The Progress of Theory in Knowledge Organization

RICHARD P. SMIRAGLIA

ABSTRACT

WE UNDERSTAND "THEORY" TO BE A SYSTEM of testable explanatory statements derived from research. In knowledge organization, the generation of theory has moved from an epistemic stance of pragmatism (based on observation of the construction of retrieval tools), to empiricism (based on the results of empirical research). In the nineteenth century, Panizzi (1841), Cutter (1876), and Dewey (1876), developed very pragmatic tools (i.e., catalogs and classifications), explaining as they did so the principles by which their tools were constructed. By 1950, key papers at a University of Chicago Graduate Library School conference on "Bibliographic Organization" recorded the role of bibliographic organization in civilization (Clapp, 1950) and deemed classification the basis of bibliographic organization (Shera, 1950). In 1961, the International Conference on Cataloguing Principles in Paris brought together key thinkers on the design of catalogs. Wilson (1968) expounded a system for bibliographic apparatus, and provided the framework for empirical theoretical development. In 2000, Svenonius asserted that knowledge organization is accomplished through a bibliographic language (or, more properly through a complex set of bibliographic languages), with semantics, syntax, pragmatics, and rules to govern their implementation. Logical positivism notwithstanding, rationalist and historicist stances have begun to come to the fore of late through the promulgation of qualitative methods, most notably those employed in classification, user-interface design, and bibliometric research.

Richard P. Smiraglia, Palmer School of Library and Information Science, Long Island University, Brookville, NY 11548
LIBRARY TRENDS, Vol. 50, No. 3, Winter 2002, pp. 330–349

INTRODUCTION

We understand "theory" to be a system of testable explanatory statements derived from research. The term is difficult, because it has a colloquial usage that is quite a lot less precise than its use in academe. Colloquially, we understand theory to mean "ideas" or "principles." We attribute vagueness and an air of indecipherability to the term. The usage in academe is quite different. Here, we mean, quite precisely, statements, derived as a result of rigorous research and testing, that explain phenomena and relationships among them. Theory does not exist in a vacuum, but rather in a system that explains the domains in which we operate, the phenomena found in those domains, and the ways in which they might be affected by manipulation or change. Theory is derived from the controlled observation of phenomena, whether this has taken place in the positivist empirical paradigm or in the qualitative paradigm. Theory is the basis of research, serving to supply hypotheses for empirical research, and to confirm observations in qualitative research. The power of theory is its explanatory capability. We can use theory to analyze, predict, and manipulate phenomena.

In knowledge organization, there is quite a lot of theory of the colloquial sort (that is, stated principles) and even a fair amount of consensus on these statements. But, there is also, increasingly, a formal theoretical base. Knowledge organization, at least as it is practiced inside the domain of library and information science, has been largely (up to now) the province of the construction of tools for the storage and retrieval of documentary entities. That is, tools, such as catalogs, indexes, and databases, have been constructed to allow the rapid manipulation of and retrieval from large collections of surrogate records that represent documents, which in turn represent recorded knowledge. Order within these tools may take a variety of forms depending on the knowledge domain (or domains) represented, the anticipated usage of the tools, and their structure. Classification uses symbolic notation to order related concepts in appropriate groupings. Controlled vocabulary is created to alleviate linguistic variation in the documents and their surrogates that might otherwise obscure relationships among concepts. So-called "known items," documents identified by some combination of creator and title, are listed in alphabetical arrays using both names of creators (subarranged by title of work and date of creation, etc.) and document titles.

All of these tools have been constructed according to bibliographical judgment and pragmatic concerns about the documents themselves and their anticipated usage. In the second half of the nineteenth century, principles were expounded for the construction of catalogs that have, more or less, governed the development of bibliographic retrieval tools to the present day. The twentieth century increasingly saw the compilation of codes of rules to govern the construction of both document surrogates (i.e., bibliographic records) and the retrieval tools themselves. Svenonius (1981)

and Smiraglia (1987), among others, called for the application of empirical research methods to describe the phenomena of knowledge organization and to inform the further development of retrieval tools. The automation of bibliographic retrieval at the end of the twentieth century was informed, to some extent, by such empirical research. At the turn of the twenty-first century, scholarship in knowledge organization has begun to embrace qualitative research methods alongside the empirical, and, in a limited way, historical perspectives have been turned to in order to comprehend the social context of knowledge phenomena. Finally, rationalism has seen the increasing use of ontological and epistemological tools to comprehend the underlying structures of knowledge.

In knowledge organization, then, the generation of theory has moved from an epistemic stance of pragmatism (based on observation of the construction of retrieval tools), to empiricism (based on the results of empirical research). Logical positivism notwithstanding, rationalist and historicist stances have begun to come to the fore of late through the promulgation of qualitative methods, most notably those employed in bibliometric research. Another major balancing force has been the introduction of epistemology and ontology into the design of classification (Hjørland, 1998; Marco & Navarro, 1993). This paper is a review of these themes. Its purpose is not so much to present an exhaustive review of theory in knowledge organization, as to demonstrate the epistemological progression from rationally derived principles, to empiricism, to historicism.

Historical Background: Pragmatism and Rationalism

Panizzi (1841), Cutter (1876), and Dewey (1876), in the nineteenth century, developed very pragmatic tools (catalogs and classifications), explaining as they did so the principles by which their tools were constructed. Their efforts were influential: The principles they expounded can still be observed in the structure of modern online retrieval systems. Each, in his own way, raised the development of pragmatic retrieval tools to the level of a professional art form, introducing the concept of bibliographic judgment into the continued maintenance and development of tools for cataloging and classifying library collections. For each, the convenience of the public was always to be held in mind, over and against the inventory of the collection, on the one hand, or the ease of the cataloger, on the other. This was a remarkable development, which when interjected into the nascent program of education for professional librarians, saw the growth of pragmaticism and rationalism in the construction of tools for knowledge organization over the next three-quarters of a century. The evolution of these objectives laid the groundwork for the research in the mid-twentieth century that would lead to better empirical understanding. This, then, can be seen as the beginning of the development of formal theory in knowledge organization.

