
Pergamon

Library & Information Science Research
25 (2003) 157–176

Library & Information Science Research

Knowledge organization in research: A conceptual model for organizing data[1]

Lisa M. Given*, Hope A. Olson

School of Library and Information Studies, 3-20 Rutherford South, University of Alberta, Edmonton, AB Canada T6G 2J4. E-mail address: lisa.given@ualberta.ca (L.M. Given).

Abstract

Organizing research data for effective analysis has been insufficiently addressed in the methodological literature. This article proposes that concepts of knowledge organization relating to relevance, precision, recall, coextensiveness, exhaustivity, specificity, and consistency offer a ready-made model that can be applied to research data. The knowledge organization (KO) model is reinterpreted for transferability to quantitative, qualitative, and textual research. In each instance, the model's applicability is illustrated with examples from the authors' research. This exploration demonstrates the model's resiliency in organizing numeric data, coding transcripts, and marking up textual statements. The limitations of the model are noted and compromises are described, providing a valuable approach to meaningful data preparation for researchers, educators, students, and reviewers of research across disciplines. The article concludes that the KO model contributes significantly to the ability of researchers to collect and organize data in a manner most likely to shed light on research problems they address.

1. Introduction

Quantitative, qualitative, and textual (QQT) methodologies are framed by formal and informal processes for organizing research data toward a common goal—effective data analysis. Although research methods texts offer some direction for developing quantitative variables, coding qualitative data, and identifying themes in texts (and although software

* Corresponding author.
[1] An earlier version of this article won the 2002 Methodology Paper Competition, awarded annually by the Association for Library and Information Science.

0740-8188/03/$ – see front matter © 2003 Elsevier Inc. All rights reserved.
doi:10.1016/S0740-8188(03)00005-7

packages such as *SPSS* and *Ethnograph* have eased the process of data management), few approaches explicitly examine the intellectual work involved in organizing research data. In information science, strategies for effective organization and retrieval of information have been honed by examining specificity, exhaustivity, relevance, and other elements central to the process of subject analysis. By linking the research process to the concepts and techniques of knowledge organization, researchers, educators, students, and reviewers of research across academic disciplines (including library and information studies) may benefit from a new and valuable approach to meaningful data preparation.

In quantitative methodological approaches, concepts, variables, and hypotheses are selected before the study begins, a priori, and remain static throughout data collection and analysis. Through a deductive process, theories and hypotheses are tested to develop generalizations that will enable the researcher to predict, explain, and understand certain phenomena. In studies that use inductive, qualitative, methodological approaches, thematic codes emerge from the information gathered during and following data collection — that is, a posteriori. These codes, analyzed with the context of the data gathered, aid in the development of patterns and theories to explain various phenomena. Textual approaches exhibit features similar to both quantitative and qualitative methodology; texts are selected a priori based on the themes or discourses being considered, and manifestations of those discourses emerge a posteriori through close reading.

This article presents a technique for preparing research data for effective analysis that is appropriate across the range of QQT methodologies. This technique makes explicit the intellectual work involved in data preparation and links the established core of concepts for knowledge organization to the process of organizing research data in three sections: (1) definition of the problems in organizing research data, (2) an explanation of a model of useful concepts for knowledge organization, and (3) application of this model to extended examples from recent QQT research in library and information studies.

2. The problem of organizing research data

In organizing data for analysis, the ideal is to turn the raw data into a beautiful data rainbow, with predetermined categories or emergent themes as distinct as the colors of the spectrum fitting into an overarching structure that makes sense given the research problem. Few projects fit this ideal, however, and categories more commonly resemble a game of pick-up sticks. Themes are identified like the colors of the sticks but need to be picked carefully from the pile during the analysis process. In a worst-case scenario, data are splattered in many directions like the colors in a Jackson Pollock painting. Potential themes may be identifiable or respondents may have selected from the predetermined categories, but overall, the data give little direction for rigorous analysis. How then do researchers prepare and manage data so that the end result looks more like a rainbow and less like splatter? This article proposes one potential solution: the application of core concepts of knowledge organization to the data analysis process in which virtually all researchers engage.

As individual researchers struggle to reflect the ideal of the data rainbow, the need to manage data effectively in this pursuit is a constant refrain in methodology texts. The lack of guidance in these texts for achieving effective data management is symptomatic of the lack of recognition of knowledge organization as a potential contributor to the creation of successful processes for data analysis. De Vaus' (1996) *Surveys in Social Research*, for example, described the process of turning research concepts into appropriate quantitative indicators for questionnaire design. He noted that whereas many indicators are quite simple to develop (e.g., marital status, education level), others are more complex and difficult to craft (e.g., social isolation). Crabtree and Miller (1992) outlined four qualitative analysis styles and associated research techniques (from content analysis to heuristics). In each of these four, the development of categories is highlighted as central to the data analysis process. In textual approaches to research, such as Barzun and Graff's (1985) *The Modern Researcher*, the same necessity for organization is well recognized: "The ways of notetaking, abstracting, indexing, and classifying are applicable to any subject of research" (p. 31). They expressed the need to categorize statements drawn from texts used as resources, particularly in relation to history, whereas theorists such as Foucault (1972) expressed the need to reject established connections between statements to reassess the roles they play in different discourses by identifying themes not previously apparent.

Although methods texts like these are intended to guide researchers through the development of quantitative categories, the construction of qualitative codes, and the identification of textual themes for later analysis, two key trends in this literature tend to complicate the process. First, methods texts tend to gloss over the process of variable categorization and data coding in favor of discussions of data collection, data analysis, and the writing process. De Vaus (1996), for example, presented only one brief section on the development of appropriate quantitative categories, by outlining three options for researchers to consider: first, examining measures developed and used in other studies to select appropriate categories (or wording of categories); second, talking with individuals ("informants") who have inside knowledge of the issue under study, to gain insight into appropriate questions and terminology; or, third, using qualitative methods (e.g., observation or unstructured interviews) as a first step, to develop appropriate categories for use in quantitative research. Qualitative methods texts are equally brief in their treatment of the code development process. For example, Crabtree and Miller (1992) noted that the goal in applying grounded theory "is to develop classifications and theory grounded in the particular social scene investigated" (pp. 26–27), but they offer little guidance for the process of developing these classifications. Textual methods, drawn from humanities research, are seldom documented as explicitly as either quantitative or qualitative methods in the social sciences. The issue of organizing "data" is raised in various contexts, however. Since the data for textual research are selections of statements drawn from texts of various types, one of the major issues is how many data to collect. The Modern Language Association's (1995) *MLA Handbook for Writers of Research Papers* raises this issue as one of the few concerns related to content that it documents. Barzun and Graff (1985) alluded to the relative exhaustivity and specificity applied in selecting statements and categorizing them (pp. 30–31) but do not explore the ramifications of those actions and how to adjust for desired results. Both of these texts follow in the tradition of (and, in fact,

document) the practice of taking notes on index cards and categorizing them with keywords or "slugs" that can subsequently be organized into an outline for analysis. This quest for structure is also present in poststructural research as in the binary oppositions (hierarchical dichotomies such as "male/female" or "reason/emotion") that are the basis for deconstruction (Olson, 1997) or the categorization of statements into themes that illustrate discourses in discourse analysis (Frohmann, 1994).

