Further Progress Toward Theory in Knowledge Organization

Vers une théorie de l'organisation des connaissances : derniers développements

Richard P. Smiraglia
Palmer School of Library
and Information Science
Long Island University
Brookville, New York USA
Richard.Smiraglia@liu.edu

Abstract: Theory is a system of testable explanatory statements derived from research. Theory 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 deliberate, self-conscious, and controlled observation of phenomena, whether this has taken place in the positivist empirical paradigm or in the qualitative paradigm. 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 three areas, research has accumulated to a degree sufficient to approach the positing of preliminary theoretical statements. These are: 1) author productivity and the distribution of name headings; 2) the work phenomenon; and 3) external validity. Research in these areas is examined, leading to the positing of three preliminary theoretical statements. These are: 1) most names in bibliographic databases will occur only once, and a very small number, which can be predicted by Lotka's Law, will occur many times; 2) the work phenomenon reflects an association with Lotka's Law, such that many works exist in only one instantiation, but a large proportion evolves steadily over time; and 3) there is a beginning of evidence that there are grounds for external validity in the examination of knowledge entities. Further domain-specific analysis in each of these areas is necessary to lead to strengthened predictive capability of theory that might emerge. Other theoretical statements might soon be possible. Such explanations could give us real predictive power for the development of sophisticated systems for the retrieval of knowledge entities.

Résumé: La théorie est un système d'exposés d'énoncés vérifiables, dérivés de la recherche. La théorie explique les milieux dans lesquels nous fonctionnons, les phénomènes présents dans ces milieux, et les voies par lesquelles ils pourraient être affectés par manipulation ou changement. La théorie est dérivée de l'observation délibérée, lucide, et contrôlée des phénomènes, effectuée dans le cadre du para-

© *The Canadian Journal of Information and Library Science / La Revue canadienne des sciences de l'infomation et de bibliothéconomie* 26, no. 2/3 2001

digme empirique positiviste ou dans le cadre du paradigme qualitatif. Dans le domaine de l'organisation de des connaissances, la production de la théorie s'est déplacée, passant d'un positionnement épistémique du pragmatisme, basé sur l'observation de la construction des outils de recherche, vers l'empirisme, basé sur les résultats de la recherche empirique. Dans trois domaines, la recherche s'est accumulée à un degré suffisant pour postuler des énoncés théoriques préliminaires. Ces domaines sont: 1)la productivité des auteurs et la distribution des vedettes de noms; 2)la notion d'œuvre; et 3)la validité externe. La recherche dans ces domaines est examinée, ce qui mène à postuler trois énoncés théoriques préliminaires : 1) la plupart des noms dans les bases de données bibliographiques n'auront qu'une seule occurrence, et un très petit nombre, prévisible par loi de Lotka, auront une très forte occurrence; 2)la notion d'œuvre s'intègre à la loi de Lotka, de manière que plusieurs œuvres existent en une seule instanciation, mais une grande proportion évolue de façon constante avec le temps; et 3)il y a un début d'évidence qu'il y a raisons d'utiliser la validité externe dans l'examen des entités de connaissance. Une analyse plus poussée de chacun des domaines, spécifique à ceux-ci, est nécessaire pour mener à la prédictibilité renforcée de la théorie qui pourrait émerger. D'autres énoncés théoriques pourraient bientôt être possibles. De telles explications pourraient nous donner une véritable puissance prédictive pour le développement de systèmes sophistiqués de repérage des entités de connaissance.

1. Introduction

"Theory" is a difficult concept, not just in information science, but in science at large today. One is bombarded at once with constructs from the traditionally Spartan positivist paradigm, the emerging qualitative paradigm, and in our own discipline we encounter terms such as "metatheory," which seem to come to us from entirely different traditions. What is a theory? And more importantly, how does theory grow? The answer is complex, but perhaps not so complex that we cannot comprehend it.

Theory, however we understand the word, refers to a system of testable explanatory statements derived from research. The term has a difficult colloquial usage that is quite a lot less precise than its use in academe. Colloquially we understand theory to mean relatively vague "ideas" or "principles." However, as scientists and scholars we use the term to mean, quite precisely, statements, derived as a result of rigorous research and testing, that explain phenomena and relationships among them.

