A model based on strikingly different philosophical assumptions from those currently popular is proposed for the design of online subject catalog access. Three design principles are presented and discussed: uncertainty (subject indexing is indeterminate and probabilistic beyond a certain point), variety (by Ashby's law of requisite variety, variety of searcher query must equal variety of document indexing), and complexity (the search process, particularly during the entry and orientation phases, is subtler and more complex, on several grounds, than current models assume). Design features presented are an access phase, including entry and orientation, a hunting phase, and a selection phase. An end-user thesaurus and a front-end system mind are presented as examples of online catalog system components to improve searcher success during entry and orientation.

The proposed model is "wrapped around" existing Library of Congress subject-heading indexing in such a way as to enhance access greatly without requiring reindexing. It is argued that both for cost reasons and in principle this is a superior approach to other design philosophies.

1. Introduction

These are times of great excitement in information systems design. With the advent of online catalogs, the nature of catalogs and access to library materials has changed more in a handful of years than it has in all the rest of the twentieth century. But the changes have been more technological than conceptual. Online catalogs to date have added powerful capabilities to the traditional catalog, yet system designs, generally, have still not gone beyond implementing the card catalog in online form, with some established online search features tacked on.

As valuable and as difficult as these accomplishments have been, they have only scratched the surface of what is possible with online systems (cf. [1]).

We might compare the current situation to that of the automobile in the earlier part of this century. At first, cars were designed like "horseless carriages," i.e., identical in shape and aerodynamics to carriages; only later did designers sort out what features should be altered to better suit the requirements of a vehicle that could move vastly faster than carriages had. So far, many of the features of the online catalog resemble the features of that horse and buggy known as the card catalog. We are just beginning to see features being designed for online catalogs which take fuller advantage of the capabilities of online systems. Designs by Doszkocs [2] (natural-language search queries and automatic term stemming and weighting), Morehead et al. [3] (weighted computations of relatedness between query and fiction records), Cochrane and Markey [4], and Markey and Demeyer [5] (novel uses of book classifications to improve subject access), Hjerppe [6] (application of the "hypertext" concept to catalogs, i.e., breaking the linearity of the traditional file and providing links in a variety of directions from displayed records), and Noerr and Bivins-Noerr [7] (linkages between many different fields in different records), among others, are early indicators of the shift to the design of catalogs truly suited to online use.

In this article the question of subject access in online catalogs is considered anew from its philosophical underpinnings. Design principles are proposed and example system components are described which express those principles. The sequence of topics of discussion is as follows: After this introduction (Section 1), a literature review is presented (Section 2), asking the question: Is improvement in online catalog subject access really needed? In Section 3 design principles differing radically in some respects from the traditional model are proposed and discussed under the rubrics of "uncertainty," "variety," and "complexity (See Fig. 1)." Implications of these
principles for the design of subject access in online catalogs are drawn in Section 4. The design features are described in Section 5, and some example system components are described in Section 6. Many system components might be possible based on the suggested design philosophy; the proposed components, an "end-user thesaurus" and a "front-end system mind" are presented to illustrate and substantiate the recommended approach. A summary (Section 7) completes the article.

While the suggested approach to subject access is significantly different from the conventional one, a design is recommended which uses existing Library of Congress subject cataloging, i.e., which wraps that cataloging into a new, sophisticated system. Not only does this approach save the vast amounts of money that would be necessary to recatalog existing records, but it also turns out to be philosophically preferable, costs aside, for reasons to be argued in Section 4. The proposed system features should provide more sophisticated intellectual access to library materials in the context of equally enhanced sophistication in the use of online capabilities.

2. Background: Is Improvement in Online Catalog Subject Access Really Needed?

As Cochrane [8] has recently noted, a shift is taking place in our assumptions about the relative importance of descriptive and subject access to materials in catalogs. Traditionally, the weight of effort and attention was given to descriptive cataloging. For example, in 1950 the University of California at Berkeley needed to reduce cataloging arrears and undertook some studies to see if the subject cataloging could be cut back to save time. Descriptive cataloging cutbacks were not examined because it was felt that no time could be saved in that area [9, p. 88]. After a study in which it was found that students averaged slightly over two subject catalog uses per year (i.e., 40,000+ student subject catalog uses per year, if there were 20,000 students in the university), it was concluded that "this use is insufficient in quantity to justify the continuance of the subject catalog in its present form" [9, p. 91]. In the article, the possibilities that the subject approach might be important to students and that the relatively low use might be due to inadequate design of the subject access were never considered.

Actually, subject access has constituted between 10% and 62% of card catalog use all along, with an average of about 40 percent [10, p. 76-77]—hardly trivial. Rather, it appears that there has been something of a bias against library subject access; it has been seen as a nice extra, rather than as something essential.

The great popularity of subject access among online catalog users has restored the subject approach in the thinking of the field to its rightful place as a major form of access to library materials. Recent articles have closely examined the strengths and weaknesses of Library of Congress subject headings, the major type of subject headings used in large libraries in the U.S. [11-13]. The Council on Library Resources study found that subject uses constituted 59% of all online catalog uses across many types of libraries and online systems [14, p. 144]. Even though the keyword matching capability of many online catalogs constitutes a great improvement over traditional card catalog subject access (and is probably responsible for much of the interest in subject searches), additional subject-related search capabilities were at the top of the list of desired further improvements named by users [14, p. 134]. In separate studies, Kaske and Sanders [15] and Larson and Graham [16] found similar enthusiasm for subject assistance. Additional subject-related catalog capabilities have also been recommended by a number of authors and study groups [10,17-25].

If we accept that subject access in library catalogs is coming to be valued now and much attention is shifting to it, the question still remains of how much of our resources should go into it. The question might be turned around to say, with all the power of online subject searching of catalogs—Boolean logic, keyword match, truncation, etc.—have we, perhaps, already given the user all the search capability that is practically necessary?

Reviews of card catalog use studies have found average satisfaction rates in subject searching of around 70% or slightly better [26, p. 162; 27, p. 217]. These results have generally been taken as indicating satisfactory performance on the part of catalogs. As Hafter said, in the context of reporting this figure: "Users have a very high success rate at the catalog" [27, p. 217]. Yet in a review of studies evaluating reference services, Rothstein found

---

**FIG. 1.** Design model: A variety enhancement model for online subject access.
rates of satisfaction on the part of the public of around 90% [28, p. 464-465]. Reviewers might as easily have looked with dismay upon this 70% figure and noted how much below reference services catalogs are rated by their users.

How have online catalogs done? In the Council on Library Resources survey across 29 libraries (academic, public, and state/federal) 46% of users found their search “very satisfactory,” and another 34% found it “somewhat satisfactory.” In terms of specific retrieval performance, 17% found more than they were looking for, 28% found all or most of what they were looking for, 40% found some, and 16% found nothing (sum of 101 percent due to rounding error) [14, p. 151]. Summing the first three gives 85%, a definite improvement over the card catalog results.

Still, even these figures could stand some improvement. To put them differently, 56% of the searchers found only some or none of what they were looking for, and an unknown additional number found only most of what they were looking for (part of the 28% figure). When we consider that most of these libraries were very large, and that in the academic libraries 60% of the users were undergraduate students [14, p. 89] whose needs would rarely fall outside the scope of a good academic collection, then we must assume that a fair number of failures, on the part of students or catalog, led to suboptimal retrieval.

There is a crucial weakness, furthermore, in all subject catalog use studies that ask users for expressions of satisfaction. The user can decide whether the materials found were satisfactory or not, but has no way of knowing what was missed. We find the same problem with the above-mentioned studies of satisfaction with reference service. A few years after Rothstein’s survey found that average 90% satisfaction rate, several studies came out showing that the actual number of correct answers given by public library reference services was on the order of 50% [e.g., 29,30]. The users were satisfied, all right, but in many cases only because they did not know what they were missing.

My research provided a comparable result for catalog use. In an experimental setting, students were asked to state what word or phrase they would use to search in the subject catalog for a book just like the one described in the title and abstract of a real book. The degree of match was then computed between the students’ terms and the headings actually used by the University of California at Berkeley library to index the books in question. This approach tested whether people would actually hit upon specific relevant material in a search rather than just whether their term would match with any heading in the catalog, whether or not that heading indexed material relevant to their query. As with the studies on the correctness of reference questions, the results were much less positive. Matching scores for student terms on the first lookup were just over 20% of the maximum possible, even when credit was given for partial term and partial word matches [31, p. 166]. Requiring match with the index terms for a series of specific relevant books is a stringent test, but one providing a truer measure of recall success than the searcher’s own guess about whether or not good material was found.