The second trend that is evident in these texts is a clear focus on the benefits of computer software packages (e.g., *SPSS* and *Ethnograph*) for data analysis. Zyzanski, McWhinney, Blake, Crabtree, and Miller (1992) noted the following:

> Computer programs now exist that make the analysis tasks much easier. Many word processors have data management functions that can do the essential tasks of data entry, data identification, and data manipulation. Finally, special-purpose software is also commercially available for text retrieval, data base management, and a variety of analysis functions. These are essential programs since qualitative research often results in large volumes of verbal text that must be coded, sorted, interpreted, and summarized using text-based analysis techniques. (p. 235)

Humanities computing offers rich potential for textual research but is still in its early stages of development. Encoding schemes (e.g., the Text Encoding Initiative [TEI]) and mark-up languages are predominantly used for the mechanics of text representation and less for analysis at this time. None of these applications is intended to provide guidance in terms of the intellectual work that is needed to develop codes that will be easy to search and retrieve in the software program and relevant to the research questions at hand. If the researcher's codes resemble data splatter, the software cannot help in the analysis.

3. The knowledge organization model

The principal concepts of knowledge organization and their interrelations, when construed broadly, offer a model for the intellectual work involved in preparing data for analysis. The following sections develop these concepts and relationships into a simple model (see Figure 1) for easy application in organizing research data. This model is unconventional in its application to organizing research data, but the fundamental concepts are the same as when an understanding of knowledge organization is used to develop and evaluate information retrieval (IR) systems.

The focus of the knowledge organization (KO) model is to retrieve, from a mass of information, that which is relevant. In applying this model to research data, relevance is determined by whether the information gathered and coded contributes to answering the research questions or hypotheses under study. This criterion resembles the concept of relevance as applied to search queries or users' needs but, in this case, specifically addresses the research problem. Similar to the determination of the relevance of documents retrieved from an IR system, relevance in relation to research data is a multifaceted

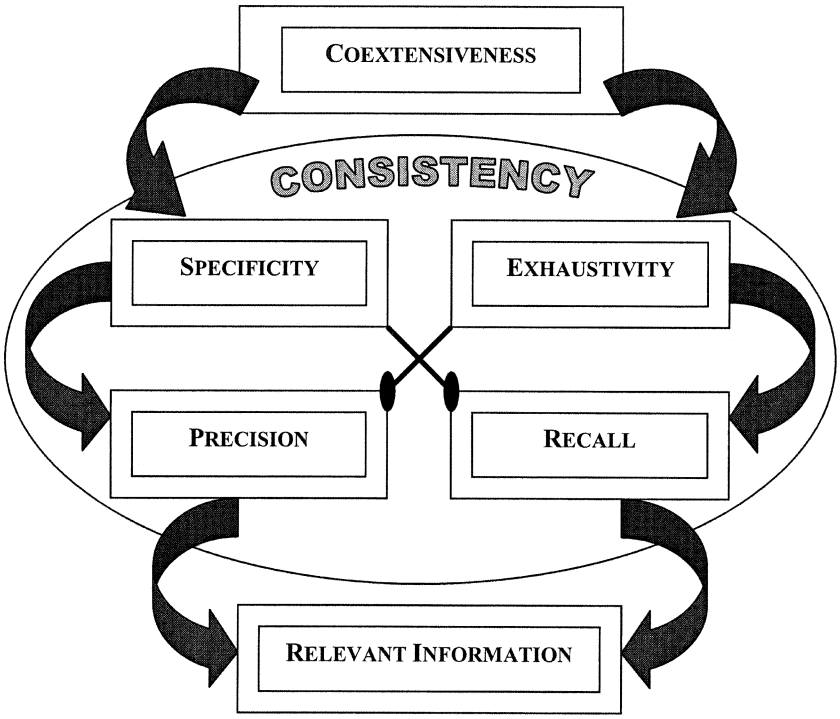

Fig. 1. Knowledge organization model.

concept. Relevance of particular data to the research question can be seen as akin to topical relevance.[2] It is similar to a simplistic view of "aboutness"—the idea that what the document (or datum) is about can be directly reflected in the indexing (or coding); Albrechtsen, 1990. A broader view of aboutness, however, can include anticipation of users' needs, just as a user-oriented view of relevance can take into consideration elements of utility and situation as well as topic in its simple sense. Similarly, the relation of data coding to the research question or questions needs to take into consideration a wide range of potential factors.

The KO model, when used for organizing information, flows from the choice of controlled vocabulary through its application via indexing to create a logically organized database ready for searching. When used for organizing research data, the model flows from the choice of variables or themes to be coded, through their application to data, to create an organized body of data ready for analysis. The following sections apply the KO model to examples of recent QQT research.

[2] See Schamber, Eisenberg, and Nilan (1990) for summary of different streams of relevance and Mizzaro (1997) for a detailed history of the concept of relevance. For the relation between relevance and aboutness, see, in particular, Hjørland (2001).

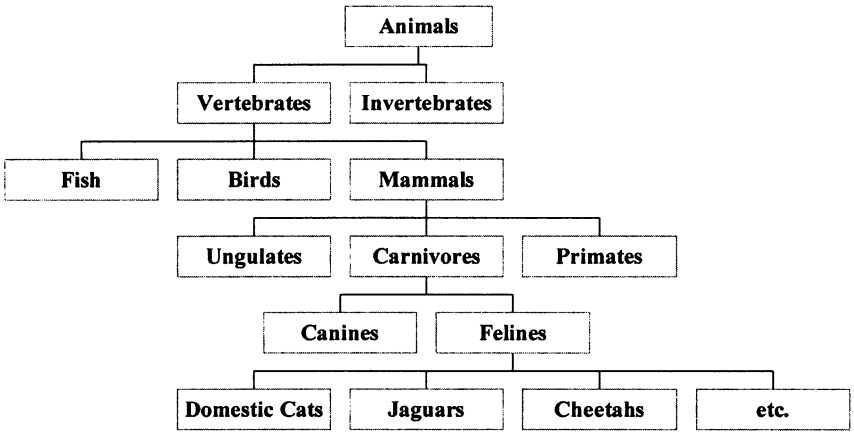

Fig. 2. Zoological taxonomy (selections).

