LIBRARY OF CONGRESS BICENTENNIAL CONFERENCE ON BIBLIOGRAPHIC CONTROL FOR THE NEW MILLENNIUM

TASK FORCE RECOMMENDATION 2.3 RESEARCH AND DESIGN REVIEW: Improving User Access to Library Catalog and Portal Information

FINAL REPORT (Version 3)

by Marcia J. Bates, Ph.D.

Department of Information Studies
University of California, Los Angeles
Los Angeles, CA 90095-1520
mjbates@ucla.edu

June 1, 2003

Contents

1. Introduction and Background 2
2. Review of Information Seeking Literature 3
   2A. General information seeking behavior 4
   2B. Card and Online Catalog Use 6
   2C. Internet, WWW, and Library Portal Use 10
   2D. Future Information Access 12
3. Review of Research Specific to the Three Issues Addressed Here 14
   3A. User Access Vocabulary 14
   3B. Grouping/Linking Bibliographic Families 18
4. Implications and Recommendations 29
   4A. User Access Vocabulary 29
   4B. Grouping/Linking Bibliographic Families 39
   4C. Staging of Access to Resources in the Interface 41
   4D. Example Implementation of Recommended Approaches 43
   4E. Drawing the Threads Together 47
5. Summary 48
   5A. Summary of Review of Information Seeking Literature 48
   5B. Summary of Review of Research Specific to the Three Issues Addressed 49
   5C. Summary of Recommendations 50
Acknowledgements 52
Bibliography 53

1. Introduction and Background

The Library of Congress Bicentennial Conference on Bibliographic Control for the New Millennium met on November 15-17, 2000 in Washington, DC (http://www.loc.gov/catdir/bibcontrol/) to consider how bibliographic control could be improved in the context of today's extraordinary new information technology capabilities (Proceedings... , 2001). At one point, the 125 conference attendees broke up into several smaller groups to consider specific sub-areas within the general question. From the material produced by these groups, an Action Plan was developed.

Point 2 of the Action Plan states: "Enhance the access to and display of records for selected Web resources across multiple systems." One of the resulting action plan recommendations was the following:

> 2.3. Explore ways to enrich metadata records by focusing on providing additional subject and other access mechanisms (e.g., front-end user thesauri) and increasing granularity of access and display (e.g., by enabling progression through hierarchy and versions and by additional description [sic] information including summaries. (Bibliographic Control…, p. 4)

Further discussion among those of us responsible for this work item resulted in the understanding that there are three distinct sub-areas to address within it:

1. User access vocabulary
2. Links among bibliographic families
3. Staging of access to resources in the interface

The author is to prepare a review of the literature on catalog users and use, and, taking into account the state of information system design, make recommendations on how the library and cataloging communities might respond.

Here is a more detailed description of these three areas and why they are of interest:

User Access Vocabulary: An extensive body of research has documented that the range of vocabulary used by information system users is extremely wide and varied, the most popular terms seldom being used in more than 20 or 30 percent of all searches, with the total number of different terms used among a group of people found to be almost always high. Traditional cross-references seldom equal the number of search terms nor match their informality or range.

Key question*:* Is there a cost-effective way to enable users to get easily from their chosen search term(s) to the richest and most relevant contents of the catalog or database they are searching?

Links among Bibliographic Families. Links among related bibliographic items have long been provided, through a variety of means, in cataloging records. However, the

costs of cataloging and the physical limitations of the card catalog and of the file structure of early online catalogs have kept these links to a minimum. The links that are provided have been physically relatively hard to follow up by the user. The availability of online links and improved technological and metadata resources now make it possible to consider the easy application and display of a much wider range of linkages among bibliographic families.

Key question: Based on what we know of user needs and the distribution of resources, is it desirable and practical to provide an enriched and extended set of links within bibliographic families?

Staging of Access to Resources in the Interface. Generations of catalog use studies have reported that people want summary information about the contents of an item in the catalog record, as well as other means of determining an item's relevance for them, before having to go to the effort to retrieve or acquire the item. The broader question is how much information should the searcher be given about records, and when, in the process of searching--that is, how should the system's responses be staged? For example, should the searcher be given a lot of very brief entries first, then more extensive information about each record in later stages of the search, or should the searcher be given fully detailed information about each record immediately? A ground-breaking Federally-funded study in the 1960's suggested reasons for this user request for a summary, and put this desire in a larger conceptual context. That context may give us the framework for better access design.

Key question: How should the presentation and quantity of catalog or database information be staged for access as the user moves through screens in an online catalog or library portal?

In the following sections, various literatures are reviewed that relate to these questions. First, in Section 2, crucial discoveries from the general information seeking literature are presented, followed by more specific information on catalog and Internet information system use. Next, in Section 3, the literatures specific to the three questions are reviewed. Finally, in Section 4, implications are drawn and recommendations made. The entire report is summarized in Section 5, Summary.

2. Review of Information Seeking Literature

The literature on direct use of online catalogs, library portals, or other World Wide Web resources is not the only appropriate literature to review in order to understand how people interact with online resources. We need to start first with the most general understandings of people's behavior in relation to information, as this infuses everything they do with specific sources. Likewise, use of online systems in general may have some relevance, as well as research on use of card catalogs. The latter has bearing here because the questions of how people understand catalog contents, independent of channel, and

how they tend naturally to search for information may be expressed in the behavior of searchers as studied in all kinds of research projects.

Thousands of articles and books might have bearing on the questions being asked here. What follows is necessarily highly selective. I will attempt to provide a balanced presentation of what has been discovered about information seeking and information system use; however, I may miss relevant material of value and would welcome further suggestions from the readers of this report.

2A. General information seeking behavior

Principle of least effort. Probably the single most frequently discovered finding on information seeking behavior is that people use the principle of least effort in their information seeking. This may seem reasonable and obvious, but the full significance of this finding must be understood. People do not just use information that is easy to find; they even use information they know to be of poor quality and less reliable--so long as it requires little effort to find--rather than using information they know to be of high quality and reliable, though harder to find. Research on this behavior dates at least as far back as the 1960's, when a major study demonstrated that physicians tended to rely on drug company salesmen for drug information, rather than consulting the research literature. (Coleman, Katz, & Menzel, 1967). Poole reviewed dozens of these studies in 1985 (Poole, 1985); Mann has a more recent review (Mann, 1992).

In almost all cases, people used the information that they found 1) easy to use, and 2) accessible (Allen, 1979), rather than higher quality information that they perceived to be harder to use and less accessible. After they get training and experience, people generally feel that resources are easier to use and more accessible. However, people rarely get sufficient training and experience to cross that bridge, whether in school or college. Library use training in school is often fragmented and repetitive--a few hours every year, without being connected to real "bottom line" consequences. Specifically, students seldom are graded on their library skills and often are taught those skills without any connection to actual school assignments. Despite heroic efforts on the part of librarians, students seldom have sufficiently sustained exposure to and practice with library skills to reach the point where they feel real ease with and mastery of library information systems.

Unself-consciousness of information seeking. For librarians, finding information is a professional challenge, and therefore stimulating and a focus of interest. However, the general public has, for the most part, never thought of information seeking per se, never thought of information seeking as a situation in which they need to strategize or plan. For most people, the closest they get to thinking about that is when they have a problem to solve. If that problem involves a need for information, they do not separate out the information need from the rest of the needs associated with solving their problem.

It was this insight that led Brenda Dervin et al. (1976), Warner et al. (1973), and Chen & Hernon (1982) to engage in the first really modern information seeking research by asking people about their problems, rather than about their information sources and strategies. When you ask the "persons on the street" what their information needs are, they are apt to look at you blankly. (This happened to me once when I went to a thriving women's center to see what kinds of women's information needs they were meeting. Though the center staff kept a log of every request that came in, it took me about ten minutes of explaining to get across to them that they were supplying *information* needs, in addition to other needs at their center. They had just never thought about it that way.)

Though the above may be the general trend, it is important to note at least two substantial exceptions to the above general rule. For one, some people become what we might call "information hobbyists." They like discovering new things and developing new techniques for finding those new things. There is a long line of research on what were first called "opinion leaders," then later "information gatekeepers." Gatekeepers are people in an organization or social group who tend to be the most active information seekers, who always like to keep their finger on the pulse of new events in the culture or industry where they find themselves. After a while, their friends or colleagues learn to go to them for the latest on the subject matter with which they are most associated. It is in this sense that they become a gatekeeper--a channel for information from the "outside," whatever that outside is.

Second, people may become intensive and active information seekers in cases either of great urgency or great interest. The large numbers of popular medical websites and busy listservs are testament to the value people place on acquiring such information in the area of health (White, 2000). Likewise, in the area of hobbies and avocations, an important part of pursuing many of these activities consists in discovering information of various kinds, whether it is the history of the antiques one buys, the location of the best fishing spots, or the latest recipe. Part of the pleasure of the avocation is the acquisition of enriched understanding, often through active information seeking.

Importance of influential figures in information seeking. In the very different environments of the public school (Blazek, 1971) and of the working world (Mick et al., 1980), it has been observed that people are responsive to what the power figures in their lives like, when it comes to information seeking. If the boss or teacher encourages information use, then information resources will be used, where the boss or teacher does not, then the use will be much less. This is one of those points that seem obvious once said, but which is nonetheless often missed.

There is a related sense in which meaningful figures are important to people's information searching. Just as with most other things in life, people learn from and model their behavior on what they see their teachers, mentors, colleagues, and friends do. In the normal course of their lives, most people have much greater exposure to these categories of people than they do to librarians or library instruction. Consequently, their information searching often quite unself-consciously develops from their experience with these important people in their lives. Children whose parents read to them grow up to be active readers and library users (Powell et al., 1984).

Academics and researchers, especially in the humanities, model their searching techniques on what their academic advisors do (Stoan,1984; Bates, 1994a). Researcher searching behavior bears little resemblance to the formal search models the library/information science field has developed. Ellis' extensive work studying the information behavior of researchers clearly illustrates this (Ellis, 1989; Ellis & Haugan, 1997).

People take what they have learned from earlier experiences into new environments. For example., beginning college students often have difficulty using large libraries on their campus, because their prior experience was solely with small browsable school and public libraries. "Browsing in the 300's" no longer works as a search strategy when you want a book on developmental cognition for your psychology class in college. In general, information seeking behavior is most often influenced by non-librarians, and because that behavior is usually quite unself-conscious, it is difficult for librarians to influence it in the short periods of time that we are generally able to work with people.

2B. Card and Online Catalog Use

Studies of catalog use have been of interest for at least fifty years, and the total must number in the hundreds. Here I will draw strongly on reviews of this literature by Karen Drabenstott (1991), Christine Borgman (1986, 1996), and myself (1977a, 1986a), as well as from a few recent example studies that seem to me to be typical of the latest findings.

Card Catalogs. It is worthwhile to look at some of the card catalog research results, as these provide independent evidence of users' experiences with the intellectual content of catalogs, without the simplifying aid of keyword searching and other online catalog features. Thus we get a sense of how well or poorly the actual cataloging content meets users' needs.

Many of the early studies in the 1950's and 1960's were done in academic libraries. A common finding was that known-item searches exceeded subject searches--the latter constituting only 20-50 percent of the total uses of the catalog. Further, as users moved up the ladder of academic expertise from undergraduate to Ph.D., the proportion of subject searches went down and known-item searches went up (Bates, 1977a, p. 162).

We now have a better understanding of why this pattern might hold. Scholars are already very familiar with key topics in their research areas, so they tend to use the catalog to locate references they are already familiar with or have come across in their reading (Bates, 1994a, 1996a,b). Undergraduates, however, new to a field, must generally start with a subject search, as they lack knowledge of the key players in a research area.

In early card catalog studies in both academic and public libraries, it was found that while about half the use of subject catalogs was to locate desired material by subject,

nearly all of the remaining half of the uses of the subject catalog were to find a call number range and then go to the stacks to browse (Bates, 1977a, p.162). Thus a high percentage of the users were using the catalog as an index to the classification scheme--a function for which it is only crudely suited. For a modern study of this sort, but under quite different conditions in a British library, see Hancock (1987). Given her method of analysis, direct comparison would be misleading, but she did continue to find searchers sometimes moving from the catalog to the shelves to browse (p. 307).

Early academic library studies found that between two-thirds and three-quarters of all subject catalog searches were one-place searches, that is, the searcher only looked under one term and then quit--yet the data also show that searchers find what they want on the first try only about half the time (Bates, 1977a, p. 162).

This result exemplifies a finding for which there is much anecdotal evidence: When people look up a term and do not find anything under it that suits them, they assume the library does not have anything on the subject. Almost never do they assume that they need to try another term. Librarians use search terms and techniques as tools. We have learned from experience that one must often try a variety of approaches to succeed. So if one "tool" does not work, we try another. The average user, however, *identifies* their search term with their whole subject query. It does not occur to them that it might be called other things by the catalog. They look up their topic, do not find it, therefore the library must not have anything on it. The figure above, of finding what they want half the time, is almost certainly high--the result of people settling for what they *do* find and not thinking to try elsewhere to find more. In my dissertation, in which the test was whether a searcher used a term that matched with the *actual assigned subject heading,* the success rate on the first try ranged between 21 and 35 percent (Bates, 1977a, p. 166). Getting a match on *some* heading, whether or not it had been applied to a test document, was much higher, 60-64 percent (p. 166).

However, the above data are for card catalogs. Carlyle (1989) carried out a rigorous test on an online catalog in the 1980's and got better results: User terms matched single Library of Congress subject headings 47 percent of the time; partial matches would have raised the figure to 74 percent (p. 44). The methods used are somewhat different, so the comparison between Carlyle and Bates is not exact. Carlyle's method is most similar to Bates' match-with-any-heading figures above, of 60-64 percent.

When people were asked if they were satisfied with their use of card catalogs, their response rate generally came in at about 70 percent or slightly better (Frarey, p. 162; Hafter, p. 217). This has sometimes been considered a satisfactory result. However, early research on satisfaction with reference services came in at about 90 percent (Rothstein, 1964, p. 464-465), suggesting that helping users through still better catalog design could be productive.

Online Public Access Catalogs (OPACs). The single largest study ever conducted on online catalog use was sponsored by the Council on Library Resources in the early 1980's, examining the use of the first widely used online catalogs. Sixteen catalogs and 29 libraries participated, including academic, public, community college,

and government libraries (Matthews et al., 1983). Other special library types were not tested. About 8,000 online catalog users and 4,000 non-users responded to carefully pre-tested questionnaires.

Non-users used the library itself and the card catalog less than OPAC users did, non-users were also slightly older than users, and had less computer experience--all not surprising results (p. 93ff). Even non-users had positive attitudes toward the online catalog, and most expected it to be easy to learn.

The key results from the study for our purposes are to be found in two tables. In Table 17, "System Interface Problems," 25 percent or more of the respondents had problems with the following nine aspects, listed in order of frequency of difficulty below, and followed by the percentage of users having that problem:

1. Increasing the result	46 percent
2. Finding correct subject term	43
3. Knowing what is in online catalog	37
4. Computer search by subject	31
5. Scanning through a long display	28
6. Entering commands when I want	28
7. Searching with truncation	28
8. Reducing the result	27
9. Interrupting or stopping the display	25 (Matthews, p. 124)

Table 20 lists the additional system features desired by users. These are the top five:

1. View related words	45 percent
2. Search table of contents/index	42
3. Determine if book checked out	26
4. Print search results	25
5. Search by subject word	24 (Matthews, p. 134)

Many of the problems and desired changes had to do with subject searching. Among the desired changes, checking circulation status and printing search results have been implemented by most library system vendors by now. Search by subject word, interpreted as keyword searching, has also been introduced in most catalogs. The top two requests, however, remain largely undone, and indeed, two of the three possible changes reviewed in this report concern those very problems--being able to view related words (other than traditional cross-references) and searching table of contents or index.

In 1986, Borgman reviewed the research on online bibliographic retrieval systems to discover implications for online catalog design. Earlier studies of online database searching had shown similar results to those mentioned above for card catalogs: "Searchers often miss obvious synonyms or fail to pursue strategies likely to be productive.... " (Borgman, 1986, p. 389). Further, Borgman noted that Fenichel had "found that in half the searches studied, the initial strategy was not modified; searchers

(even experienced ones) tended to use only the most basic techniques of selecting and combining terms" (Borgman, p. 389; compare Fenichel, 1981).

In the Council on Library Resources study, 53 percent of the searchers were trying to find information on a subject (Matthews et al., p. 91). In a 1983 article, Pauline Cochrane drew attention to the growing importance of subject searching that accompanied the advent of online catalogs (Cochrane, 1983).