Theory exists in a system—one 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 deliberate, self-conscious, and 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. In the positivist paradigm theory supplies testable hypotheses for empirical research. Hoover (1984, 76) said that theory is essentially a collection of hypotheses linked by some sort of logical, conceptual framework. In the qualitative paradigm, grounded theory is the method by which we confirm our elemental observations. The power of theory is its explanatory capability. We can use theory to analyze, predict, and manipulate phenomena.

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. The modern catalogue was constructed according to very pragmatic points of view, and the progress of theory was observed in the advancing promulgation of rules for the construction of those tools. (A review of this progress from pragmatism to empiricism can be found in Smiraglia 2002a.) The potential postmodern bibliographic retrieval environment is less hierarchical in conception and more context- (or "domain-") dependent in construction. Postmodern constructs require the input of research in both positivist and qualitative domains.

The modern catalogue is the product of nineteenth-century visionaries (Panizzi, Dewey, Cutter, among others), who developed very pragmatic tools (catalogues and classifications), explaining as they did so the principles by which their tools were constructed. By 1950, this effort had a name—"Bibliographic Organization"—and a University of Chicago Graduate Library School conference recorded the role of bibliographic organization in civilization and classification as its basis. Eleven years later, the international movement was growing, and yearning for standardization. In 1961 the famous International Conference on Cataloging Principles in Paris brought together key thinkers to inform the future design of catalogues. Many successive codes for the construction of catalogues were developed throughout this period. In their proscriptions can be observed the progress of "theoretical" thinking about the ordering of recorded knowledge.

Two key monographs seem to stand at the portals of the end of modernism and the beginning of postmodernism in knowledge organization. The first by Patrick Wilson (1968), expounded a system for bibliographic apparatus, and provided the framework for empirical theoretical development. The second, by Svenonius (2000), asserts that knowledge organization is accomplished through a bibliographic language. The introduction of epistemology and ontology into the design of classification by Hjørland (1998) preceded the challenge by Mai (1999) for the postmodern era: to look beyond fixed hierarchical models for knowledge organization and to look toward context-dependent orders.

Logical positivism notwithstanding, rationalist and historicist stances have begun to come to the fore through the promulgation of qualitative methods, most notably those employed in classification, user-interface design, and in bibliometric research. Questions of file design, record construct, and entity-relationship definition were critical to the advancement of the catalogue as a tool of the modern age. Furthermore, empirical evidence of the incidence of bibliographic phenomena, and of searching behaviour would be critical to inform the rapid development of increasingly technologically complex systems for retrieval of not only bibliographic data, but also of full document texts, archival records, surrogates for museum artefacts, and so on. Fuel, as it were, for the postmodern era.

2. The approach of theory

In three areas, research has accumulated to a degree sufficient to approach the positing of preliminary theoretical statements. These are:

(a) Author productivity and the distribution of name headings
(b) The work phenomenon
(c) External validity

Research in these areas will be examined, leading to the positing of preliminary theoretical statements.

2.1 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% (Lotka 1926, 321). That is, only 40% of authors contribute more than one paper. Lotka was concerned with the attribution of author productivity as a measure of the influence of authors in specific subject areas. Cataloguing research, however, has demonstrated an ability to observe Lotka's Law operating in the bibliographic universe.

All of these studies were conducted to examine the frequency of occurrence of name headings in catalogues. The implementation of *The Anglo-American Cataloging Rules 2nd edition* (AACR2) in the period 1978-1981 initiated a period of extreme anxiety in the cataloguing community, because of the likelihood that a changed basis for the source of name headings would result in a loss of cataloguing productivity. The research cited here was conducted to discover the degree to which specific name headings reoccur in author catalogues. If the majority of name headings was unique, for example, then the impact of the new rules would be lessened because new headings would easily intermingle with the old in existing catalogues. On the other hand, if large numbers of name headings reoccur, then it would mean changes would have to be made to existing catalogue records in order to bring them into conformity with the new code. Studies were conducted over the course of a decade by Potter (1980), Taylor-Dowell (1982), McCallum and Godwin (1981), Papakhian (1985), Taylor and Paff (1986), Fuller (1989), and Weintraub (1991). These researchers reported surprisingly consistent results concerning the proportion of names that are unique, with a couple of notable exceptions. These results are summarized in Table 1 below.