Strout (1956) told the whole history of catalogs from antiquity to modern times. Thus, we can trace developments back in time—for instance, one can postulate Hyde as Panizzi's predecessor, Maunsell as Hyde's, and so on, back to Callimachus in the great library at Alexandria. However, our point here is not to review the whole history of catalogs, but rather to establish a beginning for the theory of knowledge organization that prevails today. For this reason, we begin at this point in the mid- to late nineteenth century, when developments began to appear with great rapidity. And, of course, there were other leaders of that period, most notably Charles Coffin Jewett (1853). But here we posit the coincidence of Panizzi, Cutter, and Dewey as pragmatists as the beginning of our current backdrop of theory about the order of catalogs, relationships among subjects, and the order of knowledge itself.

Antonio Panizzi was hardly the first to develop a major catalog, nor was he even the first to develop a finding aid in the English-speaking world. That honor goes, of course, to Thomas Hyde's 1674 catalog for the Bodleian Library. Hyde's catalog has been called the first great alphabetical catalog, and was designed specifically to assist students in conducting research. Hyde's major contribution was to raise the collocating function to the level of principle, by insisting on the collocation of an author's works under a single form of name, with references from variant names and name forms. Also, in Hyde's catalog, representations of a single work that had appeared under different titles were also caused to collocate. As de Rijk (1991) has confirmed, Hyde's was a catalog in which divergent forms of names and titles of works were both expressed and reconciled.

It was Panizzi, however, for whom the construction of a catalog became more than the task of recording physical details of books. Rather, Panizzi recognized the importance of making a distinction between the retrieval and use of specific intellectual entities—that is, works—and the usual inventory of books. With Panizzi, the function of identifying and collocating works and their variant instantiations became a principle, and a very pragmatic principle at that. Panizzi was emphatic that to be useful, a catalog had to allow a reader to identify and choose among works. His famous defense of his catalog includes this very pragmatic assertion ([1848] 1985, p. 21, emphasis original):

> No catalogue . . . can be called 'useful' in the proper sense of the word, but one in which the titles [i.e. entries] are both 'accurate,' and so 'full' as to afford *all* that information respecting the real contents, state, and consequent usefulness of the book which may enable a reader to choose, from among many editions, or many copies, that which may best satisfy his wants, whether in a literary or scientific, or in a bibliographical point of view.

In other words, no catalog that merely lists items can be considered useful. Rather, it is the intellectual content—that is, the works—for which readers

consult a catalog. To be useful, then, the catalog must clearly identify works in such a way that a user is assisted in making an informed selection—a very pragmatic principle, rationally derived, which advanced the construction of the catalog from that of inventory of documents to modern tool for the retrieval of works.

Charles Ammi Cutter, librarian of the Boston Atheneum, provided rules for the construction of dictionary catalogs. The dictionary catalog was to be one in which name, title, and subject entries for books were integrated in a single alphabetical sequence. The direct successors of codes of rules by Panizzi and the Smithsonian's Charles Coffin Jewett, Cutter's rules often are seen as the direct progenitors of the modern *Anglo-American cataloguing rules.* Indeed, many cataloging practices that are encoded in today's rules for descriptive cataloging can be traced directly to Cutter's code.

Cutter's rules were originally issued as the second part of a special report of the Bureau of Education (then a division of the Department of the Interior), titled *Public libraries in the United States of America: Their history, condition, and management.* Published thusly in 1876, these rules enjoyed widespread acceptance and fueled the growth of the public library as an educational institution. As public libraries spread, Cutter's rules gave pragmatic instruction to librarians across the U.S. landscape for the construction of local dictionary catalogs. Asserting a principle of context, Cutter suggests that the given catalog might be considered short, medium, or full—depending on the level of detail considered critical to the users of the collection in question.

Cutter's rules were prefaced with a statement of the objectives of the dictionary catalog. These statements, called "Objects," frame the entire construction of the catalog within the pragmatic judgment of the cataloger. Ultimately, the cataloger is given generic directions for the creation of a description of a book, and for the selection and formation of access points that will, in many cases, lead to the collocation of entries for the work within that book. There is an expectation that, given specific instructions and a pragmatic philosophical framework, catalogers will be able to apply their own professional judgment and yield consistent results.

The popularity and widespread usage of Cutter's rules is apparent from the publication history—the fourth, and final edition was published in 1904, containing many appendices intended to inform the cataloging of nonbook materials. Ultimately, Cutter's pragmatism was expressed in his suggestion (1904, p. 6) that the cataloger always weigh local needs against the convenience of the users. While Cutter dictates that this decision must always yield to the requirements of users, still it is a critical, pragmatic instruction to take both sets of sometimes conflicting needs into account.

Melville Dewey, the father of much of American librarianship, is the third individual whose influence caused the spread of pragmatic tools for the organization of library collections. Most famous for his *Decimal Classifi-*

cation (1876), which is now in use worldwide, it is perhaps more important at this juncture for us to consider Dewey's powerful political influence on the development of the profession of librarianship. It is to Dewey that we owe the professionalization of bibliography, the beginnings of education for librarianship, the development of professional associations for librarians, and in 1908 the publication of the first joint *Anglo-American cataloguing rules.* But it is also to Dewey and his Library Bureau that we owe the spread of the card catalog utilizing 3-by-5-inch holed cards in wooden cases of standard sizes. Together with his *Decimal Classification,* the spread of the card catalog (now in dictionary form thanks to Cutter's influence) standardized the organization of knowledge in libraries all across the English-speaking world and particularly in American public libraries. This standardization ensured more than professional economies of scale. Perhaps Dewey's greatest contribution was to give generations of users the capability to find relevant materials treated in the same way in nearly any library.

As we noted earlier, the history of catalogs and cataloging has been written elsewhere. Here our point is to note the historic coincidence of the efforts of Panizzi, Cutter, and Dewey. All three were pragmatic managers of large libraries, and authors of the principles of catalog and collection management. Above all, they left a critical legacy to the practice of the organization of documents (and, thereby, of the works and recorded knowledge contained therein). They were the progenitors of the twentieth-century move toward standardization and codification. Their pragmatic guidance insisted on the judgment of the cataloger, the convenience of the user in retrieving what was sought, and the consistent ordering of bibliographic entities—be they citations for works, subject headings in the dictionary catalog, or volumes themselves ordered to facilitate browsing by the public.