After some years of use of online catalogs had passed, Larson (1991) discovered the following in a transaction log analysis of the University of California's MELVYL catalog from February 1982 through January 1988:

> The preceding analysis shows a persistent decline in the use of the subject index on the MELVYL system, with the rate of decline at about 2.2% per year over a six-year period. Title keyword access appears to be adopted as a replacement for subject index access by users. (p. 207)

So the interest in subject access had not declined, but increasingly, that need was met by keywords. More recently, Hildreth studied OPAC use in a university and concluded: "users of this online catalog search more often by keyword than any other type of search, their keyword searches fail more often than not, and a majority of users do not understand how the system processes their keyword searches" (Hildreth, 1997, p. 52). Subsequent research on end-user online systems, to be discussed shortly, also demonstrates the persistent popularity of short, simple queries.

Borgman also noted evidence from a variety of sources that searchers have persistent difficulty with Boolean logic (1986, p. 392). The Getty Online Searching Project found these problems particularly severe for humanities researchers (Bates et al., 1993; Siegfried et al., 1993). Connaway et al. (1995) found that searchers still seldom took advantage of the capability to do Boolean searching.

In 1991, after online catalogs had improved considerably in design, and after a number of additional studies on catalog use had been done, Drabenstott published an excellent review of the state of knowledge at that time regarding use of online catalogs. Following is a summary listing of her key findings (Drabenstott, 1991, p. 67-74). (All bullets are direct quotations, drawn from the headers of her review.)

- Users like online catalogs. (p. 67)

- A lot of subject searching is being performed by online catalog users. (p. 68)

- Subject searching in online catalogs using the Library of Congress Subject Headings is difficult. (p. 68)

- Users want subject searching improved in online catalogs. (p. 68)

• Most users do not know that the catalog has a controlled vocabulary, and, as a consequence, enter queries that express subjects that come to mind. (p. 68)

• The highest percentage of access points producing zero retrievals are subject access points. (p.69)

• About half of user queries for topics or for geographical names match the catalog's controlled vocabulary but they produce excessively high retrievals. (p.69)

• The utility of alphabetical lists of subject headings that systems produce in response to subject queries is not recognized by some online catalog users. (p. 69)

• A lot of queries for known-items and personal names would have retrieved cataloging records had they been entered correctly. (p. 73)

• Users want the online catalog to provide them with access to much more than the library's book collection. (p. 74)

2C. Internet, WWW, and Library Portal Use

As Drabenstott noted in her 1991 review, at least as early as the late 1980's library users were clamoring to have access to other types of information besides traditional catalog data in their OPAC (p. 74). Especially since the World Wide Web emerged as a powerful force in the mid-1990's, libraries have transitioned to providing Web access to their catalogs, as well as enlarging their websites to become true portals to library-based information. (For working purposes, a portal is defined here as a website, generally produced by an institution or organization, intended to provide access to a variety of types and/or sources of information, built around a coherent purpose.)

As is usually the case, however, evaluation of, and studies of the use of, these new resources have lagged well behind their development. In a 2000 descriptive review of a variety of Web OPAC Interfaces, Babu & O'Brien (2000) found just one evaluation of Web OPAC features, by Lombardo & Condic (2000). The latter evaluated student reactions to a new online catalog at a medium-sized university. The catalog was available in three forms, Web, Windows, and Telnet.

Drawing on the latter study, Babu & O'Brien state: "...users rate most highly the ability to access it remotely, select, mark and download results from their searches and integrate these references into their own personal documentation. Advanced features such as hyperlinks, limiting and more flexible keyword searching are also valued but to a lesser extent" (Babu & O'Brien, p. 325). Lombardo & Condic also note: "Respondents in the first survey commented on their confusion in generating appropriate Library of

Congress subject headings and keywords" (p. 139). Further: "...although most students found [the catalog] easy to use, many of them were unable to take advantage of more sophisticated searching techniques because these features are not intuitive" (p. 139). It would seem that the more things change, the more they stay the same.

(As one commentator on an earlier draft of this report noted, and I concur, we should not assume that all search problems can be solved by good system and interface design. There is a body of more general analytical and search skills needed by the searcher, in order to have optimal success searching for information in the complex documentary structures available today.)

Just as there is a dearth of Web OPAC use studies, so also there have been few studies of searching on the Internet or World Wide Web, according to Jansen & Pooch in a 2001 article (2001). They concentrated on three major studies of Web searching. These were all transaction log studies done on large search engines, so they were context-free; only the actual search terms were available. There was no information about the full query the user brought to the session and why. (However, one of these studies had the largest sample size this writer has ever heard of--just under a billion queries!)

Generally, the search queries were short, almost always one or two words. Less than 10 percent of the queries used Boolean operators, and searchers generally viewed ten documents or fewer (p, 241).

More recently, Cothey studied the Web searching behavior of 206 college students over a ten-month period (Cothey, 2002). Interestingly, "the users adopted a more passive or browsing approach to Web information searching and became more eclectic in their selection of Web hosts as they gained experience" (p. 67).

Finally, one more recent study compared user success with three types of Web-based searching: query-based (Google), directory-based (Yahoo), and phrase based query reformulation-assisted search (via the Hyperindex browser) (Dennis et al., 2002). "Results indicated directory-based search does not offer increased relevance over the query-based search (with or without query formulation assistance), and also takes longer. Query reformulation does significantly improve the relevance of the documents through which the user must trawl, particularly when the formation of query terms is more difficult (p. 120).

Taking a different tack, we next review Martha Yee's draft guidelines, which she developed for a Task Force on Guidelines for OPAC Displays for the International Federation of Library Associations and Institutions (IFLA) (Yee, 1999). (Revised guidelines are expected to be presented by the Task Force shortly.) Some of Yee's draft guidelines relate to the issues being discussed in this review. Specifically:

• regarding subject indexing and display:

Principle 9: Integrate cross references in displays.

Principle 23: Display the hierarchical relationship between headings and their subject subdivisions.

Principle 26: Display the hierarchical relationship between a classification number and the entire classification.

• regarding bibliographic families:

Principle 19: Display works about an author or corporate body with the works of the author or corporate body.

Principle 20: Display works about a work, or related to a particular work with the work.

Principle 21: Display works about a particular genre or form with examples of the genre or form.

Principle 22: Create clear displays of serial works that have changed title.

Principle 24: Display the hierarchical relationship between a corporate body and its subordinate bodies.

Principle 25: Display the hierarchical relationship between a work and its parts.

• regarding staging of access:

Principle 3: Effective and efficient displays of large retrievals should be available.

Principle 14: provide compact summary displays.

2D. Future Information Access

The discussion to this point has concentrated on existing systems and means for their improvement. However, several people have argued that we now are in a position to "think outside the box," that information technology has advanced to the point that we now need to think in a more creative and visionary mode about the future of information systems.

In 1989, Layne stated:

> I would like to suggest that it might be more useful…if we take the view that we as catalogers provide access to the *catalog*, that we can limit our thinking dangerously by concentrating on providing access to particular *records.*
> ….

> Access points should be points at which the user gains access to the *catalog,* not merely access to a particular *record* in that catalog. (Layne, 1989, p. 189)

Borgman makes the following argument:

> Online catalogs should be judged by their success in answering questions rather than by their success in matching queries. In the long term, we need to design systems that are based on behavioral models of how people ask questions. Such a design model could assist in the question-negotiating process, allowing the searcher to pursue multiple avenues of inquiry by entering fragments of the question, exploring vocabulary structures, capturing partial results, reformulating the search with the assistance of various specialized intelligent agents, retaining elements of a search for future sessions, and even transferring elements to other systems. (1996, p. 500)

A number of approaches intended to better reflect real-world searching had indeed been proposed earlier, yet, as Borgman notes, "Very little of this body of research has informed the design models of the commercial online catalogs that are in general use around the world" (1996, p. 501).

In 1989, I suggested the idea of "berrypicking" as a truer model for how people actually search than the conventional model of the single unvarying query matching with index terms in a single database, as assumed by much research (Bates, 1989a). Berrypicking differs from the conventional model for searching in that the searcher picks up bits of information here and there, just as one plucks berries from various bushes in real-world berrypicking. Further, with each bit of information, the searcher 1) modifies and adapts a query based on the information gathered to date, 2) uses a variety of search techniques, rather than the single approach of subject searching, and 3) searches in different resource domains (Bates, 1989a, p. 409).

The paper further proposed that information systems should support users not only in subject searching, but also in other characteristic information searching behaviors, such as footnote chasing, citation searching, area scanning, reviewing journal contents lists, and author searching. Design features to achieve each of these things were proposed for future information systems.

In another paper, I argued that users want to maintain control over their searching, but be supported by the information system in making the kinds of moves that are natural for the searcher. For example, a searcher should be able to input a command that says "broaden query" rather than have to know the sequence of specific moves that make it possible to reformulate the query in the desired way. In other words, the interface should present capabilities that mesh with the searching process that the searcher is thinking about, rather than the searcher having to adapt behavior to the design of the information system itself (Bates, 1990a).

Charles Hildreth has long advocated for an "enhanced, expanded, and extended" catalog (1995). Discussing an unpublished paper by Kevin Cox, Hildreth (1995) states:

"According to Cox, the user best searches through browsing, recognition and discovery, rather than by a formal process of explicit query formulation, entry and modification"(p. 66).

Hildreth concludes:

> To break out of the query-oriented, Boolean mind-set, we need to turn the conventional query-first-then-browse paradigm upside down. Searching by exploration, recognition, and discovery in a well-structured bibliographic space should be the primary search interface provided to information seekers, augmented by secondary query expansion methods and a choice of similarity operations. (p. 72)

Finally, I have recently taken this browsing-oriented position a step further, arguing that browsing may in fact be the dominant and most natural form of searching, and that systems that make information discovery *feel* like browsing, whatever their actual structure, will attract more users and help those users to be more effective information seekers (Bates, 2002c).

3. Review of Research Specific to the Three Issues Addressed Here

3A. User Access Vocabulary

How people really use vocabulary in searching. As noted throughout the previous sections' discussion of catalog user research, subject searching is a persistently problematic area. Match rates with search terms vary across studies, but few exact match rates top 50 percent, and many are lower. Zero match cases are high (Markey, 1988). Title searching is popular, almost certainly because it is easier to get *some* match (Schabas, 1982), but we know that uncontrolled vocabulary fails to group related materials together and much valuable material may be missed. Users seldom alter their initial search terms, despite the fact that the search terms frequently either fail to match at all or match with terms that do not, in fact, index the material of interest to the searcher.

I have long been advocating that matching and lead-in terminology be made available for information searchers to help them in their search process (Bates, 1986a). Such an end-user thesaurus would recognize the many variants, informal terms and other terms that users actually input when searching. The thesaurus would be designed to link directly with whatever database the searcher wanted to use, so that the searcher could be led to the "legitimate" indexing terms. By dramatically increasing these often more informal lead-in terms, the searcher should have a higher hit rate with initial search terms and should far more reliably be led to useful material to meet their needs. Instead of frequent zero-hit situations and frequent cases of marginal relevance, they might have a better chance of homing in much more directly on the core of the information sources

relevant to their interests. I have elsewhere called this the "Side of the Barn Principle." That is, the searcher should only need "hit the side of the barn" in an initial query input, i.e., start with a reasonable term or phrase, even if not the best, in order to be launched into the materials in the database (1986a). The searcher should not need to hit a knothole in the side of the barn!

There are several arguments for providing this front-end information:

1. Quoting from a recent paper:

> *In study after study, across a wide range of environments, it has been found that for any target topic, people will use a very wide range of different terms, and no one of those terms will occur very frequently.* These variants can be morphological (forest, forests), syntactic (forest management, management of forests) and semantic (forest, woods). (Bates, 1998, p. 1188)

And from another, earlier paper:

> ...the average likelihood that any two people will use the same term for a concept or a book, or that a searcher and an information system will use the same term for a concept, is in the range of 10 to 20 percent. The total number of terms generated by a group of people for a given topic is almost always very large. (Bates, 1989b, p. 409)

Here are some example research studies supporting the above statements:

• 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 (Lilley, 1954). The most frequent term suggested for each book by Lilley's students averaged 29 percent of total mentions across the six books (my calculation). (Most of Lilley's examples were simple, the easiest being *The Complete Dog Book,* for which the correct heading was "Dogs.") (Bates, 1989b, p. 408)

• 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. They did several studies, which produced similar results. They concluded: "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." (Furnas et al., 1982, p. 252)

• I had done my dissertation on the matching rates between assigned library catalog subject headings and subject search terms that students would use to find a book just like a real book described in an abstract. Upon seeing the above data, I returned to the dissertation and did a calculation on the first of the abstracts given to the students. I found that 71 students responded to that abstract;

> they produced 46 different headings (some varying by singular/plural only), no one of which was suggested by more than six people. (Bates, 1977a,b)
>
> • Saracevic & Kantor studied 40 real queries submitted by real users for online database searching by skilled intermediaries. Five intermediaries searched each of the 40 questions. The authors compared each pair of searchers' queries to determine how much overlap there was among these skilled searchers in approaching the same queries. In total, there were 800 pairwise comparisons. In fully 94 percent of the comparisons the overlap in terms used was 60 percent or less. In only 1.5 percent of the cases were the formulations identical. (Saracevic & Kantor, 1988, p. 203-4)

See also Thomas Mann (1997) for a still more extensive discussion on this matching problem.

2. The traditional cross-reference structure has far too few access terms to meet the above need. Yet providing large numbers of additional access terms within the Library of Congress Subject Headings would clutter the listing and add extra labor for catalogers.

3. It is a truism in the field of psychology that people can *recognize* information far easier than they can *recall* it. The typical library catalog functions as a black box for the searcher. That is, the searcher has to produce a search phrase with no direct help from the system. The phrase is entered, then the delphic system responds with a match or a failure, seldom with any guidance on what to search for instead.

4. It can be surprisingly difficult to come up with an alternative term if one's first try fails. Once we have produced a name for something we have in mind, a kind of cognitive interference sets in; it is hard to *re-name* the thing.

5. Context helps immensely in understanding index terminology. Valid, closely related headings placed in the middle of a wide range of terms remind the searcher of the many other meanings that a term or its neighbors may have. When provided a structured layout, those terms can also show relationships among related terms that help the searcher clarify just what sense they have in mind in their search.

Finally, there is considerable anecdotal evidence of the need to be exposed to multiple search terms. In the heyday of online database searching, experienced searchers soon learned to OR together multiple terms for high-recall searches, some from the thesaurus of the database being searched, to be sure, but they also added many other terms from other thesauri, as well as from the general vocabulary.

This was such a common phenomenon that Sara Knapp, an experienced searcher, developed a database called TERM for the old BRS search vendor, that searchers could access to find large numbers of search terms. Eventually, she published her thesaurus, which has now come out in a second edition (Knapp, 2000). Note that Knapp's searcher

thesaurus is designed significantly differently from conventional indexer thesauri. Knapp's thesaurus will be discussed again in a later section.

Attempted solutions to the multi-search-term problem. In a valuable recent review article, Shiri et al.(2002a,b) review dozens of articles dealing with information searching and various attempts to build what they call "thesaurus-enhanced search interfaces." Two key points come out of their review:

1. Almost all effort to enhance interfaces with thesauri are working to include conventional *indexer thesauri,* not *searcher thesauri.* Thus, while the thesauri may show searchers other possible index terms, they do not contain the large number of end-user access terms to be recommended here.

2. Shiri et al. (2002a,b) found a number of both experimental and commercial efforts to make thesauri available to searchers. However, out of all systems they reviewed, they found that very few actually evaluated the interfaces "in terms of the ways in which they support query formulation and expansion" (2002b, p. 120). Further, most of the evaluations they did find, such as the ones done on the Okapi system in Great Britain (Beaulieu, 1997), were done with information system designs and/or bodies of text so unlike that of typical online catalogs, that results cannot reasonably be generalized to most OPACs.

A large number of attempts to supplement searcher vocabulary are based on automatic indexing and automatic query expansion (Efthimiadis, 1996), which will not be reviewed here. However, one experimental effort pays closer attention to how the user can be supported to do his or her own searching. Brajnik et al. (2002) have developed a system they call FIRE, which provides hints and advice to searchers at various points of the search. Within the system is a "Terminological Aid Module" that contains three thesauri, the INSPEC Thesaurus (science and engineering), as well as two specially made thesauri, one drawing on the hierarchical relationships of the INSPEC Thesaurus and the other based on co-occurrence data.

At the time of publication, they had not yet evaluated the system in any rigorous manner. Based on early, more informal evaluations, they state:

> The participants judged positively the quality of interaction with the system. In particular, they appreciated the wide variety of search activities proposed, their proposal without explicit help requests and without interrupting users [sic] activity, and the control of the interaction kept by them. (p. 355)

In the late 1980's, as a consulting sub-contractor for companies designing a modern online records management system for the Los Angeles Department of Water and Power, I designed and oversaw the construction of a "cluster vocabulary" and a thesaurus interface for the several indexing vocabularies used by various divisions of the DWP. The system was to accommodate both searching and indexing.