Table 1: Proportions of unique author names in library catalogues

Study	Proportion
Potter 1	63.5%
Potter 2	69.33%
McCallum and Godwin	66%
Papakhian SR	47.6%
Fuller	61%
Weintraub	36.7%

Taylor-Dowell studied changes to headings in catalogue records; her results, though quite different in scope, support the consistent data in Table 1. In particular, she reported the proportion of headings that had been caused to change form under AACR2 but that would not conflict when entered into the catalogue, because they would be unique. These proportions were (1982, 101) 51.6%, 50% and 29.7% respectively in small, medium, and large library catalogues. (Taylor and Paff [1986] did not report similar data, but they did confirm the validity of Taylor-Dowell's other projections about changes in name headings, suggesting vaguely that some universal factor is at work in the distribution of names in catalogues.) These data are clearly of a different kind altogether than those in Table 1 above, so direct comparisons are not possible. However, these data do support the contention that in most catalogues large proportions of headings are unique, and that the degree of multiple-occurrence of headings is greater in large, research collections. In other words, the pattern observed in Table 1, demonstrating that in most homogeneous collections between 61% and 69% of names occur only once, likely represents a reality about the distribution of author names in academic library catalogues. And that pattern is clearly similar to the distribution of author productivity noted by Lotka.

Papakhian found a clearly different proportion, 47.6%, because he was looking at a collection of sound recordings, in which the occurrence of multiple entries (many recordings of major works of Western art music) is associated with an increase in multiple occurrence entries. Weintraub's very low figure is attributed to the presence of large numbers of works by prolific writers; note the similarity of this proportion to the low

proportion of changed headings Taylor reported for her "large" library (University of North Carolina at Chapel Hill).

These results, both the set of those that are consistent across studies and the two anomalies (Papakhian and Weintraub) come together to tell a story about the distribution of author names. That is, these two sets of data point to a preliminary theory—a system of explanations. 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. We will return to these phenomena below, when we compare proportions of works with multiple instantiations.

2.2 The work phenomenon

The pragmatic, document-based approach to the design of bibliographic retrieval systems has proven short-sighted in one key respect. That is, catalogue design failed to account for the interrelatedness of documents and their contents. Tillett (1987) was the first to conduct empirical research for the purpose of classifying and quantifying the range and complexity of bibliographic relationships. Following her lead, a number of authors have pressed forward with empirical and qualitative research on bibliographic relationships. We have begun to see clear evidence of the importance of the role of works—a prime form of document content—in the bibliographic universe.

Collectively, this research has begun to create a clearer picture of the nature of the work phenomenon. The patterns by which related instantiations are seen to multiply across the bibliographic universe, have been described in a number of ways, but there is beginning to be unity among formerly disparate descriptions (Smiraglia 2002b). Much of the empirical research is summarized by Smiraglia (2001), and more recent research, both empirical and qualitative, will appear in Smiraglia (2002b). Epistemological analysis of the work phenomenon appears in Smiraglia (2002c) as does a similar analysis of musical works in Smiraglia (2002d). My recent epistemological analysis yields the following definition of a work:

> A work is a signifying, concrete set of ideational conceptions that finds realization through semantic or symbolic expression. That is, a work embraces a set of ideas that constitute both the conceptual (signified) and image (signifier) components of a sign. A work functions in society

> in the same manner that a sign functions in language…. If a work enters a canon then its signifying texts may derive and mutate. Derivations may take one or more forms: 1) simultaneous editions; 2) successive editions; 3) amplifications; or, 4) extractions ... Mutations may take one or more forms as well: 1) translations; 2) adaptations; or 3) performances.

Here we see the key components of the phenomenon. We begin with an abstract conceptual point of origin, from which a variety of related instantiations might be generated. Key is understanding that the role of the work is more than pure documentation. There is a cultural function that is in some way determinative of the degree to which a given work will be seen to evolve across time.

From the most recent empirical analysis (Smiraglia 2002e) comes confirmation of three compelling aspects of the work phenomenon. In these, lie the seeds of a nascent theory of the work. Taking them one at a time, we can examine the accumulated empirical evidence. The studies presented here include a 1992 analysis of works from Georgetown University Library, a 1999 analysis of works from the OCLC WorldCat, two 1999 analyses of theological literatures (all summarized in Smiraglia 2001), and a 2002 analysis of bestsellers (reported in Smiraglia 2002e).

