Exploring Categorization

Undergraduate Student Searching and the Evolution of Catalogs

By John Budd

Debate about the future of library catalogs and cataloging has been, and continues to be, featured in the literature of librarianship. Some research into the ways undergraduate students at one institution assign subjects to selected works provides insight into the cognitive elements of categorization. The design of catalogs can be informed by this research as well as work currently being done on alternative means of organization, such as information systems ontologies.

Some debates within librarianship endure even as particular aspects of the issue change. The future of cataloging is one such debate. Thomas points out that Osborn wrote of a crisis in cataloging in 1941.1 She goes on to review changes in cataloging as an operation (including moves to outsource processes) with an eye to the quality of bibliographic control. It is apparent from her review that the idea of “quality” has not been static. Recently, writings on cataloging have focused on the catalog as an effective access mechanism, amenable to users locating the items that they will find relevant. To that end, some alterations to cataloging practices have been proposed. For example, Ortiz-Repiso and Moscoso urge, “The traditional distinction between main and added entries must be banished.”2 Their principal point is that the Web allows much more flexibility than the physical catalog ever could, so the possibilities opened by the technology should be explored with greater alacrity.

In 1994, Franz et al. published the results of a study of end-users’ agreement with a group of catalogers on the meaning of subject headings. The interpretations of the meaning of current subject headings were consistent between the two groups about 40 percent of the time.3 These results suggest a question, though—is 40 percent agreement bad or good? In other words, is such agreement higher or lower than diverse groups’ agreement on the interpretation of the meaning of anything? The question related closely to the concern regarding catalogs’ effectiveness at helping searchers find relevant works.

Any conceivable solution to the problem of cataloging (assuming that a problem of cataloging exists) is complicated by the sheer amount of “stuff” that is being produced. The publication of books proceeds unabated; the production of journal articles grows and grows; the less formal creation and dissemination of texts, images, and sounds increases at tremendous rates. Cataloging, since its formal and operational inception, has been intended primarily to be used to describe and provide subject access to physical items. A physical item can be described (partly) in physical terms, including size, length, and other attributes. Items can also, however, be described according their creation (the individual, group, or body that is responsible for bringing it into being) and the entity that publishes or produces...
them. The latter elements of description may transcend the physicality of the work; that is, knowing who created a Web site, and under the auspices of what body, could be very useful. There is a tradition of disclosure of such data when it comes to physical items (not always observed, of course). The disclosure is one means of assisting searchers.

Virtual works do not follow standards of responsibility and production statements with nearly as much regularity as physical ones. That information seekers find elements of responsibility and production of utility is enforced by schooling at all levels. Teachers want students to acknowledge and evaluate such sources; in higher education, the requirement of citing sources of some discernible authority is frequently explicit. The subject access component of cataloging is more problematic than other kinds of description. The subject—the “aboutness”—of a work is only partially determinate insofar as a work employs explicit and unitary lexical semantics. As that rarely, if ever, occurs, a necessary interpretative act accompanies both subject description and information seeking.

A digital work does, without question, broaden the possibilities for searching the entirety of the text. In some instances (such as works employing a technical language), searching the text is a boon; in other instances (such as works employing figures of speech and idiomatic language), the searchable text may embody some self-obfuscation. The debate also has included discussion of the efficacy of controlled vocabularies, such as thesauri, as aids to locating useful works. Miller defines “thesaurus” in a way that is intended to allow for much flexibility, yet imposes as much stability as possible.4 He writes that it is “a lexico-semantic model of a conceptual reality or its constituent, which is expressed in the form of a system of terms and their relations, offers access via multiple aspects and is used as a processing and searching tool of an information retrieval unit.”5 The semantic, lexical, and relational aspects of a thesaurus of any kind contribute to its utility as a finding aid.

The debate includes a practical issue for libraries—the cost of cataloging as it entails the above functions. Having physical items, such as books, available to users as soon as possible after receipt is important to achieving service objectives. On the other hand, Harmon writes:

No library director in his or her right mind would dare tell the dean or the college president that “henceforth, the university library will be providing a lower quality of service for all its patrons and proud of it.” But this is exactly what they are saying when speaking of lean records, less bibliographic data, less authority control, "mark it and park it." No machine, no matter how advanced, can extract data that are not there.6

According to Harmon, a minimal approach signifies a dilemma for libraries, particularly with regard to providing access for users. Some others, including Marcum (Associate Librarian for Library Services, Library of Congress), are of a different opinion. As an anecdote, she relates an experience of conducting a Google search on “President Fillmore Foreign Policy”:

Never mind that the first five references include articles from Encarta and LookSmart that come with commercial advertisements. Never mind that the second reference is a sketch about Fillmore by, quote, “Caroline,” last name not given, who turns out to be a Pocantico Hills School fifth grader. And never mind that the fifth reference gives some information on Fillmore from a decade-by-decade outline of events, provided by some unidentified individual. . . . So, is it any surprise that many students just go Googling instead of to the library, virtual or physical, and use whatever turns up first in the keyword search?7

She goes on to use Google as an example of an effective search tool for digital or digitized works.