Searchers would be able to enter whatever search term they wished. The system would match that term with the vocabulary clusters, which contained both indexing terms and related uncontrolled terms. If the searcher's term matched with any term in a cluster, then the whole cluster would be brought up on screen. In this way, whatever term the searcher used, it would be likely to match with some cluster or another. One or more clusters containing their term would then appear on the screen. The searcher would then have the option of having the system search on all of the terms in the cluster (implicit OR), or check off terms of interest, which the system would then search with an implicit OR. In practice, this approach generally did not produce an excess of records, because searchers usually also input other specifying features, such as document type, e.g., "memo," sender or receiver, or date range. The system design for this project was described in Bates (1990b).

In the early 1990's, California had an economic depression all its own. A new Los Angeles mayor decided to cut the LADWP's budget drastically. Ultimately, the entire online records information system (cost in the tens of millions), of which the thesaurus module was a part, was scrapped by remaining staff--and any opportunity to evaluate the thesaurus and online search system was gone.

Somewhat later, the Getty Information Institute developed a front-end vocabulary system rather similar in intent to the above system, called *a.k.a.* (Busch, 1998). *a.k.a.* linked three of the Getty vocabularies, the *Art & Architecture Thesaurus,* the *Union List of Artist Names,* and the *Thesaurus of Geographic Names* (added last) with several bibliographic databases, such as the Getty's own *Bibliography of the History of Art.* The searcher could use the vocabularies to identify useful terms to search with, could have the system search the vocabularies for them, or could bypass the vocabularies. I was commissioned by the Getty to evaluate this system in 1997.

Serious design problems were uncovered that literally thwarted the chief purpose of the *a.k.a.* system, of providing improved vocabulary access (detailed in Bates, 2002a). Some improvements were subsequently made; then new leadership at the Getty Trust decided to abolish the entire Getty Information Institute, along with many of its programs, including the development work on *a.k.a.*. The so-called "creative destruction" of the modern American system proved to be mostly destructive in these cases!

In sum, it would appear that searcher vocabulary systems that support the users in conducting their searches have not yet been tested in a substantial way.

3B. Grouping/Linking Bibliographic Families

Intellectual Issues. In nineteenth and early twentieth century catalogs, there were no cheap and easy-to-follow electronic linkages; creation of the links that were possible was costly; and cross-references had to be followed up physically by the catalog user. Consequently, catalog design was relatively atomistic; the emphasis was on individual records more than groups of records.

Now we have the technical capability to make links easily and to enable the user to follow those links easily. Now various thinkers in the field are seeing the possibility of more fully realizing Charles Cutter's second objective of enabling the user to see what the library has by an author, on a subject, and in a given kind of literature. Further, in line with the popularity of browsing that was discussed in a prior section, the capability for the user to browse along linkages may now be a particularly important capability to provide for catalog and portal users.

The idea of a hypertextual catalog has roots going back to the nineteenth century concept of the syndetic catalog. The idea of a new-style catalog in the modern sense of hypertext goes back at least as far as the 1985 proposal for a "HYPERCAT" by the visionary Swedish researcher, Roland Hjerppe (1985; see also Bertha, 1993). However, the technology for such a catalog has been practical only in the last few years.

In the meantime, however, the cataloging world has turned to consider the implications of a richer linkage structure for the intellectual relationships within the catalog. At the heart of all information organization lies the question of what shall be grouped and what shall be separated. Some things are grouped but also linked to something else that is separated. We see this in a typical subject catalog, for example. Topics on a subject are grouped under a subject heading. At some point in the creation of the subject heading list, a decision is made on whether to group two closely related concepts under a single heading or to keep them as separate headings. Within the heading list and in resulting catalogs, the user is directed by see references from potentially separate terms to the heading that includes that see-from term, and is directed from one grouping of terms to another, separate, group by see also references.

A similar process holds true in descriptive catalog. When does a text become different enough to be considered a different work and not just a variant of an existing one? Smiraglia has produced a book-length consideration of the nature of "the work" (2001). Much of this is about defining in or out what constitutes a bibliographic individual.

To clarify these relationships, several researchers have proposed ways of conceptualizing relatedness in descriptive catalog. Barbara Tillett brought new rigor to this question when she identified seven types of bibliographic relationship between records (Tillett, 1991b): equivalence, derivative, descriptive, whole-part, accompanying, sequential, and shared characteristics relationships (p. 156).

Smiraglia broke out the derivative relationships into seven: simultaneous derivations, successive derivations, translations, amplifications, extractions, adaptations, and performances (2001, p. 42). See also O'Neill & Vizine-Goetz (1987) and Yee (1994) for other such hierarchies.

Within this rich intellectual development of the discussion of bibliographic relationships, it is helpful to keep in mind the UNIMARC distinction among horizontal, vertical, and chronological relationships. Vertical relationships are hierarchical, as in the

relationship of a serial to its subseries. Horizontal relationships express links between versions of an item in different formats, media, etc., and chronological relationship express changes through time in a record (Tillett, 1991b, p. 153; UNIMARC…, 1980).

The most dramatic departure from traditional cataloging approaches can be seen in the Functional Requirements for Bibliographic Records (FRBR) (IFLA Study Group…, 1998), which presents an entity-relationships approach to the description of resources. The objectives of description are defined as enabling the user to *find, identify, select, obtain* (p. 7-8), then goes on to describe the record elements that will make these types of access possible. The clarity and simplicity of these objectives and of the means to reach them represent a major step forward in the analytical rigor of cataloging. Finally, Lagoze (2000) proposes a different underlying database structure and approach to digital document description that gives far more importance to the sequencing of versions and variations through time of records.

Now we turn to consider what the use of these complex relationships is like for the user. Allyson Carlyle has made it her objective to find ways to simplify searching for users in the online interface. She has suggested a number of ways existing information can be used to help the user, without additional cataloging. Drawing on the forms of grouping created by filing rules and by the sorts of bibliographic relationships identified by others, she states:

> This new scheme, the organized display scheme, combines the strengths of both of the earlier schemes to give users a precise indication of the nature of items retrieved and the relationships among them by taking into account both the types of relationship present among items as well as the distance of an item from the original. It also acknowledges the presence of peripheral and unlinked items retrieved in a keyword environment. (Carlyle, 1997, p. 96)

She provides schemes for both work grouping and author grouping. Here is her work grouping:

Editions:

- Books
- Recordings
- Large print, Braille, ...
- Illustrated editions, editions with commentary, ...
- *Work name* published with other works

- Revisions, updated editions, ...
- Translations

Adaptations & Related Works:

- Abridgments, simplified versions, summaries...
- Sequels, supplements, ...
- Videos, motion pictures
- Musical versions

- Pictures and other graphic versions
- Computer versions, CD-ROMs, ...
- Indexes, concordances, ...
- Miscellaneous

Works about *Work name*

Items probably related to *Work name*

Items that may or may not be related to *Work name*

Other works by *Author name* (Carlyle, 1997, p. 96)

Moving down the list, one goes from the most closely related to the most distantly related records--what Tillett has recently called the "content continuum" (Tillett, 2001).

It is important for the user to understand these relationships. Thus the records should not only *be* in the above order, but also *labeled* in the above order. For large collections of related records, the above listing should also precede the actual records like a contents list, so that the user can see immediately what types of references to target.

Statistical Underpinnings. The researchers discussed above are devoting a lot of time to considering how to handle records with large clusters of related records around them. How common are these groups anyway? Does it really matter to consider them? In four large studies, Smiraglia (1992), Smiraglia & Leazer (1999), Vellucci (1997), and Smiraglia (1999) studied the incidence of bibliographic families in a variety of environments: an academic library catalog, OCLC's WorldCat union catalog, a music library, and theological libraries, respectively. Smiraglia (2001) summarizes the results:

> Obviously, bibliographic families are prevalent in large numbers in the bibliographic universe, demonstrating clearly the tendency of works to mutate over time. Half of the works in an academic research library, a third of the works in a bibliographic utility [OCLC), between one-half and two-thirds of theological works, and four-fifths of the works in a music library were members of bibliographic families. (2001, pp. 87-8)

Overall, he found that the mean size of bibliographic families in the several test catalogs ranged between 3.5 to 8.4. Note: these are means across several samples, not the range of size of individual bibliographic families. The full range of size of bibliographic families across the several samples was 2 to 322 members.

The OCLC WorldCat sample had the largest proportion of singleton entries; only 30.2 percent of progenitor works possessed derivative works (Smiraglia, 2001., p. 87). Smiraglia hypothesized that this result may be due to the fact that the utility includes works of all types from scattered libraries around the world, while the academic libraries may concentrate on works that form the core canon of research and scholarship in our society. Because so much scholarly attention has been devoted to these works, many

derivative works have appeared as well. (More recently, Hickey et al. (2002) have experimented with automatic algorithms to identify what they call work-sets, groups of records above works, which reflect all the various formats--book, film, etc.--that a work may appear in.)

These statistics are reviewed at some length, because they are important for our understanding of this phenomenon and for the development of solutions. What is almost certainly operating underneath these patterns is what is known as a "power law." Though he ignores the entire bibliometric literature, Barabási nonetheless provides a helpful layperson's explanation of these laws (2002, p. 65-78). The power law variant of importance in our field is Bradford's Law (Bradford, 1948; Brookes, 1977). This law tells us that wherever there are numbers of information phenomena, whether books, catalog records, databases, or whatever, they will be distributed in certain characteristic ways.

Power laws contrast with the common "bell" curve (also known as Gaussian or normal curve), in which the center of the distribution holds the largest number of individuals. For example, if we make a record of the heights of all the individuals in a population, we will find that most of the individuals bunch around a mean at the center of the bulge, and the remaining individuals are to be found at the tails of a relatively narrow distribution. So, typically, adults may range between four feet and seven feet, with the vast majority of people being between five and six feet tall. With normally distributed populations, one does not find one-foot tall adults or 100-foot tall adults.

With power law distributions, however, the center of the distribution is relatively small, and the ends stretch out very far. Here in Figure 1 is a typical Bradford-type curve:

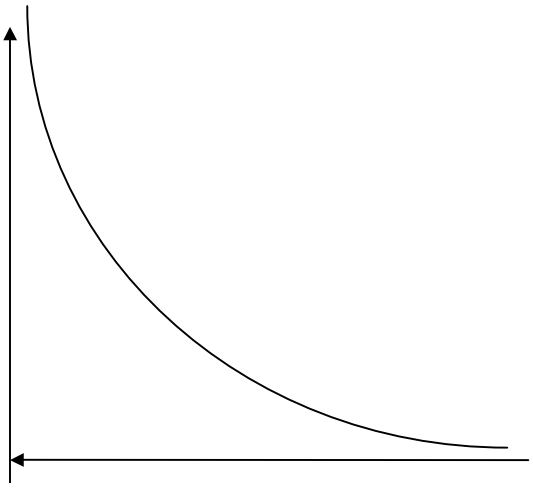

Figure 1: Rough approximation of power law curve
The actual ends of the distribution, if shown in their entirety, might continue well off the page.

With Bradford's Law, there are a few things that are very frequent, and many things that are very rare. Examples: some search terms are used very often, and most search terms are used rarely; some books circulate constantly, while many do not circulate for decades; some World Wide Web sites are visited, or linked to, very intensively, most are rarely visited or linked to; some catalog records have dozens of related records, most have few or none. I do not have the mathematics to prove that the bibliographic family data fits a power law, but the figures provided by Smiraglia look very much like one.

Each individual, whether a cataloger or a website visitor, makes his or her own decisions about what to cluster or where to go, but, when viewed from a population-wide perspective, these power law patterns appear over and over again in social and information-related phenomena.

If we could draw a two-dimensional map of the clustering of bibliographic records, it would probably look very much that of Figure 2. The figure is drawn from citation mapping research (Small & Garfield, 1985). It demonstrates how records cluster when the form of linkage studied is citations, rather than the bibliographic relationships being discussed here. In both cases, however, the bunchings and groupings almost certainly follow Bradford's Law.

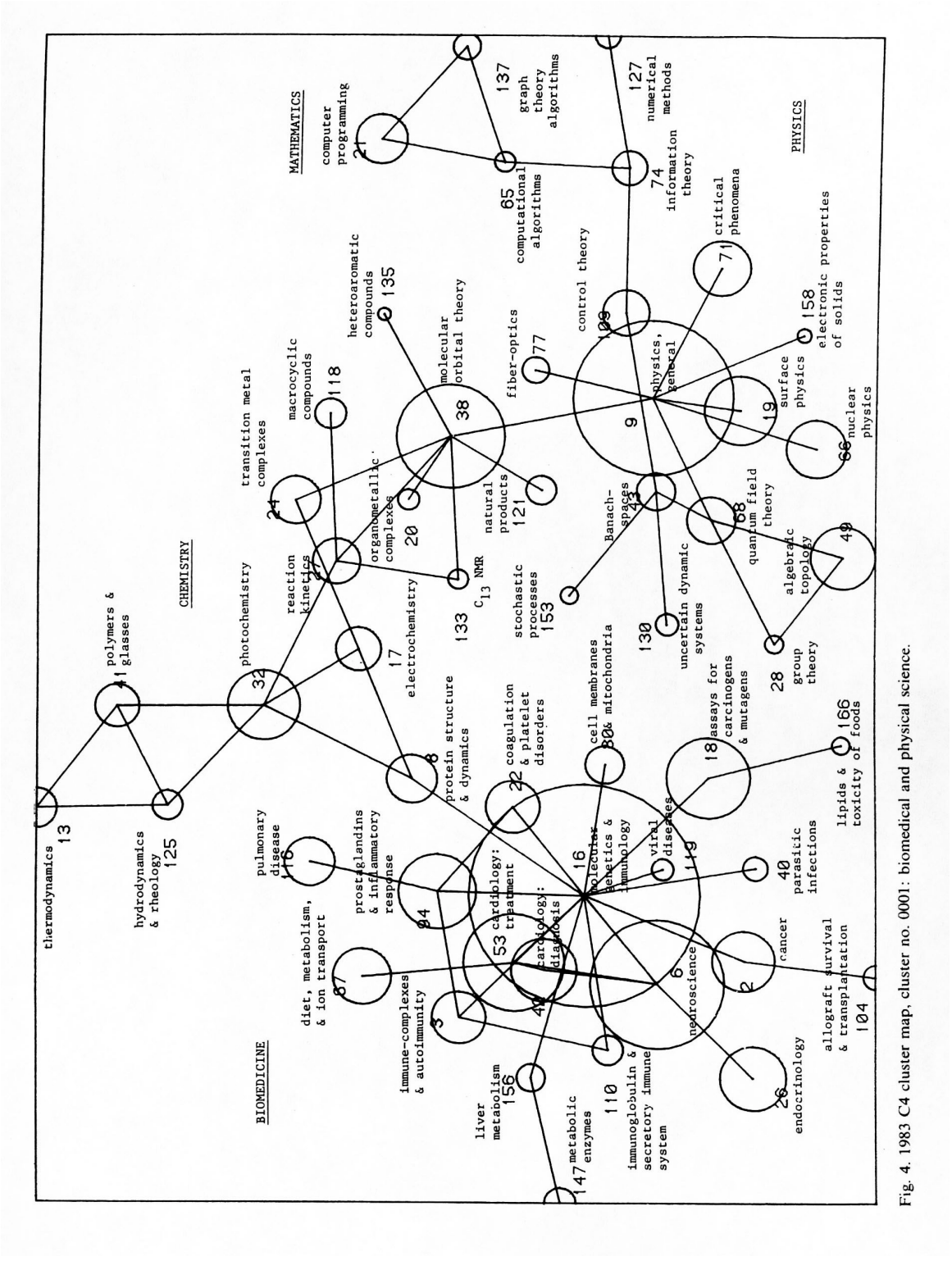

Fig. 4. 1983 C4 cluster map, cluster no. 0001: biomedical and physical science.

Figure 2: Citation Cluster Map for the Sciences
[Small & Garfield, 1985, p. 154, by permission]

One commentator on an earlier draft of this report asked whether the increase in the number of versions in bibliographic families on the Web would affect the size and distribution of family clusters or the statistical curve, thus requiring different ways of handling these families in the Web environment. As I understand these Bradford-style distributions, the answer is "no." Million-item and billion-item distributions will be of the same general pattern. Thus Figure 2 could be representing bibliographic universes of dramatically different sizes.