For all these reasons, it appears that we may still find much improvement possible in library catalogs. Now, while we are in the midst of designing online systems, and both our attitudes and the systems are open and flexible, is the time to investigate ways of bringing about those improvements.

3. Design Principles

A. Indexing and Access According to “Uncertainty Principle”

In 1974, in an article entitled “The Scientific Premises of Information Science,” Rosenberg argued that “Most of the research done to date in information science has been . . . in what we can broadly call the tradition of Newtonian mechanics” [32, p. 264]. According to Zukav, “Classical [Newtonian] physics is based on the assumption that our reality, independently of us, runs its course in space and time according to strict causal laws. Not only can we observe it, unnoticed, as it unfolds, we can predict its future by applying causal laws to initial conditions” [33, p. 134]. Rosenberg again: “The method [in Newtonian physics] for discovering the mechanism or cause is to reduce all phenomena to their basic component parts, to dissect, to simplify. Also essential to the methodology is the notion of objectivity—the removal of the observer from that which is observed” [32, p. 264].

But in the modern physics of quantum mechanics and Einstein’s relativity, unsettling discoveries were made that had ramifications beyond the physical laws being discovered. In studying the subatomic world of quantum mechanics, Heisenberg demonstrated that it is impossible to measure both the position and the momentum of a particle at a given instant. This was not a conclusion based on the inadequacy of measuring instruments. Rather, the physical nature of the subatomic world is such that the very act of measuring changes the thing measured. The most that physics could do was state probabilities, but it could never state with certainty the position and momentum of a particular particle at a particular moment. This discovery of Heisenberg’s is known as the “uncertainty principle” [33, pp. 132-136].

In recent years other scientific fields, particularly the social sciences, have been recognizing that the ideal of the perfectly objective observer is simplistic and naive, and that the observer and the observed may interact in extremely complex ways that are very difficult to sort out. Further, it has been observed that human social systems may operate according to certain common patterns with certain probabilities, but may be indeterminate at the
level of specific behaviors at specific moments. Here I wish to argue that an uncertainty principle is similarly appropriate for study of some of the behavioral aspects of information science.

Our Newtonian/mechanistic assumption has been that somewhere there is an ideal indexing system or language that will enable us to produce the one perfect description or set of descriptions for each document. These ideal descriptions will, in turn, produce the best possible match with users' needs as expressed in queries. Each improvement in human or machine indexing is to take us closer to that ideal.

But suppose instead that that ideal is impossible in principle, because both indexing behavior and information searching behavior are at least in part indeterminate and probabilistic? Wilson argued persuasively a number of years ago that it is practically impossible to define what "subject of a document" means, that is, to define what it is that a person looks for in identifying the subject of a document [34, p. 69-92]. By implication, therefore, it is practically impossible to instruct indexers or catalogers how to find subjects when they examine documents. Indeed, we cataloging instructors usually deal with this essential feature of the skill being taught by saying such vague and inadequate things as "Look for the main topic of the document."

There is more to this issue than arguments in principle, however. We also have an extensive body of empirical data that suggests the impossibility of the Newtonian ideal, viz., indexer consistency studies. A great number of these studies were conducted in the 1960's, and no matter what measure of consistency was used, rates of inconsistency in indexing were found to be very high [35, 36]. Two indexers well trained in an indexing system (interindexer consistency) would frequently index a given document differently, and even the same indexer (intraindexer consistency) would use different terms at different times on the same document. For example, a typical result is found in a study by Jacoby and Slamecka [37], reported in Stevens [35]: The interindexer consistency was found to be 20% and the intraindexer consistency 50% [35, p. 159]. These results were puzzling and disturbing, but have been largely ignored in subsequent years, perhaps because our mechanistic view of information retrieval has no explanation for these results.

To demonstrate this point further, let us note the results of extensive association studies done in psychological research. The patterns of association in the human mind are extraordinarily complex and multifarious. In these studies people are given a term and asked to state the next term that comes to mind. The research shows that even the simplest terms like "river" or "whistle" evoke a variety of responses in people. In repeat studies many of the same terms will come up, with similar frequencies as in earlier studies, so probabilities can be stated for common associations. For example, among men the term "whistle" evokes the words "train," "blow," "girl," and "noise" with frequencies, respectively, of 14%, 13%, 11% and 6% [38, p. 19]. But there is no way to say which of these common terms (or the occasional "far-out" term) a particular individual will name in a specific instance. Furthermore, the effort to examine the thought processes to determine what was behind the mentioned term are so intrusive (the observer disturbing the observed processes) that we are unlikely ever to know what was involved. It may even be that at times these human thought processes are in some part random in that the potential to associate to two given terms from "whistle" may be virtually identical and only some random chemical process in the mind trips the association in one direction or another. Whatever the mechanism, we as observers must note that associations are various and many to even the simplest of stimuli.

Similar results have been found closer to home. Furnas et al. were interested in identifying the best names to use for text-editing operations so that these names could be used in the design of automated text-editing systems. So they conducted several related experiments on this question. For example, in one, 48 secretarial and high-school students were given a sample manuscript with author's corrections and asked to "prepare a typed list of instructions for someone else who was actually going to make the changes but did not have the author's marks" [39, p. 251]. Here one might expect the range of terms used to be much smaller, since each individual is asked to describe a specific operation, rather than free associate. Yet the authors note: "The most striking result from the verbal production data was the great diversity in people's descriptions... The average likelihood of any two people using the same main content word in their descriptions of the same object ranged from about .07 to .18 for the variety of stimulus domains studied [including the above mentioned experiment]" [39, p. 252].

It seems quite likely, then, that indexer consistency studies reflect the same variability in associations that have been found in these psychological and office automation studies. In other words, these are fundamental human traits, and show up wherever human beings have the opportunity to make mental associations. In one reading of a document certain features are noted, in the next reading other features are. The associations indexers will make to the words of the document differ and so the weight the indexer attaches to different topics or aspects of the document's contents will also differ. Hence the terms chosen will differ. We must live with the fact that while certain patterns may recur, we usually cannot predict indexing behavior in a specific instance.

It is appropriate at this point to recall that Bush's dream of a "memex," written up in what is generally considered to be the founding article of information science [40], contained features closer in spirit to the approach being discussed here than to that of the traditional information retrieval system. Bush's dream was of a system
that had the rich and sometimes random connections found in the human mind. He envisioned a system in which the user could set up connections between any two thoughts or entries he or she wished. He accepted without dismay that these connections might be irrational, illogical, or inconsistent. He accepted that the associations people make in their minds between one thought and another or one word and another go along on an extraordinary variety of lines. Bush's memex, in effect, mimicked the dense interconnectivity of human thought processes. (In what follows, the searcher will not be able to set up connections, but may follow up any of a huge variety of connections provided by the system.)

Now, if, in given instances, indexers are inconsistent from one to another or from one time to another, and we accept that this is characteristic of human beings, might we not also assume that search terms and phrases put to an information retrieval system will also show the same variety and unpredictability? That is, if a group of 50 searchers all had the same problem to search, would they not also show the same inconsistency in terms used as found among the indexers?

In general, it is harder to answer this question, because there have not been searcher consistency studies in the way that indexer consistency studies were performed, although recent research by Fidel on trained, experienced online intermediaries suggests that on questions of any complexity, the agreement between searchers on search terms is low [41]. Lilley and I, in separate card catalog studies, found similarly low overlap in terminology used in searching the same queries in card catalogs. Lilley asked 340 students to give subject headings that they might search on to find six books. An average of 62 different headings were suggested for each book [42, p. 42]. In my study [43], students were asked to state the search term they would use to find a book just like the one described in an abstract. The study was not designed to examine intersearcher consistency, but a scan of the responses of undergraduate and graduate students reveals the same enormous variety found by Lilley. For example, 71 students responded to the first of the books in my study; they produced 46 different headings (some varying by singular/plural only), no one of which was suggested by more than six people.

Let us assume, then, that there is a certain indeterminacy associated with information description and retrieval that is rooted in the nature of the human mind. Given a document, it is impossible (and is likely to remain so) to predict exactly what description a trained indexer will give to it. Given a topic of interest to a searcher, it is impossible to predict what specific aspect of the topic a searcher will pursue and which specific terms or phrasings of terms the searcher will use. Rather than viewing this human characteristic as an obstacle to overcome, let us instead work with this complexity and variety, and design information systems that accept and use this trait.

Having said that, then what do we do to meet this situation? An answer is to stop trying to design systems that will target the desired information through perfect pinpoint match on the one best term; rather, design systems to encompass the answer by displaying and making it easy to explore a variety of descriptive terms.

Show searchers a wide range of descriptive terms and thereby implicitly educate them on the need to produce variety. Do not worry about whether any one term is the best to search with; rather, get the searcher in the habit of using a number of terms (for recall searches), or at least exploring various terms until the most descriptive ones are found (for precision searches).