3.1. Coextensiveness

The first element in developing a controlled vocabulary is to define its area (i.e., what it will cover) and then decide how that area will be divided into categories. The ideal is to make the categories coextensive with the aspects of the area to be covered to promote relevant retrieval (Milstead, 1984; Taylor, 1999). For example, a vocabulary to represent animals might follow the traditional zoological taxonomy based on certain physical characteristics, such as absence or presence of a backbone, diet, and means of reproduction (see Figure 2). In this taxonomic order, cats are categorized as felines, felines as carnivores, carnivores as mammals, and mammals as vertebrates. This defining of categories is useful if searchers are likely to have questions about carnivores or felines. If searchers want to know about pets, however, this arrangement is useless. They would benefit from an organization based on functions of animals in relation to humans (see Figure 3). In this arrangement, animals might be divided into domesticated and wild categories. Domesticated animals might be divided

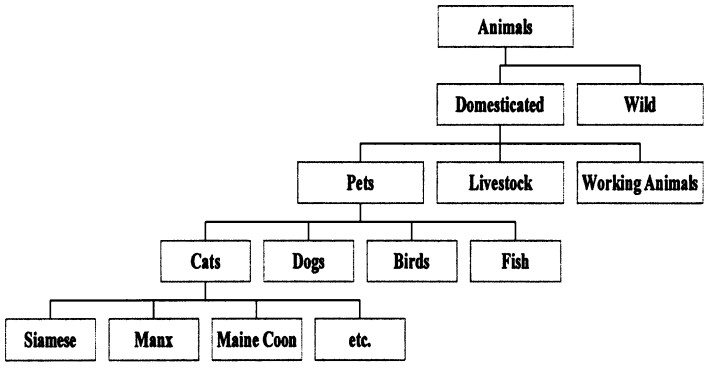

Fig. 3. Animals in relation to humans.

into pets, livestock, working animals, and so forth. In the class of pets, selected felines, canines, birds, fish, and even ants might be included. These animals would be scattered all over the zoological taxonomic organizational structure. There would be no category coextensive with the concept of "pets."

In organizing research data, the definition of variables and thematic codes must be coextensive with the concepts represented or implied by the research questions and/or hypotheses. For example, the concept of "library" may be defined as a physical place in one study (e.g., a user-needs analysis of the holdings of the local public library) or in relation to constructing discourses in another study (e.g., the library as democratizing agent—providing information access for all citizens). Depending on the goals of the particular study, the categories that are selected may differ substantially from another study, to be coextensive with the research problem at hand. Two other factors govern the vocabulary or encoding scheme and its application: specificity and exhaustivity.

3.2. Specificity and precision

Specificity can be viewed in at least two ways: as a characteristic of the vocabulary and as a factor in the application of a vocabulary (the process of indexing). In the first sense, specificity is the relative detail within the vocabulary—the number of hierarchical levels defined. As shown in Figure 3, a scheme that stops at differentiating only domesticated from wild animals would have very low specificity. The specificity increases with each level as the hierarchy becomes deeper, ending with the one differentiating particular breed of particular animals (e.g., Siamese cats), which has fairly high specificity. As with coextensiveness, the level of specificity should serve the needs of users of that organizational scheme; a veterinary index needs much higher specificity in regard to animals than a sociological index. In approaching research data, the research questions and/or hypotheses determine the level of specificity required for organizing the categories and codes assigned by the researcher. Once the level of specificity is defined, it must be applied to take full advantage of that specificity—the second way of viewing specificity. Specificity of application is standard indexing practice; each item must be coded at the most specific level available. Returning to Figure 3, this scheme would dictate that Siamese cats be indexed as "Siamese" and not as "cats," "pets," or "domesticated animals." Similarly, a scheme that categorizes different types of "libraries" might include "academic libraries," "public libraries," "prison libraries," and so on and would be applied in the indexing process at the most specific level.

Precision is one standard way of measuring how effectively a system retrieves relevant information. It refers to how much of the information retrieved in a search is relevant compared with how much is irrelevant. If precision is high, then all of the information retrieved is relevant and little or no irrelevant information is retrieved. For research data, this would mean that all of the data gathered for a particular variable or coded as having a particular attribute do actually have that attribute. It implies that categories can be mutually exclusive and that data can readily be assigned to those categories. To approach this ideal, it is important to remember that precision is enhanced by specificity: generally, the higher

the specificity, the higher the level of precision.[3] High specificity means that data are indexed (or categorized or coded) at a very precise level — the selected terms are finely grained, using detailed levels of categorization. Low specificity tends to result in high recall (Wellisch, 1991).

For example, if researchers are coding data about cats, they might assign only the code "cats" or they might assign codes for individual breeds if that level of specificity is desired. A problem arises, however, because this practice presumes that the chosen categories can be mutually exclusive. If particular breeds are categorized or coded, each cat must fit a breed. Cats that are crosses between breeds or do not belong to a recognized breed would fall between categories, even if these cats were the majority in the data set. Achieving precision may not be quite as obvious a task as it first seems.

3.3. Exhaustivity and recall

Exhaustivity is also an issue in both the development and application of the vocabulary. Exhaustivity is defined as the breadth of representation — the number of factors indexed or encoded. As with specificity, exhaustivity is a characteristic of both the vocabulary and its application. At the level of the vocabulary or encoding scheme, exhaustivity is concerned with the different aspects or facets included. For example, the structures in Figures 2 and 3 might be combined with facets to represent functions (e.g., digestion and reproduction), environment (e.g., aquarium, barn, and house), and type of contact with humans (e.g., food and companionship). Each facet that is added to the scheme raises the exhaustivity of the vocabulary. Returning to a categorization of types of libraries, exhaustivity would raise questions about the nature of each category within that grouping. For example, should libraries in publicly funded academic institutions be categorized separately from libraries in privately owned academic institutions? Should public libraries that have separate children's facilities be distinguished from those without? In a given research project, the nature of the research questions and/or hypotheses must drive these decisions.