2.2.1. Proportion of works with multiple instantiations

The majority exist in only one edition, but substantial proportions generate bibliographic families through mutation and derivation over instantiations over time.

Table 2: Works with multiple instantiations

Site	Proportion	Confidence interval	Confidence level
Georgetown	49.9%	±4%	.095
WorldCat	30.2%	±4%	.095
Burke Theology	52.9%	±10%	.090
Bobst Theology	57.9%	±6%	.095
Bestsellers	98%	±2%	.095

What we see in the above table is the tendency of heterogeneous sampling frames to include works with multiple instantiations, with larger numbers of such works in more specialized collections (the difference between the low figure in the WorldCat, and the higher figure in theology). The bestsellers all exist in multiple instantiations, but even there the pattern is confirmed. Preliminary results suggest that approximately two-thirds of the works have extensive networks of instantiations; the others are published in multiples (two or three simultaneous editions, for instance) but no new instantiations appear over time. A tentative conclusion is that some cultural influence (other than sheer popularity) plays a role in determining which works will evolve across time.

2.2.2. Types of instantiation

Simultaneous and successive editions, and translations, predominate among the types of mutation and derivation observed. Thus, a distinct pattern emerges in which works that are likely to spawn large networks of instantiations are published simultaneously. With the proper cultural influence (as yet undiscovered), evolution begins with successive editions and translations. Other types of instantiation are much less prevalent. Data from four studies are compiled in Table 3; preliminary analysis of the bestsellers indicates the presence of a similar distribution among their networks of instantiations.

Table 3: Distribution of relationship types

Site	Simultaneous	Successive	Translation	Other
Georgetown	4.3	52.1	23.3	>7
WorldCat	7.1	55.5	6.8	>3
Burke Theology	34	76	26	>4
Bobst Theology	18	80	26	>4

2.2.3. Age of progenitor

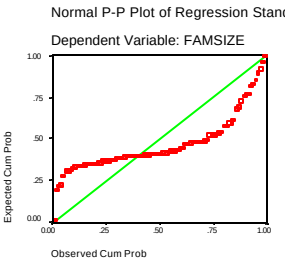
Older progenitors are associated with the largest bibliographic families. Regression coefficients demonstrate this effect in each study. This suggests that there is a relatively weak growth rate among instantiation networks such that, for evolving networks, the longer the time-span the greater the number of instantiations will appear.

Table 4: Age of progenitor and rate of evolution

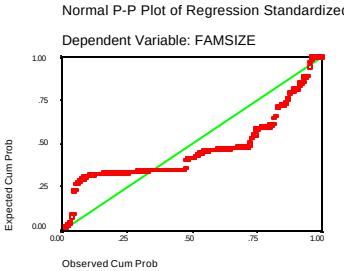
Site	Regression Coefficient	Constant
Georgetown	.123	.84
WorldCat	.002	.239
Burke Theology	.01	2.2
Bobst Theology	.006	2.3
Bestsellers	.002	.039

The coefficient indicates the probability of evolution, and the constant indicates the rate of evolution over time. Similar studies of music by Vellucci and films by Yee support these observations. Comparative plots are demonstrated in Figure 1.

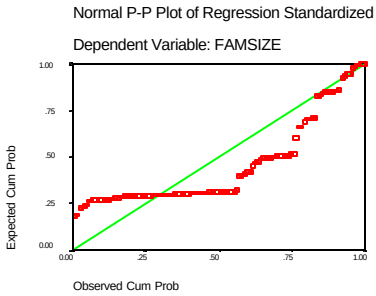
Union Theological Seminary



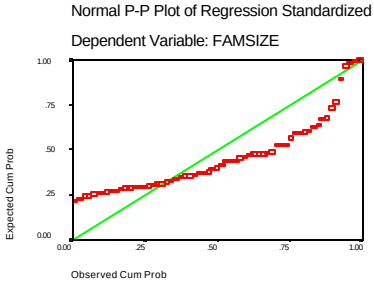
Bobst Library—NYU—Theology Collection



OCLC WorldCat



Bestsellers (mutating instantiation networks >7)



In sum, large proportions of works will generate multiple instantiations over time. Most instantiations will be simultaneous and successive editions, and translations. The longer the life-span of the work, the more likely there will be a large number of instantiations. And, the work phenomenon is broader than simply the control of documents in libraries. That is, there appears to be an as yet undisclosed cultural phenomenon that influences the selection of works that evolve. It is likely that similar patterns of evolution can be discovered among other sorts of documentary entities (web sites, for instance).