From time to time, the pendulum would swing away from their pragmatic guidance, but ultimately, pragmatism was the theoretical norm through the twentieth century. For example, Panizzi had called for the entry of pseudonymous works under the authors' pseudonym, so as to yield a direct result for the searcher. The pragmatism of the idea is clear—a user should be able to seek a work under the citation by which it is popularly known in the marketplace (or in the culture). However, a more academic approach was used—entry under the real name—from Cutter's time until the second revision of the second edition of the *Anglo-American cataloguing rules* in 1988. At last, at the end of the twentieth century, the flood of romance fiction written by authors using several pseudonyms at once resulted in a compromise measure that allows for collocation of works under an author whose real name has become synonymous with his/her pseudonyms, but for entry under the pseudonyms (even under several) for those that have not.

Key papers at a 1950 University of Chicago Graduate Library School conference on "Bibliographic Organization" recorded the role of bibliographic organization in civilization (Clapp, 1950) and deemed classifica-

tion the basis of bibliographic organization (Shera, 1950). Clapp defined bibliographic organization as: "The pattern of effective arrangements which results from the systematic listing of the records of human communication" (p. 4). Asserting the social role of the organization of knowledge, Clapp set about to list the areas in which empirical research would be critical for developing the discipline. These were: (1) Types (suggesting the taxonomic study of kinds of bibliographies); (2) Gaps (where possible these should be closed); (3) Duplication (which should be eliminated); (4) Informativeness (it would be necessary to combine comprehensive and selective lists); (5) Physical location; (6) Cooperation, or the coordination of energies; (7) Classification (the tools of library organization should be generalized to all bibliography); and (8) Mechanical devices (a challenge to develop cheaper, more compact, and more flexible bibliographical apparatus) (pp. 17–21). Similarly, Shera asserted the importance of classification as the very basis of bibliographic organization. However, he also pointed to the failure of a century of library classification to resolve the key problems of organizing knowledge, saying: "There can no longer be any doubt that library classification has failed, and failed lamentably, to accomplish what it was designed to do" (p. 72). Shera outlined four basic historical assumptions about the utility of classification: (1) There exists a universal order of nature that should reveal a permanent conceptual framework of the entirety of human knowledge; (2) Schematization of that universal and permanent order is a hierarchy; (3) There is a principle of differentiation derived from likeness or unlikeness of the properties of phenomena; and (4) These properties partake of the substantive nature of the pheomena. He related what he calls the "failure of traditional approaches to classification" to the lack of social epistemology, or social context of a given knowledge domain (pp. 72–73). Like Clapp, Shera also posited a research agenda, which includes: (1) Studies of existing classifications; (2) Development of new schema, based on new principles; (3) Experimentation in the construct of conceptual frameworks; (4) Content analysis of research literatures; (5) Careful scrutiny of subject headings; (6) Measurements of effectiveness; (7) Analysis of dispersion; and (8) Precise measurement of costs (p. 93). As though to demonstrate Shera's point, the Chicago conference also witnessed the introduction of Ranganathan's Colon Classification, from which the notion of faceted indexing would be derived and expanded. The 1961 International Conference on Cataloguing Principles in Paris brought together key thinkers on the design of catalogs. Lubetzky (1961) provided the impetus for restating Cutter's principles in a way that would begin to shift the focus of the catalog from its role as inventory of books to a new role as pathfinder among works. Verona's concept (1961) of literary unit vs. bibliographical unit would underlie this shift in roles, as would Osborn's pragmatic approach (1961) to the construction of tools for bibliographic retrieval. Hickey summarized much of this theory in 1977, at the brink of the paradigm

shift from paper-based systems to electronic, automated systems. Taken together, these key statements of rules and principles can be seen to constitute a core for theory of knowledge organization.

Wilson (1968) was the first to analyze and summarize these accomplishments in a single text, expounding a system for bibliographic apparatus, and providing the framework for empirical theoretical development. Wilson stated underlying philosophical points, for example, descriptive and exploitative domains, in which the bibliographical apparatus (as created by Panizzi, Cutter, Dewey, et al.) plays a key role. According to Wilson, the descriptive domain (in today's parlance the word "domain" might better be rendered as "concept space") is the domain in which descriptive bibliographic activity takes place. In the descriptive domain, catalogers, bibliographers, and indexers strive to create listings of various depths and degrees of detail to record the existence of writings available to searchers. In the exploitative domain, scholars seek answers to their questions, and especially they seek to make the best possible use of recorded knowledge. That is, they seek to exploit what is already known, so as to create new knowledge.

Here Wilson provided, for the first time, a means by which the efficacy of the bibliographical apparatus can be measured. Whatever in the descriptive domain facilitates activity in the exploitative domain can be said to be efficacious. Likewise, whatever hinders activity in the exploitative domain can be said to be detrimental. By inserting specific activities (e.g., searching) or entities (e.g., access points) and measuring retrieval success, researchers could operationalize variables, and begin empirically to test such theoretical statements as had heretofore had the status of "principles." This contribution moved the field of knowledge organization forward as a research discipline, allowing practice to be informed by the results of scientific investigation, and paving the way for an accumulation of observations over time that might contribute to true theory.