Incidentally, in this discussion the literature drawn from psychology on association patterns is dated from the 1960's because once psychologists exhausted the limits of the research model of tallying associations, they went on to try to develop models for the internal mental processes that lead to these association patterns and other manifestations of thought [44, 45]. In what follows, however, it is not being proposed that we mimic human thought within the system in an artificial intelligence manner; rather it is proposed that we make available the means in the information system to facilitate the searcher's own rich associative tendencies.

B. High Variety: Redundancy.

It is appropriate at this point to introduce Ashby's law of requisite variety, which he described in his book An Introduction to Cybernetics [46]. The law of requisite variety holds that for a system (whether machine or organism) to function successfully, it must generate as much variety in its responses to the environment as the environment generates as input to the system [46, p. 202-212] (with the exception of the cases where the same response can be used for more than one disturbance). Variety may come in the form of a physical disturbance or of information. Any time the homeostatic stability of a system is threatened by some environmental input, whether a rise in temperature detected by a thermostat or news that a relative has died (or, alternatively, received the Nobel Prize), the system must produce some response which restores stability and enables it to continue functioning normally. In the case of the thermostat, its response may be to kick on the air conditioner. In the case of the human being, the appropriate response may be to mourn (or celebrate) until the emotional stress of the news has been discharged and stability returns. If the human being responds with inadequate variety, then impairment in functioning may appear. For example, in the case of the loss of a loved one, suppose the individual denies the loss and the associated grief, and tries to go on as normal, with the usual set of responses, instead of generating a new, appropriate, response. The grief then "goes underground," may ultimately cause disruptions in other relationships, and can even lead to depression and death.
We can see countless instances in our own and others’ lives where people have responded with inadequate variety to environmental variety. For example, a worker in an industrial city may provide successfully for his family for many years—meeting a great variety of challenges, but when a major new one comes along, like the closure of the main manufacturing plant in town, may be unable to produce an adequate response—such as looking for work in another city or industry—to cope successfully with it. Then, too, there is no law that says that there even exists a possible response to certain challenges. Ashby’s law only says that to function successfully, a system’s variety must match the variety of environmental inputs, that is all.

Now let us take the law of requisite variety and apply it to information retrieval (IR) systems. With information systems the situation is the reverse from those described above: We want the variety (information) and we initiate the interaction by producing the “response” (search formulation) in order to provoke the variety (information).* We cannot get the full, desired information unless the variety of our search formulation is as great as the variety in the information. If, as I have argued in the previous section, indexers produce great variety in their indexing, then in order to cope successfully the searcher must produce an equal variety in formulating a search on any given topic.

There are two logical strategies in the design of IR systems that have been used to promote a good match in the variety of these two systems, the IR system and the human searcher. The strategies are (1) reduce the variety of the IR system (specifically, document description), and (2) increase the variety of the searcher (specifically, searcher’s query). Both of these strategies are used currently in the design of index languages. Strategy 1 is accomplished by vocabulary control. The principal purpose behind controlling vocabularies is to reduce the variety in natural language expressions that would otherwise appear in the indexing, particularly, the variety in word forms (e.g., singular/plural, verb conjugations, etc.), syntactical variations (e.g., different word orders), and synonymy.

This strategy has had considerable success in improving information retrieval performance—but it has its limits. The variety that is wreaking havoc in indexer consistency studies is not the variety mentioned above in word forms, synonyms, etc., because those are controlled in the indexing languages used by the indexers. Rather, indexer inconsistency must be due to selecting different topics or features to emphasize, or to making a different decision in a close judgment call where the document’s topic might fall within two areas. We cannot reduce that sort of variety, because that is not variety of expression, but rather variety of meaning. We do not want to merge conceptually distinct topics in an indexing language; to do so would be to impoverish the language’s ability to describe the distinct topics we want to be able to retrieve on. Operationally, such merging of distinct topics would be disastrous for precision. To put it differently, we can reduce the variety in the language used to describe information, but we cannot reduce the variety in the information itself without defeating the point of an IR system. Thus the strategy of reducing the variety in the information system has been pushed to its limit.

The second strategy, of increasing the variety of the searcher’s query, is used also in controlled vocabularies, but to a much smaller extent. The principal mechanism is the provision of cross references. See references refer from unused to used terms for a topic, and see also references refer from one topic to related topics. Note that increasing the searcher’s variety through the use of cross references does not cancel out the first strategy; the variety provided the searcher upon entry is then channelled to the more limited set of accepted terminology.

Despite the enormous variety in expression in the phrases for a topic that people bring to an information system, index languages seldom have more “see from” (unused) terms in them than used terms. In other words, the index language provides a little variety of expression for the searcher, but not much. Variety in topics as provided by see also references is generally viewed by practitioners as the least important part of an index language. Some or all of the see also reference terms may not be added to catalogs in times of staff shortage. In online systems, cross references are the last to be added to online catalogs, and only a few online databases contain the cross references used in thesauri.

It is a fundamental premise of this article that the second strategy, that of increasing the searcher’s variety, has been underutilized and contains great potential for improving information retrieval. Many of the approaches proposed in this article deal with various ways of implementing the second strategy. Online capabilities make it much easier to implement these than was formerly the case, and we are only just beginning to see the potential. Most searchers do not realize the great variety that exists in the target information. They use a reasonable term for the topic and stop there. The system must not only help the searcher generate the variety, but also first show the searcher that the variety is there.

Another closely related way to look at this matter is through the concept of redundancy. In Shannon’s information theory, redundancy can be described, loosely, as the difference between the maximum efficiency possible in coding a message and the actual efficiency [47, p. 13]. For example, in English, the phrase, “I am tired”, contains some redundancy because the word “am” is the first
person singular, and only that, for the verb to be. Thus "I" really tells us nothing here that we do not already know from the word "am"; it constitutes redundancy in the coding of the message. Of course, English grammar requires that we use the word "I"; that is simply one of the many ways grammar builds redundancy into language.

Despite its popular connotation of waste, redundancy is very valuable in communication, which is no doubt one of the reasons why redundancy is so common in language. Shannon pointed out that noise is very common in communication channels, from the static of radio broadcasts to the background chatter of a cocktail party to the typographical errors of a computer printout. Redundancy in speaking or writing provides insurance that the message gets through. For example, if either the I or the am of the above message were missing, the receiver could still guess the meaning intended.

A subject catalog is a communication channel as well. The receiver (searcher) needs information about books, which is conveyed by catalog entries. This communication channel functions in a particular way, however. The messages (catalog entries) are labeled by subject headings, and the message is not transferred until and unless there is a match between subject heading and term used by the searcher. To increase the chances of matching, there should be a number of these labels for each book.

Another way to state this variety/richness desideratum is to say that information systems should be generous. (See also Hildreth's recommendations for the user interface [48, p. 59].) An information system is the gateway into the realms of learning and knowledge. The gate currently provided by typical catalogs is tight and narrow. Most people with an ounce of intelligence and curiosity enjoy exploring through books and other carriers of information. I will even go so far as to suggest that catalogs and other information retrieval systems should be fun. Whoever talks about catalogs being fun? But why should they not be? People enjoy exploring knowledge, particularly if they can pursue mental associations in the same way they do in their minds. People of all ages enjoy computer games, some of them quite intellectually demanding. Should that not also carry over into enjoying exploring an apparatus that reflects knowledge, that suggests paths not thought of, and that shows relationships between topics that are surprising? We might turn it around and say that any indexing/access apparatus that does not stimulate, intrigue, and give pleasure in the hunt is defective.

It is appropriate now to look at the Library of Congress subject heading (LCSH) approach [49] in the context of the terminology being used in this section. In traditional cataloging practice, economy has been an overwhelming consideration. In a library with a million books in it, to add just one additional card per book to the catalog would enlarge the catalog by one million cards, a formidable addition to catalog bulk. So the pattern has been to purge redundancy in every way possible. The average number of subject headings assigned per document by the Library of Congress and large academic libraries is very low: 1.3 in a study at LC for the years 1950-64 [50, p. 189], and 1.4 in a 1979 study of OCLC records [51, p. 78]. Such minimal redundancy is accomplished by the following means.

**Precoordinate, whole document indexing.** The cataloger working according to Library of Congress principles (at LC or elsewhere) is trained to index the whole document, not parts or concepts within it. This practice leads to the assigning of a single, sometimes long, subject heading—often with attached subheadings called subdivisions—which is, in fact, a good reflection of the contents of the entire document. Closely linked to this practice is the concept of precoordination. Indexing systems designed for coordination (or combination) of descriptive terms by the searcher are called "postcoordinate," and those designed for combination by the indexer/cataloger are called "precoordinate" (see Foskett [52, pp. 433-434]). Library of Congress practice is precoordinate, so the subject concepts appear in that single long heading instead of in separate entries. Additional headings are assigned only if there is no one heading available which reflects the contents of the entire document. Thus, in a card catalog there is typically only one subject access point per document. Precoordinated, whole document indexing contrasts with concept indexing systems (as used in many abstracting and indexing services), which may use ten, fifteen, or more terms per document, each constituting a different entry and reflecting a distinct concept.