2.3 External validity

As in much of information science, the lack of comparative data that might provide the grounds for external validity hampered research in knowledge organization. In particular, it has been difficult to know the extent to which generalization could be made from empirical research describing phenomena that underlie the construction of catalogues. There are now indications that catalogues containing bibliographic records for similar collections of materials exhibit similar characteristics. The research reported in 2.1 above demonstrated the degree to which several researchers discovered similar proportions of single-occurrence name headings in research library catalogues. These studies support the contention that catalogues 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.

Countless other studies, notably those examining bibliographic relationships, have gathered data on the inherent characteristics of the documents in specific library collections. For methodological reasons, these data are too diverse to be

compiled in as formal a manner as we have used in 2.1 and 2.2 above. However, the essential bibliographic "demographics" reported in six studies demonstrate remarkable similarity among catalogues representing a variety of libraries and one bibliographic utility. Here, briefly, is a summary of those reports.

POTTER (1980)—Potter sampled from two academic library catalogues, the University of Wisconsin at Whitewater (170,000 titles, or medium-sized) and the University of Illinois at Urbana–Champaign (3 million titles, or large). He reported that, at Whitewater, the bulk of the collection is post-1965 imprints, with then-current acquisitions of approximately 10,000 titles per year of recent imprints. Illinois added at the time 85,000 titles per year both retrospective and current.

MCCALLUM AND GODWIN (1981)—McCallum and Godwin reported statistics on the size of the Library of Congress' MARC files in September of 1981. They reported 1.3 million records, of which 1.1 million were monographs, 90,000 serials, 60,000 maps, and 51,000 films. They reported at the time the acquisition of approximately 225,000 titles per year.

TAYLOR-DOWELL (1982)—Taylor-Dowell took random samples from three library catalogues: Greensboro College, The University of North Carolina at Greensburg, and the University of North Carolina at Chapel Hill. It was important to her methodology to discover the difference (if any) between distribution of name headings in small, medium, and large libraries. In this case, the sample was not taken at random from the catalogue at large, but rather from the catalogue records created during a given year. Thus, the bibliographic "demographics" reported are representative of the acquisitions of a single year in each library. Some summary figures were reported (43 ff.). In all three libraries the majority of acquisitions were monographic (>95%), US imprints predominated except in the large library, where they represented slightly less than half of acquisitions, and the majority were recent imprints (54%–74%).

TILLETT (1987)—Seeking to discover bibliographic relationships, Tillett studied the automated catalogue of the Library of Congress in 1987. She analyzed all catalogue records created between 1968 and July of 1986—2,854,252 bibliographic records. She reported the presence of: "nearly all languages and nearly all types of bibliographic items: books, serials, audiovisual materials, music, and maps" (1987, 126). The constitution by documentary medium was (1987, 127) predominantly monographic (81.6%).

WEINTRAUB (1991)—Replicating Potter, McCallum and Godwin, and Taylor-Dowell, Weintraub studied the catalogue if the University of California at San Diego Libraries. At the time of the sample the catalogue held 721,000 records. "Records for materials of all formats, time periods, and subjects held by UCSD libraries ... were included" (1991, 218).

SMIRAGLIA (1992) AND SMIRAGLIA AND LEAZER (1999)—These studies utilized random samples of bibliographic records, from which was extracted random samples of works. Thus, the initial samples provide the opportunity, with certain probabilities, to make representative statements about the contents of the collections. The 1992 study used a sample from the automated catalogues of the Georgetown University Libraries; the 1999 study used a sample from the OCLC WorldCat. A summary as reported in Smiraglia (2001, 83 ff.) indicated that the majority represented monographs (> 90%), were in English (> 66%), half were North American imprints (ca. 53%), and the majority (>73%) were published after 1960.