The Beginnings of Empiricism

Clapp (1950) and Shera (1950) posited research agendas, essentially marching orders for the world's scholars in bibliographic retrieval and classification. Other calls to action were to follow, in particular papers by Gorman (1980, 1982) and others, at the time of *AACR2*'s first edition being published. In 1981, Svenonius reviewed current research in bibliographic control and found it wanting, particularly in regard to problems of heading integrity and file structure:

> Questions of efficient file design need researching, such as how is linkage information to be accessed, should all linkage information be contained in an authority file, and how are authority and bibliographic files to be interfaced? (p. 101)

Gorman (1982) called for similar research, suggesting a design schema for the online catalog in which physical items would be represented by unique

bibliographic records, and all access points (names, works, subjects, etc.) would be represented in unique authority records. Explicit links could then be created in several directions, both among related authority records and between authority records and the bibliographic records that represent bibliographic entities. Similarly, Taylor, in a 1988 review of progress in authority control research, pointed out the need for continued research in bibliographic relationships:

> The questions Svenonius asked about how linkage information is to be accessed, whether all linkage information should be contained in an authority file, and the means for interfacing authority and bibliographic files have been examined to some extent, although the answers are not yet clear. (p. 51)

Taylor suggested further study of file design, concluding:

> Perhaps these questions remain unanswered because Svenonius's remaining question, that of efficient file design, has yet to be examined. . . . The conflicts we now have of some linkage information being held in the authority file and the remainder being held in the bibliographic file [should] be resolved. (p. 51)

In a 1992 review Svenonius stated:

> Library catalogs . . . must be able to distinguish uniquely bibliographic entities at a variety of aggregate levels. . . . Further experimenting is needed to identify the necessary and sufficient data elements needed to distinguish various kinds of bibliographic entities. . . . (p. 11)

She went on to say:

> A library catalog in addition to distinguishing unlike bibliographic entities must also collocate and otherwise relate like entities. The failure to do so is a failure in recall. . . . An entity in the bibliographic universe is not an island unto itself but is connected to other entities in a variety of constellations and relationships. In order for a user to navigate the bibliographic universe to a desired end, a map is needed to show how entities are clustered and where the pathways are between and among them. Such a map would depict the collocating relationships specified by the second objective of the catalogue and it would show other bibliographic relationships as well. (pp. 11–12)

These papers represent a call to arms from the major scholars of bibliographic control in the last quarter of the twentieth century, issued to the up and coming researchers in the field. Questions of file design, record construct, and entity-relationship definition were critical to the advancement of the catalog as a tool of the modern age. Furthermore, empirical evidence of the incidence of bibliographic phenomena, and of searching behavior would be critical to inform the rapid development of increasingly technologically complex systems for retrieval of not only bibliographic

data, but also full document texts, archival records, surrogates for museum artifacts, and so on. Empiricism, represented by scientific research in the positivist paradigm, was clearly called for if the cause of knowledge organization was to advance. And, chief among the problems of empirical researchers, therefore, was the lack of comprehension of the extent to which external validity (the ability to generalize a research result from one collection of documents to another, which would depend on the degree to which collections of documents were inherently alike or different) was key.

Many took up the challenge, and the research journals are filled with reports of research that examined the problems posed by these pivotal scholars. In four areas, to be described below, research has accumulated to a degree sufficient to posit theoretical statements. Let us now turn to these four areas to understand the role of positivism and pragmaticism in the growth of theory in knowledge organization.

Author Productivity and the Distribution of Name Headings

In 1926 Lotka asserted an inverse relationship between the number of authors writing in a given subject area and their productivity. Known as "Lotka's Law," this relationship can be stated thus: The total number of authors y in a given subject, each producing x publications, is inversely proportional to some exponential function n of x. The practical result of Lotka's observation was to demonstrate that the total number of authors contributing a single publication would be just over 60 percent (p. 321). That is, only 40 percent of authors contribute more than one paper. Lotka was concerned bibliometrically with the attribution of author productivity as a measure of the influence of authors in specific subject areas. But research by Taylor, Potter, Papakhian, and others has demonstrated an ability to observe Lotka's law operating in the bibliographic universe.

These studies were conducted to examine name headings' frequency of occurrence in catalogs. Potter (1980) examined this frequency in two general catalogs, and discovered that roughly two-thirds (63.5 percent and 69.33 percent respectively) of all names occur only once (p. 9). Fuller (1989, p. 81) found a similar proportion, 61 percent, in the catalog of the University of Chicago. McCallum & Godwin (1981, p. 198) found that 66 percent of names occurred only once in the Library of Congress machine-readable catalog. Papakhian (1985, p. 285), replicating Potter's design in a sound recordings catalog, found that fewer than half (47.6 percent) of names could be said to occur only once, concluding that the presence of nonbook materials could be associated with an increase in multiple occurrence entries. This research was conducted to help the community understand the impact of changes in cataloging rules. Collectively, these results demonstrate a theoretical assumption that underlies the infrastructure of bibliographic databases. That is, most names will occur only once, and a very small number, which can be predicted by Lotka's Law, will occur many times.

Bibliographic Relationships

No document is an island, and the interrelatedness of documents and their contents, as well as the complexity of these relationships, has prevented the increasing sophistication of online retrieval systems. Beginning with Tillett (1987), who sought to classify and quantify the entire range of bibliographic relationships in the Library of Congress catalog, research has demonstrated the efficacy of comprehending bibliographic relationships. Smiraglia (1992) investigated the derivative relationship, which holds among all versions of a work, refining its definition to include several different categories of derivation. Leazer and Smiraglia studied the presence of derivative relationships in the OCLC WorldCat (Smiraglia & Leazer, 1995, 1999; Leazer & Smiraglia, 1996, 1999), affirming the taxonomy of derivative relationship types. Yee (1993) examined problems of relationships among moving image materials, including the substantial problems of associating bibliographic records for varying instantiations of films. Vellucci (1994, 1997) examined musical works and found that the categories of work relationships that Tillett (1987) and Smiraglia (1992) had suggested were present, and in large numbers. Smiraglia (1999) demonstrated the effectiveness of the taxonomy of relationship types by analyzing the extent of derivation among entities in theological collections. Research in bibliographic relationships reinforced the observation of Lotka's law, exploded unitary concepts of bibliographic entities by demonstrating their complexity and interrelatedness, and confirmed the importance of the role of works in the bibliographic universe.

Entity-Relationship Design

Traditional catalogs and indexes were conceived as linear files of bibliographic records (i.e., citations). However, with the introduction of syndetic structure from Panizzi onward, catalogs grew increasingly complex. Translation to the online environment yielded the early (unfortunately misnomered) "online card catalog." Research that would apply the principles of database construction to the infrastructure of the catalog was needed. Authors examined catalog data conceptually to identify independent entities. Fidel & Crandall (1988) described the *Anglo-American cataloguing rules* from a generalized database approach, using the entity-relationship model to suggest a problem-based typology of rules that might underlie a theoretical framework of rules for bibliographic database design. Leazer (1992) documented intra-record data redundancy, as well as the apparent absence of a conceptual schema, for the MARC-based online catalog. Leazer (1993, 1994) described a conceptual schema for the explicit control of works in catalogs, taking into account both Tillet and Smiraglia's taxonomies of relationship types. Green (1996) presented a conceptual design for a full-scale bibliographic database based on entity-relationship modeling. The

1998 report of the IFLA Study Group on the Functional Requirements for Bibliographic Records presented a framework that identified and defined the entities of interest to users of bibliographic records, the attributes of each entity, and the types of relationships that operate between entities. Collectively, this research has demonstrated the utility of the entity-relationship approach to the design of bibliographic databases.