Further Context

With the foregoing as something of a context, the task of the project reported here is an examination of the possible search strategies of a group of undergraduate students. Before describing and providing the results of the study, a more detailed painting of the contextual picture is revealed in two exchanges that have appeared in the literature of librarianship. In 1991, Gregor and Mandel argued that some quite drastic changes to cataloging practice should occur.8 They began with a premise that “It is now possible to give up some time-consuming practices that were never based on knowledge of the worth to users.”9 They then offered some evidence on which they would base their conclusion.

Bates cites studies that, taken together, indicate that the average likelihood that any two people will use the same term for a concept or a book is 10 to 20 percent.10 Based on that evidence Gregor and Mandel stated, “It is necessary to recognize what is unachievable, and to understand that the nature of online subject searching will compensate.”11 A few years after the appearance of that article, Mann examined the evidence that Gregor and Mandel employed.12 For example, he pointed out that many of the studies Gregor and Mandel invoke actually studied that agreement of the interpretation by non-librarians. People may use inconsistent terminology as they try to infer meaning via subjects. Mann maintained:
Bates’s overall point is that there is a general pattern to the way the human mind works, and that this pattern of inconsistent use of terminology in describing a subject is probably valid because it shows up in so many different contexts. . . . What this suggests to me as a reference librarian is that, given such a human tendency, we as librarians need to find ways to solve these problems of inconsistent use of terminology.”

The second exchange began with Calhoun’s publication of the report, “The Changing Nature of the Catalog and Its Integration with Other Discovery Tools.” The report, commissioned by the Library of Congress, stated the Association of Research Libraries members spent about $239 million in 2004 on technical services labor. On the face of it, that datum does give one pause. Also, it suggests that there may be some unnecessary duplication of effort and resources. A portion of Calhoun’s report was based on twenty-three interviews of librarians and other information professionals, including vendors. The interviewees tended to affirm her position. She concluded that “The catalog is in decline, its processors and structures are unsustainable, and change needs to be swift; today, the online catalog is losing appeal for students and many scholars,” and “The declining demand for today’s catalogs reflects diminishing interest in already low-use research library collections, at least as they are currently housed, managed, and delivered.”

Some inferences can be drawn from the report: research libraries are placing resources where users’ needs will not be served effectively; cataloging and access should change according to the work habits of faculty and students; and automatic categorization (classification and subject headings) can result in cost savings. Once again, Mann responded. He took issue with Calhoun’s emphasis on an inherent business aspect to cataloging and other library operations. In other words, he disagreed with the attention to efficiency, perhaps at the expense of effectiveness. His principal point in his rebuttal was that the measures adopted to cater to quick searching are inimical to serious scholarship, but the reverse is not the case. He brought home his point by saying, “Left to their own devices—i.e., without any prior instruction or education—they will always find only ‘something’ rather than an overview of the full range of material available to them.”

The Study

During the fall 2005 and spring 2006 semesters, students enrolled in the Library Research course offered at the University of Missouri-Columbia (MU) were surveyed. The surveys were administered at the beginning of each semester. The course used a Blackboard Web site to supplement discussion and materials, such as tutorials. The Blackboard site also has a survey function that was used to disseminate the present survey. At the time the survey was dissemination, students had been introduced to the library’s catalog, so they were considering the structure and use of the catalog. The students—all undergraduates, and most freshmen—were asked the following:

You are looking for a book in Ellis Library [MU’s main library]. While you are given the title here, you do not know this title when you search the library’s catalog. List two ways you would categorize this book (that is, provide two subjects for the book) based on the title below. The subjects should be ones that you could then use to search the library’s catalog. In short, two things that describe what the book is about.

The purpose of the request was to investigate how students, as novice searchers, conceive of descriptions of the content of works. The students had approximately two weeks to respond, and they were free to use any resources at their disposal to select the two ways of categorization, including the catalog. The Blackboard site also allowed the students to respond over a period of time before they officially submitted their responses (after which no changes could be made). A total of 405 students responded. The titles (along with the subject headings as they appear in the MU Library’s catalog) that the students were asked to categorize are listed in figure 1. While there is some artificiality to the exercise, it does garner information about the cognitive strategies that some undergraduate students employ when categorizing works.