The practical value of this approach for keeping down the bulk of card catalogs is evident. Unfortunately, however, in a card catalog the content words of this long heading can be accessed only through the first word in the main heading, which is the word by which the entry is filed in the subject catalog. The searcher must use that one word or a cross reference to it or fail utterly to find books indexed there. In an online catalog with keyword matching, the searcher can get access to other words buried in the heading behind the initial word.

**Uniform Heading.** The principle of uniform heading holds that for any particular description there is to be one and only one heading reflecting that description. For example, "hypnosis" should not also be expressed by the synonymous term "hypnotism." One or the other of these terms must be chosen to be the one uniform heading for the topic, with a cross reference from the other term.

An even more interesting implication of this rule is that the same description may not be expressed in different word orders. So apart from a few exceptional cases where this sort of thing is allowed, if, say, the heading "Capitalism—United States" is approved, then by the rule of uniform heading, the cataloger is expressly forbidden from applying the heading "United States—Capitalism" to any document. It can be seen that keyword matching capabilities in an online catalog will overcome
some of the limitations of the uniform heading practices of LC. However, keyword matching will not help with the next feature of LC heading practice:

**Specific Entry.** The rule of specific entry is the single most important rule of LCSH subject cataloging and dates to Cutter’s writings [53]. (See also [49, 54-55].) It holds that each book is to be entered under a heading which is specific to the content of the book, that is, which is neither broader nor narrower in scope than the scope of the book’s contents. These headings are, in turn, alphabetically arranged, thus producing what is known as an “alphabetico-specific catalog.” The resulting catalog was to enable the searcher to find books on exactly the topic in mind, without having to worry about whether the topic was part of some broader category.

This approach contrasted with that of classified catalogs popular in the nineteenth century, which were arranged by classified hierarchies. The searcher interested in memory, for example, might have to look under “Psychology—Cognition—Memory” in a classified catalog. None of these terms would be alphabetically arranged; the entire catalog would be arranged in a classified sequence, so Psychology would be next to Social psychology, then Sociology, etc. Consequently, the searcher would usually have to use an alphabetical index to the catalog.

Cutter’s direct alphabetical approach sounds much preferable, but there were some unanticipated consequences from it (at least the way it was developed by LC). By the rule of specific entry, a book on all of psychology should be given a heading like “Psychology,” one on all of cognition the heading “Cognition,” and one on memory, the heading “Memory.” LC interpreted the rule of specific entry to mean that there could be no posting upward. That is (but for a few exceptions—see Boll [11]), the cataloger is expressly forbidden from entering the memory book under cognition or psychology. Thus the searcher who looks up “Cognition,” thinking to find a book on memory, will instead find only books that are on the whole topic of cognition. A book on just one aspect of cognition, such as memory, will be indexed only under “Memory” and not under “Cognition.”

The problem with this, of course, is that material of interest to the searcher may be found in books at all levels of specificity. Both the broader and narrower terms are, in an alphabetical file, distributed all over the file. All the different subtopics within cognition besides memory will be spread all over the catalog, as will all the subtopics within psychology besides cognition, and so on. Cross references were intended to make up for this alphabetical splitting apart of conceptually related hierarchies, but in LC practice see also references upward to broader terms are forbidden. Thus the searcher is only directed to terms at the same or more specific levels. Usually, there are not very many of even these latter references anyway.

It is the rare user who has ever heard of the rule of specific entry, let alone thought out the implications of it. My research showed that end users frequently entered the catalog under terms that were broader than the subject they were actually interested in [43, pp. 370, 371]. So it is likely that the average user expects to find books on memory under “Psychology” or “Cognition;” when no such books are found, or only books on all of cognition with some mention of memory, the searcher assumes the library has no other more relevant books and thus misses all the books wholly devoted to memory. But by limiting entries for each book to only the level of specificity found in that book, the LC system eliminates the redundancy that would be introduced by posting a book to other levels of specificity as well. This is one major area where LC’s low-redundancy practices must be compensated for by system assistance in online catalogs. The searcher interested in memory who uses “Cognition” will not be helped by current online catalog search capabilities such as keyword matching, Boolean logic, or truncation.

**Limited Syndetic Structure.** The “syndetic structure” of a catalog is its cross-reference structure, usually consisting of see and see also references. Though the study was on a limited number of cases, my research found LCSH see references frequently outdated and limited to just a few of the likely search terms used by searchers [43, p. 369]. Likewise, in a study of online catalog use, Markey found that 5% of online accesses were exact matches with LCSH cross references [10, p. 66]. Further, as noted above, see also references to broader terms are excluded from LCSH catalogs (though they appear in coded form in the LC subject heading list).

**Access Vocabulary Interfiled with Catalog Entries.** The tradition in American libraries is that the apparatus for finding good terms and moving from illegitimate to legitimate terms is built into the catalog, that is, interfiled with the catalog entries. The searcher goes directly to the catalog and looks up a term; it either matches with a subject heading actually indexing book(s) or with a cross reference—or nothing. Generally, no separate access guide is provided to the user. In such a system, then, each additional cross reference added to help the searcher find the legitimate terms adds to the bulk of the catalog. It is not surprising, then, that cross references, like the rest of LC indexing, were kept to a minimum. (This tradition might be questioned; after all, might it not be easier for the searcher to follow cross references and check term relationships in the pages of a book, rather than having to walk twenty feet to another part of the catalog for each cross reference? But to my knowledge, the provision of thesauri or other user aids specifically designed for the end user of a catalog has never received serious consideration until recently. Some libraries make the LCSH list available to patrons, but that list is designed for indexers and is hard for end users to follow.)

All of these various devices in the design of Library of Congress subject heading practice add up to an indexing system with extremely low redundancy—just over the one
heading per book that would be the lowest possible redundancy in such an environment. What happens to this low-redundancy system in an online environment and how access in LC-based catalogs might be improved will be discussed in Section 4.

C. Complexity: Access to and Interaction with System Complex and Subtle.

We have seen from the argument heretofore that the user, rather than being possessed of a simple straightforward need requiring an equally straightforward match with documents in the system, may instead be a wonderfully complicated creature with associations firing in all directions and information interests to match. To do right by this user we will have to design a system that helps the user take advantage of these powerful associative propensities.

But rich associations are not all. There are at least three other senses in which interaction with an information system, especially at the beginning of a search, may be complicated. A well-designed system should help with these needs too.

Getting in in the First Place. In both manual and online catalogs the user must launch the search with a subject term. For the search to be successful, the term must not only match with some term in the system, but must match, either directly or through cross references, with a term describing relevant material. As I have argued earlier, LCSH uses so few terms for indexing each document and provides so little assistance to the searcher that the latter is hard to do. In an online catalog with keyword matching, matching with some term is not hard to do, but then the searcher is often stalled if the term is a poor one for the information need. The cross-reference network is quite small (if it has even been built into the online catalog) and scope notes or other explanations or suggestions are sparse or nonexistent. I therefore propose the Side-of-a-Barn Principle: In a properly designed system, to get into the system and to get going searching effectively, the searcher need only hit the side of a barn, i.e., any reasonable English language word or phrase should get the searcher started and linked to explanatory, guiding information to assist in the search.

Formulating and Articulating the Query. Belkin has argued persuasively that we are demanding too much when we expect a client to approach an information system with a reasonably well-articulated query. To state their needs, people have to describe what they do not know. In effect, people do not naturally have "queries;" rather, they have what he calls an "anomalous state of knowledge" [56, p. 62].

In addition "A document, after all, is supposed to be a statement of what its author knows about a topic, and is thus assumed to be a coherent statement..." [56, p. 64]. He goes on to argue: "Thus, the document is a representation of a coherent state of knowledge, while a query or other text related to an information need will be a representation of an anomalous, or somehow inadequate or incoherent state of knowledge" [56, p. 64]. So traditional systems which require a match between systematic information (a coherent document) and a query representing ignorance and incoherence may be requiring a match between two fundamentally dissimilar sorts of text.