External Validity

A lack of comparative data that might provide the grounds for external validity has hampered research in knowledge organization. However, there are now indications that catalogs containing bibliographic records for similar collections of materials exhibit similar characteristics. Potter (1980), McCallum & Godwin (1981), Papakhian (1985), and Fuller (1989) all discovered similar proportions of single-occurrence name headings in research library catalogs. These studies support the contention that catalogs of similar materials exhibit similar characteristics. That is, there is reason to believe that there are grounds for generalizing research results from studies conducted in a specific library to other similar library environments. Taylor & Paff (1986) found that changes of name and title headings required by the implementation of *AACR2* in the catalog of a medium-sized academic library were in line with projections made by Taylor in her 1980 study of a similar library (Dowell, 1982). The replication tested proportions of change in the new catalog against the proportions reported in the 1980 study and found no statistically significant difference in the proportions from the two independent samples:

> The fact that there was no significant difference between the projections . . . may indicate that samples of the collections of libraries (at least of academic libraries) are drawing from essentially the same universe. (Taylor & Paff, 1986, p. 280)

Further, they found that certain patterns of headings occurrence were comparable in the two independent samples:

> Is it possible that various types of heading occur in predictable proportions in the bibliographic universe? . . . It can be noted that, although the exact proportions varied somewhat, the pattern . . . found in all three libraries in the Dowell study . . . was repeated at ISU. This is not simply a representation of the relative proportions of these types of headings in the cataloging as a whole. (pp. 280–281)

Countless other studies, notably those examining bibliographic relationships, have gathered data on the inherent characteristics of the documents in specific library collections. These data have yet to be compiled, but taken together with the studies cited here, there is evidence that theoretical predictability about bibliographic phenomena might be possible.

Historicism

Epistemology is the division of philosophy that investigates the nature and origin of knowledge. Poli (1996) contrasted the tools of ontology and epistemology for knowledge organization, suggesting that while ontology represents the "objective" side of reality, epistemology represents the "subjective" side. Ontology ("being") provides a general objective framework within which knowledge may be organized, while epistemology ("knowing") allows for the perception of the knowledge and its subjective role. Olson (1996) used an epistemic approach to comprehend Dewey's classification, asserting a single knowable reality reflected in the topography of recorded knowledge. Dick (1999) described epistemological positions in library and information science. He suggested that experience (i.e., empiricism) provides the material of knowledge, and reason (i.e., rationalism) adds the principles for its ordering. Rationalism and empiricism supply the basic platform for epistemological positions. They have been the primary modes of theoretical development in knowledge organization to this point. At the turn of the twenty-first century, the field of knowledge organization has begun to turn increasingly to the tools of qualitative analysis to explain the complexities of phenomena surrounding knowledge and its documentary record. This can be seen as an attempt to move beyond the strictures of empiricism, to bring a historicist epistemology to bear on the problems of the organization of knowledge.

Hjørland's Epistemological Framework

Hjørland (1998) asserted a basic epistemological approach to base problems of information retrieval, particularly to the analysis of the contents of documentary entities. He began from a basic metaphysical stance, stating that ontology and metaphysics describe what exists (basic kinds, properties, etc.), whereas epistemology is about knowledge and ways in which we come to know. Hjørland listed four basic epistemological stances:

- Empiricism, derived from observation and experience;
- Rationalism, derived from the employment of reason;
- Historicism, derived from cultural hermeneutics; and,
- Pragmatism, derived from the consideration of goals and their consequences.

Hjørland described a domain-analytic approach to subject analysis, recognizing that any given document may have different meanings and potential uses to different groups of users. Hjørland & Albrechtsen (1999) delineated recent trends in classification research, demonstrating the utility of Hjørland's epistemological framework for deriving categories.

Marco & Navarro (1993) described contributions of the cognitive sciences and epistemology to a theory of classification:

> The study of epistemology is, therefore, essential for the design and implementation of better cognitive strategies for guiding the process of documentary analysis, particularly for indexing and abstracting scientific documents. The ordering and classifying of information contained in documents will be improved, thus allowing their effective retrieval only, if it is possible to discover the conceptual framework (terms, concepts, categories, propositions, hypotheses, theories, patterns, and paradigms) or their authors from the discursive elements of texts (words, sentences, and paragraphs). (p. 128)

Epistemology, then, is concerned with the theory of the nature of knowledge.

Knowledge organization has been too long enamored of the rationalistic and pragmatist approaches. Indeed, rationalism expounds detail, and some of the hallmarks of knowledge organization theory are the rationalist works on descriptive cataloging. Most notable among these are the groundbreaking works of Seymour Lubetzky, who first sought to explain *rationally,* the purposes and construction of the modern catalog (summarized in Lubetzky, 1969). Domanovszky (1974) and Carpenter (1981) also offered rationalist constructs that advance the theory—that is, the system of principles that govern the construction—of the dictionary catalog.

However, the problem remains that too few conceptual arrays are based on either empirical knowledge of what exists in the universe of documentary knowledge entities, or on essential understanding of the cultural importance, historic origins, or social roles, of the entities we propose to systematize. Knowledge organization, as Hjørland (1998) and Hjørland & Albrechtsen (1999) have suggested, must proceed from more finely developed epistemological positions, and these are the empiricist and historicist points of view.