Findings

For the first book by Johnson, only two students suggested “Knowledge—Theory of” and two suggested “Philosophical anthropology.” While none mentioned “Objectivity,” 174 of them did say that they would search “objective” as a subject heading. Students did offer some other possibilities; the most frequently mentioned were history, culture, philosophy, sociology, psychology, mind, and thinking. With the Gaddis book, the subject headings “History—Philosophy” and “History—Methodology” were each listed by one student; “Aesthetics—History” did not appear in any student’s suggestions. Other subjects mentioned were history, landscape, past, map, and geography. The Hartman title presented at least some clarity for students, 309 of whom assigned “Ethics” as a subject. None, however, came up with “Business ethics” or “Corporate culture.”

Other subjects occurring included organization(al), good life, and morality. Likewise, the work by Coyne
lent itself to the mention of an assigned subject heading, “Information technology,” as a subject (103 times). Another 209 respondents listed “technology” alone. One included “System design,” but eighty-six mentioned “design.” The other subject listed with some frequency was “postmodern.” The book by Greenblatt was assigned several recurring subjects, but few of the ones that were in the MU catalog record. Seven students listed “Reading,” but there was only one mentioned each of “Literature—Philosophy” and “Criticism.” “Literary theory” and “transactional theory” were listed most frequently. Also, theory, literature, and poetry appeared in respondents’ lists.

**Analysis**

Taken as a whole, the kinds of subjects listed by the respondents demonstrate a limitation—many of the terms are taken from the titles of the books. Other terms that are mentioned with some frequency can usually be inferred from the titles. The limitation is essentially a cognitive one. As was mentioned earlier, one component of categorization is the representation of what is real, or ontology. It is the ontological aspect of categorization that can offer stability to any controlled vocabulary. To the extent that ontological categorization is possible (and the consistency of the lexical semantics of a work contributes to the possibility), descriptive subject terms represent works. The “discovery” of subject terms (as opposed to the “invention” of subject terms) necessitates a way of thinking that presumes at least a somewhat stable reality.

Another way of stating this point is that, as the students categorize the five works, they appear to be seeking a particular kind of disclosure. Disclosure is defined as the interpretability of meaning from reading apparent linguistic evidence. The simplest disclosure may occur in the form of terms in titles. This kind of disclosure constitutes a logical process, albeit a rather naive logic. The naiveté becomes evident, and can result in error, when “geography” is listed as a potential category for the Gaddis book, based on the title word “map.” The categorization seems to represent what might be known about the works, as such knowledge is disclosed through terms in the titles. It is through the inferred disclosure that the students engage in a cognitive process based on discovery; the titles reveal ideas, knowledge, and reality to be discovered.

“Reality” appears in the preceding sentence, but it is different from “idea” and “knowledge,” at least insofar as many conceptions of idea and knowledge are separate from those of reality. There is by no means anything resembling universal agreement on the connection, or lack thereof, of reality with knowledge and idea. Returning to ontology, one sees that another way of describing reality is by attempting to grasp the essential substance of something (this is an important component of Aristotle’s *Categories*). Those essential substances are the features upon which other, secondary substances follow. This is a complicated notion, but a hint of its application is given in the structure of Library of Congress Subject Headings. For example, a subject heading assigned to the work by Rosenblatt is “Literature—History and criticism—Theory, etc.” (although this is not included in the MU record). The logic of the subject heading’s structure is that “Literature” is essential, and “History and criticism” and “Theory, etc.” follow from it (or are secondary to it). The notion of categorization is more complicated still. Attention in librarianship and other information fields is presently
given to ontological description and categorization. This is a promising development, and it will be revisited in the next section. Ontological categorization includes some requirements, though.

Librarianship cannot ignore what philosophers have had to say about ontology, especially if it is to provide a basis for categorization in systems, perhaps including catalogs in the future. For example, almost a century ago Husserl expounded on the logical processes and necessities of ontology. For one thing, material ontology is the study of the essences of physical things. Linnaean taxonomy is an instance of a kind of material ontology. According to Husserl, there is also formal ontology, which is the study of the essences of any thing, including abstract ideas. Husserl suggests that material ontology relies on “eidetic reduction” (from the Greek eidos, meaning essence). This essence is not bound to any single temporal representation (or token) of a thing, which means that a particular triangle embodies an essence that is defined by the properties of “triangle.” Husserl goes so far as to say, “Every factual science (empirical science) has essential theoretical essences in eidetic ontologies.”