However, human beings are finite processors of information. We are willing to look at 30 records, but not 3,000. So we may have to find a different solution with larger bodies of information, just as earlier small classifications of the nineteenth century were replaced with subject headings in order to provide finer-grained access to ever-larger collections. Size does affect the solutions we develop--not because the pattern or distribution changes, but because our finite processing capacity necessitates different solutions with size growth, each new solution restoring equilibrium and making effective use of the Bradford-distributed collection again possible. (See further discussion of this matter in Bates [2002b].)

Why does it matter what the distribution is? If we understand, first, that there are remarkable regularities underlying all the individual records and individual behaviors, then we are in a position to *work with* these statistical patterns in ways that help users and reduce cataloging effort. In the 1970's, the so-called Pittsburgh Study (Galvin & Kent, 1977) demonstrated that book circulation followed this Bradford pattern. But commentators at the time apparently had no understanding of the underlying statistics, and the study results produced many shocked reactions. How could libraries possibly own so many books that were seldom circulated? Had we had a better understanding of the statistics of all information collections, we could have responded much better.

Second, there are particular characteristics of power law distributions that we can take advantage of. First of all, these distributions are what is variously known as "self-similar" or "scale-free" (Barabási, 2002). That is, sub-sets within the collections *also* follow a power law distribution. So, for example, when librarians seek to reduce the number of records under a very popular subject heading by subdividing the heading, the resulting set of heading-plus-subdivisions will also follow a power law, that is, there will be a lot of records under a few popular subdivisions of the heading, and few records under most of the subdivisions under that heading.

Here is one more example of the value of understanding Bradford distributions of information resources. In 1995, Jennifer Younger argued for a different approach to authority control in this century (Younger, 1995). Though she did not mention Bradford's Law, it is clear that she had an intuitive understanding of it, and was seeking to use that law to the benefit of cataloging processes. She says:

> The concept of utility is emerging as a new goal in catalog construction and authority control. Utility, in conjunction with comprehensiveness, calls for

> focusing our attention on those personal names for which authority control is most likely to prove significant in effective information retrieval. This focus will be achieved by an iterative authority control process wherein name headings move from an uncontrolled state to a fully controlled one as the need arises. (p. 140-141)

Younger emphasized the particular circumstances under which people search for names, but also present in her discussion is an awareness that some people are more important to researchers, generate more interest (and therefore more related publications). These names represent the popular end of the Bradford Distribution. In cases of limited resources, why not give more of scarce resources to addressing the high-demand end of the distribution than the low-demand end? In line with the "80/20 rule" (a crude expression of a power law), if 20 percent of the work effort can take care of 80 percent of the user needs by attending to the popular end of the distribution, is it always necessary to invest as much effort in each item in the other 80 percent?

3C. Staging of Access to Resources in the Interface

As noted earlier, two related issues are considered under this rubric: Should catalog users be provided summaries or abstracts of the books or other materials they are selecting from, and, more generally, how should the presentation of the information in and about the resource be sequenced for the user when moving through screens while searching?

When people are asked what additional features they want from a catalog entry, a request for a summary, abstract, or other additional content information is one of the most commonly mentioned (Cochrane & Markey, 1983; Matthews, 1983, p. 134, and discussion in Drabenstott, 1991). The idea has floated around in the field for some time; the costs of adding non-machine-readable information of this type has generally retarded efforts to follow through with the necessary changes in cataloging.

It turns out, however, that there exists a statistical theory in information access that strongly supports the value of such additional information for users. In fact, the addition of summaries is just one component of a larger theory about the nature of optimal human access to information. Just as we found in section 3B. that the Bradford Distribution might underlie the grouping and distribution processes of related documents, so also here we may draw up underlying statistical patterns to provide insight on the question of staging of presentation of information. This material is treated in detail in Bates (1998). The following draws extensively on that discussion.

Under a large Federal grant, Howard Resnikoff and James Dolby researched the statistical properties of information stores and access mechanisms to those stores (Dolby & Resnikoff,1971; Resnikoff & Dolby, 1972). Over and over again they found values in the range of 28.5:1 to 30:1 as the ratio of the size of one access level to another. For

mathematical reasons, they pegged the precise number at 29.55, but used 30:1 for simplicity's sake in most cases. They found the following in their research data:

> • A book title is 1/30 the length of a table of contents in characters, on average (Resnikoff & Dolby, 1972, p. 10).
>
> • A table of contents is 1/30 the length of a back of the book index, on average (p. 10).
>
> • A back of the book index is 1/30 the length of the text of a book, on average (p. 10).
>
> • An abstract is 1/30 the length of the technical paper it represents, on average (p. 10).
>
> • Card catalogs had one guide card for every 30 cards on average. Average number of cards per tray was 30^2, or about 900 (p. 10).
>
> • Based on a sample of over 3,000 four-year college classes, average class size was 29.3 (p. 22).

These ratios may seem improbable, but are based on real empirical data. But even if we accept that these results are valid, what do they mean? We can be pretty confident that no one started out with the intent to create these ratios deliberately. The results are surprising and unpredicted through any conscious mechanism. Thus, the fact that there are such remarkable regularities suggests the operation of some underlying statistical and/or cognitive process. People, as we have so often noted in this report, tend toward least effort. *The persistence of these ratios suggests that they represent the end result of a shaking down process, in which, through experience, people became most comfortable when access to information is staged in 30:1 ratios.*

No publishers' council mandated the above ratios for the contents of books. Yet, unaware of Resnikoff & Dolby's results, I discovered in 1986 that the structure of book contents was remarkably stable over hundreds of years, despite the extraordinary changes during that time (Bates, 1986b). Over that long a period, we would surely have broken out of old patterns if the pre-existing ones did not meet our needs.

To say that these averages are consistent is certainly not to say that every book, technical article, or classroom has exactly a multiple of 30 in its countable units. These are all averages, representing the central tendency of a range. However, Resnikoff & Dolby did find that the data clustered fairly tightly around these means (1972, p. 92). The largest item at one level seldom exceeded the smallest item at another (p. 90).

Resnikoff & Dolby (1972) suggested that the line between two access levels should be drawn in the following manner: "Mathematically, the natural way to define a boundary between two values on an exponentially increasing scale is to compute the geometric mean of the two values" (p. 12). The figures that they computed for the range

around an average of 30 goes from 5 to 161 (remember these are *geometric means*, not the conventional average). The range around 29.55^2 , or 873, is 161 to 4,747; the range around 29.55^3, or 25,803, is 4,747 to 140, 266, and so on (Bates, 1998, p. 1198).

What do these figures matter for modern catalogs? Let us look at the data in a study done independently of Resnikoff & Dolby's research. Wiberley, Daugherty, and Danowski (1995) carried out two studies of user persistence in displaying online catalog postings, first on a first-generation online catalog, and later on a second-generation online catalog. In other words, they studied how many postings users actually examined when presented with a hit rate of number of postings found when doing a search in an online catalog. They summarized the results of both studies in the abstract of the second: "The findings suggest that given sufficient resources, designers should still consider 30 to 35 postings typical persistence, but the findings also justify treating 100 or 200 postings as a common threshold of overload" (1995, p. 247).

In other words, people would most commonly look at about 30-35 postings in a search, and they might search as many as 100-200 postings, but anything beyond that was overload. Note the striking parallels with Resnikoff & Dolby's data. When the users submit queries to the catalog, they are willing to look at about 30 postings on average, and will sometimes look at fewer or more--but almost never over 100-200, which is an approximate match with Resnikoff and Dolby's range of five to 161. Over the point of about 161 postings, the users feel they are in overload, and stop.

In other words, whether we believe Resnikoff & Dolby's theory or not, independent evidence shows that people tend to act in a way that produces results conforming to their model. (And, in fact, we have no particular reason to doubt their model. Both men are highly qualified, Dolby was a professor of mathematics, and Resnikoff went on to head a section of the National Science Foundation. However, these regularities seem *too* absurdly regular at first encounter, and the reaction of many people has been to reject them out of hand. I think they are worth a closer look.) It is likely that statistical processes, arising from human cognitive characteristics, are responsible for these striking results. We are not ordinarily aware of those processes, but tend to act in ways that produce these results.

These results have clear implications for staging of the presentation of material to users in any information processing situation, including information retrieval. As noted earlier (Cochrane & Markey, 1983), many user studies produce results in which people state that they want more information, an abstract or contents list, on the catalog record. Resnikoff and Dolby discuss at some length questions of the role the catalog plays, in terms of levels of access to the collection. Both bodies of data converge on the idea that people need another 30:1 layer of access between the catalog entry and the book itself. This need shows up in the request for an abstract or summary, presumably about 30 times as long in text as the average book title or subject heading.

To use an analogy, people need to be able to move down stepwise into large bodies of data, just as they step down a staircase. Imagine a staircase with some steps two inches high and others two feet high. Such a staircase would cause many accidents

and lead its users to dread using it. In contrast, a consistently 30:1 stepping ratio for access to information stores should feel easy and natural to people.

Staging issues have also been recognized in various efforts in the field to compress redundant listings of subject headings in the OPAC interface when people search by subject. Allen (1993), Drabenstott & Weller (1996), and McGarry & Svenonius (1991) have all proposed ways of reducing the on-screen overload for subject searchers.

In the early days of research on human-computer interaction, a common debate concerned how wide or deep on-screen (tree-structured) menus should be. Shneiderman (1998, chapter 7) reviews a number of these studies. The consensus of the various studies is that menus should be wider than deep (p. 247 ff). That is, it is better to offer the searcher more options at each level, and end up with fewer levels down the hierarchy, than to offer few options with many levels deep.

Interestingly, none of the tests Shneiderman describes ever offered *as many as* 30 options at one level; the highest number tested was 16. Resnikoff and Dolby's work, produced much earlier, suggests that users can comfortably handle 30 or so as the breadth for a single level. It is not surprising, then, that the tests that were done confirm the value of menu breadth.

4. Implications and Recommendations

Though catalogers and metadata experts must concern themselves with all the design and access particulars for the bibliographic universe, we can also look at these specific decisions in a larger context, as a part of a unified and sophisticated bibliographic description and access universe. Standardization and cooperative efforts over many decades have been devoted to just such a unified vision.

However, the discussion in this report also brings to the fore both the underlying statistical unity of the bibliographic universe we manage, as well as the reasonably stable human behavior that can be seen to interact with that bibliographic universe in characteristic ways. The goal is both good bibliographic control and optimal access for users.

What, then, do we learn from the above review that is applicable to the three issues being considered?

4A. User Access Vocabulary

In 1994, in a consulting report to the Council on Library Resources, I analyzed the possibilities of creating an expanded entry vocabulary for the Library of Congress Subject Headings (Bates, 1994b). With the current state of the Internet and of networked library and information resources, *I believe a much better goal for an entry-vocabulary project today would be to develop a general end-user entry vocabulary that can be used anywhere across the Internet, as well as in libraries, intranets, and other information environments, and by any agency.*

Under the aegis of some organization such as the Library of Congress, or ALA's Association for Library Collections and Technical Services Subject Analysis Committee, which would monitor additions and changes in the vocabulary, the vocabulary could be licensed for a maintenance fee by any catalog or database developer.

How would such a vocabulary work, and how could it be created at little cost? Here are the key design features I propose:

• The vocabulary is intended for two principal purposes:

1) Help the user by suggesting additional related terms than may be ORed in to a query to improve recall (proportion of relevant items in the database that are retrieved).

2) Help the user find the best vocabulary (most precise, most accurately representative of the subject of interest) to describe a search topic to improve precision (proportion of retrieved items that are relevant).

• Use of the entry vocabulary would be an option in the interface--the searcher could either go through the vocabulary or directly to the database to search. (Under no circumstance should the user have to enter one database, write down vocabulary, withdraw and enter another database. Such an approach violates the Principle of Least Effort. The vocabulary has to feel easier than this to use, or it will not be used.)

• The design would function at the *database level*, not the *individual record level.* Thus it would not be necessary for catalogers to assign entry terms to individual records. Rather, the searcher explores in the vocabulary, then selects desired legitimate index terms or subject headings, and enters, or clicks on hyperlinks, of these legitimate terms in order to search in the usual fashion.

• The basic design of the vocabulary would consist of human-made (with computer support) *clusters of terms.* The cluster would contain all those terms that relate closely to a core concept.

• The entry vocabulary is used in this manner: The searcher enters a term of interest and the cluster database is searched for that term. *When any search term matches with a term in one or more clusters, the whole cluster is brought up to the screen.* The objective would be to have a system in which any reasonable word or phrase (including even popular misspellings) would match with some cluster or other. Thus a searcher

vocabulary is structured differently from an indexer vocabulary. (See these websites for listings of vocabulary resources of all types: INFORUM, Lutes, Middleton.)

In each cluster, both accepted "legitimate" indexing terms or headings appear in the cluster (marked as such), as well as the many sorts of entry terms that people use in information systems. These entry terms would be morphological, syntactic, or semantic variants on the core concept. The clusters would also contain the more informal and often ambiguous multi-meaning terms that might match with several clusters. Figure 3, drawn from the 1994 CLR report (Bates, 1994b), illustrates what I mean by clusters in the case of Library of Congress Subject Headings.

Figure 3: Example Cluster

Cluster Title: Equality

Your search word(s) relate to these subject headings used for indexing:

- Equality
- Equality before the law
- Democracy
- Liberty
- Individualism
- Social justice
- Equal pay for equal work
- Discrimination in employment
- Women--Employment
- Sex discrimination--Law and legislation
- Affirmative action programs
- Educational equalization
- Hate crimes
- Equal rights amendments
- Pay equity
- Sex discrimination against women--Law and legislation
- Civil rights
- Justice
- Race discrimination--Law and legislation
- Reverse discrimination--Law and legislation
- Sex discrimination against men--Law and legislation
- Women's rights

These terms are not used for subject indexing but may be useful for a title search:

- Inequality
- Social equality
- Job bias
- ERA
- Social equity
- Equal
- Equal protection of the law
- Equal pay for work of comparable value
- Equal opportunities
- Social justice
- Equalization

Related clusters:

- Egalitarianism
- Justice
- Civil rights
- Women's rights

The CLR example is hypothetical. Real clusters for a general-purpose vocabulary would probably have even more of the uncontrolled terms for keyword searching. However, this cluster approach was fully and successfully implemented in the Los Angeles Department of Water and Power vocabulary. (Linda Rudell-Betts was chief lexicographer.) The Department was not a small entity; the staff was 10,000 strong, and the cluster vocabulary had to cover water and power engineering, construction, real estate, customer relations, and a number of other areas. Ultimately, there were about 4700 clusters in the vocabulary, including clusters for popular proper names. Figure 4 illustrates the very different look to this more engineering-oriented vocabulary.

```
LA DWP CLUSTER THESAURUS, 2nd Edition                    PAGE 305

RS BILLING RATES                  Raw Water Chemical       -----
RS HEAT RATES                        Treatments
RS INSURANCE RATES                Raw Water Cooling     CT REACTORS
RS UTILITY RATES                     Towers             TE  Reactors  *
RS RATE SCHEDULES                 Raw Water Facilities
RS RATE STRUCTURE                 Raw Water Feed Systems RS ELECTRIC REACTORS
RS PRICING                        Raw Water Pipes        RS NUCLEAR REACTORS
RS CONSUMPTION RATES              Raw Water Pumps        RS SMOOTHING REACTORS
RS RATINGS                        Raw Water Systems  *   RS SHUNT REACTORS
   -----                       RS WASTEWATER             RS WATER COOLED REACTORS
                               RS WATER                  RS CAPACITORS
   CT RATINGS                     -----                  RS VAR
   TE  Ratings  *                                        RS FILTER REACTORS
       Bipole Overload Power   CT REACTIVE POWER            -----
           Ratings             TE  Reactive Power  *
         Filter Ratings            Reactive  Voltamperes  CT READINGS
         Merit Ratings             Wattless Power         TE  Readings  *
         Nameplate Rating  *       Kilovars  *               Meter Readings  *
         Overcurrent Ratings       Kvar                      Ultrasonic Readings
         Power Ratings             Reactive Power Balance    Readouts
         Transient Ratings         Reactive Power            Meter Readers  *
           Voltage Ratings           Controllers          RS METERS
           Productivity Charts     Reactive Power         RS MEASUREMENTS
       RS EFFICIENCY                 Requirements           -----
       RS EVALUATIONS              Reactive Power Swings
       RS GRADING                  Reactive Compensation  CT REASSIGNMENTS
       RS PERFORMANCE EVALUATIONS  Reactive Compensation  TE  Reassignments  *
       RS PERFORMANCE                Controls               Management
       RS RATES                    Fast Reactive Power        Reassignments
          -----                      Controllers            Work Reassignments
                                    Fast Reactive Power     Lateral Transfers
   CT RAW WATER                       Control              Change of Work Location
   TE  Raw Water  *              RS POWER SYSTEMS          Change of Address
         Black Water             RS ELECTRIC REACTORS      Change of Telephone
         Flush Water  *          RS VAR                    Change of Work
         Gray Water  *           RS POWER                     Locations
         Grey Water              RS ELECTRIC POWER         Change of Work Periods
         Untreated Water         RS POWER FACTOR        RS ASSIGNMENTS

   CT = Cluster Title   TE = Term    * = Preferred Term    RS = Related Subject
```

Figure 4: Los Angeles Dept. of Water and Power Cluster Thesaurus
For searchers, a match with any term in a cluster would bring up that cluster. The thesaurus was never used in paper form; all users saw it only online within the information system. Related Subjects are other clusters. Underlined title refers to High Frequency Cluster, which contains only other more specific clusters within.