Research Moves Away from Empiricism

To inform our cognitive structures with epistemological perspectives from the historicist point of view requires new analytical tools. A few examples will demonstrate the power of the historicist perspective. For instance, cocitation analysis, reviewed extensively by White & McCain (1997), has demonstrated the complex relationships that exist among authors working within and between disciplines. Beghtol (2000, 2001) demonstrated the centrality of key concepts, such as "Genre" and "A Whole and its Parts." Mai (2000a, 2000b) brought the tools of semiotics to bear on problems of indexing and classification. Smiraglia (2000, 2001) used semiotics to comprehend the social role of works and Hjørland's epistemological stances to derive an expanded definition of the work. By understanding from an empirical perspective what has been observed from a historicist perspective, we can begin to rationally and pragmatically derive appropriate constructs for systems for information retrieval. The potential uses of epistemology for documentary analysis, then, are many; a few have been attempted. Whereas ontology

may be relied upon to frame the organization of knowledge, epistemology provides us with key perceptual information about the objects of knowledge organization. Each perspective can contribute to understanding; collectively, a balanced perspective can be achieved. To begin, empiricism can lead us to taxonomies of knowledge entities. Rationalism can demonstrate the cultural role of, and impact on, knowledge entities.

Svenonius

Svenonius (2000) represents, like Wilson (1968), a milestone summary and analysis of all that has come before. Svenonius asserted that knowledge organization is accomplished through a bibliographic language (or, more properly through a complex set of bibliographic languages), with semantics, syntax, pragmatics, and rules to govern their implementation. She cumulated the historical record of research in knowledge organization, and brought ontological tools to bear on the problems of the definition of phenomena. Like Wilson, she drew together the results of empirical research in every aspect of knowledge organization, stating principles where appropriate, and demonstrating lacunae in the empirical record. Also, like Wilson, she contributed a tool that may come to be used as a theoretical benchmark for future research. This is her set theoretic model "that regards the bibliographic universe as consisting of documents, sets of these (formed by attributes . . .), and relationships among them" (p. 32).

Theory in Knowledge Organization: Concluding Remarks

"Theory," then, remains a system of testable explanatory statements derived from research. In knowledge organization, the generation of theory has moved from an epistemic stance of rationalism (construction of retrieval tools based on reasoned principles), to pragmatism (based on observation of the phenomena of knowledge entities), to empiricism (based on the results of empirical research). After nearly two centuries of formal work on the construction of catalogs and classifications, we are blessed with a wellspring of rationalist thought and large codes of pragmatic rules. At the same time, three decades of advancing formal, empirical research have yielded the beginning of a set of formal theories for the organization of recorded knowledge.

Two key contributions are those of Wilson (1968) and Svenonius (2000). Each expounded an entire system for the knowledge domain and its retrieval apparatus. Given the similarities between their approaches, one can also view these systematic presentations as standing at two points on the epistemological spectrum. That is, Wilson's system followed a century of pragmatism, and seems to arise at the beginning of what would be the most intense period of empirical research into knowledge phenomena. Svenonius' system arises at the point where research seems to have turned toward the historicist stance.

And so there is no single, formal statement of theory of knowledge organization. However, we can posit, based on this review, three simple theoretical statements:

1. A theoretical assumption underlies the infrastructure of bibliographic databases, such that most names will occur only once, and a very small number, which can be predicted by Lotka's Law, will occur many times.

As noted above, Lotka's law has been observed in a variety of bibliographic environments. We are not certain why this law holds, or what, exactly, it represents. Smiraglia & Leazer (1999) have suggested that canonicity plays a role in this function. That is, some works enter an academic canon, and thereby gain value for the academic community, which in turn causes them to be variously translated, edited, and reproduced, thus contributing to the frequency of occurrence of author names in databases. It is also likely that some larger number of works are published, consumed by the culture, and then discarded (in a sense, such works are "digested"). However, it is equally likely that Lotka's law reflects phenomena that are as yet unobserved. In sum, the pragmatic influence of this distribution is that 60 percent of records (names, etc.) in a file will be unique; another 40 percent will require extra effort to delineate the relationships among the knowledge entities they represent.

2. Bibliographic relationships reinforce the observation of Lotka's law, exploding unitary concepts of bibliographic entities by demonstrating their complexity and interrelatedness.

Bibliographic relationships are complex. These are the relationships among bibliographic entities, such as the equivalence relationship (that holds among copies of an item, e.g., a book and its microform reproductions) or the derivative relationship (that holds among variations on a work, e.g., editions and translations). Research has shown that for a small proportion of works in catalogs (about 40 percent, in line with Lotka's law) there will be a complex set of interrelated entities that require explicit linkage to facilitate efficacious retrieval.

3. There is a beginning of evidence that there are grounds for external validity in the examination of knowledge entities.

That is, we have begun to observe similar distributions from one collection to another among the bibliographic characteristics that describe knowledge entities. This means that empirical research can advance secure in the knowledge that results can be generalized from one subset of the bibliographic population to another.

Other theoretical statements, of course, might soon be possible. These will come to light as a result of the combined use of all four epistemological stances. For instance, much research has been undertaken on the na-

ture of subject searching in library catalogs. This research suggests that cognitive aspects of user behavior are at least as important as the subject characteristics of the documents represented. One might expect research to soon provide theoretical statements in this area. Another area rife for theoretical development is the extensive work of cocitation and coword analysis. This work describes relationships among scholars, essentially mapping intellectual relationships within knowledge domains as represented by citations and abstracts. What is needed are sociological (i.e., cognitive) explanations of the behaviors that lead to these intellectual relationships. Such explanations could give us real predictive power for the development of sophisticated systems for the retrieval of knowledge entities.

One thing is clear: A variety of epistemic stances are required to advance the pursuit of theory. Where pragmatism could only suggest what to do, and empiricism could only describe unique phenomena in isolated contexts, rationalism and historicism can help us uncover the ineluctable truths of the natural order of knowledge entities.

REFERENCES

Anglo-American cataloguing rules (1998). 2nd ed. rev. Chicago: American Library Association.

Beghtol, C. (2000). A whole, its kinds, and its parts. In C. Beghtol, L. C. Howarth, & N. J. Williamson (Eds.), *Dynamism and stability in knowledge organization* (Proceedings of the 6th International ISKO Conference, 10–13 July 2000, Toronto, Canada). Advances in knowledge organization (vol. 7, pp. 313–319). Würzburg: Ergon Verlag.