Not everyone agrees with Husserl's reduction to essences, but it does offer a beginning. It is difficult, if not impossible, to apply only material ontology to the content of works such as the five books used in this study. A full understanding of things, for Husserl, requires reflection; the experiencing of things (including abstract ideas) adds a layer of richness to the perception of the reality of those things. Reflection rounds out what Husserl calls a phenomenological analysis, and this analysis is what many people have in mind regarding information systems ontologies. Reading a book necessitates interpretation of possible meaning, so the experience of reading transcends the physical book or the physical act of viewing images or hearing sounds. The reflective act is transformative; “every category of 'reflexion' has the character of a modification of consciousness.”

Both the ontological reduction and the transcendental reduction are necessary to realize the intended outcome of categorization, or the understanding of the meaning of something. Further, the reductions are promising practices in the development of systems that can assist searchers as they seek works that relate to their queries. More fundamentally, instruction in the practices of reduction can help searchers formulate their queries.

An extrapolation from the categories provided by the respondents in this study helps to illustrate the complexity related to achieving understanding. A limitation of the study is that the instruction in the reductions, just mentioned, had not occurred at the time the students were surveyed. The limitation was intentional; the purpose of the survey, as stated previously, was to gain understanding of the ways undergraduates conceive of categories early in their programs. The extrapolation entails searching MU Library's catalog using the terms suggested by the students. The results of the searches can then be examined. Table 1 presents the numbers of hits for the terms, searched as subjects and as keywords.

The numbers of hits for the listed terms tend to be high. Moreover, the sorting options for the display of the hits are limited. A searcher could limit by date and by language (among some other options), but the display is in alphabetical order. As subjects headings are not assigned hierarchically (e.g., the first listed heading being the one that most clearly or strongly designates “aboutness”), there is no ranking of displayed hits. When hits are numerous and there is no form of ranking, students—especially undergraduate students—may be frustrated by the results. Recalling Marcum’s anecdote, the results of a Google search may suggest some relevant items among the first page of listings. Even when valid and assigned subject headings are used there can be a large number of hits. “Reading” is assigned to the Rosenblatt work, but a subject search yields 3,805 hits.

The cognitive limitations of undergraduate students are not likely to be alleviated by structural elements of library catalogs. Instruction in the nature of catalog structures may help students overcome some initial limitations (for example, the kind of phenomenological analysis Husserl details can enhance students’ abilities to reflect upon searching and finding possibilities) and may assist students in locating and using potentially relevant works, but systemic obstacles remain.

Discussion

The implications of the present study’s results should be taken in combination with some recent work in categorization, especially inasmuch as categorization is a practical tool for finding concrete and abstract ideas related to a searcher’s needs and wants. Much of the recent work centers on conceptions of ontology. It is important to define “ontology,” as it tends to be used in libraries and the information world. Jacob provides a definition that is a useful starting point: “an ontology can be defined as a partial, simplified conceptualization of the world as it assumed to exist by a community of users—a conceptualization created for an explicit purpose and defined in a formal, machine-processable language.”

In practice, ontologies may exhibit some characteristics of “uncontrolled vocabularies,” but only to a limited extent.

Some exercise of control in the construction of ontologies (as distinguished from ontology in the sense of the study of essences) could be possible. The control may be exercised communally, rather than centrally. That is, an ontology might be an emerging structure that includes contributions from a number of individuals; sometimes the individuals are
members of a disciplinary community. This kind of communal categorization may result in efficient and effective categorization, especially if some standards of consistency and requirements of ontological and phenomenological contribution are applied.

A detailed critique of current information systems ontologies by Fonseca and Martin suggests some reasons why present practice has some shortcomings, though. They point out limitations of any purely instrumental approach to the design of ontologies (that is, focus on how searchers may employ terms, rather than any effort at representing reality or truth, broadly defined), as only narrow intellectual, technical, or practical domains would be able to reach some agreement on common linguistic representations that would be useful. Any broad domain would experience disputes over the categorizations of particular works or events. Drawing from Kuhn, Fonseca and Martin maintain that the broader the intended domain of users, the more likely it is that some incommensurability (or categorical inconsistency resulting from differing worldviews) will be present. They conclude, “Perhaps the key point is to see that an ontology editor is distinct from ontologies. We conceive it to be a ‘place’ where persons assuming different conceptual schemas may come to learn from one another through interaction with each other and with their texts.”