In application, exhaustivity is related to the level of indexable matter; that is, how much of a given topic must be covered by an information-bearing document before it is represented in the scheme (Milstead, 1984). If half of a book is devoted to a topic, should that topic be represented? What if only a quarter of the book deals with that topic? What if only one chapter (or one subsection of a chapter) deals with the topic? What if the content is limited to one paragraph? Choices need to be made regarding the level of representation appropriate to the research problem under study. A related question is as follows: how many concepts will be represented for any particular item? Here, exhaustivity meets specificity. If a discussion on pets includes cats and dogs it seems reasonable to be

[3] For a clear explanation of the relationship between specificity and precision, see Cleverdon (1972). More recent work, such as the massive Council on Library Resources Online Public Access Catalog [CLROPAC] studies of the 1980s (Lawrence, 1984; Markey, 1984b; Matthews, Lawrence, & Ferguson, 1983), has generally reinforced Cleverdon's findings.

both specific and exhaustive in categorizing those concepts. If the discussion includes cats, dogs, fish, birds, rodents, rabbits, and llamas, however, it may not include enough information about any one type of animal to justify highly specific representations. In this case, "pets" might be the more appropriate choice. Returning to the first-mentioned library study (i.e., a user-needs analysis of public library holdings), the concept of "library" may be represented by highly specific categories for the analysis of holdings—including collections of "books," "audiotapes," "software," "videotapes," and so on. The researcher must make decisions based on the exhaustivity and specificity of the scheme; for example, if a children's book includes an audiotape of music to accompany the story, will the item be categorized under both "books" and "audiotapes"? In the second study (i.e., the discourse of library as democratizing agent), the categories may not make any reference to the physical attributes of libraries but may instead be represented by categories related to the conceptual ideal of the library—including "community involvement," "freedom of information," "universal access," and so on.

Exhaustivity is closely related to recall, the measure of how much of available relevant information or data is retrieved from the total available relevant information. Maximum recall means retrieving every last instance of a theme or variable. In achieving high recall, however, it is unlikely that one can retrieve all relevant information and no irrelevant information. That is why precision and recall tend to show an inverse correlation to each other and this factor will have an impact on the construction of data categories and codes for analysis as discussed later. One way to enhance recall is with high exhaustivity.[4] If exhaustivity is high, more codes are used, which allows more data to be retrieved and analyzed. Every last theme will be identified and coded. Each time another element is coded, it becomes more likely that that piece of data will be retrieved. Hence, each search or gathering of data for analysis will be larger and likely to contain a larger quantity of both relevant and irrelevant information.

4. Complications in the knowledge organization model

The notions of relevance, precision, recall, specificity, and exhaustivity seem like they could produce perfect categories and codes—the data rainbow—but as indexers know, there are several potential problems in applying these organizational concepts. Two issues particularly complicate pursuing ideal data organization: (1) the relationship between precision and recall and (2) problems of consistency.

[4] As with the relation between specificity and precision, the relation between exhaustivity and recall was clearly defined by Cleverdon (1972) and has since been confirmed in the CLR OPAC studies and by other researchers, such as Boyce and McLain (1989). Sparck Jones (1973) linked the level of exhaustivity in indexing inversely to the level of exhaustivity required in search formulation to achieve similar results in terms of precision and recall.

4.1. Precision versus recall

The first complication is the inverse relationship between precision and recall (as denoted by the crossed hammers in the center of Figure 1).[5] High exhaustivity tends to lower precision, because the addition of more and more codes results in the retrieval of irrelevant data alongside the relevant. Conversely, high specificity results in low recall. Since high specificity uses narrower categories, it produces fewer data in each category than does low specificity. Theoretically, it is possible to have an ideal level of both precision and recall, but in practice, this rarely occurs. When developing categories and codes, a researcher must decide which tendency is most important to the data analysis process. In some information-seeking studies, for example, the general category of "library" may be a sufficient marker of one type of location where individuals locate information to solve their information needs (in addition to "friends," "school," "family," "television," and other categories of information sources). In other studies, however, the research problem may dictate a much higher level of specificity and therefore multiple codes for the different types of libraries in which individuals find information (e.g., "academic," "corporate," "special," "school"). If the research demands this more specific approach, the researcher's job in analyzing the data is made all the more difficult; quantitative categories must include all types of libraries or qualitative coding must reflect each type as it emerges during data collection. In either case, the retrieval of multiple codes for analysis is more time consuming than for the singular term "library."

4.2. Consistency

Interindexer consistency, the second complication in the KO model, refers to solving problems of inconsistency in the application of terms and concepts. If indexers of a document use different levels of specificity or exhaustivity when creating and/or applying the vocabulary scheme, the end result will be documents that are difficult or impossible for searchers to retrieve (i.e., low recall and/or precision). If different categorizers or coders (or the same categorizers or coders at different times) use different levels of specificity or exhaustivity when preparing research data, the analysis will produce potentially misleading results. In a quantitative survey of library use, for example, inconsistent survey questions asked of patrons may produce conflicting results: one interviewer may ask, "Do you visit the library at least once a month?" whereas another may ask, "Do you visit the public library at least once a month?" Where the "library" terms are used at different levels of specificity, the answers to each question may vary considerably and produce misleading results. Both recall and precision will be affected because both depend on accuracy of categorization whether the emphasis is on specificity or exhaustivity. In this case, the use of standard interview questions will increase the level of consistency in the research data. In

[5] Precision and recall have been commonly considered to be inversely related at least since the Cranfield studies (Cleverdon, 1962) and the confirmation of Cleverdon's data as reworked by Swanson (1965). More recently, Buckland and Gey (1994) further reinforced this relationship. Fugmann (1994), however, made a cogent argument that this inverse relationship is a tendency, not a law.

qualitative research, inter- and intracoder reliability are commonly used to ensure the consistent application of thematic codes. Inconsistency introduces noise into the categorization and/or coding process and has the potential to yield irrelevant research results. Unfortunately, consistency is extremely difficult to achieve. The indexing literature has long been replete with studies that demonstrate considerable inconsistency, even among experienced professionals using familiar well-documented systems.[6] The three factors that most often increase consistency are as follows: documentation to aid application of vocabularies or encoding schemes, low specificity, and low exhaustivity. Obviously, the solution of consistency problems may create other inadequacies. Using low specificity and exhaustivity to achieve consistency will lead to low precision and low recall. Again, a balance is required, and this must be grounded in the research problem at hand. If the generic term "library" is sufficient for the purposes of the study, the researcher may be assured of more consistent results in using this as a research code or category; however, if the research dictates higher levels of specificity and exhaustivity, due diligence will be required to achieve consistency as closely as possible and implement checks (e.g., intercoder reliability) to enhance the rigor of the project. The same kind of distinction must be made in textual analysis, interpreting individual statements at appropriate levels of specificity and exhaustivity for the research problem while maintaining a consistency of interpretation sufficient to avoid inappropriate generalizations. Further, in textual research, it is important to allow enough freedom of interpretation for the serendipity of creativity.