Earlier, Taylor had demonstrated that even with the traditional understanding of "query," users frequently go through four stages of developing and articulating their query (Q1-Q4): all the way from a visceral, felt need (Q1), through a conscious need (Q2), to an expressed, formalized need (Q3), and finally to a need compromised to meet the user's assumptions about the requirements of the information system being used (Q4) [57, p. 182]. He was dealing with people approaching reference librarians. If a client arrives when still in one of the earlier stages, librarians can adapt and assist people to articulate their need. Similarly, if the client has already overadapted the query to meet mistaken assumptions about the nature of the information system (Q4), the information specialist can help the end user work back to the true need. There is no built-in human intermediary for people when they use an online catalog, however—unless we find a way to design in that assistance.

Docking. Any time a person approaches an information retrieval system, whether that system is another human being or a machine, an initial phase of orientation or "getting a feel for" must be gone through before settling down to do searching proper. Awareness of this need has developed slowly over the years in the writing on the reference interview and negotiation. Reference librarians have long sighed with frustration over the propensity of patrons to make very general requests of librarians at the desk, rather than stating their specific need directly and immediately. Schiller argued that the reason patrons so often do this is not perversity but a real need to establish a connection with the librarian, to get a feel for the rules of the game and to begin the interaction in some common topic area [58, p. 58]. The user, after all, does not know if a question on "metatarsals" will make any sense to the librarian. If the librarian does not understand this query, embarrassment could ensue. So, according to her argument, the patron first makes a request at a broad level of generality that the librarian is sure to know: "Can you tell me where the medical books are?", then if the librarian shows willingness to help further, goes into the detailed question.

Eichman took this argument further. Using the linguist Leech's writings on functions of language use, he identified a "phatic" function that generally takes place initially between two individuals [59, pp. 217-218]. The purpose of phatic speech is to open channels of communication. Such speech occurs frequently in various contexts in human communication, and Eichman argued that it is common to the initial phase of the reference interview.
Eichman’s analysis thus provides a theoretical basis for the anecdotal argument made by Schiller.

Research on card catalogs suggests that people make the initial approach there on a general level as well. I found search terms generated by end users to be more commonly too broad than too narrow [43, p. 370]. This use of broad entry terms is a good strategy, statistically, for the searcher. Comparing the half of the subjects in my study who tended to use broad terms against the half who tended to use narrow terms, I found a substantially higher percentage of matches with some term in the catalog for the former than for the latter [43, pp. 372, 373]. In short, those who take the same approach to the inert system as they might to another human being, i.e., using broad, general terms, have a better chance of finding common ground and entering the system. In LCSH catalogs, unfortunately, the rule of specific entry guarantees that while they may match with some term, it is often not the best term for the information need because it is at the wrong level of specificity (see discussion under “variety” principle).

There are other senses in which finding common ground might be necessary for the searcher entering an information system. When a human being is the respondent, rapid adaptations can be made on both sides. The librarian who receives the halting request of the seven year old can immediately adapt in a way different from what she or he does when approached by the busy physician wanting the latest research data. Questions can be asked which immediately pinpoint vital parameters of the query so the entire transaction can be relatively brief. But even the best online systems lack most of the adaptive features of human respondents. Almost all of the adaptability therefore necessarily resides in the human searcher in a searcher-machine interaction. What the machine must do is provide the information to enable the human being to make the best adaptation to and use of the system.

Let us call this process of adaptation “docking.” Just as the skipper of a boat cannot pilot the boat at 40 miles an hour straight toward the slip, then suddenly slam on the brakes and come to a complete halt just an inch from the dock, but must instead slow down well in advance and maneuver into the slip, so also must the information system user do some initial maneuvering before hitting the ideal selection of information.

There are at least two senses of docking in the discussion above. One is the maneuvering based on an understanding of how the system works, not necessarily its inner workings, but how to interact as an end user (cf. Borgman [60]). The second sense is that based on an understanding of the conceptual and linguistic world of the subject retrieval system in use in the system.

4. Implications of Principles for Design of Subject Access in Online Catalogs

The thrust of the section on the uncertainty principle is that we cannot predict the term an indexer will use to describe a document or a searcher will use to describe a query topic. The selection will not be totally random; both will probably draw from a fairly small cluster of terms (though even that cannot be guaranteed), but we have no way of knowing which of these terms will be selected in a given case. So we must assume that even with a carefully controlled indexing language, a book can be described by many terms and both indexers and searchers are likely, collectively, to use a variety of these terms.

The individual searcher, however, is usually unaware of the many terms that might be used and certainly does not appreciate the complexity of information description and retrieval. In the section on the variety principle this point was reinforced in the discussion of the law of requisite variety: In order to succeed in information retrieval the searcher needs to generate as much variety in a search formulation as there is variety in the indexing of the topic of interest. Current systems do little to help the searcher generate the necessary variety.

In the section on the complexity principle it was argued that entering an information system, getting a feel for it, and figuring out which terms to search on are complex and subtle processes. Every system has implicit assumptions and patterns of operating which the searcher can pick up on only when exposed to the “thinking” of the system. Searchers must go through an orientation phase where they get a feel for how the system handles problems of document access, a feel for the index language, and a deeper sense of the meaning of given terms by seeing them in the context of a semantic network.

All of these points underline the difficulty of entering and matching with terms in the system and argue for much more powerful means of assisting the searcher, particularly in the early phases of the search process. In the next section some specific recommendations will be made for helping the searcher.

But, first, we need to consider how to relate the current system, i.e., Library of Congress subject headings in manual and online catalogs, to the proposed approach. It would certainly be highly desirable to make use of existing indexing, rather than reindexing, while still improving access. It turns out that there is indeed a way to utilize current indexing in the context of the recommended changes—a way which may be the best approach anyway, cost considerations aside.

Earlier it was noted that the design of Library of Congress subject headings stringently purged redundancy from the subject description of books. On the face of it then, the current Library of Congress subject indexing would seem to be a poor candidate to be a part of an improved system relying on variety and redundancy to help searchers enter the system and find desired materials. This turns out not to be so, however, for several reasons.

(1) The problems that have been emphasized in the discussion have to do with getting into the system (i.e., matching one’s initial term with some term in the cata-
log), getting oriented, and finding good terms for one’s topic of interest. All of these things are fairly independent of document indexing. It is possible, for example, to have a huge entry vocabulary—much larger than is found in the set of cross references in the current Library of Congress subject headings list—with references to used terms without increasing the size or character of the legitimate vocabulary. As we shall see, much can be done to aid access that is independent of the indexing of individual documents.

(2) Though Library of Congress indexing is done with few headings—usually only one—those headings often contain several, even numerous, content-bearing words. In a card catalog the searcher can get at the heading only through the first word, which is the basis for the sort in the linear card file. But in an online catalog with keyword matching capability, the user can search on any of the significant words contained in the heading. Such keyword matching thus opens up Library of Congress subject heading indexing to much richer retrieval possibilities. Clever use of the Library of Congress Classification and Dewey Decimal Classification numbers \([4,5]\) opens up even more possibilities for controlled vocabulary access. Additional free text access on titles and other subject information bearing elements of the record completes the picture. The typical catalog record has now become a rich source of subject information in the online environment. Of course, many of these features are already being used; the question now is how to use these features in a unified and powerful approach to subject access.

(3) Recent research by Schabas \([61]\) sheds light on the question of whether better indexing would make a difference in retrieval performance. She compared the retrieval performance of Library of Congress subject headings and PRECIS in a Canadian selective dissemination of information system. PRECIS, a system developed by Austin \([62–64]\), has many advantages in principle over LCSH. It follows more consistent principles than LCSH, is better founded on linguistic principles, and by its design it is capable of having more up-to-date vocabulary than LCSH. It was surprising, therefore, that Schabas’s results showed the two systems’ retrieval performance to be similar, with PRECIS being somewhat superior to LCSH only in the area of social science terms.

There was another striking statistic in her study as well. While the performance differences were minimal between these dramatically different indexing systems, performance was very significantly improved by adding title terms to either LCSH or PRECIS. Specifically, recall was significantly improved with very minimal decline in precision. So the addition of the simple mechanism of free text searching on title terms made a vastly greater improvement in retrieval performance than all the subtle, complex, changes in indexing techniques to be found in PRECIS over against LCSH.

The answer to these puzzling results lies in the interaction of the online retrieval system and the original indexing. The small difference between PRECIS and LCSH indexing performance can probably be explained by the fact that the very characteristics which most distinguish these two indexing systems are ignored by the retrieval system. In Schabas’s study retrieval was done by matching SDI profile terms against the PRECIS indexing, LCSH indexing, and title. Yet PRECIS and LCSH are both whole-document indexing systems. The effort in each case is to come up with a single well-formed descriptive phrase describing the whole book. The ingenuity of PRECIS lies in the well-developed grammar for creating several multiword entries, each featuring a different entry word, out of that one descriptive phrase. In a manual system, each of these created entries would be found at a different point in an alphabetical index. Since the online system matches on individual words in the descriptive phrase anyway, much of that ingenuity is wasted. On the other hand, LCSH lacks this grammar for multiple entries; in fact, as discussed earlier, the rule of uniform heading precludes the creation of alternate entries out of the same descriptive phrase in most cases. However, since the online system matches with individual words in the heading, multiple entries are created de facto by online systems anyway. Thus, the strengths of PRECIS are ignored, and the weaknesses of LCSH vis-à-vis PRECIS are largely overcome when these indexing systems are used in conjunction with typical online search system capabilities—hence the similarity of outcome in Schabas’s study.