One way to interpret their conclusion is that categorization could be of enhanced usefulness if there could be some informed dialogue about the process and its outcomes. Such a suggestion is quite distinct from application of folksonomies that can be more anarchic in application and in results. Noting that “folksonomy tags are not merely ‘messy,’ they can be inaccurate,” Peterson states, “A traditional classification scheme based on Aristotelian categories yields search results that are more exact. Traditional cataloging can be more time consuming, and is by definition more limiting, but it does result in consistency within its scheme.”

The disputes that have involved Gregor et al. should be revisited within the context of the work being conducted on information system ontologies and other alternative structures. There are opportunities for a considerable amount of empirical inquiry into the searching, retrieving, and use practices of information seekers. It may be that future systems design can build upon the successes (material and phenomenological) of several ideas of categorization. Library catalogs may not be dead, but there could be some informed research that can contribute to their evolution. While full- or free-text searching has some benefits, it is by its nature unstructured. Some kind of conceptual categorization will almost inevitably be used by information seekers. Following Lakoff, a couple of points should be remembered: (1) people are very likely to use differing conceptual systems (that is, people may see the same things in different ways), and (2) “To change the concept of category itself is to change our understanding of the world. At stake is our understanding of everything from what a biological species is to what a word is.”

Lakoff’s observations are reflected in the results of the present study. While categorization is common in human action, it can be quite variable. The subject organization of a library’s catalog necessarily imposes some structure on the contents of the works (necessary because the controlled vocabulary can incorporate definitional application that is used in the entirety of the catalog). The findings and analysis here in no way suggest abandonment of subject cataloging; rather, the findings point to the need to make the concept and use of a controlled vocabulary central to formal and informal instruction in the use of the catalog. For libraries in any educational setting, an objective is not merely to help searchers locate something, but to help them find something meaningful. The design of future catalogs can

<table>
<thead>
<tr>
<th>Term</th>
<th>No. of hits (subject)</th>
<th>No. of hits (keyword)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Object</td>
<td>221</td>
<td>1,198</td>
</tr>
<tr>
<td>History</td>
<td>11,632</td>
<td>32,000*</td>
</tr>
<tr>
<td>Culture</td>
<td>1,476</td>
<td>32,000*</td>
</tr>
<tr>
<td>Philosophy</td>
<td>2,573</td>
<td>32,000*</td>
</tr>
<tr>
<td>Physiology</td>
<td>1,362</td>
<td>32,000*</td>
</tr>
<tr>
<td>Sociology</td>
<td>641</td>
<td>12,292</td>
</tr>
<tr>
<td>Landscape</td>
<td>962</td>
<td>5,456</td>
</tr>
<tr>
<td>Past</td>
<td>524</td>
<td>10,802</td>
</tr>
<tr>
<td>Map</td>
<td>1,590</td>
<td>13,802</td>
</tr>
<tr>
<td>Organization</td>
<td>124</td>
<td>10,071</td>
</tr>
<tr>
<td>Good life</td>
<td>35</td>
<td>262</td>
</tr>
<tr>
<td>Morality</td>
<td>See Ethics</td>
<td>2,969</td>
</tr>
<tr>
<td>Technology</td>
<td>965</td>
<td>32,000*</td>
</tr>
<tr>
<td>Design</td>
<td>4,701</td>
<td>29,998</td>
</tr>
<tr>
<td>Postmodern</td>
<td>370</td>
<td>2,263</td>
</tr>
<tr>
<td>Literary theory</td>
<td>603</td>
<td>919</td>
</tr>
<tr>
<td>Theory</td>
<td>3,718</td>
<td>32,000*</td>
</tr>
<tr>
<td>Transactional theory</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Poetry</td>
<td>3,663</td>
<td>32,000*</td>
</tr>
<tr>
<td>Literature</td>
<td>16,982</td>
<td>32,000*</td>
</tr>
</tbody>
</table>

* The system default maximum is 32,000.
combine formal categorization with a form of ontology so as to create useful access to the contents of libraries. The authors of the Indiana University report affirm this observation: “Catalogers need to look beyond the online catalog for places to apply their knowledge and skills. Cataloging departments must adopt a more holistic approach that broadens the concept from ‘cataloging’ to the ‘organization of information.’”27

References

5. Ibid., 489.
9. Ibid., 46.
13. Ibid., 5.
15. Ibid., 5, 10, 15.
17. Ibid., 15.
19. Ibid., 57.
20. Ibid., 200.
24. Ibid., 55.