5. Applications of the knowledge organization model to research data

The following sections of this article apply the KO model to quantitative, qualitative, and textual (QQT) data gathered and analyzed by the authors. Although the examples, in this case, are drawn from library and information studies, readers are encouraged to generalize these applications of the KO model to data collected in other areas of social sciences and humanities research.

5.1. A quantitative application of the knowledge organization model

In a quantitative approach, decisions about specificity and exhaustivity come at the point where variables and their values are defined. Exhaustivity relates to what and how many variables are chosen. The old and sound wisdom that mandates gathering only the data that are needed to satisfy the study's hypotheses is really a matter of exhaustivity, thus variables

[6] Lack of consistency in indexing has been established as a given for decades. It was well documented in the 1970s by researchers such as Preschel (1972) and Leonard (1977). Markey (1984a) reviewed earlier studies to find similar results. More recently, researchers such as Bertrand and Cellier (1995) continue to assess the variables involved in varying levels of consistency; and researchers such as Collantes (1995), Iivonen (1990), and Iivonen and Kivimäki (1998) have explored more ramifications of consistency that led to a better understanding of the way in which Zipfian distributions are linked to greater inconsistency with higher exhaustivity and specificity.

should be chosen to gather data relevant to those hypotheses. Since the best quantitative studies justify their variables on the basis of reasonable assumptions and prior knowledge regarding the research problem—often the result of earlier descriptive research—relevant variables can be defined with some confidence. The result is a level of exhaustivity and thus recall appropriate to the particular hypotheses. The values of each variable will determine the level of specificity of the data and, therefore, the level of precision. For example, many library-related studies explore patrons' use of library resources through quantitative survey design. Here, "use" may be categorized in many different ways (e.g., borrowing materials, reading materials in-house, attending storytime), and these categories inform both the development of appropriate survey questions and the values assigned in analysis packages (e.g., *SPSS*). Determining these values on the basis of sound assumptions, preliminary explorations, and the research problem will make collection of relevant data more likely to have an optimal level of precision, setting a framework for meaningful analysis.

The following example is from Given's (2000) dissertation research on the ways that mature undergraduates come to terms with being socially identified as "students" and the impact of this identification on these students' academic information behaviors. The study was comprised of two parts: (1) the manipulation of Canadian Census and university data to examine the demographic and academic characteristics of adult students engaged in formal education in Canada; and (2) in-depth qualitative interviews with 25 mature undergraduates at one Canadian university, to explore their personal perceptions of "student" life and their information-seeking activities. One of the first steps in designing the quantitative approaches used in the study was to operationalize the concept of "mature university students." This category was then applied to the examination of several variables coded in the Census and university data, including demographics and the major fields of study undertaken by mature university students in Canada. Although this process of operationalization was only one of many similar decisions made in the study, it illustrates the ways that the concepts of knowledge organization apply to quantitative data.

Several questions were addressed to categorize these students: Which "mature students" will be studied—those enrolled in undergraduate programs, graduate programs, or professional degree programs (e.g., law)? Is age the only relevant marker of a "mature student" or should other characteristics (e.g., marital status) be considered in identifying this group? How do other organizations (e.g., Statistics Canada, Canadian universities) define mature students—and are these definitions appropriate for the research problem at hand? First, high recall was an important consideration for the study: that every student meeting the profile of a mature university student could be accounted for in the data collection. At the university under study in the interview phase of the research, for example, students who had not applied for admission under the institution's special "mature admission" procedures were excluded from the university's statistics. A student who met the technical criteria set by the university (e.g., over age 21 and out of formal schooling for at least three years), but who had applied under standard admissions procedures, was not classed as a "mature student" and could not be tracked using the university's data-gathering methods. In this case, recall of relevant data was potentially very low, given the fact that many mature students on campus were not being properly identified. For those students who were tracked, however, the

university's criteria reflect a highly precise definition that is common to universities across Canada. These precise markers (e.g., over age 21 and out of formal schooling for at least three years) were chosen as the initial points of definition for the quantitative phase of the study. In addition, a range of degree programs was selected for the study but these were limited to undergraduate programs.

Specificity was another important consideration for the study—how fine the distinctions needed to be between types of mature students. Adults engage in many different types of educational endeavors; these include degree-credit and noncredit courses, design workshops, informal reading groups, leisure activities, and many others. It was important that the definition of mature students reflect the type of educational pursuit that matched the ultimate goal of full study, so the definition was limited to those students engaged in degree-credit courses and, more specifically, those enrolled in undergraduate degree programs. However, as this was an exploratory study of these students' experiences an exhaustive list of academic disciplines (e.g., science and visual arts) and a range of demographic categories were used to identify a cross-section of mature students.

The result is the quantitative data analysis model (see Figure 4), in which mature students were included across a range of disciplinary and demographic backgrounds (exhaustivity) and identified in such a way that students meeting the "mature student" criteria were not mistakenly excluded during data collection (high recall). Fine distinctions between types of adult learners were used to narrow the study to students engaged in formal degree-credit study (high specificity), and the criteria used to identify students as "mature" were relevant to the hypotheses under study (high precision). In addition, all of these decisions were made in light of the research problem guiding both phases of the study.

5.2. A qualitative application of the knowledge organization model

This section of the article uses an example from the qualitative interview phase of Given's (2000) research. The data analysis followed a grounded theory approach, where themes were coded as they emerged from the data in an ongoing and iterative process. The interview phase of this study included the following research question: What are the academic information behaviors in which mature university students engage? Information behaviors refer to any activity related to students' quests for information for their academic careers, from visiting a library, to asking a spouse for advice, to obtaining essay topics from the television news. Although this research question was only one of many addressed in the study, it offers an illustration of the ways that the concepts of knowledge organization can apply to real qualitative data.