It is for these reasons that I would argue that designing a sophisticated up-front system and wrapping its new features around an old indexing system with well-documented weaknesses (LCSH) may nonetheless be a very powerful approach to take to online catalog retrieval. PRECIS is, to my knowledge, the most powerful whole-document indexing system available today, the best candidate to replace LCSH. Yet even if the enormous effort were made to convert Library of Congress practices to PRECIS, the improvement in retrieval performance would, if Schabas’s study is any guide, be minimal. Instead, retrieval improvement must be gained in a different way—through something like the front-end system mind to be described later.

So, to summarize the implications of the arguments in the first part of this article for the design of online subject catalog access:

- There will invariably be variety in the indexing of any given topic.
- To cope successfully with an information system, the searcher must generate as much variety in search formulation as the system produces in description of a given topic.
- Searchers need help getting into and getting oriented to information systems.
- Searchers are likely to be unaware of the need for variety in searching and have difficulty generating variety even when they do become aware of the need.
A. Access

Between stages or cycle through them in a regular pattern—the options are, and exploring among the possible subject terms. The access stage most emphasizes natural language semantics, i.e., the meaning and conceptual relationships among words and index terms. These components reflect the major stages of the user's search. The dividing line between these stages is not always clear cut, and the searcher may move back and forth between stages or cycle through them in a regular pattern—perhaps without awareness that the system design contains these elements. It is argued, however, that our thinking for design purposes will be clarified if we identify and work with these distinct components. In the following discussion emphasis is placed on the access phase as the area needing the most change from current systems. Changes in access will in turn have implications for hunting and selection which cannot be developed in this article.

A. Access

The searcher has first the problem of getting into the system, getting a feel for how the system works and what the options are, and exploring among the possible subject terms. The access stage most emphasizes natural language semantics, i.e., the meaning and conceptual relationships among words and index terms.

1. Entry. The first job of the user is to get into the system in the first place. Though this is a very brief phase of the search, it is a crucial one, because if too many difficulties (sometimes any difficulties) are encountered, the user will not persist. Here we exercise the Side-of-the-Barn Principle and make it possible for the searcher to get into the system with any reasonable English language word or phrase. An "end-user thesaurus" is proposed to help the user with entry and, to a lesser extent, with the later stages of the search as well.

2. Orientation. Of all the stages this is the one most ignored in current online catalogs; hence it will be strongly emphasized here. (Such attention as has been given tends to be more in online bibliographic system research, particularly in gateway systems and the like, e.g., [65-67]).

The system should be designed to assist users to articulate queries and formulate them in good terms—including helping them generate the variety necessary to cover the many ways indexers and authors have talked about topics of interest. The system helps searchers get a feel both for how to interact with the system and for the intellectual world of the system through exploration of vocabulary and relationships between terms. Finally, the system provides searchers with links and associations between terms that may be surprising and stimulating of further thought and information seeking.

It is proposed that both entry and orientation be facilitated by a system component known as a "front-end system mind," or FSM. The FSM, to be described below, contains the end-user thesaurus (mentioned under "Entry") as well as a variety of linkages among terms and document indexing of a sort that encourages awareness of other search terms, exploration of unanticipated possibilities, the making of mental associations, and the development of a "feel" for the system.

B. Hunting

Hunting features are the best developed aspects of current systems. Many of the current features such as keyword match, implicit Boolean AND, and searching on various bibliographic fields may remain in the new model—though it will be a while before the full impact of the availability of an FSM on the hunting phase can be worked out. Some features would need to be rethought, given the availability of a powerful access component. Linguistically, the emphasis in the hunting phase is on command language syntax, i.e., most of the power of the system at this stage is in system capabilities exercised through command language syntax.

C. Selection

At this stage the searcher is assisted in selecting documents (i.e., document citations) to take away. Here the emphasis is on natural language syntax. That is, definitions of terms do not matter so much as their use in the operational context of document texts. While a searcher initially pursues a topic in general, it is usually particular aspects, attitudes, or approaches to the topic that finally interest the searcher and help him or her decide to select a particular document. This subtle information can usually only be found in the text of the document itself, so it is valuable at this stage to provide portions of the documents—contents lists, introductions, or other sections of text.

6. Some Suggested System Components

A. Access

Throughout this article a number of needs have been identified that the access apparatus should help the searcher meet: (1) getting into the system in the first place, (2) following up a wide variety of mental associations, (3) generating as much variety in search terms as exist in the indexing of a topic, (4) articulating and formulating a query, (5) "docking" with the system in both the operational and conceptual/linguistic senses. I propose to meet these needs with a front-end system mind, or FSM. The FSM is a dense semantic network. The net-
work relationships may be of almost indefinitely many types, only some of which will be suggested here. It is called front-end because it is the part of the system the searcher encounters first, and while the FSM can and will be used throughout the search, its heaviest use is expected to come at the beginning and early stages of the search. It is called a system mind because it reflects the thinking and organization imposed on the data by systems designers and catalogers. By working with the FSM the searcher is implicitly shown how to deal effectively with the main product of that thinking, the book cataloging in the subject catalog proper. This FSM facilitates entry with an end-user thesaurus and supports orientation efforts with an even more powerful cluster of capabilities to be described below.

**Entry Assisted with an End-User Thesaurus.** First, let us distinguish an end-user thesaurus (ET) from an indexer thesaurus (IT). “Thesaurus” is defined for our purposes here as a controlled vocabulary used for document description. An indexer thesaurus is a thesaurus designed for use by indexers. The product of thesaurus use, the indexing, is of course intended for the end user, but the thesaurus itself is designed primarily for indexer use and only secondarily or not at all for the searcher, whether intermediary or end user. Most thesauri in use today are indexer thesauri, although some are beginning to show some features intended to help the searcher or end user [68]. The following lists of features of indexer and end-user thesauri are not exhaustive; rather, they highlight contrasting features.

An indexer thesaurus

1. Excludes many terms that would actually be used by indexers in indexing. Examples are those terms that are to be established according to some set pattern, such as personal names as subjects, or terms that are constructed out of standard parts. Many of the headings produced by Library of Congress subject heading rules are of both of these types. For example, the so-called “floating subdivisions,” i.e., standardized subdivisions, may be added to main headings at cataloger discretion, and do not appear in the main body of the LCSH list where the user might seek them.

2. On the other hand, an IT includes some terms not used in a particular library or database, if the library or database does not yet have any documents on the topic.

3. Gives scope notes for only those terms indexers are likely to have trouble with, not end users. Scope notes also often only elucidate problem areas of the term’s usage, and do not describe all aspects of its application, a point which may be lost on the end user.

4. Often uses terms or codes (“xx,” “BT,” etc.) known only to indexers.

5. Provides cross references within the grammar of the thesaurus. That is, it gives terms which are phrased in typical index term fashion, viz., noun and noun adjective phrases predominantly, rather than the loose grammar of the end user. Also, an IT does not provide cross references from highly colloquial terms and terms that indexers would not be likely to look up. On the whole, size of cross-reference set is limited.

An end-user thesaurus (See also Piternick [68]):

1. List all terms in use in catalog or database at any given time.

2. Carefully distinguishes terms actually used in a given catalog or database from those not used in it.

3. Gives scope notes for problems likely to be encountered by end users and even provides some definitions.

4. Uses self-explanatory names for terms or relationships.

5. Provides a vast entry vocabulary, geared to end-user propensities. This fifth feature is particularly important and marks the most noticeable difference between end-user and indexer thesauri. The entry vocabulary should be several times as large as the set of legitimate terms. This entry vocabulary is then linked through see references to legitimate terms. Very colloquial terms are included, and the grammar of entry terms may be loose. Note that the looseness of the entry terms does not prevent having legitimate terms that meet high standards of coherence and good indexing grammar. In an online ET word-form variations can also be included. It might be possible to replace certain common variations with carefully designed automatic truncation routines to reduce bulk. Extensive see also references are also provided.

In manual systems the propensity of IT’s to give all legitimate terms, not just the ones used in a particular library, can be confusing for the searcher who looks in vain in the catalog for a listed word. Thus an ET for a manual catalog should clearly designate which terms have actually been used in the catalog. In online systems, however, there is more flexibility in handling this problem. For example, if a user inputs a legitimate, but as yet unused term, a match may be made against title terms.