Table 5 includes a rough summary of these data—readers should bear in mind that the purpose of this table is to reveal trends, methodological problems leave no room for statistical comparison across these studies.

Table 5: Bibliographic "Demographics"

Study Author (Date)	Characteristic	Small	Medium	Large	Utility (WorldCat)
Taylor-Dowell (1982)	Monographs	98%	95.2%	95.7%	
	Serials	2%	4.8%	4.3%	
(ONE YEAR'S ACQUISITIONS)	Foreign imprints	12%	30.4%	53.5%	
	US imprints	88%	69.6%	46.5%	
	Retrospective 6 yrs.	44.4%	25.2%	26.3%	
	Current <6 yrs.	54.9%	74.8%	73.5%	
Tillett (1987)	Books			81.6%	
	Serials			11.3%	

(LIBRARY OF CONGRESS MARC RECORDS)	Maps	3.5%	
	Visual	2.8%	
	Music	.8%	
Smiraglia (1992)	Monograph	99%	
	English language	74.7%	
(RANDOM SAMPLE FROM GEORGETOWN U. CATALOGUE)	North American	53.5%	
	Before 1800	.4%	
	1800-1899	1.9%	
	1900-1959	19.9%	
	1960-	77.7%	
Smiraglia and Leazer (1999)	Monograph		90.6%
	English language		66.5%
(RANDOM SAMPLE FROM OCLC WORLDCAT)	North American		53.2%
	Before 1800		1.3%
	1800-1899		6%
	1900-1959		19.7%
	1960-		73.1%

It will come as no surprise to most that these studies all reveal the shape of most North American academic library collections. That is, most documents are monographs, most are published in English in North America, and most have been published since 1960. Nonetheless, it is an elementary indication that there is likely evidence that collections that are sufficiently alike in these demographic characteristics provide grounds for external validity in the application of research into individual bibliographic or documentary phenomena. Taylor-Dowell, Smiraglia, and Smiraglia and Leazer also report disciplinary, form, and medium categories, indicating that most of these monographs are narrative texts in the liberal arts and sciences. As the RLG Conspectus project has demonstrated, it is possible to compare collection strength in quite specific disciplinary catalogues (see for example, White [1995]). Taken together, it would seem that theoretical predictability about bibliographic phenomena might be possible.

3. Conclusions

Scholars working in the domain of knowledge organization are engaged in understanding the phenomena of knowledge and knowledge-artefacts. In particular, we seek to comprehend the natural orders of knowledge phenomena, and to impose order on knowledge artefacts so as to facilitate their storage and retrieval. The focus of diligent research for most of the past century, knowledge organization can be seen to have begun to generate preliminary theoretical statements. The research presented here demonstrates not only the beginnings of accumulation that can lead to theory, but also the utility of continued research in these dimensions.

Hoover (1984, 18) has suggested that the function of theory is to give meaning and motivation to scientific method by enabling interpretation of observed phenomena. Theory, then, is at one and the same time the responsibility of each researcher in a domain, as well as the responsibility of the domain as a collective community of scholars. The individual research is methodologically responsible—each observation must be undertaken with the conscious intention of allowing its result to contribute to theory. The community, on the other hand, is contextually responsible—the whole realm of observations must be monitored, compared, and contrasted so as to allow theoretical statements to emerge.

As we noted in the introduction above, all theory exists in a system. Useful theory provides contextual requirements and posits cause and effect among phenomena based on multiple empirical observations. Nagel suggested that formal theories have a tripartite structure, such that:

> (a) an abstract calculus ... is the logical skeleton of the explanatory system, ... that "implicitly defines" the basic notions of the system;
>
> (b) a set of rules ... in effect assign an empirical content to the abstract calculus by relating it to the concrete materials of observation and experiment; and,
>
> (c) an interpretation or model for the abstract calculus ... supplies some flesh for the skeletal structure in terms of more or less familiar conceptual or visualizable materials. (1979, 90)

It is clear that it is only with difficulty that any theoretical formulation could emerge from single acts of observation. Rather, it is up to the domain acting as a community to provide the "abstract calculus" that will serve as the logical, contextual skeleton of any explanatory system. But this, taken together with the empirical content supplied in the many individual observations will allow the emergence of useful explanatory theory. And useful explanatory theory will itself fuel further investigation into the underlying parameters of the domain itself.