The coding process for this question involved several considerations. First, the recognition of an information behavior theme in a transcript (e.g., reading a book) and coding that theme (e.g., "reading," for the act of reading; "book," for the item itself). Then, in examining additional pages of the transcript (or reviewing the next transcript), the question arose as to whether highly specific, finely grained codes were needed (e.g., to distinguish library books from books that the student owned) and how these very specific codes might come together under a higher-level category (e.g., "material sources," which could include books of all

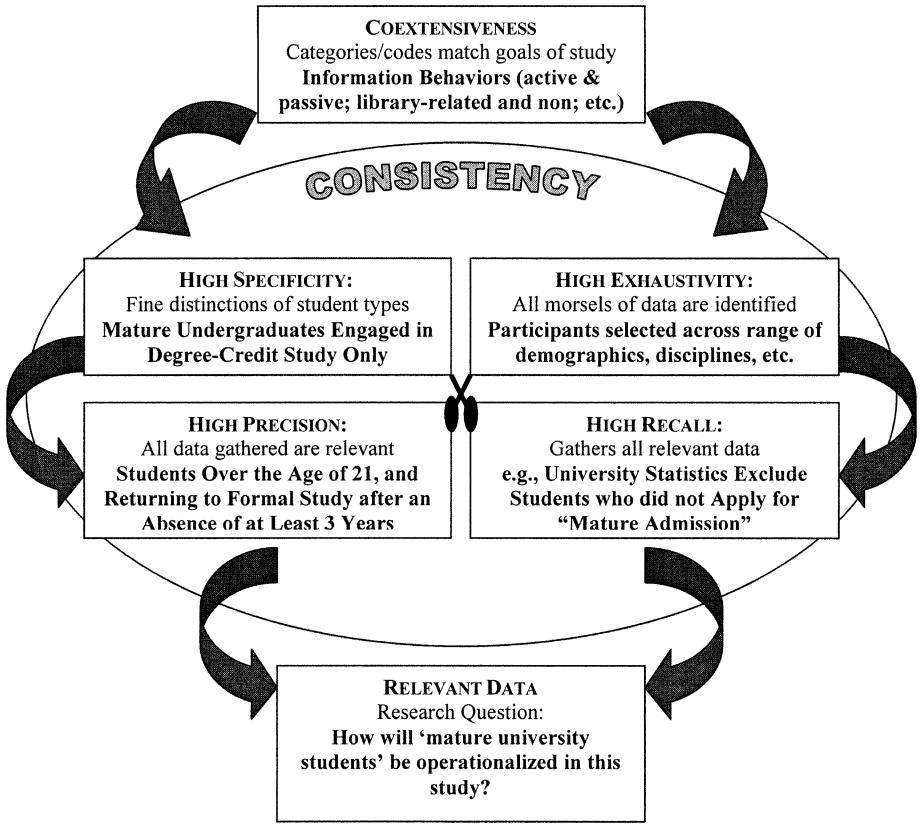

Fig. 4. Quantitative data analysis model.

types, as well as other material sources, such as computers). The key was to make choices about the level of specificity to ensure optimal precision. For example, instances of reading a book that were not related to the student's academic life (e.g., where they mentioned reading a bedtime story to a child) were not normally coded because they were not relevant to the research question. Where the process of reading a bedtime story brought to mind a potential research topic for a class paper, however, this instance of reading became relevant to the research question. Whether or not this single instance is sufficiently relevant to be assigned a separate code is a question of both specificity and exhaustivity.

It is the other half of the model that relates to exhaustivity, or this question: How many themes are needed to address the research questions? A completely exhaustive codebook is not practical because it would take a long time to develop all themes and code these across all transcripts. How exhaustive, then, should the codebook be? First, in considering each research question, the researcher must decide how many themes will contribute to identifying relevant data. Will the reading of a textbook also be coded for the time of day it was read (e.g., late at night, after the bedtime story) or where the student did the reading (e.g., at the kitchen table)? Second, to achieve optimal recall for each new theme or code that is assigned,

the researcher must go through each transcript (often many times) to code all instances. The more exhaustive the coding, the more iteration that is required. This iteration parallels the well-recognized device of iteration in query formulae (see, e.g., the series of Text Retrieval Conference [TREC] projects exploring iteration and relevance feedback; Belkin et al., 2001) or in query reformulation in the IR literature (Spink, Jansen, & Ozmultu, 2000). It is important to remember, however, that overcoding leads to extreme levels of exhaustivity and specificity and thus to low precision and low recall. The problem of overcoding occurs when researchers code beyond the research questions and include interesting themes that are simply not relevant. This problem can be difficult to avoid in a grounded theory approach, because the data captured in qualitative research are extremely rich and filled with engaging details. Also, one does not want to restrict coding to the point of missing relevant emergent themes (particularly those that were not anticipated when the study was first designed). The key, then, is to exert restraint when tempted to code "interesting" themes that have nothing to do with the original research questions.

It is also important to address the preiteration problem that occurs when new themes or specific codes for existing themes are added to the coding manual but the transcripts have not been thoroughly reviewed to code all instances of those themes. If thorough iteration is not followed for every theme in every transcript, low recall and precision will result. So, if a new code is added to represent "the place a book is read," another reading of transcripts is required to ensure that this theme has been coded exhaustively. Similarly, if "books" were originally coded, but the data suggest that more specific codes for "textbooks" and "library books" are required, recoding of earlier mentions of "books" is necessary. In either case, the researcher must ensure thoroughness in coding, or it will be impossible to effectively retrieve relevant data for analysis. A balance must always be struck to address the inverse relationship between precision and recall, and the fact that inconsistency tends to lurk in the research data. The key for qualitative coders is to approach consistency as closely as possible and to bear inconsistencies in mind when drawing conclusions from the data. Many qualitative texts refer to processes for testing inter- and intracoder reliability, which can enhance the level of consistency in the assigned codes.

The result is the qualitative data analysis model (see Figure 5), in which new themes are coded as they emerge from the data (exhaustivity) and the data are checked and rechecked in an iterative fashion to apply these new codes to all instances of the relevant themes (high recall). All themes chosen to be coded are relevant to the research questions (high precision) and decisions are made about the levels of specificity needed for each theme according to the research questions being addressed.