The use of an ET should lead to a greatly increased hit rate upon first entering a term on a topic, i.e., should thus constitute the implementation of the Side-of-the-Barn Principle. Such an improvement would in itself constitute a major contribution to catalog success rates. In my research on card catalogs, I found matching scores of only about 60% of the total possible score with any term in the catalog (whether or not it was a good match with the topic of interest or at the right level of specificity), even when partial credit was given for partial matches [31, p. 166].

Further, in a study of the Syracuse University online catalog, SULIRS, Markey categorized 859 accesses made by searchers during the course of 188 online searches. Known item access points and errors constituted 27% of the accesses, various forms of exact and partial matches...
with Library of Congress subject headings and cross references 37%, and a category she called "whatever popped into the searcher's mind." 36% [10, p. 66]. The accesses in this last category consisted overwhelmingly of words containing subject content, but which did not resemble LC headings, for instance, "Painter Goya," and "Development projects in Uganda" [10, p. 70].

The second category, the exact and close LCSH matches, had a 21% "no retrievals" rate (output sets of size zero), i.e., a 79% match rate with something, while the third ("whatever") category had a 65% no retrievals rate, or only a 35% positive match rate. (The combined average positive match rate across both categories is 57%.) [10, p. 66] With a failure rate of over three times that of the second category, the third category, at 65%, is clearly ripe for improvement. Yet it is just these colloquial, non-index-language terms that are missing from the entry vocabulary of current thesauri. If the searcher could get a match with these more colloquial phrases which would then lead to a display of various controlled vocabulary terms better expressing the searcher's interests, then great improvements in searcher success rates can be expected. Furthermore, if the searcher can thus hit some term virtually every time upon first entry to the system, rather than 57% of the time, we will have nearly doubled the frequency with which the searcher has a positive experience right off with the system. The searcher will be involved in an interactive exchange with the system, rather than emotionally thrown off by a zero response, and hooked into a network of terms after the first word input nearly every time.

**Orientation Assisted through Additional Powerful Features of the FSM.** As discussed above, the FSM contains the vast entry vocabulary of an end-user thesaurus. Once the entry vocabulary has helped the searcher match with a term, the searcher is then connected to a dense semantic network. That network contains the end-user thesaurus, i.e., the entry vocabulary and the legitimate terms, plus a network of associations between terms and concepts that is much richer and more varied than that found in the typical thesaurus or authority file. The kinds of relationships that are possible are almost unlimited; only some of them will be suggested here.

Before describing those relationships, however, it is important to make clear the relationship between the FSM and the document indexing. The FSM is independent of but linked with the document indexing. All the goals of the FSM—to help people get into the system, explore, make their own mental associations, discover the many topics related to their interests and the many terms under which material might be found—can be met by the searcher without having to plunge directly into searching of documents. Once we have freed ourselves from the traditional assumption that access vocabulary has to be interfaced with the indexing as it was in a card catalog, then we can develop an enormously rich and complex FSM, and not have to load all that depth and complexity onto the indexing of individual documents. At the same time, the FSM can be different, and more than, an online subject authority file, which is a kind of indexer thesaurus.

It has been assumed throughout this discussion that there are many links among the terms in the FSM. As long as (1) every term in the FSM is linked to some document index term(s), in many cases by being linked to some other FSM term which is the one actually connected to the document indexing, and (2) every document has some term or terms which is linked to terms in the FSM, then much of the FSM can be independent of document indexing, while the user can still search documents at any time.

For example, suppose we have an FSM linked to existing LCSH book records. The many variant terms made available in the FSM can be linked to one or more appropriate LCSH headings or directly matched with title terms. At any time during the process of the search, the searcher may ask to have specific documents listed out for any particular term. A very high percentage of the terms the searcher might use would not be actual LCSH headings (though a number of them might be title terms). But LCSH headings could be linked on a one-to-many basis to colloquial and other variant terms in the FSM. Where the natural language term is ambiguous or multi-meaning the searcher could be shown the two or more LCSH terms which might stand for that term and be asked which is meant. The searcher responds with the term number and document records are brought up on the screen.

It can be seen that, in principle, an FSM can be linked to any document indexing, that is, indexing done by any thesaurus. The FSM may itself be very sophisticated, and change through time, while the original document indexing stays the same. While the FSM is conceptually distinct from the book indexing, the internal file structure of the FSM should be designed so that the searcher may ask at any time to see individual documents indexed by search terms (or ones linked to them by cross references), i.e., the searcher should not have to withdraw from one file, the FSM, and enter another, documents file, to see document indexing.

Now to the types of relationships possible: Alphabetical lists of terms, legitimate and not, and including word-form variations, may be displayed, as well as all the conventional thesaurus relationships—narrower, related, and broader terms (the Library of Congress subject headings list does not currently distinguish between these types of relationships). If the FSM is properly designed, the searcher should be able to follow chains of association indefinitely far in the FSM by following references to parents/children and related terms.

These relationships would be much more clearly understandable if displayed on the screen as trees, with the search term in boldface at the heart of the tree. Not infrequently, a given term is a part of several different trees. For example, the term "bond" may be a part of trees in...
finance, law enforcement, anthropology, chemistry, and adhesives. If all of these were to be shown at once or in short succession on a screen, the searcher may be helped in at least two ways: (1) if exploration or browsing is of interest, the searcher would be stimulated by the prospect of a number of quite different lines of search, and (2) if a more directed search is wanted, the searcher will realize that the one meaning the searcher had originally in mind may be confounded with other meanings in searching unless something is done.

But conventional thesaurus relationships are only the start. The nice thing about the FSM is that many relationships can be shown, some of them in parallel, overlapping, or, as mentioned above, in multiple hierarchies. Redundancy in access is good; it encourages exploration and orientation. Redundancy in access does not imply redundancy in indexing. Providing greatly enhanced and sometimes redundant access does not mean that we are improving recall at the expense of precision. We are insuring that the searcher finds the best terms up front, so both recall and precision should be improved.

All sorts of relationships can be developed using the Library of Congress (LCC) and Dewey Decimal Classifications (DDC). I will not review them all here; they have been well described by others [4,5,69]. For example, trees expressing the well-organized hierarchies of the DDC can be displayed, or see also references can be made up out of the relationships in the DDC hierarchies. Alternatively, the hierarchical layers above and below the entry term can be shown in outline form for those who prefer that sort of display.

Another kind of linkage perhaps most closely approximates some of the strange associations we make normally in thought. Upon entering a term, the searcher may receive a sampling of other terms, both subject and bibliographic, which co-index documents indexed under the entered term. In an online catalog, that means that when a term matches one or more words in one of the subject headings (or possibly also title) of a book, the user is then shown the other tracings (added entries), plus perhaps main entry, in the record. Some of the added entries will be closely related, as with a conceptually close subject term, or an author who is a frequent writer in the chosen subject field; other entries will be puzzling and stimulating to further thought, as with a subject heading indexing a totally different aspect of the book. Such a technique might be particularly valuable for the searcher in the humanities or history, because mental associations in these fields often move back and forth among topics, important individuals, and writers on certain topics.

A very similar co-indexing capability has already been developed in the ESA/RECON system. Called “Zoom,” the feature rank orders by frequency all descriptors (or terms in other designated fields) in all records (up to 200) in a retrieved set. The resulting output is a list of the entry term and all co-indexed terms in the document set, with the added benefit that these terms are rank ordered by frequency in the set. As Ingwersen notes, the Zoom feature may help the searcher identify preferred (high-frequency) search terms, new terminology, appropriate natural language terms, synonyms, serendipitous associations, and different spellings, among other things [70, p. 482].

Closely related to the idea of co-indexing is a proposal by Mischo that title terms of book records be linked as see references to the first topical subject heading appearing in those records. Searchers using a title term will then be referred to applicable subject headings. Mischo notes that such cross references would be particularly valuable in cases where the most current terminology has not yet been converted into a subject heading [71, pp. 11, 12].

So the searcher, upon entering a single word or phrase initially, may be shown by the FSM an extensive network of related terms in a single or successive screensful. See Figure 2 as an example of the system's response to the input term of “Hysteria.” Terms in the figure represent only some of the possibilities in the design of an FSM—ones easily accessible to this author.

To keep down the FSM vocabulary size a little, sometimes this matching may be accomplished by keyword match on significant words in the entry phrase, or sophisticated stemming algorithms on individual words. Any such word or phrase the searcher uses will in turn always be linked to some other part of the network. Selective definitions and scope notes will be available. Relationships between terms will be explained in lay language, and extensive help screens will be available.