With the LADWP, we found in practice that it worked for the vocabulary to consist of two levels--clusters, and clusters of clusters (High-Frequency Clusters). If the search term matched with the high-frequency cluster, then the searcher saw a listing of term clusters and was asked to select among them. People tend to enter fairly broad terms (see discussion in Bates, 1977b). Searching on very broad terms is often not helpful because of the general practice of indexing at a specific level (Rule of Specific Entry in LCSH). Thus by being guided down to more specific terms, the searcher was often enabled to narrow the search appropriately. The system was designed so that the searcher could always nonetheless search on a broader term if desired.

• Fortunately, the cluster vocabularies would probably not have to be developed *de novo.* Earlier, the searcher vocabulary developed by Sara Knapp (1993) was mentioned. She has now enlarged that vocabulary to cover all of the social sciences and humanities (Knapp, 2000). She and her publisher might be amenable to licensing the vocabulary, or to receiving royalties for its licensing by organizations using the entry vocabulary ultimately created.

Knapp's vocabulary could form the core of the entry vocabulary. It is arranged alphabetically by the titles of term clusters. Within each cluster are listed numerous natural language term variants for the core concept. See Figure 5.

unipolar disorder, ecstatic symptom(s). *See also* Affective disorders; Agitation; Cyclothymic disorder; Mania.

Manifesto. Manifesto(es,s). Public declaration(s). Proclamation(s). Pronouncement(s). Announcement(s). *See also* Public relations.

Manipulation. Manipulat(e,ed,ing,ion,ive). Ingratiat(e,ing). Machiavellian(ism). Exploit(ive,ation). Triangulat(e,ed,ing,ion). Extort(ing) sympathy. Hidden agenda(s). Brinksmanship. Double talk. Double think. Con(ned,ning). Main(s) syndrome. *Consider also:* impression management, propagand(a,ize), cunning, schem(e,ed, ing), conniv(e,ed,ing). *See also* Exploitation; Interpersonal relations; Machiavellianism; Social behavior.

Mankind. See *Humanity.*

Manners. Manners. Courtes(y,ies). Nicet(y, ies). Well mannered. Rules of conduct. Custom(ary,s). Tradition(al,s). Mores. Social code(s). Formalities. Polite(ness). Civilit(y,ies). Propriet(y,ies). Decorum. Social convention(s). Protocol. Folkway(s). Norm(s). Etiquette. Socially prescribed. Unwritten law(s). *See also* Etiquette; Norms; Social skills.

Manpower. Manpower. Womanpower. Personpower. Personnel. Workforce. Operator(s). Employee(s). Staff. Worker(s). Supervisor(s). Work force. Office force. Sales force. Subordinate(s). Labor force. Labor supply. Functionar(y,ies). Crew(s). Troop(s). Hired help. Manager(s). Coworker(s). Occupational group(s). Laborer(s). Practitioner(s). Paraprofessional(s). Professional(s). Nonprofessional(s). Air traffic controller(s). Attorney(s). Administrator(s). Agricultural personnel. Blue collar worker(s). Caseworker(s). Church worker(s). Civil servant(s). Caregiver(s). Clerk(s). Consultant(s). Computer programmer(s). Crafts(man,men,women,woman). Craftsperson(s). Craftspeople. Designer(s). Driver(s). Director(s). Domestic servant(s). Editor(s). Faculty. Firefighter(s). Guard(s). Government employee(s). Health personnel. Hairdresser(s). Home economist(s). Librarian(s). Machine operator(s). Military personnel. Nurse(s). Police officer(s). Pilot(s). Programmer(s). Sales(men,man,women,woman,people, person,persons). Secretar(y,ies). Stevedore(s). Teacher(s). Technician(s). Trainer(s). Typist(s). White collar worker(s). *Consider also:* employed, civil service, working, hired *with:* woman, women, men, man, people, group(s), SS, males, female(s), person(s), civilian(s). *See also* Allied health personnel; Dislocated workers; Employee turnover; Health manpower; Human resources; Labor market; Labor supply; Nonprofessional personnel; Personnel termination; Professional personnel; Workers.

Manpower, health. See *Health manpower.*

Manslaughter. See *Homicide.*

speech. *See also* Communication (thought transfer); Communication skills; Deafness; Oral communication; Sign language.

Manual labor. See *Blue collar workers.*

Manual workers. See *Blue collar workers.*

Manuals, sex. See *Sex information.*

Manufacturing. See *Factories; Industry.*

Manumission. See *Emancipation.*

Manuscripts. Manuscript(s). Letter(s). Text(s). Diar(y,ies). Typescript(s). Pre print paper(s). Handwritten text(s). Draft(s). Correspondence. Notebook(s). Memoranda. Handwritten script(s). Author's copy. Rough draft. Unpublished. Working papers. *See also* Authors; Letters (correspondence); Publications; Written communications.

Mapping, career. See *Career goals.*

Mapping, cognitive. See *Cognitive mapping.*

Maps. Map(s). Mapp(ed,ing). Cartogram(s). Cartography. World globe(s). Projection(s) of the world. Chartbook(s). Chart(s). Navigational log(s). Navigational chart(s). Atlas(es). Street guide(s). Gazetteer(s). Plat(s). Grid(s). Guide(s). *See also* Geography.

Maquiladora. See *Offshore production (foreign countries).*

Marathon group therapy. Marathon group therap(y,ies). Marathon group(s). *Choose from:* marathon *with:* encounter(s), group(s), experience(s), therap(y,ies), workshop(s), growth group(s). *Consider also:* accelerated interaction, time extended therap(y,ies). *See also* Group psychotherapy; Human relations training; Psychotherapy; Sensitivity training.

Marginality, social. See *Marginality.*

Marginality (sociological). Marginal(s,ism, ity,ization). Marginal man. Outsider(s). Cultural hybrid. *Consider also:* bicultural(ism), bilingual(ism), peripheral, emargination, underclass, deviance, eccentric(ity). *See also* Alienation; Center periphery; Centrality; Deviance; Minority groups; Ostracism; Social distance; Social isolation; Social status; Social structure; Underclass.

Mariculture. Mariculture. Marine aquaculture. Sea farm(s,ing). Ocean kelp farm(s). *See also* Agriculture; Farming; Fisheries.

Marihuana. See *Street drugs.*

Marijuana. See *Street drugs.*

Marijuana abuse. See *Substance abuse.*

Marijuana laws. Marijuana law(s). *Choose from:* cannabis, marihuana, marijuana, hashish *with:* law(s), legaliz(e,ed,ing, ation), legal status, Uniform Controlled Substances Act, decriminaliz(e,ed,ing). *See also* Drug and narcotic control; Drug enforcement; Legalization of drugs; Street drugs.

Marijuana legalization. See *Legalization of drugs; Marijuana laws.*

Marine, merchant. See *Shipping industry.*

shore, sea(s) *with:* mammal(s). *See also* Animals; Mammals; Marine resources.

Marine personnel. Marine personnel. *Choose from:* marine(s) *with:* corps, personnel, enlistee(s), recruit(s), officer(s), enlisted, non commissioned officer(s), force(s), trainee(s), general(s), colonel(s), lieutenant(s), sergeant(s). *See also* Armed forces; Military personnel.

Marine resources. Marine resource(s). Fisher(y,ies). *Choose from:* aquatic, archipelagic water(s), coastal water(s), continental shel(f,ves), fishing water(s), international strait(s), international water(s), marine, maritime, nautical, ocean(s), offshore, sea(s), seabed(s), seafloor, underwater, Atlantic Ocean, Pacific Ocean, Mediterranean Sea, Caribbean Sea, Red Sea, South China Sea, Indian Ocean, Arctic Ocean, Antarctic Ocean, Baltic Sea, North Sea, Irish Sea, Adriatic Sea, Ionian Sea, Aegean Sea, Black Sea, etc. *with:* resource(s), reserve(s), conservation, mineral(s), mining, deposit(s), petroleum, oil, gas, geolog(y,ical), living resource(s), ecosystem(s), vegetation, plant(s), fish(es), mammal(s), peaceful use(s), exploit(ed, ing,ation). *See also* Fisheries; Fishes; Geopolitics; International law; Mariculture; Marine mammals; Maritime industry; Maritime law; Natural resources; Oceans; Offshore oil; Shipping industry; Territorial waters.

Marital adjustment. See *Marital conflict; Marital relationship.*

Marital conflict. Marital conflict(s). *Choose from:* marital, marriage(s), married, husband wife, spous(e,es,al), conjugal, domestic, couple(s) *with:* trouble(s,d), stress(es), distress(ed,es), dissension, problem(s), conflict(s), role strain(s), sex conflict(s), pathology, difficult(y,ies), argument(s), competetive(ness), competition, frustration(s), aggression, breakup(s), fail(ed,ure,ures), dissatisf(ied,action,actions), unhapp(y,iness), discord, resent(ful,ment,ments), hostil(e, ity,ities), impasse(s), misunderstanding(s), fight(s,ing), clash(es), incompatib(le,ility), dysfunctional, instability, unstable. *See also* Arguments; Battered women; Family conflict; Family crises; Family relations; Family stability; Family violence; Marital disruption; Marital relationship; Spouse abuse.

Marital counseling. See *Marital therapy.*

Marital disruption. Marital disruption. *Choose from:* marriage(s), marital, married, conjugal, couple(s), spous(e,es,al), family, parent(s,al), wife, wives, husband(s) *with:* disrupt(ed,ion), breakdown, problem(s), conflict(s), dissatisfaction, violence, dysfunction(s, al), unhapp(y,iness), burnout, instabil:ty, separat(e,ed,ing,ion), leaving, uncoupling, breakdown, dissolv(e,ed,ing), dissolution, estrange(d,ment), sever(ed,ing), break up, desert(ed,ing,ion). *Consider*

Figure 5: Page from Searcher Thesaurus. Sara D. Knapp, *The Contemporary Thesaurus of Social Science Terms and Synonyms: A Guide for Natural Language Computer Searching.* Phoenix, AZ:

Set up in an end-user interface, such vocabulary could be adapted to be better understandable by non-professional searchers. Terms might be presented on a list, or in a relationship diagram, with each term variant spelled out, instead of presented in the parenthetical method seen in Figure 5. The searcher might click on terms to search on, which would then be combined in a Boolean OR.

Note that Knapp has already identified core concepts and their associated search term variations. Thus, using her set would eliminate the need to select cluster concepts from scratch. The experience of online database searchers suggests that the social sciences are the most problematic when it comes to concepts having various names. Thus it would make sense to start the project by developing an entry vocabulary based on the social sciences.

If funding could be found for just one or two lexicographers to manage the vocabulary centrally, participating libraries and Web organizations could contribute additional entry terms from transaction logs of zero-hit searches and other sources. The vocabulary could be licensed and tested by various organizations, and its structure refined with experience. Once the baseline vocabulary was developed, additional work on it would consist of continual updating, for the most part, rather than dramatic changes.

The entry vocabulary would thus become a universal entry vocabulary, enabling people to find terms useful for their purposes in a wide variety on online information searching situations. Individual organizations could merge their own indexing vocabulary into their copy of the universal cluster vocabulary, so that when searchers match a term in a cluster, they can be shown the legitimate search terms to use for their own organization. It may be possible to use the Topic Maps standard for data interchange (ISO/IEC 13250 [19 May 2002]). This proposal is also consonant with the objective, proposed by Chan (2000), to broaden our conception of subject description and access for Web-based information.

Commentators on an earlier draft of this report have urged 1) the use of other vocabularies also, including LCSH, and 2) the utilization of various types of software to create vocabularies automatically wherever possible. These are both highly desirable goals. But they would need to be implemented in a particular way. Let us address each of them in turn.

Knapp is being emphasized because hers is the only vocabulary I am aware of that is specifically designed for searching, as opposed to indexing, *and* comes from a deep knowledge of what kinds of vocabulary is effective for online searching. Knapp based the work on her own lifetime of experience, as well as the input from a number of other experienced online database searchers. However, having proposed that we start with her vocabulary, it is by no means being suggested that we stop there.

Ultimately, the library world may create a website that becomes a kind of Vocabulary Headquarters (VHQ). Vocabularies from many different sources and with a

wide variety of purposes could be mounted. The Library of Congress Classification--perhaps mapped to the LCSH--could also be a part of the VHQ. These could include descriptive information as well as subject (author name variants, place name variants, etc.), more specialized subject vocabularies, all manner of non-textual information (thumbnail images, sounds, etc.), cross-language retrieval (e.g., MACS, [2003]), different world views, and a wide variety of types of classification scheme access.

Further, the *contents* of still and motion visual information resources, as well as audio resources, represent a very wide range of new issues that have to be addressed in subject access. As a practical matter, it is suggested that the better-understood textual vocabulary be the starting place, but with the explicit intention to move quickly into the less traditional, but increasingly core, visual and sound media. A commentator also suggested that improved ways to help searchers locate information in a particular graphic or genre form be included in the access vocabulary. As a variety of terms are used in Web resources for each of these forms, having a cluster of terms for each type of form would constitute another excellent means of improving access for searchers.

Some of these vocabulary resources would be designed to promote searching effectiveness, and can be listed as a place to start on an entry screen. *Within* the searcher vocabulary, the use of additional sources of terms may add to the richness of possible terms in the clusters, beyond what is present at the beginning. It is fitting that such a website be associated with the library field, and its very presence on the Web will promote understanding of our special expertise.

One reviewer of this report asked if the heterogeneity of types of vocabularies used on the Web might detrimentally affect the nature or functionality of an access vocabulary. It is precisely that heterogeneity that an access vocabulary composed of a very wide range of terms should be well suited to match. Even misspellings, common misunderstandings (confusing "gantlet" and "gauntlet"), and zero-hit searches would be good sources of terms to add to the clusters.

One difference between access vocabulary and regular indexing vocabulary is that a perfectly self-contained and complete vocabulary is not needed to start with. A good working set of access terms can be used at the beginning, later to be supplemented by other forms of enrichment.

Though I attempted to design a way to use Library of Congress Subject headings in the Council on Library Resources report (Bates, 1994b), I believe there would be both psychological and operational difficulties in starting with LCSH to build such a vocabulary. LC headings have a certain conventional syntactic structure that experienced librarians are familiar with, but which most users are not (Bates, 1977a). In developing an access vocabulary, it would be all too easy to fall into the verbal rhythms of LCSH, and end up producing an access vocabulary that looked quite like the LCSH. Further, the lengthy structures of heading plus one or more subdivisions are quite unnatural for end users. OCLC's effort to produce segmented Library of Congress headings, known as FAST subject headings (OCLC, Library of Congress, 2002), could, however, be a source of terms closer to those typically used for access purposes.

The second question concerns the use of available software for generating access terms. Anything that can be well done automatically should be. We should not assume, however, that all such software is well designed for the purposes here. I have observed that, with many commercial applications as well as with research experiments by linguists, artificial intelligence experts, and computer scientists, there is frequently no recognition that information searching and retrieval is a distinctive cognitive and linguistic activity, and that general language dictionaries and thesauri, such as Roget's, do not work well for information retrieval. So, for example, a lot of attention is being given to Wordnet (2003), which is a highly developed general English vocabulary, displaying all sorts of relationships between the terms. But the vocabulary is not well suited for, nor intended for, information retrieval. This can be quickly seen when comparing almost any indexing thesaurus with the structure of Wordnet.

Thus, several types of vocabularies relevant to our concerns can be identified:

a) general English-language dictionaries and thesauri
b) indexer thesauri for use by people indexing documents
c) searcher thesauri to support the act of searching itself

Among uncontrolled vocabularies there are:

d) words/phrases taken from spoken or written English
e) words/phrases used at the moment of search

So the first question that could be asked about the commercial and experimental systems out there is: Do the creators of the software understand the specific character of information searching/retrieval? Are they designing their vocabulary to support the actions of (c) above, incorporating (for matching purposes) the vocabulary of (e) above? Vocabulary processing software is getting more and more powerful, to be sure, but we sometimes underestimate the specific expertise we have as vocabulary experts. The best result is almost certainly some mix of automatic computer processing combined with human vocabulary design and editing.