Hjørland has called for meta-theoretical explanations of phenomena studied in our larger discipline of information science, incorporating particularly the philosophical bases that underlie research in knowledge organization. He writes:

> Meta-theoretic assumptions are thus broader and less specific than theories. They are more or less conscious or unconscious assumptions behind theoretical, empirical, and practical work. Meta-theoretical assumptions are connected to philosophical views, and are often parts of interdisciplinary trends. (1998, 607)

More recently, Hjørland (2001 and 2002) and White (2002) have written about the importance of meta-analysis as a tool for information science. Essentially the task of combining and summarizing the results of multiple studies, meta-analysis can yield meta-theoretical constructs that comprehend phenomena in a context much broader than the original domain. Along these lines, Mai (1999) has challenged the domain of knowledge organization to seek contextual explanations of knowledge phenomena. His intent is both that we must look beyond modernist

attempts to uncover an unyielding universal order as the engine to fuel information storage and retrieval, and that to do so we must study knowledge phenomena in many and more specific contexts. At the same time, we must begin to engage in meta-analysis by comparing and summarizing results across our whole spectrum, and by seeking explanations outside our own, limited, domain.

Of course, there is as yet no single, formal, statement of theory of knowledge organization. However, we can posit, based on this review, three simple theoretical statements. These simple statements are the result of limited meta-analysis. They are:

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. Recently Huber (2002) has suggested that the rate of publication and the career duration of authors follow an exponential distribution for a wide range of fields, and that these exponential distributions are the result of continuous distributions of talent and tenacity. In other words, tautologically put, authors who write more, write more. Huber also suggests from his model that most prolific authors produce at a constant rate over the course of their careers. But there is more to it than mere author productivity; there are cultural factors at work as well. Smiraglia and 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"). Analysis of bestsellers, in an attempt to demonstrate the role of "popularity" in the generation of large instantiation networks, suggests that some as yet undiscovered cultural phenomenon plays a causative role in the evolution of works. 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% of records (names, etc.) in a file will be unique; another 40% will require extra effort to delineate the relationships among the knowledge entities that they represent.

2. The work phenomenon reflects an association with Lotka's Law, such that many works exist in only one instantiation, but a large proportion evolves steadily over time.

Research has shown that for some proportion of works there will be a complex set of interrelated entities that require explicit linkage to facilitate retrieval. Large proportions of works will generate multiple instantiations over time. We still do not know what makes the essential difference between works that spawn multiple instantiations and those that do not. However, it appears that cultural influences, such as canonicity, are at work.

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

We have begun to observe similar distributions from one collection to another among the bibliographic characteristics that describe knowledge entities. In addition, the material reported in this paper suggests that these phenomena (name occurrence and work evolution) are interrelated. That is, most authors write only one work, but some authors are prolific. Most works appear in only one instantiation, but some evolve to a large extent over time. Homegeneity of any given collection can be associated with low rates of name occurrence and work evolution, but special collections report more exaggerated patterns. Bibliographic characteristics such as language, date, and place of publication serve as demographic data for knowledge artefacts. Most collections used for the research reported here have very similar demographic characteristics. 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.

These three statements all point to directions for further research. In particular, domain-specific analysis should be undertaken in each of these areas. More comparative data that can be compiled and evaluated through meta-analysis will lead to strengthened predictive capability of theory that might emerge from these research strains. And, meta-analytical statements must be in turn analyzed from perspectives external to our domain as Hjørland suggests in order to provide useful theory.

Other theoretical statements might soon be possible as well. For instance, research on the nature of subject searching in library catalogues

has suggested cognitive aspects of user behaviour are at least as important as the subject characteristics of the documents represented. Another area rife for theoretical development is the extensive work of co-citation and co-word 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 behaviours that lead to these intellectual relationships. It is possible that the undiscovered cultural phenomena cited in the present paper could be informed by explanations developed through co-citation or co-word analysis. Finally, semiotic analysis holds promise for increasing our understanding of seemingly anomolous phenomena. Such explanations could give us real predictive power for the development of sophisticated systems for the retrieval of knowledge entities.