After seeing the initial screen(s), the searcher may follow up these associations by using “action codes” in association with terms or document numbers. (Where many terms are displayed on the screen at once, they may be automatically numbered so the searcher need only input the action code number plus term number or numbers.) Some possible action codes might be the following:

**Action codes**

(1) Show me other words for the same subject [i.e., synonyms]
(2) Show me other related topics [i.e., related terms]
(3) Show me the subject classification for this topic area [i.e., trees]
(4) Show me broader topics in this subject area
(5) Show me narrower topics in this subject area
(6) Show me other terms indexing this (these) books [i.e., co-indexing]
(7) Show me the definition of this term
(8) Show me how this term is used to index books [i.e., scope notes]
(9) Show me some book titles on this subject
(10) Show me other books like this one

The tenth code may be used for any of a variety of interesting algorithms in which relatedness is measured...
and top-ranked books printed out on screen. For example, the algorithm might be that among the set of books which have a subject heading that matches a subject heading in the indicated book record, the most recent twenty titles will be printed out which have the most non-trivial words in common with the initial book, i.e., a quorum match.

Many of the relationships suggested in these action codes have already been developed and would not require the great labor of developing classification schemes and thesauri from scratch—though the work to transform these source materials into FSM use would still be extensive. I am proposing that a cluster of these techniques be brought together in the FSM so that the searcher sees a great variety of relationships.

As long as appropriate linkages are made between terms, the FSM can be updated with the latest terminology without updating the LCSH indexing (or not as frequently)—an important money-saving feature. Library of Congress Classification and Dewey Decimal Classification numbers already appear in almost every entry. Thesaurus terms for the FSM could be drawn from many sources—already existing technical thesauri from many fields, the relative index to the Dewey Decimal Classifica-
tion (DDC), or the subject terms in the schedules of the DDC or the Library of Congress Classification. As online searchers are discovering, one can use all sorts of thesauri, not just the one for the database being searched, to generate search terms for free text (as opposed to controlled vocabulary) searching. Online catalog users would have a similar freedom in using the FSM for access.

B. Hunting

Now we may consider the hunting component of the system. With the availability of an FSM, the hunting component will need to be rethought. I can explore only a few of the possibilities here. The searcher will have the option of jumping over the FSM entirely and going straight into term matching as in current systems (probably not a good idea in most cases), doing FSM exploration only (i.e., technically, no hunting at all), or, most commonly, some mix of the two. The searcher can pull up example documents the way online searchers do now to see whether they have found terms that index relevant materials.

Alternatively, the searcher may do what is in effect a number of minisearches, each trial with particular terms producing more or fewer relevant documents. Currently, with searches in online bibliographic systems, we have a tendency to think of the search as a matter of successive iterations in modifying the search formula until we find, finally, the one best formulation of the query—then we have those results printed out offline. But with an FSM, it is easy for the searcher to follow up a number of variant but closely related lines of thought. The search that started on Canadian Parliamentary politics might spread into relations between the federal and province levels, relations with language minorities, and so on. Each step of the way the searcher selects a few items and, in an ideal future catalog, has them printed out at that moment. No one of these minisearches is a better, more refined search formulation than the others; rather they represent different stages or parts of a multifaceted search.

In Figure 1 the new model is said to permit “docking” match. By that is meant that the searcher does not have to produce immediately the one (and usually only one) subject heading used by the library in a perfect, “pinpoint” match. Instead, the searcher may explore a variety of terms at various levels of specificity, look at sample document records, and, like the sailor coming up to the dock, gradually dock into the best heading(s) or title terms for the search in question.

C. Selection

Evidence from many studies over the years has shown that the average catalog user makes little use of some of the more obscure elements of standard book description such as size of book and many of the notes [72, p. 19]. At the same time, there have been recommendations from catalog studies that more subject information be added to records to improve search success [73, p. 109]. Cochran has picked up on this result to suggest that subject information be added to the catalog record from contents lists, index, introduction, etc. She has experimented to check for costs and retrieval effectiveness [74, 75]. As discussed earlier, added subject information about individual books would be most useful at the selection stage of catalog searching, though ability to search on an expanded record may be valuable earlier in the search too. In a system such as that proposed in which the old cataloging information would be retained, additional subject information could be added in documents indexed from the present time on. With growth, records with added subject information would soon constitute a substantial part of the file.

Problems and Potential with the New Model. A catalog designed according to the new model should enable searchers to get into the system easily, i.e., get help immediately for pursuing their search almost no matter what initial term they use, enable them to “dock,” i.e., get a feel for the system in several senses, help them to generate as much variety in their search formulations as exists in the desired information, enable them to explore the knowledge represented in the catalog, and, finally, provide them extensive information about candidate documents to help them decide which to select. All this would be done with the help of an end-userthesaurus, a front-end system mind linked to existing Library of Congress subject heading indexing, powerful system search capabilities, and additional information in the document records to help in selection.

Cost. A system which does not require reindexing of existing documents, coupled with an up-front system which can be altered while individual book indexing remains the same, is likely to be far cheaper than a number of other alternatives that would produce a comparable degree of improvement in the existing setup. There are many different kinds of network relationships possible, and they need not all be put up at the same time. Staged development would permit immediate benefit to be gained from each segment of development work on the FSM. Furthermore, the FSM can be essentially universal. Once developed, it can be used with fairly minor modification by all libraries using Library of Congress headings.

But Will Searchers Use it? Traditional card catalog studies found that users often did not look in more than one place, even when they found nothing under the first term. Three studies found percentages of one-place searches ranging from 66% to 77% [76-78]. If people will not do simple followup, how can we expect them to use a powerful system such as the one proposed here? First, early evidence is that people do persevere more in online catalogs: Markey found that just 26% of the searchers in the Syracuse online catalog study made one-
place (single access) searches [10, p. 66]. In a card catalog one must walk elsewhere to try another drawer, whereas in an online catalog one can make another try with just a few keystrokes.

Matthews and Lawrence contrasted online catalog systems that automatically display headings as the first result of a search with systems that display only records as the result of a search request. They found that "systems displaying headings are particularly effective in improving user satisfaction with subject-searching features and in reducing causes of user impatience with the system" [79, p. 366 (italics added)]. So browsing first for headings, though it might appear to cause delay, actually appears to be preferred by users. It appears that once it is a little easier to explore variant terms, people are willing to do it. We also know from information-seeking research, however, that people have a powerful preference for information easily acquired and will bypass information sources that are known to have good information if they are perceived as the least bit difficult to use [80,81].

What all these results suggest is that on the one hand, the user is not aware of the complexity and variety necessary in subject searching and the system should therefore show the user other possibilities through display of terms related in the many ways suggested for the FSM. On the other hand, the searcher does not want to have to do any extra work. The way to reconcile these two trends is to show the searcher FSM relationships automatically upon searcher entry of terms, and give them simple one- or two-character codes to use to explore further. If the searcher does not want to use this information or follow up in any way, the searcher need only ignore the contents of the screen and input the next term or command—no action whatever need be taken on the displayed information.

In sum, if an FSM is going to be used, it must feel very easy to use. Searchers do not have to be aware of the mechanics of searching; the system makes it completely clear and simple. Searchers should be able to concentrate on the content of the search topic and the system responses. Those responses in turn should be generous and the exploration in the system mind should feel fun. Then they will use it.

7. Summary

It has been argued that indexing is fundamentally indeterminate beyond a certain point, that searcher variety must equal indexing variety, and searching is complex and subtle. For all these reasons, the subject searcher must be assisted in ways that have conventionally not been available in catalogs. Now that online capabilities have been added to catalogs, various new forms of assistance to the searcher can be provided, vastly improving searcher success. An end-user thesaurus and a front-end system mind have been described as example system components to help the searcher.

There are many other ways of helping the searcher in the design of online catalogs but, it has been argued, whatever else they do, those methods must incorporate some means of enabling the searcher to enter the catalog more easily with a matching term, get oriented readily to the "thinking" of the system, explore terminology, ideas, and book records with pleasure, generate variety in the search formulation, hunt in powerful ways, and have more information available about each book to enable more effective selection. A system expressing these features will more fully realize the possibilities inherent in online systems, and will, above all, enable the searcher to perform truly powerful, pleasurable, and easy searches.

References


42. Lilley, Oliver L. "Evaluation of the subject catalog." American Documentation. 2(1):41-60; April 1954.


58. Schiller, Anita R. "Reference services: instruction or information?" The Library Quarterly. 35(1):52-60; January 1965.


72. Markey, Karen. Research report on analytical review of catalog use studies. Columbus, OH: Research Department, Office of Planning and Research; February 1980; OCLC/OPR/RR-80/2; ED 186041.