4B. Grouping/Linking Bibliographic Families

As Tillett (1991a) and Carlyle (1997) have shown, there are already many forms of linkage prescribed in standard AACR2 cataloging practice. The links do not yet form a part of a general theory or rule set for creating bibliographic families; however, the FRBR model holds great promise for future creation of Web-based bibliographic families.

It was suggested earlier that these clusters of bibliographic families almost certainly follow a Bradford Distribution. That is, there are probably few very large families, say, over 100 members in size, a moderate number of medium and small families, say, 10 to 100 members, and a very large number of small families, under 10. In

fact, the single largest category of works in most cases consists of those with only one member in the family.

Thus, where there is a desire to create cataloging that recognizes bibliographic families, such data provide a clear indicator of where to start--with the largest families. As a rule, large families arise in the first place because progenitor items are of great interest in research and/or in society generally. Many editions are published, translations, adaptations, films made, etc. Thus it is almost certainly the case that most bibliographic families are not only large themselves but are also of interest to disproportionate numbers of catalog users.

There may be some variation in use relative to family size. For example, major works currently out of fashion will certainly receive fewer searches than more popular ones. But on the whole, when human beings generate large bibliographic families, there are likely to be lots of other human beings also interested in consulting the works in those families. As noted earlier, the "80/20 rule" is a crude description of power laws such as the Bradford Distribution. It is likely that amount of use tracks fairly well with size of family. Thus, attention to a few large families may satisfy many users quickly.

Further, we may guess that these large families, simply because of their size and complex internal relationships, may cause more confusion for searchers--and thus be good candidates for clarifying attention, regardless of the numbers of searchers they attract. On the other hand, since there is a wide range in family size, catalog users will understand intuitively when they see some large families with extra links or labeling, while smaller families or singletons do not yet have such information.

All this is by way of saying that whatever is chosen to be done to these families, it should be possible to start with selected large families, without unduly confusing catalog users, and proceed as resources permit, down the chain to smaller and smaller families.

So what can be done with these families? All of the several writers reviewed above have contributed valuable and helpful perspectives on this question: Tillett, Carlyle, Smiraglia, Leazer, Vellucci, Yee, and others. They are much better qualified than I to address the specific descriptive cataloging issues. However, I will suggest a general framework within which those cataloging decisions might function.

• First, some agreement is reached on a relationship model to be used in managing these families. The powerful FRBR appears increasingly to be the model to build on. Recently, the vendor VTLS developed an integrated library system that supports FRBR (VTLS…2002).

• Next, on an experimental basis, qualified people or libraries identify families that they are willing to be responsible for. *Their responsibilities lie solely at the family level*--all record-level cataloging continues in the same cooperative manner it does now.

• Bibliographic families are listed in an auxiliary database available online to all catalogers (and perhaps eventually to searchers too). The database will be small at first,

because only a few large families are being tackled. So the database can be funded easily on an experimental basis.

• Those cooperating in the experiment track down all the related records they can find for their chosen family, then assign linkage types to the records. This may involve identifying existing markers for relationships in MARC fields, as well as making new judgments on links.

• Using record categories derived from FRBR, a family is listed in the test database in its entirely, using the individual item records already created for OCLC or other bibliographic utilities.

• The above process would demonstrate the implications for descriptive cataloging theory, impacts on MARC, etc.

• There is another development, however, that we can anticipate and which may make this work appear much more meaningful to the average catalog user. Suppose, as online display capabilities get better and better, that the bibliographic family is laid out on the screen in two dimensions. Thus, the various groupings that catalogers identify could be listed under standard icons or labels dotted around the screen. The progenitor is at the center of the screen, and the various relationships are arrayed around it.

As Lagoze (2000) pointed out, the ease of creating ever new versions in the digital environment will probably lead soon to the necessity of developing even more means of distinguishing these different versions and displaying them for the searcher. More versions will be just one of the many ways in which the size of the bibliographic universe grows explosively. It is likely that what now seems like a nice addition to bibliographic control--grouping families of records together--may become a virtual necessity, as users will need to determine quickly just which version/edition/adaptation etc. etc. they need out of ever growing numbers of items.

• After the experimental effort is completed with a few families, decisions can be made about further steps, with the goal of the ultimate incorporation of fully linked families within bibliographic utilities and catalogs. The ultimate target design may be one in which all records that are part of families will have a “related records” link when brought up on an OPAC screen. If the searcher clicks on the link, then the sort of two-dimensional display described above appears for the user. Ultimately, responsibility for bibliographic families could be shared in a similar manner to the way original cataloging is shared today.

4C. Staging of Access to Resources in the Interface

This writer has been critical in many ways of the often poorly-executed information systems in websites (Bates, 2002a). But one commercial website has been a striking exception. The online bookstore site, amazon.com, has essentially implemented every recommendation to come out of decades of information system and catalog

research and design. Whether they have actually read that literature is another question--perhaps their staff are uniquely attentive to user needs--but the end result is superb. (One qualification: The descriptive information in the book records is not as complete as the subject and reviewing information.)

One of the things amazon.com has mastered is the provision of subject information coupled with excellent staging of access. Initially, they provided reviewer comments, then subsequently added yet another 30:1 layer by providing as many as 10, 20 or 30 pages of the actual text of the book. So now, the interested shopper can search by title, author, or subject, then see reviewer and other-reader comments--roughly another 30 times as much text as the title--then if desired, access another 30 times as much information by reading the text sample pages.

The only remaining layer of information is the full text of the book itself. Thus amazon.com ensures an easy stepping down into the material--every 1:30 ratio is accounted for in their interface--title, summary, long chunk of text, leaving only the full text for purchase. Libraries are beginning to load the full text of books, but even then, users are likely in many cases to want to see a summary without having to maneuver around the full text to get a feel for the book. The Library of Congress' experiment to include summaries via ONIX records from publishers constitutes a good effort in this regard.

Contents lists vary in their informativeness; for this reason the jacket material may be routinely desirable as well. To be sure, publishers are touting the book in the jacket material, but such text also is often very informative. If we really want to help catalog users find what they want (and, by the way, use the library more), then provision of this additional information is highly desirable.

Amazon's assistance to the shopper does not stop with a 1:30 ratio in dropping down into the text of a given book. Other forms of relatedness are exploited as well. Users are shown other books on the same topic, as well as other books bought by people interested in the current book under view. These sorts of linkages create a kind of subject-based bibliographic family. The end result is that:

- On any one book, the searcher may drop down 1:30 layers as far as desired to explore the book.

- One is provided links to other books based on a variety of types of relatedness. Amazon.com does not make formal distinctions among these relationship types the way catalogers do. However, the information is often implicitly available to the searcher.

My point, however, in reviewing the staging and linking patterns found in amazon.com, is to demonstrate that there already exists an information system that has implemented many of the things being considered in this review, and, to judge by the popularity of amazon, com, the results have been shown to be very positive.

4D. Example Implementation of Recommended Approaches

In this section, a simplified example implementation of the recommended design features is provided in screen shots. Many of the specifics of the proposed screen shots could be changed to adapt to the needs of specific library systems. However, the key features that would implement the proposals in Sections 4A, B, and C of this report will be noted in the discussion.

Figure 6, Example Initial Library Catalog Screen, shows what the user might first see when approaching a library's online catalog. The user has the options of 1) browsing the higher levels of the Library of Congress Classification, 2) browsing through vocabulary clusters, or 3) directly searching on a term or phrase. Regarding the first option, we know that users frequently want to size up the conceptual organization of a library or website by browsing the classification. (Though not strictly a part of the charge for this report, there are interesting design possibilities that could be developed here. For example, users could be shown the top two or three layers of the LCC --preferably, no more. Major sub-categories within each layer could be hot-linked to appropriate subject headings, classification number ranges, and/or vocabulary clusters.)

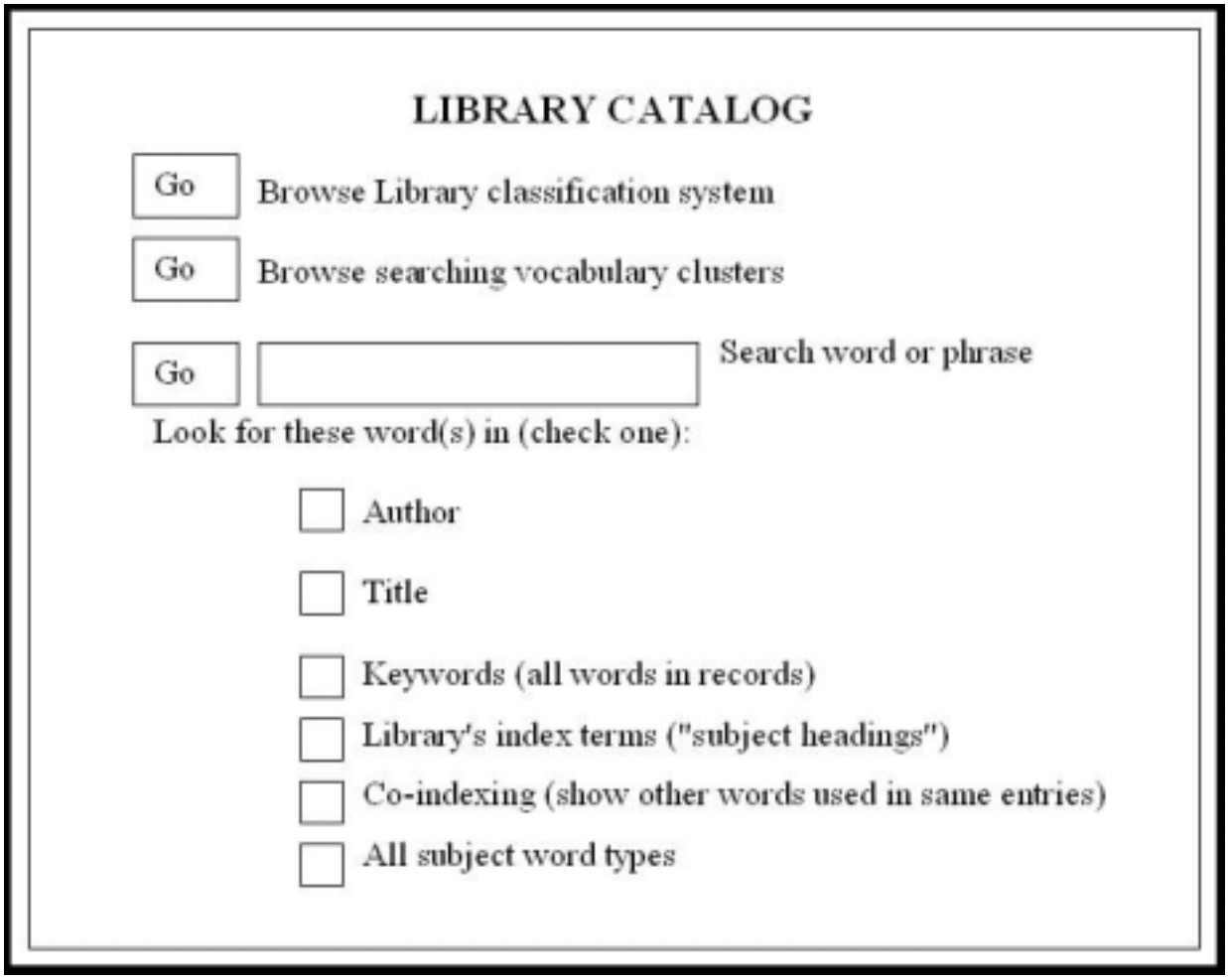

Figure 6: Example Initial Library Catalog Screen

If the searcher clicks on the vocabulary cluster box, he or she will see a display like that in Figure 7. At first, the searcher sees only the top part of the figure--a search box and the statement, "Find me search terms like these." Suppose that the searcher enters the term "equality." After clicking "Go," the searcher sees the bottom of Figure 7 on the screen. The system first states the number of matching clusters, then lists the cluster titles. After the cluster titles have been listed, the actual full clusters are to be found below on the screen. (See Figure 3 for an example cluster.) When the searcher clicks on a hyperlinked cluster, the system automatically takes the searcher to the part of the cluster list where the full text of the cluster appears.

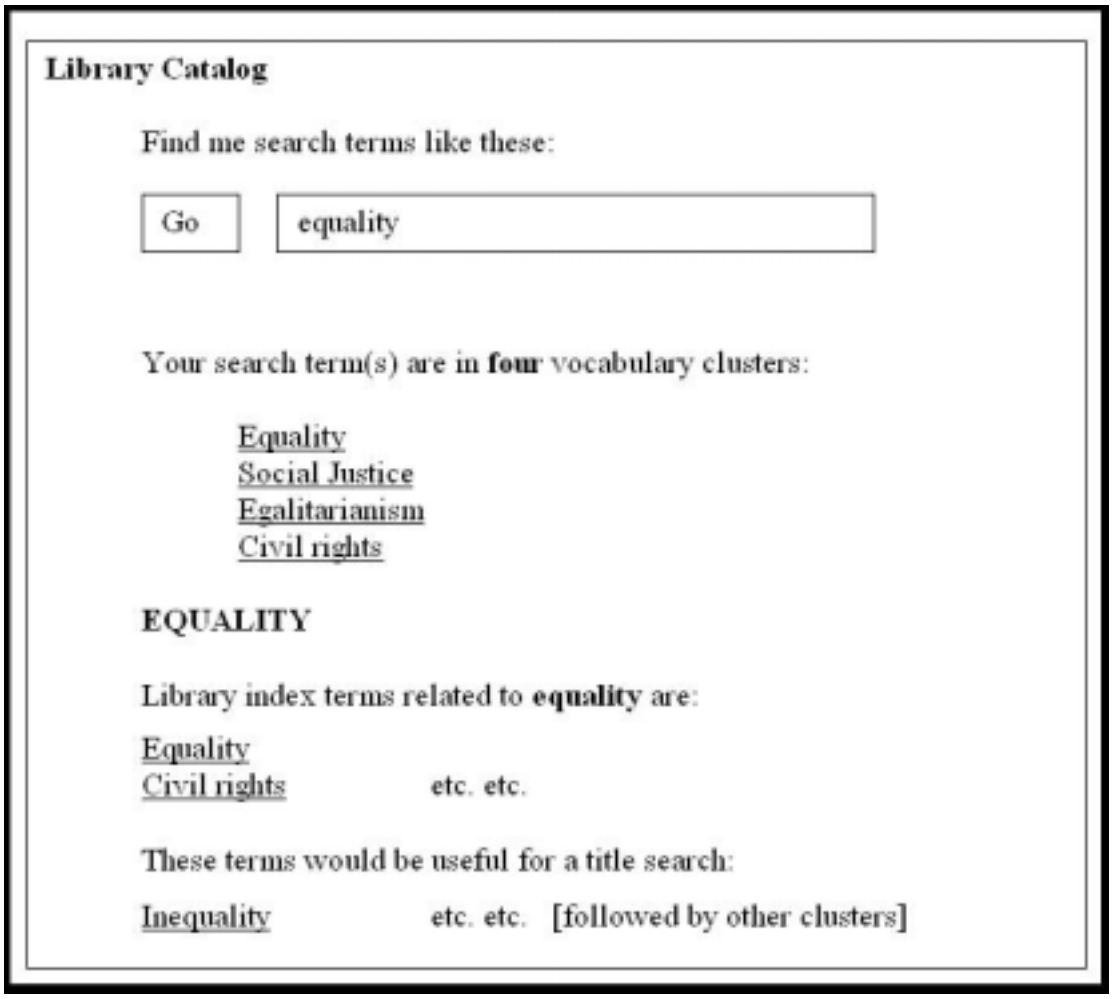

Figure 7: Next Screen If Click on "Browse Searching Vocabulary Clusters"
Screen at first displays top question and search box. When searcher enters term in Search box and clicks "Go," the rest of the screen contents appear.

It would be easy to use a simple algorithm to determine the order of the clusters in the display for the searcher. For instance, where the cluster title is an exact match with a user search term, then that cluster comes up first. Exact and partial matches with terms within the clusters could then be ranked by numbers of term matches within each cluster, and so on.

If the searcher instead enters a search term and clicks "Go," the search will be done on the desired word or phrase against the chosen index. (The requirement that the searcher check one box could also be handled by having a pull-down menu of the various index options. See the new "MELVYL-T" interface if the University of California catalog http://melvyl.cdlib.org). Options given are Author, Title, Keywords, Subject headings, Co-indexing, and All subject word types.