Beghtol, C. (2001). The concept of genre and its characteristics. *Bulletin of the American Society for Information Science and Technology, 27*(2), 1–5.

Carpenter, M. (1981). *Corporate authorship: Its role in library cataloging.* Westport: Greenwood Press.

Chaplin, A. H., & Anderson, D. (Eds.) ([1961] 1981). *International Conference on Cataloguing Principles Report.* London: IFLA International Office for UBC.

Clapp, V. W. (1950). The role of bibliographic organization in contemporary civilization. In J. H. Shera & M. E. Egan (Eds.), *Bibliographic organization: Papers presented before the 15th annual conference of the Graduate Library School July 24–29 1950* (pp. 3–23). Chicago: University of Chicago Press.

Cutter, C. A. (1876). *Rules for a printed dictionary catalog.* 1st ed. Washington: U. S. G. P. O.

Cutter, C. A. (1904). *Rules for a printed dictionary catalog.* 4th ed. Washington: U. S. G. P. O.

de Rijk, E. (1991). Thomas Hyde, Julia Pettee and the development of cataloging principles; With a translation of Hyde's 1674 Preface to the reader. *Cataloging and Classification Quarterly, 14*(2), 31–62.

Dewey, M. (1876). *A classification and subject index for cataloguing and arranging the books and pamphlets of a library.* Amherst: M. Dewey.

Dick, A. L. (1999). Epistemological positions and library and information science. *Library Quarterly, 69,* 305–323.

Domanovszky, A. (1974). *Functions and objects of author and title cataloguing: A contribution to cataloguing theory* (Anthony Thompson, Trans.). Budapest: Akademiai Kiado.

Dowell, A. T. (1982). *AACR2 headings: A five-year projection of their impact on catalogs.* Littleton, CO: Libraries Unlimited.

Fidel, R., & Crandall, M. (1988). The *AACR2* as a design schema for bibliographic databases. *Library Quarterly, 58*(2), 123–142.

Fuller, E. (1989). Variation in personal name beadings and title page usage. *Cataloging and Classification Quarterly, 9*(3), 75–96.

Gorman, M. (1980). *AACR2* main themes. In D. H. Clack (Ed.), *Making of a code* (pp. 41–52). Chicago: American Library Association.

Gorman, M. (1982). Authority control in the prospective catalog. In M. W. Ghikas (Ed.), *Authority control: The key to tomorrow's catalog* (pp. 166–180). Tucson: Oryx Press.

Green, R. (1996). The design of a relational database for large-scale bibliographic retrieval. *Information Technology and Libraries, 15*(4), 207–221.

Hickey, D. J. (1977). Theory of bibliographic control in libraries. *Library Quarterly, 47*(3), 255–273.

Hjørland, B. (1998). Theory and metatheory of information science: A new interpretation. *Journal of Documentation, 54*(5), 606–621.

Hjørland, B., & Hanne, A. (1999). An analysis of some trends in classification research. *Knowledge Organization, 26*(3), 131–139.

IFLA Study Group on the Functional Requirements for Bibliographic Records (1998). *Functional requirements for bibliographic records.* München: K. G. Saur.

Jewett, C. C. ([1853] 1985). Smithsonian catalogue system. Reprinted from *On the construction of catalogues of libraries and of a general catalogue and their publication by means of separate, stereotyped titles with rules and examples.* 2nd ed. In M. Carpenter & E. Svenonius (Eds.), *Foundations of descriptive cataloging* (pp. 51–61). Littleton, CO: Libraries Unlimited.

Leazer, G. H. (1992). An examination of data elements for bibliographic description: Toward a conceptual schema for the USMARC formats. *Library Resources and Technical Services, 36,* 189–208.

Leazer, G. H. (1993). A conceptual plan for the description and control of bibliographic works. Unpublished Ph.D. thesis, Columbia University.

Leazer, G. H. (1994). A conceptual schema for the control of bibliographic works. In D. L. Andersen, T. J. Galvin, & M. D. Giguere (Eds.), *Navigating the networks* (Proceedings of the ASIS mid-year meeting) (pp. 115–135). Medford, NJ: Learned Information.

Leazer, G. H., & Smiraglia, R. P. (1996). Toward the bibliographic control of works: Derivative bibliographic relationships in an online union catalog. In *Digital libraries '96* (1st ACM International Conference on Digital Libraries, March 20–23, 1996, Bethesda, Maryland) (pp. 36–43). New York: Association for Computing Machinery.

Leazer, G. H., & Smiraglia, R. P. (1999). Bibliographic families in the library catalog: A qualitative analysis and grounded theory. *Library Resources and Technical Services, 43,* 191–212.

Lotka, A. J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences, 16,* 317–323.

Lubetzky, S. (1969). *Principles of cataloging. Final report, phase I: Descriptive cataloging.* Los Angeles: Institute for Library Research.

Lubetzky, S. ([1961] 1981). The function of the main entry in the alphabetical catalogue—One approach. In A. H. Chaplin & D. Anderson (Eds.), *International Conference on Cataloguing Principles Report* (pp. 139–143). London: IFLA International Office for UBC.

Mai, J. E. (2000a). Likeness: A pragmatic approach. In C. Beghtol, L. C. Howarth, & N. J. Williamson (Eds.), *Dynamism and stability in knowledge organization* (Proceedings of the 6th International ISKO Conference, 10–13 July 2000, Toronto, Canada). Advances in knowledge organization (vol. 7, pp. 23–27). Würzburg: Ergon Verlag.

Mai, J. E. (2000b). The subject indexing process: An investigation of problems in knowledge representation. Unpublished Ph.D. thesis, Univ. of Texas, Austin.

Marco, F. J. G., & Navarro, M. A. E. (1993). On some contributions of the cognitive sciences and epistemology to a theory of classification. *Knowledge Organization, 20,* 126–132.

McCallum, S. H., & Godwin, J. L. (1981). Statistics on headings in the MARC file. *Journal of Library Automation, 14*(3), 194–201.

Olson, H. A. (1996). Dewey thinks therefore he is: The epistemic stance of Dewey and DDC. In R. Green (Ed.), *Knowledge organization and change* (Proceedings of the 4th International ISKO Conference, July 15–18, 1996, Washington, DC). Advances in knowledge organization (vol. 5, pp. 302–303). Frankfurt/Main: Indeks Verlag.