5.3. A textual application of the knowledge organization model

The kinds of textual interpretation that library and information studies are increasingly drawing from the humanities resemble qualitative approaches in their use of a posteriori encoding, but also share the a priori definition of themes characteristic of quantitative research. Thus, it is not surprising that the concepts of knowledge organization also apply to textual research. For example, a deconstruction is based on the notion of binary

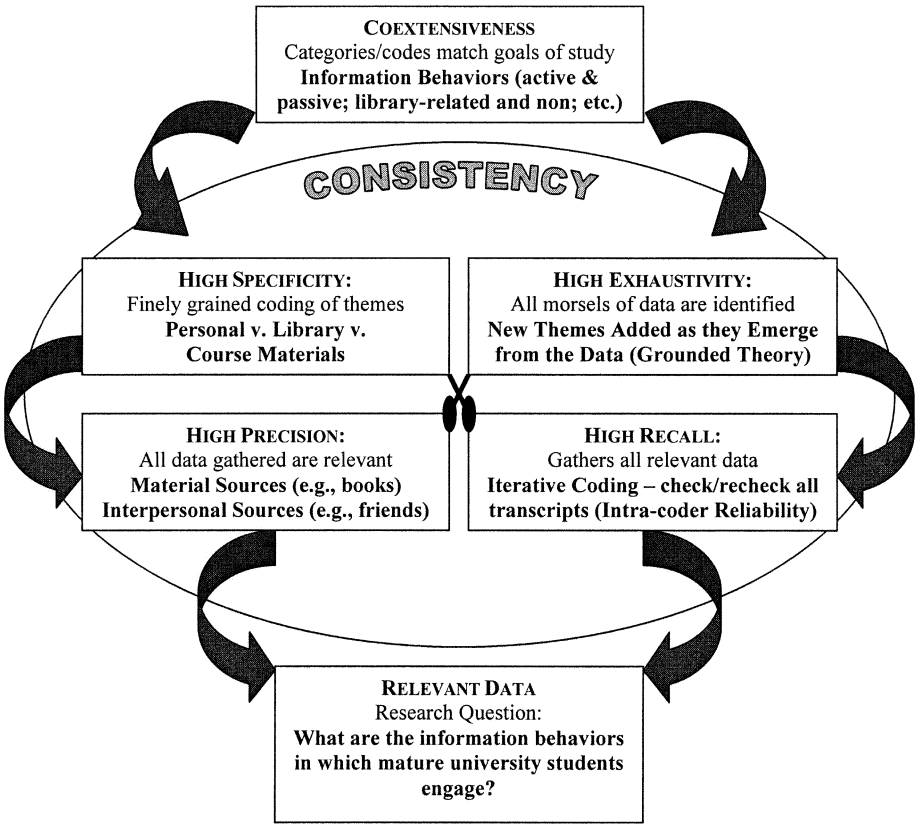

Fig. 5. Qualitative data analysis model.

oppositions—dichotomies that have one aspect subordinate to the other with the two aspects being mutually defining (Olson, 1997). Such binary oppositions may be at various levels of abstraction, so it is fruitful to look at concrete binary oppositions such as "male/female" at the same time as more abstract issues such as "mind/body" or "reason/emotion." Deconstruction shows that the boundary between the two elements of a binary opposition is constructed and can only address the research problem in a meaningful way if the binary and related binaries are at an appropriate level of specificity. Exhaustivity in deconstruction relates to the closeness of reading the texts. Passages that effectively demonstrate the binary opposition need to be sifted from passages that do not add to the analysis or the deconstruction will wander from its focus. Deconstruction shows that our realities are constructed.

Discourse analysis shows the factors that have constructed a particular reality or discourse formation. As mentioned previously, Foucault (1972) suggested rejecting the conventional structures of and relations between statements and beginning with as few preconceptions as possible to see how statements (written or otherwise) construct and are constructed by the powerful discourses of our society and culture. His *The Archeology of Knowledge* might be

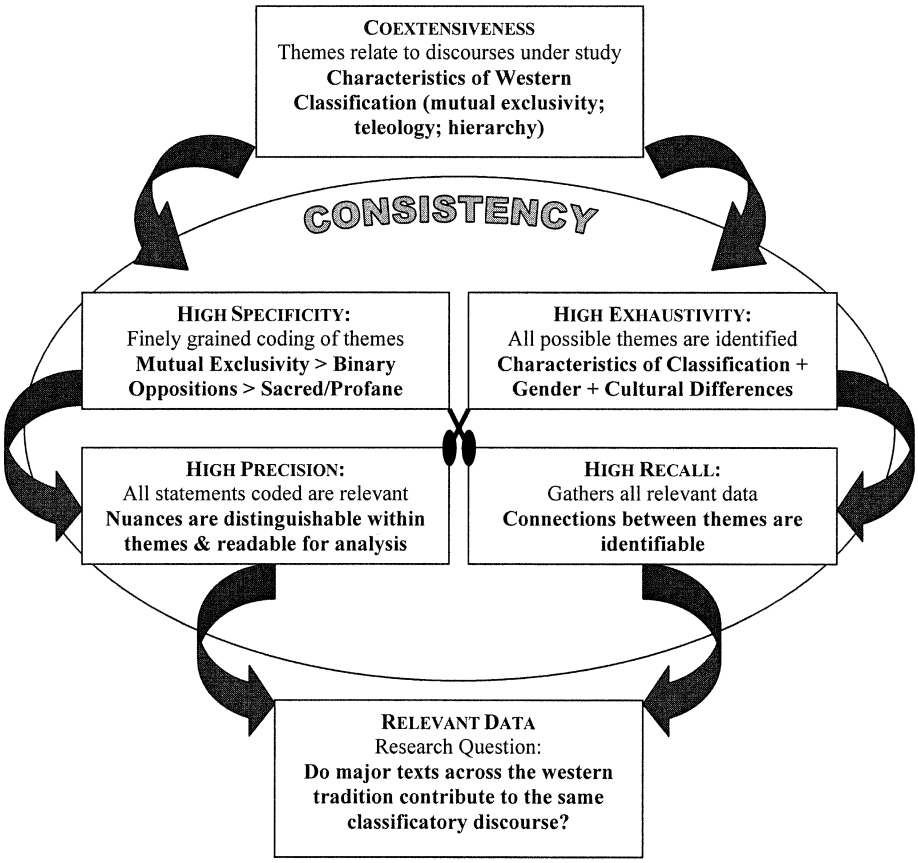

Fig. 6. Textual data analysis model.

described as a methodology for uncovering the knowledge organization systems that underpin discourses. Again, these discourses may be at various levels, so defining the themes is affected by both specificity and exhaustivity. One way of thinking about this issue is to concretize the discourses in particular texts by using an application of the TEI to markup the themes representing those discourses. Text may be implicit or explicit in its manifestations of particular discourses. In either case, the relevant passages must be identified to examine the construction of our realities.