Co-indexing refers to identifying and counting the frequencies of other words in the first 20-50 records that respond to the stated search word. This is a way of expanding the search by *syntactically related* terms, whereas the search that pulls up vocabulary clusters is showing groupings of *semantically related* terms. Syntactically related terms

are those that appear in use in sentences near the chosen term. The assumption is that such terms may be semantically different, but, since they touch on topics that co-occur with the chosen term in document records, they may be productive and helpful for the searcher. (Again, co-indexing is not a part of the charge for this report, but it does represent an effective additional search feature to provide for users.)

If the searcher, when at the screen in Figure 6, instead enters an author name and clicks on the "Author" box, the screen in Figure 8 ("Next Screen If Click on Author Box…") may appear. In Figure 8, the system tells the searcher how many hits there are in the catalog. In accordance with the discussion in section 4B, whenever a bibliographic family has been organized and arrayed into groupings, the catalog screen will give the searcher the option of seeing all these related records pulled together and organized, whenever any one of the included records is pulled up in a search. The actual categories by which the works would be grouped would consist of a set such as suggested by Carlyle (see Section 3B) or other authors. For the vast majority of records--those that are singletons or those in families that have not yet been explicitly grouped--there would be no mention of groupings, and this screen would look like any other screen resulting from a conventional author search in an OPAC.

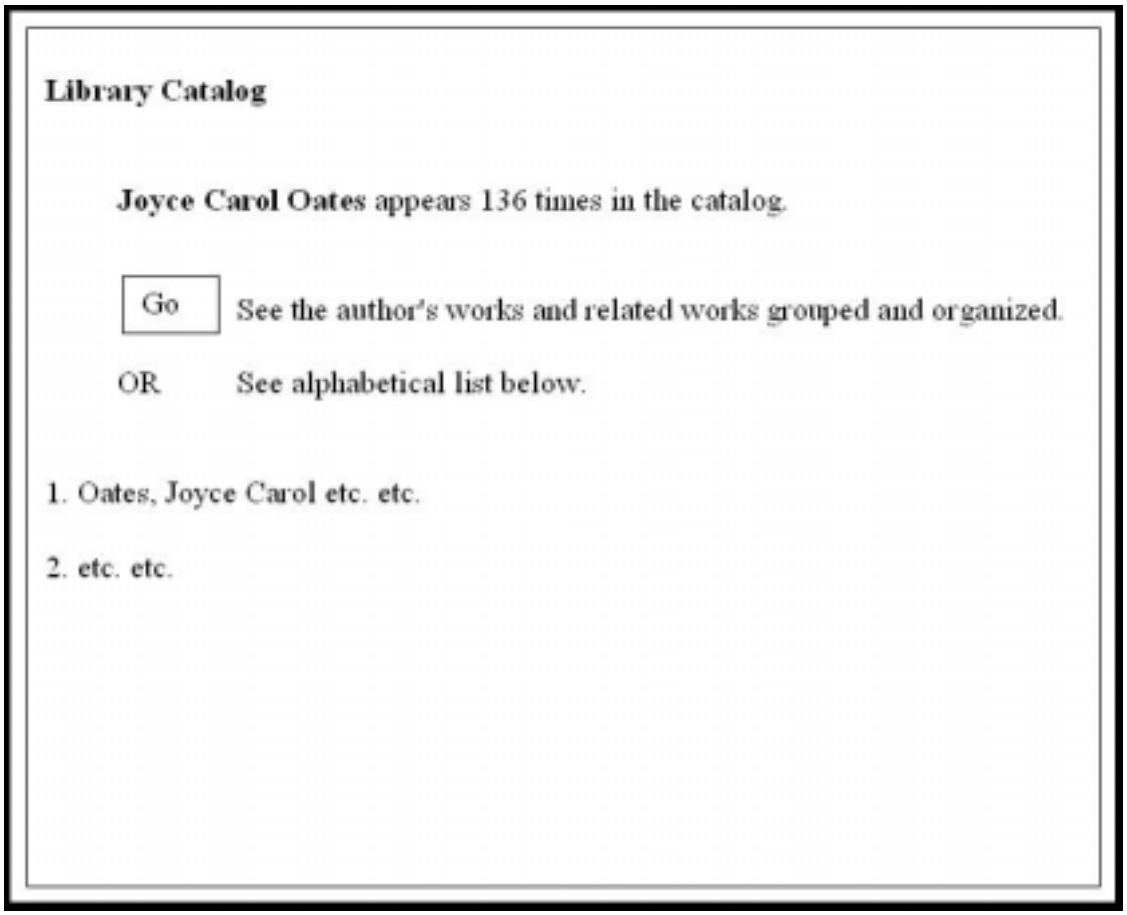

Figure 8: Next Screen If Click on Author Box and Enter Author Search Name
The option to see the author's work grouped and organized appears on the screen only for those authors whose work has been treated as a bibliographic family.

Finally, the principles discussed in Section 4C, regarding the sequencing of screen contents should be kept in mind in any design plan. Does the amount of text about any one meaningful record go up from screen to screen by roughly a factor of 30? For instance, in the Joyce Carol Oates example, by grouping large bibliographic families by their constituent sub-groupings, the searcher may not feel so overwhelmed by the bulk of material from voluminous authors. Thus, in the Oates case, the searcher who clicks the box to see the works grouped and organized, may go from entering her name on one screen, to seeing on the next screen, not the first of a list of 136 records, but rather something like the set of types of groups suggested by Carlyle or others (see list in Section 3B). These would fit on one page, and could be hyperlinked to pages providing the grouped records under each hyperlink. Such an intermediate step would not be necessary for the typical singleton or doubleton record. But in the cases where there are large numbers of records, the jump from the starting page to long lists of undifferentiated works violates the 1:30 ratio suggested by Resnikoff and Dolby (1972).

Thus, in these simplified example screen shots, we can see how the ideas expressed throughout section 3 could be implemented--clusters of related vocabulary, clusters of related records in voluminous authors and works, and following the 1:30 sequencing of Resnikoff and Dolby.

4E. Drawing the Threads Together

In discussing the Internet, Barabási (2002) is at pains to demonstrate that the arrangement of links and nodes is not random. Rather, there is a power-law tendency to create a wide range of sizes of websites and of numbers of links to any one website. In different ways, all three topics discussed in this report involve the recognition of underlying statistical patterns, with important implications for human activities in relation to information.

Throughout the three areas of discussion, there is evidence of a need to recognize a level of grouping of bibliographic and other information records that lies above the information individual. Searchers can best search (it is argued) when they are shown the context of other related terms around their concept, i.e., the group of terms of which their chosen term is a part. Traditionally, cataloging practice has been based on providing a few links between records or individual subject headings. What we are seeing here, perhaps, is the beginning of a way of looking at subject terms that recognizes a larger group, a "society" as it were, of terms that are the fellows to any initial search term.

Further, the sections above that discuss bibliographic families recognize the growing need to process records at the family (group) level, as well as the individual level. Now that bibliographic and technological means are making it easier to group related materials, and, as the size of information resources continues to grow explosively, we are finding it increasingly valuable to recognize groupings of closely related records. Once the searcher finds one, all may be presented in the interface. The searcher no longer has to follow up dozens of individual leads, but can go to one place where the related

items can be found collected together. Somewhere, Cutter must be smiling at the thought.

The sections on staging of access to information suggests that in the sequence from brief entry to full text, the ratio of amount of information should be on the order of 1:30 in size with a range of 5 to 161. Where there are more than two layers of information, the further steps down into the full text should also follow the 1:30 ratio. Though these ratios may sound improbably regular, they probably also follow from some of the same or--similar--underlying statistical regularities that the power laws express. Resnikoff and Dolby's research (1972) provides a number of supports for why this is a good ratio.

Finally, several commentators noted the importance of thoroughly user-testing the interfaces that would manifest these various improvements for the searcher. Having devoted much of my professional life to questions of improving the design of information systems, I could not agree more. However, one concern expressed was that the development and testing of suitable interfaces would be a costly and time-consuming process. I would reply by saying that interface design *can* be done relatively quickly, effectively, and inexpensively, provided the design process is appropriately constituted.

The biggest problems usually come when a system is evaluated after being completely developed. By this time in the process, changes are extremely costly, egos are invested, and a satisfactory result is difficult to achieve. Instead, systems should be user-tested throughout the development process, before expensive investments have been made. In this way, an excellent system evolves organically with the overall design/development process. Effective testing does not require vast numbers of human subjects. Jakob Neilsen, a well-known user-testing guru, recommends as few as three to five test subjects for some conditions (Neilsen,1993).

5. Summary

5A. Summary of Review of Information Seeking Literature

Here are summary points from sections 2A-2D:

• People use the Principle of Least Effort, preferring easy-to-get information over harder-to-get information, no matter how high the quality of the latter, as a rule. They like browsing best of all, and, for the most part, are quite unself-conscious about their information seeking behaviors.

• In matters of great urgency or of great interest, people will sometimes invest a great deal of energy and interest in finding information, and will become skilled at doing so.

• People respond to what the power figures in their lives encourage, and tend to model their information seeking strategies after that of trusted or esteemed people in their lives, rather than responding to brief educational encounters with reference librarians.

• With few exceptions, people like online public access catalogs a lot.

• The most problematic aspect of OPAC searching is generally subject searching. People have a lot of no-match or poor-match hits when searching for subject, and have learned to use keyword searching as a substitute for difficulties matching up with relevant subject headings. Yet they still like to do subject searching online.

• Subject searching has grown with the advent of online catalogs, generally constituting over half of the uses. Many of the improvements people want in OPACs concern subject searching.

• Searching methods, even by highly educated professionals, tends to be simple--one or two word queries, use of Boolean logic rare, modification of terms rare.

• People want their library catalogs to provide much more than just a catalog; they want library portals, in effect.

• Research on Web-based OPACs and Web searching in general is in its early stages. There is some indication that people like to be able to access catalogs remotely and download those results to their own computers.

• Proposed IFLA OPAC display guidelines include many principles that conform to design possibilities in the three areas being considered here (Yee, 1999).

• It is time to be open to dramatically different design models for catalogs and portals--well-controlled and -structured bibliographic environments that enable searchers to feel like they are simply browsing and encountering a lot of useful and interesting material relevant to their queries.

5B. Summary of Review of Research Specific to the Three Issues Addressed

Here are some summary points from sections 3A-3C:

• An extensive body of research done in a wide variety of contexts confirms that people use a very large number of terms for any given subject/topic. On average, no one of those terms is used by more than 20-30 percent of the test group--and the number of overlaps is often still lower.

• People can *recognize* information much easier than they can *recall* it. The meaning of a single term is enriched and clarified for they user by being placed in the context of related terms.

• There is a lot of activity in adding indexer thesauri to various types of online systems; testing of these systems, however, is almost non-existent.

• The idea of the end-user front-end vocabulary has been designed and developed a couple of times, but has essentially not been user-tested, except through the practical experience of numerous search intermediaries.

• With the advent of easy-to-make and easy-to-follow electronic links, there has been growing interest in the cataloging world in the development and presentation in OPACs of larger descriptive cataloging groupings, of "bibliographic families."

• These linked groupings of bibliographic families have been defined and conceptualized by Tillett, Smiraglia, Leazer, Yee, and the Functional Requirements for Bibliographic Resources, among others. Carlyle has proposed ways in which several types of these groupings can be related in the OPAC interface.

• The statistical pattern underlying these groupings is almost certainly Bradford's Law. As Bradford's Law has well-known characteristics ("self-similar" distribution with very long tails), we may be able to take advantage of those characteristics to achieve relatively inexpensive improvements in the catalog.

• There is a lot of evidence that catalog users would like additional subject information to be provided in the cataloging record, specifically, a summary or abstract of the item.

• This expectation on the part of users falls within a larger statistical pattern identified by Howard Resnikoff and James Dolby some years ago. They identified many situations in which the steps from briefer down into more extensive information fell into ratios of 1:30.

• Independent research by librarians and human-computer interaction researchers confirms or conforms with the Resnikoff & Dolby research.

5C. Summary of Recommendations

It is recommended that with regard to access vocabulary:

• A cluster vocabulary be created, based on the searcher vocabulary developed by Sara Knapp (1993, 2000), if she and her publisher agree.

• For the price of a share of the maintenance of the database, libraries and commercial firms may subscribe to the searcher vocabulary database, and install it in their catalogs, portals, and websites.

• With experience, other types of clusters are added--for names, works, geographical locations, etc.

• Access to catalogs and portal information should be available both directly through and around the vocabulary database. In this way, searchers may choose to use the database or not, and, if they do choose it, they do not have to enter and exit a separate database (a violation of the ever-present Principle of Least Effort).

• Institutional users may link the searcher vocabulary with their own controlled vocabulary. As a result, users of these sites may input their search term(s), be shown a cluster of terms, including "legitimate" controlled terms, and use the clusters as a basis for selecting terms for either controlled vocabulary or keyword searching.

• With this vocabulary as a core, one or two lexicographers are hired cooperatively to maintain the searcher vocabulary, adding popular new terms as they come along, and adding terms found by cooperating organizations in "zero hit" searches. As changes are made in the vocabulary, rather than in millions of individual cataloging records, cultural and research changes can be accommodated much more rapidly and cheaply.

• These vocabularies become part of a "Vocabulary Headquarters" (VHQ) website, supported by the library community or organizations therein.

It is recommended that with regard to bibliographic families:

• Preliminary agreement be gained on what shall constitute bibliographic families at the work level, probably based on the work of Tillett, Smiraglia, Hickey, and others. It may be found that work-sets, as described by Hickey et al. (2002), should also be considered.

• As these bibliographic families probably follow the Bradford Distribution, there will be some few that are very large, and many that are very small or singletons. As the larger families are much more likely to cause difficulties for searchers, and as they are also often around canonical works that attract a great deal of research and cultural interest, the larger families should be grouped first.

• At first on an experimental basis, individual libraries or other institutions offer each to do the work to collect just one large family (from records already created at the individual level). The results of these experiences are shared at conferences and other meetings.

• Based on these experiences, criteria are finalized for the creation of bibliographic families. Libraries may acquire the cataloging information for the families in a manner similar to the currently existing cooperative cataloging arrangements.

• Further experience will also provide enlightenment regarding just how far down the chain of family size the cooperative effort should go.

• Eventually, with further technological advances, it becomes possible that whenever a searcher happens on a record that is part of a bibliographic family, the searcher may click on a "related records" link and see displayed on the screen the progenitor record plus links to all the different types of bibliographically related records arrayed around the core record.

It is recommended that with regard to staging of access to records:

• Libraries and other information institutions take as an objective the approach of providing staged access to information that drops down into the information in a 1:30 ratio. For example, in a catalog a book has a title of a few words, and an abstract of about 30 times the number of words in the title. With this ratio specifically in mind, the effectiveness of catalogs so designed can be tested.

• Current cooperation with publishers can be extended, including use of book flap and contents information that is already in electronic form for catalog records.

• The online bookstore, amazon.com, contains within it many of the design features that have been recommended by catalog and database user studies over the years. Amazon.com can be seen as a source of ideas and prior testing of design features.

Acknowledgements

I wish to thank the commentators for their careful review and thoughtful comments on an earlier draft of this report. The report is much improved as a result. These included the members of the ALA ALCTS Metadata Task Force, Judith Ahronheim, Task Force Leader; Priscilla Caplan, Mary Woodley, Amy Tracy Wells, Sara Shatford Layne, Shelby Harken, Carolyn Larson, Barbara Tillett, Shirley Hyatt, as well as the Library of Congress Reactors Group.

Bibliography

Allen, T. J. (1979). Managing the Flow of Technology: Technology Transfer and the Dissemination of Technological Information within the Research and Development Organization. Cambridge, Mass.: MIT Press.

Allen, B. (1993). Improved browsable displays: An experimental test. Information Technology and Libraries, 12(2), 203-208.

Babu, B. R., & O'Brien, A. (2000). Web OPAC interfaces: An overview. The Electronic Library, 18(5), 316-327.

Barabási, A.-L. (2002). Linked: The New Science of Networks. Cambridge, Mass.: Perseus.

Bates, M. J. (1977a). Factors affecting subject catalog search success. Journal of the American Society for Information Science, 28, 161-169.

Bates, M. J. (1977b). System meets user: Problems in matching subject search terms. Information Processing & Management, 13, 367-375.

Bates, M. J. (1986a). Subject access for online catalogs: A design model. Journal of the American Society for Information Science, 37, 357-376.

Bates, M. J. (1986b). What is a reference book? A theoretical and empirical analysis. RQ, 26, 37-57.

Bates, M. J. (1989a). Design of browsing and berrypicking techniques for the online search interface. Online Review, 13, 407-424.

Bates, M. J. (1989b). Rethinking subject cataloging in the online environment. Library Resources & Technical Services, 33, 400-412.

Bates, M. J. (1990a). Where should the person stop and the information search interface start? Information Processing & Management, 26, 575-591.