References

Cutter, Charles Ammi. 1876. *Rules for a printed dictionary catalog*. Washington: USG.P.O.

Dewey, Melvil. 1876. *A classification and subject index for cataloging and arranging the books and pamphlets of a library.* Amherst, MA: M. Dewey.

Fuller, Elizabeth. 1989. Variation in personal name headings and title page usage. *Cataloging & classification quarterly 9*(3), 75–96.

Hjørland, Birger. 1998. Theory and meta-theory of information science: a new interpretation. *Journal of documentation 54*: 606–621.

—. 2001. Why is meta analysis neglected by information scientists? Letter to the editor, *Journal of the American Society for Information Science and Technology* 52: 1193–1194.

—. 2002. Meta-analysis should also be visible inside information science. Letter to the editor: rejoinder, *Journal of the American Society for Information Science & Technology* 53: 324.

Hoover, Kenneth R. 1984. *The elements of social scientific thinking*. 3rd ed. New York: St. Martin's Press.

Huber, John C. 2002. A new model that generates Lotka's Law. *Journal of the American Society for Information Science and* Technology 53: 209–219.

International Conference on Cataloging Principles. 1961, 1981. *Report*, ed. By A.H. Chaplin and Dorothy Anderson. London: IFLA International Office for UBC.

Lotka, Alfred J. 1926. The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences* 16: 317-23.

Mai, Jens-Erik. 1999. A postmodern theory of knowledge organization. *Proceedings of the 62nd annual meeting of the American Society for Information Science*, ed. Larry Woods, 547–556. Medford, NJ: Information Today, 1999.

McCallum, Sally H. and Godwin, James L. 1981. Statistics on headings in the MARC file. *Journal of library automation 14*: 194-201.

Panizzi, Antonio. 1841, 1985. Rules for the compilation of the catalogue. In *Foundations of descriptive cataloguing*, ed. Carpenter, Michael and Svenonius, Elaine, 3-14. Littleton, CO: Libraries Unlimited.

Potter, William Gray. 1980. When names collide: Conflict in the catalog and AACR2. *Library resources & technical services 24*: 3-16.

Smiraglia, Richard P. 1992. Authority control and the extent of derivative bibliographic relationships. PhD dissertation, University of Chicago.

—. 2001. *The nature of a work*. Lanham, MD: Scarecrow Press.

—. Forthcoming 2002a. The progress of theory in knowledge organization. *Library trends*.

—. Forthcoming 2002b. *Works as entities for information retrieval*. Binghamton, N.Y.: Haworth Press. (Also issued in *Cataloging & classification quarterly* 34.)

—. Forthcoming 2002c. Works as signs, symbols, and canons: The epistemology of the work. *Knowledge organization 28:* 192-202.

—. Forthcoming 2002d. Musical works and information retrieval. *Notes: The quarterly journal of the Music Library Association 58:* 747-64.

—. Forthcoming 2002e. Crossing cultural boundaries: Perspectives on the popularity of works. Paper presented at the 7th International ISKO Conference 2002 Granada, Spain.

Smiraglia, Richard P. and Leazer, Gregory H. 1999. Derivative bibliographic relationships: The work relationship in a global bibliographic database. *Journal of the American Society for Information Science 50:* 493–504.

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

Taylor, Arlene G. and Barbara Paff. 1986. Looking back: Implementation of AACR2. *Library quarterly* 56: 272-85.

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

Tillett, Barbara Ann Barnett. 1987. Bibliographic relationships: Toward a conceptual structure of bibliographic information used in cataloging. PhD dissertation, University of California, Los Angeles CA.

Weintraub, Tamara S. 1991. Personal name variations: Implications for authority control in computerized catalogs. *Library resources & technical services* 35: 217–228.

White, Howard D. 1995. Brief tests of collection strength: A methodology for all types of libraries. Westport, CT: Greenwood Press.

—. 2002. Library and information science (LIS) in aid of meta-analysis. Letter to the editor, *Journal of the American Society for Information Science and Technology* 53: 323.

Wilson, Patrick. 1968, 1978. *Two kinds of power: an essay in bibliographical control*. Berkeley: California Library Reprint Series, University of California Press.