Osborn, A. D. ([1961] 1981). Relation between cataloguing principles and principles applicable to other forms of bibliographical work. In *International conference on cataloguing principles report* (pp. 125–37). London: IFLA International Office for UBC.

Panizzi, A. ([1841] 1985). Rules for the compilation of the catalogue. In M. Carpenter & E. Svenonius (Eds.), *Foundations of descriptive cataloging* (p. 314). Littleton, CO: Libraries Unlimited.

Panizzi, A. ([1848] 1985). Mr. Panizzi to the Right Hon. the Earl of Ellesmere, British Muse-

um, January 29, 1848. Reprinted from *Appendix to the report of the commissioner appointed to inquire into the constitution and management of the British Museum.* In M. Carpenter & E. Svenonius (Eds.), *Foundations of descriptive cataloging* (pp. 18–47). Littleton, CO: Libraries Unlimited.

Papakhian, A. R. (1985). The frequency of personal name headings in the Indiana University Music Library card catalogs. *Library Resources and Technical Services, 29*(3), 273–285.

Poli, Roberto (1996). Ontology for knowledge organization. In R. Green (Ed.), *Knowledge organization and change* (Proceedings of the 4th International ISKO Conference, 15–18 July 1996, Washington, D.C.). Advances in Knowledge Organization (vol. 5, pp. 313–319). Frankfurt/Main: Indeks Verlag.

Potter, W. G. (1980). When names collide: Conflict in the catalog and *AACR2. Library Resources and Technical Services, 24*(1), 3–16.

Shera, J. H. (1950). Classification as the basis of bibliographic organization. In J. H. Shera & M. E. Egan (Eds.), *Bibliographic organization: Papers presented before the 15th annual conference of the Graduate Library School July 24–29, 1950* (pp. 72–93). Chicago: University of Chicago Press.

Smiraglia, R. P. (1987). Bibliographic control theory and nonbook materials. In S. S. Intner & R. P. Smiraglia (Eds.), *Policy and practice in bibliographic control of nonbook media* (pp. 15–24). Chicago: American Library Association.

Smiraglia, R. P. (1992). Authority control and the extent of derivative bibliographic relationships. Unpublished Ph.D. thesis, University of Chicago.

Smiraglia, R. P. (1999). Derivative bibliographic relationships among theological works. In L. Woods (Ed.), *Proceedings of the 62nd annual meeting of the American Society for Information Science* (pp. 497–506). Medford, NJ: Information Today.

Smiraglia, R. P. (2000). Works as signs and canons: Toward an epistemology of the work. In C. Beghtol, L. C. Howarth, & N. J. Williamson (Eds.), *Dynamism and stability in knowledge organization* (Proceedings of the 6th International ISKO Conference, July 10–13, 2000, Toronto, Canada). Advances in knowledge organization (vol. 7, pp. 295–300). Würzburg: Ergon Verlag.

Smiraglia, R. P. (2001). *The nature of a work.* Lanham, MD: Scarecrow Press.

Smiraglia, R. P., & Leazer, G. H. (1995). Toward the bibliographic control of works: Derivative bibliographic relationships in the online union catalog. *OCLC Annual Review of Research 1995.* Dublin, OH: OCLC Online Computer Library Center, Inc.

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.

Strout, R. F. (1956). The development of the catalog and cataloging codes. *Library Quarterly, 26*(4), 254–275.

Svenonius, E. (1981). Directions for research in indexing, classification and cataloging. *Library Resources & Technical Services, 25*(1), 88–103.

Svenonius, E. (1992). Bibliographic entities and their uses. In R. Bourne (Ed.), *Seminar on Bibliographic Records* (Proceedings of the Seminar held in Stockholm, August 15–16, 1990, sponsored by the IFLA UBCIM Programme and the IFLA Division of Bibliographic Control). UBCIM Publications, New Series (vol. 7, pp. 3–18). München: K. G. Saur.

Svenonius, E. (2000). *The intellectual foundation of information organization. Digital libraries and electronic publishing.* Cambridge, MA: MIT Press.

Taylor, A. G. (1988). Research and theoretical considerations in authority control. *Cataloging and Classification Quarterly, 9*(3), 29–56.

Taylor, A. G., & Paff, B. (1986). Looking back: Implementation of *AACR2. Library Quarterly, 56*(3), 272–285.

Tillett, B. A. B. (1987). *Bibliographic relationships: Toward a conceptual structure of bibliographic information used in cataloging.* Unpublished Ph.D. thesis, University of California, Los Angeles.

Vellucci, S. L. (1994). Bibliographic relationships among musical bibliographic entities: A conceptual analysis of music represented in a library catalog with a taxonomy of the relationships discovered. Unpublished Ph.D. thesis, Columbia University.

Vellucci, S. L. (1997). *Bibliographic relationships in music catalogs.* Lanham, MD: Scarecrow Press.

Vellucci, S. L. (1998). Bibliographic relationships. In Jean Weihs (Ed.), *The principles on future*

of AACR (International Conference on the Principles and Future Development of AACR, Toronto, October, 23–25, 1997) (pp. 105–146). Ottawa: Canadian Library Association.

Verona, E. ([1961] 1981). The function of the main entry in the alphabetical catalogue—A second approach. In A. H. Chaplin & D. Anderson (Eds.), *International Conference on Cataloguing Principles Report* (pp. 145–157). London: IFLA International Office for UBC.

Verona, E. ([1959] 1985). Literary unit versus bibliographical unit. In M. Carpenter & E. Svenonius (Eds.), *Foundations of cataloging* (pp. 155–175). Littleton, CO: Libraries Unlimited.

White, H. D., & McCain, K. W. (1997). Visualization of literatures. *Annual Review of Information Science and Technology, 32,* 99–168.

Wilson, P. ([1968] 1978). *Two kinds of power: An essay in bibliographical control.* California library reprint series. Berkeley: University of California Press.

Yee, M. M. (1993). *Moving image works and manifestations.* Unpublished Ph.D. thesis, University of California, Los Angeles.