group]. Oslo, Norway: Riksbibliotektjenesten; 1973. 11-61. (Skrifter fra Riksbibliotektjenesten, Nr. 2). ISBN: 82-7195-001-0.

SPARCK JONES, KAREN. 2000. Further Reflections on TREC. Information Processing & Management. 2000 January; 36(1): 37-85. ISSN: 0306-4573; CODEN: IPMADK.

SPARCK JONES, KAREN; WALKER, STEPHEN; ROBERTSON, STEPHEN E. 2000. A Probabilistic Model of Information Retrieval: Development and Comparative Experiments, Part 1. Information Processing & Management. 2000 November; 36(6): 779-808. ISSN: 0306-4573; CODEN: IPMADK.

SPARCK JONES, KAREN; WALKER, STEPHEN; ROBERTSON, STEPHEN E. 2000. A Probabilistic Model of Information Retrieval: Development and Comparative Experiments, Part 2. Information Processing & Management. 2000 November; 36(6): 809-840. ISSN: 0306-4573; CODEN: IPMADK.

SPINK, AMANDA. 1995. Term Relevance Feedback and Mediated Database Searching: Implications for Information Retrieval Practice and Systems Design. Information Processing & Management. 1995 March/April; 31(2): 161-171. ISSN: 0306-4573; CODEN: IPMADK.

SPINK, AMANDA; SARACEVIC, TEFKO. 1997. Interaction in Information Retrieval: Selection and Effectiveness of Search Terms. Journal of the American Society for Information Science. 1997 August; 48(8): 741-761. ISSN: 0002-8231; CODEN: AISJB6.

TAUCHERT, WOLFGANG; HOSPODARSKY, JÜRGEN; KRAUSE, JÜRGEN; SCHNEIDER, CHRISTINE; WOMSER-HACKER, CHRISTA. 1991. Effects of Linguistic Functions in Information Retrieval in a German-Language Full-Text Database: Comparison between Retrieval in Abstract and Full-Text. Online Review. 1991 April; 15(2): 77-86. ISSN: 0309-314X; CODEN: OLREDR.

TENOPIR, CAROL. 1984. Retrieval Performance in a Full Text Journal Article Database. Urbana-Champaign, IL: University of Illinois, Graduate School of Library and Information Science; 1984. 264p. Available from: UMI, Ann Arbor, MI. (UMI order no. 85-02315).

TENOPIR, CAROL. 1985a. Full Text Database Retrieval Performance. Online Review. 1985 April; 9(2): 149-164. ISSN: 0309-314X.

TENOPIR, CAROL. 1985b. Searching Harvard Business Review. Online. 1985 March; 9(2): 71-78. ISSN: 0146-5422.

TENOPIR, CAROL; RO, JUNG SOON. 1990. Full Text Databases. New York, NY: Greenwood Press; 1990. 252p. (New Directions in Information Management no. 21). ISBN: 0-313-26303-5; LC: 89-25683.

TIBBO, HELEN R. 1993. Abstracting, Information Retrieval and the Humanities: Providing Access to Historical Literature. Chicago, IL: American Library Association; 1993. 276p. ISBN: 0-8389-3430-7.

TURNER, THOMAS P.; BRACKBILL, LISE. 1998. Rising to the Top: Evaluating the Use of the HTML META Tag to Improve Retrieval of World Wide Web Documents through Internet Search Engines. Library Resources & Technical Services. 1998 October; 42(4): 258-271. ISSN: 0024-2527.

VAN KUPPEVELT, J. 1997. Topic and Comment. In: Lamarque, Peter V.; Asher, R.E., eds. Concise Encyclopedia of Philosophy of Language. Oxford, UK: Pergamon; 1997. 191-198. ISBN: 0-08-042991-2; LC: 97-28781.

VAN RIJSBERGEN, C. J. 1986. A New Theoretical Framework for Information Retrieval. In: Rabitti, Fausto, ed. SIGIR '86: Proceedings of the Association for Computing Machinery Special Interest Group on Information Retrieval (ACM SIGIR) 9th Annual International Conference on Research and Development in Information Retrieval; 1986 September 8-10; Pisa, Italy. New York, NY: ACM Press; 1986. 194-200. ISBN: 0-89791-187-3.