Bates, M. J. (1990b). Design for a subject search interface and online thesaurus for a very large records management database. Proceedings of the 53rd ASIS Annual Meeting, 27, 20-28.

Bates, M. J., Wilde, D. N., & Siegfried, S. (1993). An analysis of search terminology used by humanities scholars: The Getty Online Searching Project Report No. 1. Library Quarterly, 63(1), 1-39.

Bates, M. J. (1994a). The design of databases and other information resources for humanities scholars: The Getty Online Searching Project Report No. 4. Online & CDROM Review, 18(6), 331-340.

Bates, M. J. (1994b). Expanded Entry Vocabulary for the Library of Congress Subject Headings: A Final Report (CLR Grant # 891). Washington, DC: Council on Library Resources.

Bates, M. J. (1996a). Document familiarity in relation to relevance, information retrieval theory, and Bradford's Law: The Getty Online Searching Project Report No.5. Information Processing & Management, 32, 687-707.

Bates, M. J. (1996b). Getty end-user Online Searching Project in the humanities: Report No.6: Overview and conclusions. College & Research Libraries, 57, 514-423.

Bates, M. J. (1998). Indexing and access for digital libraries and the Internet: Human, database, and domain factors. Journal of the American Society for Information Science, 49(13), 1185-1205.

Bates, M.J. (2002a) The cascade of interactions in the digital library interface. Information Processing & Management, 38, 381-400.

Bates, Marcia J. (2002b) "Speculations on Browsing, Directed Searching, and Linking in Relation to the Bradford Distribution," Emerging Frameworks and Methods: Proceedings of the Fourth International Conference on Conceptions of Library and Information Science (CoLIS4), Edited by Harry Bruce, Raya Fidel, Peter Ingwersen, and Pertti Vakkari. Greenwood Village, CO: Libraries Unlimited, 2002, pp. 137-150.

Bates, M.J. (2002c) Toward an integrated model of information seeking and searching. The New Review of Information Behavior Research, 3, 1-15.

Beaulieu, M. (1997). Experiments on interfaces to support query expansion. Journal of Documentation, 53(1), 8-19.

Bertha, E. (1993). Inter- and intrabibliographical relationships: A concept for a hypercatalog. In A. H. Helal & J. W. Weiss (Eds.), Opportunity 2000: Understanding and Serving Users in an Electronic Library (pp. 212-223). Essen, Germany: Universitätsbibliothek Essen.

Bibliographic Control of Web Resources: A Library of Congress Action Plan. (Dec. 24, 2002) http://lcweb.loc.gov/catdir/bibcontrol/actionplan.pdf.

Blazek, R. (1971) Teacher Utilization of Nonrequired Library Materials in Mathematics and the Effect on Pupil Use. Ph.D. Dissertation. Champaign-Urbana: University of Illinois.

Borgman, C. L. (1986). Why are online catalogs hard to use? Lessons learned from information-retrieval studies. Journal of the American Society for Information Science, 37(6), 387-400.

Borgman, C. L. (1996). Why are online catalogs still hard to use? Journal of the American Society for Information Science, 47(7), 493-503.

Bradford, S.C. (1948) Documentation. London: Crosby Lockwood.

Brajnik, G., Mizzaro, S., Tasso, C., & Venuti, F. (2002). Strategic help in user interfaces for information retrieval. Journal of the American Society for Information Science and Technology, 53(5), 343-358.

Brookes, B.C. (1977) Theory of the Bradford Law. Journal of Documentation, 33, 180-209.

Busch, J. A. (1998). Building and accessing vocabulary resources for networked resource discovery and navigation. In P. A. Cochrane & E. H. Johnson (Eds.), Visualizing Subject Access for 21st Century Information Resources (pp. 148-156). Urbana-Champaign, Ill.: Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign.

Carlyle, A.(1997) Fulfilling the second objective in the online catalog: Schemes for organizing author and work records into usable displays. Library Resources & Technical Services, 41(2), 79-100.

Carlyle, A. (1989). Matching LCSH and user vocabulary in the library catalog. Cataloging & Classification Quarterly, 10(1-2), 37-63.

Chan, L.M. (2000) Exploiting LCSH, LCC, and DDC to Retrieve Networked Resources. Bicentennial Conference on Bibliographic Control for the New Millennium. Available: http://lcweb.loc.gov/catdir/bibcontrol/chan.html [June 1, 2003]

Chen, C., & Hernon, P. (1982). Information Seeking. New York: Neal Schuman.

Cochrane, P. A. (1983) A paradigm shift in library science. Information Technology and Libraries, 2(1), 3-4.

Cochrane, P. A., & Markey, K. (1983). Catalog use studies--Before and after the introduction of online interactive catalogs: Impact on design for subject access. Library & Information Science Research, 5(4), 337-363.

Coleman, J., Katz, E., & Menzel, H. (1967). Medical innovation: A Diffusion Study. New York: Bobbs-Merrill.

Connaway, L. S., Budd, J. M., & Kochtanek, T. R. (1995) An investigation of the use of an online catalog: User characteristics and transaction log analysis. Library Resources & Technical Services, 39(2), 142-152.

Cothey, V. (2002). A longitudinal study of World Wide Web users' information-searching behavior. Journal of the American Society for Information Science and Technology, 53(2), 67-78.

Dennis, S., Bruza, P., & McArthur, R. (2002). Web searching: A progress-oriented experimental study of three interactive search paradigms. Journal of the American Society for Information Science and Technology, 53(2), 120-133.

Dervin, B. et al. (1976) The Development of Strategies for Dealing with the Information Needs of Urban Residents: Phase I; Citizen Study. Final Report. ERIC ED 125 640.

Dolby, J.L., & Resnikoff, H.L. (1971). On the multiplicative structure of information storage and access systems. Interfaces; The Bulletin of the Institute of Management Sciences, 1, 23-30.

Drabenstott, K.M. (1991). Online catalog user needs and behavior. Think Tank on the Present and Future of the Online Catalog: Proceedings. RASD Occasional Papers, No. 9. Chicago: American Library Association.

Drabenstott, K. M., & Weller, M. S. (1996). The exact-display approach for online catalog subject searching. Information Processing & Management, 32(6), 719-745.

Efthimiadis, E.N. (1996) Query expansion. Annual Review of Information Science and Technology, 31, 121-187.

Ellis, D. (1989). Behavioural approach to information retrieval system design. Journal of Documentation, 45(3), 171-212.

Ellis, D., & Haugan, M. (1997). Modelling the information seeking patterns of engineers and research scientists in an industrial environment. Journal of Documentation, 53(4), 384-403.

Fenichel, C.H. (1981) Online searching: Measures that discriminate among users with different types of experience. Journal of the American Society for Information Science, 32(1), 23-32.

Frarey, C. J. (1953). Studies of use of the subject catalog: Summary and evaluation. In M. F. Tauber (Ed.), Subject Analysis of Library Materials (pp. 147-166). New York: Columbia University, School of Library Service.

Furnas, G. W., Gomez, L. M., Landauer, T. K., & Dumais, S.T. (1982). Statistical Semantics: How Can A Computer Use What People Name Things to Guess What Things People Mean When They Name Things? Proceedings of the Human Factors in Computer Systems Conference, Gaithersberg, Maryland.

Galvin, T. J., & Kent, A. (1977). Use of a university library collection: A progress report on a Pittsburgh study. Library Journal, 102, 2317-2320.

Hafter, R. (1979). Performance of card catalogs: A review of research. Library Research, 1(3), 199-222.

Hancock, M. (1987). Subject searching behaviour at the library catalogue and at the shelves: Implications for online interactive catalogues. Journal of Documentation, 43(4), 303-321.

Hickey, T.B., O'Neill, E.T., & Toves, J. (2002) Experiments with the IFLA Functional Requirements for Bibliographic Records (FRBR). D-Lib Magazine, 8(9). http://www.dlib.org/dlib/september02/hickey/09hickey.html.

Hildreth, C. R. ((1995) Online Catalog Design Models: Are We Moving in the Right Direction? : Council on Library Resources, 85 pp.

Hildreth, C. R. (1997). The use and understanding of keyword searching in a university online catalog. Information Technology and Libraries, 16(2), 52-62.

Hjerppe, R. (1985). Project HYPERCATalog: Visions and preliminary conceptions of an extended and enhanced catalog. Paper presented at the IRFIS 6 Conference, Frascati, Italy.

IFLA Study Group on the FRBR. (1998) Functional Requirements for Bibliographic Resources. München: K.G.Saur.

INFORUM: Subject Analysis Systems Collection. Faculty of Information Studies, University of Toronto. Available: http://www.fis.utoronto.ca/resources/inforum/sas.htm [October 1, 2002].

ISO/IEC 13250 (19 May 2002). Topic Maps: Information Technology Document Description and Processing Languages. 2nd ed. Available: http://www.topicmap.com [January 19, 2003]

Jansen, B. J., & Pooch, U. (2001). A review of Web searching studies and a framework for future research. Journal of the American Society for Information Science and Technology, 52(3), 235-246.

Knapp, S.(Ed.) (1993) The Contemporary Thesaurus of Social Science Terms and Synonyms; A Guide for Natural Language Computer Searching. Phoenix, Ariz., Oryx.

Knapp, S. (Ed.) (2000) The Contemporary Thesaurus of Search Terms and Synonyms: A Guide for Natural Language Computer Searching. Phoenix, Ariz.: Oryx.

Lagoze, C. (2000) Business unusual: How “event-awareness” may breathe life into the catalog? Library of Congress Bicentennial Conference on Bibliographic Control for the New Millennium, November 15-17, 2000. Available: http://lcweb.loc.gov/catdir/bibcontrol/lagoze_paper.html.

Larson, R. R. (1991). The decline of subject searching: Long-term trends and patterns of index use in an online catalog. Journal of the American Society for Information Science, 42(3), 197-215.

Layne, S.S. (1989) Integration and the objectives of the catalog. In E. Svenonius (Ed.), The Conceptual Foundations of Descriptive Cataloging (pp. 185-195). San Diego: Academic Press.

Lilley, O. L. (1954). Evaluation of the subject catalog. American Documentation, 5(2), 41-60.

Lombardo, S. V., & Condic, K. S. (2000). Empowering users with a new online catalog. Library Hi Tech, 18(2), 130-141.

Lutes, B. (1999). Web Thesaurus Compendium. Available: http://www.darmstadt.gmd.de/~lutes/thesalpha.html [October 1, 2002]

MACS Multilingual Access to Subjects Available: http://infolab.kub.nl/prj/macs/prototypee.html [January 17, 2003]

Mann, T. (1992). Library Research Models. New York: Oxford University Press.
Mann, T. (1997). 'Cataloging must change!' and indexer consistency studies: Misreading the evidence at our peril. Cataloging & Classification Quarterly, 23(1), 3-45.
Markey, K. (1988) Integrating the machine-readable LCSH in online catalogs. Information Technology and Libraries, 7, 299-312.
Matthews, J. R., et al (Eds.). (1983). Using Online Catalogs: A Nationwide Survey. New York: Neal-Schuman.
McGarry, D., & Svenonius, E. (1991) More on improved browsable displays for online subject access; Using records from ORION. Information Technology and Libraries, 10, 185-91.
Mick, C., Lindsey, G., & Callahan, D. (1980). Towards usable user studies. Journal of the American Society for Information Science (JASIS), 31, 347-356.
Middleton, M. (2001). Controlled Vocabularies. Queensland University of Technology. Available: http://www2.fit.qut.edu.au/InfoSys/middle/cont_voc.html [October 1, 2002].
Neilsen, J. (1993). Usability Engineering. Boston: AP Professional.
O'Neill, E.T., & Vizine-Goetz, D. (1989) Bibliographic relationships: Implications for the function of the catalog. In E. Svenonius, ed., The Conceptual Foundations of Descriptive Cataloging (p. 167-79). San Diego: Academic Press.
OCLC, Library of Congress. (January 8, 2002). Changes for FAST Subject Headings. Discussion Paper 2002-DP03. Available: http://www.loc.gov/marc/marbi/2002/2002-dp03.html [January 17, 2003]
Poole, H. L. (1985). Theories of the Middle Range. Norwood, New Jersey: Ablex.
Powell, R. R., Taylor, M. T., & McMillen, D. L. (1984). Childhood socialization: Its effect on adult library use and adult reading. Library Quarterly, 54, 245-264.
Proceedings of the Library of Congress Bicentennial Conference on Bibliographic Control for the New Millennium. (2001) Washington, D.C.: Library of Congress Cataloging Distribution Service.
Resnikoff, H.L., & Dolby, J.L. (1972) Access: A Study of Information Storage and Retrieval with Emphasis on Library Information Systems. Washington, D.C.: U.S. Department of Health, Education, and Welfare, Office of Education, Bureau of Research. (ERIC No. ED 060 921)
Rothstein, S. (1964). Measurement and evaluation of reference service. Library Trends, 12, 456-472.
Saracevic, T., & Kantor, P. (1988). Study of information seeking and retrieving. III. Searchers, searches, and overlap. Journal of the American Society for Information Science, 39(3), 197-216.
Schabas, A. H. (1982). Postcoordinate retrieval: A comparison of two indexing languages. Journal of the American Society for Information Science, 33, 32-37.
Shiri, A.A., Revie, C., & Chowdhury, G. (2002a). Thesaurus-assisted term selection and query expansion: A review of user-centred studies. Knowledge Organization, 29(1), 1-19.
Shiri, A.A., Revie, C., & Chowdhury, G. (2002b). Thesaurus-enhanced search interfaces. Journal of Information Science, 28(2), 111-122.
Shneiderman, B. (1998). Designing the User Interface: Strategies for Effective Human-Computer Interaction (3rd ed.). Reading, Mass.: Addison-Wesley.

Siegfried, S., Bates, M. J., & Wilde, D. N. (1993). A profile of end-user searching behavior by humanities scholars: The Getty Online Searching Project Report No. 2. Journal of the American Society for Information Science, 44(5), 273-291.

Small, H., & Garfield, E. (1985). The geography of science: Disciplinary and national mappings. Journal of Information Science, 11(4), 147-159.

Smiraglia, R. P. (1992). Authority Control and the Extent of Derivative Bibliographic Relationships. Unpublished Ph.D. dissertation, University of Chicago.

Smiraglia, R. P., & Leazer, G. H. (1999). Derivative bibliographic relationships: The work relationship in a global bibliographic database. Journal of the American Society for Information Science, 50(6), 493-504.

Smiraglia, R. P. (1999). Derivative bibliographic relationships among theological works. Proceedings of the 62nd Annual Meeting of the American Society for Information Science, 36, 497-506.

Smiraglia, R. P. (2001). The Nature of "A Work": Implications for the Organization of Knowledge. Lanham, Md.: Scarecrow Press.

Stoan, S. K. (1984). Research and library skills: An analysis and interpretation. College & Research Libraries, 45, 99-109.

Tillett, B. B. (1991a) Summary of the treatment of bibliographic relationships in cataloging rules. Library Resources & Technical Services, 35(4), 393-405.

Tillett, B. B. (1991b). A taxonomy of bibliographic relationships. Library Resources & Technical Services, 35(2), 150-158.

Tillett, B.B. (2001). Bibliographic relationships. In C.A. Bean and R. Green (Eds.), Relationships in the Organization of Knowledge. Boston, Kluwer Academic.

UNIMARC Format. (1980) 2nd ed. London: IFLA International Office for UBC.

Vellucci, S. L. (1997). Bibliographic Relationships in Music Catalogs. Lanham, Md.: Scarecrow Press.

VTLS Inc. Announces FRBR Implementation. (2002) Available: http://www.vtls.com/Corporate/Releases/2002/20020514b.shtml [January 18, 2003]

Warner, E. S., et al. (1973). Information Needs of Urban Residents (ERIC ED 088 464). Rockville, Maryland: Regional Planning Council, Baltimore, and Westat, Inc.

White, M. D. (2000). Questioning behavior on a consumer health electronic list. Library Quarterly, 70(3), 302-334.

Wiberley, S. E., Jr., Daugherty, R. A., & Danowski, J. A. (1995). User persistence in displaying online catalog postings: LUIS. Library Resources & Technical Services, 39(3), 247-264.

Wordnet. Available: http://www.cogsci.princeton.edu/~wn/ [January 18, 2003]

Yee, M. M. (1994). Manifestations and near-equivalents: Theory, with special attention to moving-image materials. Library Resources & Technical Services, 38(3), 227-255.

Yee, M. M. (1999). Guidelines for OPAC displays. ALCTS Online Newsletter, 10(6). Available: http://www.ala.org/alcts/alcts_news/v10n6/gateway_pap14.html

Younger, J. A. (1995). After Cutter: Authority control in the twenty-first century. Library Resources & Technical Services, 39(2), 133-141.