VOORBIJ, HENK J. 1998. Title Keywords and Subject Descriptors: A Comparison of Subject Search Entries of Books in the Humanities and Social Sciences. Journal of Documentation. 1998 September; 54(4): 466-476. ISSN: 0022-0418.

VOOS, HENRY; DAGAEV, KATHERINE S. 1976. Are All Citations Equal? Or, Did We Op. Cit. Your Idem? Journal of Academic Librarianship. 1976 January; 1(6): 19-21. ISSN: 0099-1333.

WANG, PEILING; SOERGEL, DAGOBERT. 1993. Beyond Topical Relevance: Document Selection Behavior of Real Users of IR Systems. In: Bonzi, Susan, ed. ASIS'93: Proceedings of the American Society for Information Science (ASIS) 56th Annual Meeting: Volume 30; 1993 October 24-28; Columbus, OH. Medford, NJ: Learned Information, Inc. for ASIS; 1993. 87-92. ISSN: 0044-7870; ISBN: 0-938734-78-4; CODEN: PAISDQ.

WELWERT, CLAES. 1984. Läsa eller lyssna?: redovisning av jämförande undersökningar gjorda åren 1890-1980 rörande inlärning vid auditiv och visuell presentation samt ett försök till utvärdering av resultaten. Malmö, Sweden: CWK Gleerup; 1984. 233p. (Studia psychologica et paedagogica. Series altera; LXX). ISBN: 91-40-05065-3.

WINDSOR, DONALD A. 1995. Abstract Concerns. Journal of the American Society for Information Science. 1995 October; 46(9): 717-718. ISSN: 0002-8231; CODEN: AISJB6.

WORMELL, IRENE. 1985. Subject Access Project—SAP: Improved Subject Retrieval for Monographic Publications. Lund, Sweden: Lund University; 1985. 174p. (Doctoral thesis at Department of Information and Computer Science, Lund University). OCLC: 19763326.

WRIGHT, LAWRENCE W.; NARDINI, HOLLY K. GROSSETTA; ARONSON, ALAN R; RINDFLESCH, THOMAS C. 1999. Hierarchical Concept Indexing of Full-Text Documents in the Unified Medical Language System(R) Information Sources Map. Journal of the American Society for Information Science. 1999 May; 50(6): 514-523. ISSN: 0002-8231; CODEN: AISJB6.

YITZHAKI, MOSHE. 1992. The Variation in Informativity of [Titles of] Research Papers with Time and Field. In: Neelameghan, A.; Gopinath, M. A.; Raghavan, K. S.; Sankaralingam, P., eds. Cognitive Paradigms in Knowledge Organisation: 2nd International ISKO Conference; 1992 August 26-28; Madras, India. Madras, India: Sarada Ranganathan Endowment for Library Science; 1992. 401-418. OCLC: 28511718.

YITZHAKI, MOSHE. 1996. Informativity of Journal Article Titles: The Ratio of "Significant" Words. In: Ingwersen, Peter; Pors, Niels Ole, eds. Proceedings of the 2nd International Conference on Conceptions of Library and Information Science (COLIS): Integration in Perspective; 1996 October 13-16; Copenhagen, Denmark. Copenhagen, Denmark: The Royal School of Librarianship; 1996. 447-458. ISBN: 87-7415-260-2.

YITZHAKI, MOSHE. 1997. Variation in Informativity of Titles of Research Papers in Selected Humanities Journals: A Comparative Study. Scientometrics. 1997 February; 38(2): 219-229. ISSN: 0138-9130.

ZOBEL, JUSTIN; MOFFAT, ALISTAIR; WILKINSON, ROSS; SACKS-DAVIS, RON. 1995. Efficient Retrieval of Partial Documents. Information Processing & Management. 1995 May/June; 31(3): 361-377. ISSN: 0306-4573; CODEN: IPMADK.