An example of such research (shown in Figure 6) is currently being implemented by Olson[7] (1999, 2000, 2001) to explore the presumptions underlying practices of classification in a Western context. In this project, texts by Parmenides, Plato, Aristotle, Hugh of St. Victor, Bacon, Diderot, Coleridge, Darwin, Durkheim, and others are being used to document mutual exclusivity of categories, the teleological progression from concrete to abstract, and

[7] Hope Olson acknowledges the support of the Social Sciences and Humanities Research Council of Canada for this research.

hierarchical order as themes characterizing and enforcing the Western discourse of classification. This identification of themes is the a priori definition of the limits of the project (the coextensiveness in Figure 6). Where copyright permits, electronic versions of the texts (either already available or scanned as part of this project) are being encoded using a customized Extensible Markup Language (XML) Document Type Definition (DTD) augmented by TEI Lite.[8] The DTD defines themes relevant to the discourse under examination. The three themes formed a basis for developing the DTD but have been made more specific to meet the needs of analysis. For example, the theme of mutual exclusivity (the discreteness of categories) has the more specific code "binary oppositions," which can be further broken down into particular dualities such as "sacred/profane," "male/female," and so forth. The result (as shown in Figure 6) is greater precision for understanding the many possible nuances of a particular theme.

Analysis will include a general critical perspective as suggested by Foucault's technique but will also address concerns specific to feminist and postcolonial theory. Therefore, codes indicating statements related to gender and/or cultural difference are also included in the DTD, thus increasing the level of exhaustivity. Many statements within these texts are coded for multiple themes. For example, one passage from a text might be coded for its distinction between what is sacred and what is profane, the gendered nature of that distinction, and the implication that such a distinction is "civilized" since it is typical of Western culture. The result, as shown in the model, is a high level of recall reflecting the complexity of the texts. In the analysis, this mirroring of complexity will allow identification of connections between themes.

The texts thus coded will be amenable to reading as interactions of the various themes and also searchable for instances of particular themes. Both the individual occurrences and the concatenations of multiple themes will contribute to the analysis of these data.

6. Conclusion

The KO model is robust, growing from concepts rooted in centuries of practice and research. This article has synthesized these concepts and their relations to each other into a visual model and demonstrated that model's potential applicability to QQT methodologies. Concepts of coextensiveness, specificity, and exhaustivity may be consciously used in any type of research to expand or focus the data and their analysis. Relevance to the research problem is the guiding principle, with the organization of data being adapted to that end. This tailoring of the data through gathering and encoding provides relevant information for addressing research problems just as high-quality indexing and classification offer a means for obtaining relevant information from a knowledge organization system. Our development and exploration of the KO model offers a methodical conceptual approach to the

[8] The DTD defines codes unique to this purpose. The TEI provides codes common to other electronic text projects, most of which reproduce the text rather than analyzing its content.

organization of research data. The model can serve as a prism to transform undifferentiated data (like white light) into an orderly spectrum—an artificially created rainbow. As a rainbow is limited to the wavelengths that humans can perceive, so the spectrum of data is limited by the available codes, being more or less specific and exhaustive. Like a rainbow's colors blending into each other, research data may not fit into watertight categories. However, the data can be organized so that blues are grouped together, as are violets, whether or not the added specificity of indigo is coded as an intervening category. The nuances must be left for the subtleties of data analysis. Although not a panacea, the KO model offers researchers a means of organizing data that is both flexible and rigorous because of its conceptual foundation. Ultimately, the KO model can contribute to the overall quality of research.

References

Albrechtsen, H. (1990). Subject analysis and indexing: From automated indexing to domain analysis. *The Indexer*, *18*, 219–224.

Barzun, J., & Graff, H. F. (1985). *The modern researcher* (4th ed.). San Diego: Harcourt, Brace, Jovanovich.

Belkin, N. J., Cool, C., Kelly, D., Lin, S.-J., Park, S. Y., Perez-Carballo, J., & Sikora, C. (2001). Iterative exploration, design and evaluation of support for query formulation in interactive information retrieval. *Information Processing & Management*, *37*, 403–434.

Bertrand, A., & Cellier, J. (1995). Psychological approach to indexing: Effects of the operator's expertise upon indexing behaviour. *Journal of Information Science*, *21*, 459–472.

Boyce, B. R., & McLain, J. P. (1989). Entry point depth and online search using a controlled vocabulary. *Journal of the American Society for Information Science*, *40*, 273–276.

Buckland, M., & Gey, F. (1994). The relationship between precision and recall. *Journal of the American Society for Information Science*, *45*, 12–19.

Cleverdon, C. W. (1962). *Report on the testing and analysis of an investigation into the comparative efficiency of indexing systems* (Cranfield, UK). Unpublished report.

Cleverdon, C. W. (1972). On the inverse relationship of recall and precision. *Journal of Documentation*, *28*, 195–201.

Collantes, L. Y. (1995). Degree of agreement in naming objects and concepts for information retrieval. *Journal of the American Society for Information Science*, *46*, 116–132.

Crabtree, B. F., & Miller, W. L. (1992). Primary care research: A multimethod typology and qualitative road map. In B. F. Crabtree, & W. L. Miller (Eds.), *Doing qualitative research* (pp. 3–28). Newbury Park, CA: Sage.

De Vaus, D. A. (1996). *Surveys in social research* (4th ed.). London: UCL Press Limited.

Frohmann, B. (1994). Discourse analysis as a research method in library and information science. *Library & Information Science Research*, *16*, 119–138.

Foucault, M. (1972). *The archaeology of knowledge and the discourse on language (A.M. Sheridan Smith, Trans.)*. New York: Pantheon Books.

Fugmann, R. (1994). Galileo and the inverse precision/recall relationship: medieval attitudes in modern information science. *Knowledge Organization*, *21*, 153–154.

Given, L. M. (2000). *The social construction of the 'mature student' identity: Effects and implications for academic information behaviours*. Unpublished dissertation, The University of Western Ontario, London, Ontario.

Hjørland, B. (2001). Towards a theory of aboutness, subject, topicality, theme, domain, field, content, and relevance. *Journal of the American Society for Information Science*, *52*, 774–778.

Iivonen, M. (1990). Interindexer consistency and the indexing environment. *International Forum for Information and Documentation*, *15*, 16–21.

Iivonen, M., & Kivimäki, K. (1998). Common entities and missing properties: Similarities and differences in the indexing of concepts. *Knowledge Organization*, *25*, 90–102.

Lawrence, G. S. (1984). Lessons from the CLR public online catalog study. In M. Gorman (Ed.), *Crossroads: Proceedings of the First National Conference of the Library and Information Technology Association, 1983* (pp. 84–93). Chicago, IL: American Library Association.

Leonard, L. E. (1977). *Inter-indexer consistency studies, 1954–1975: A review of the literature and summary of the study results (occasional papers)*. Chicago, IL: Graduate School of Library Science, University of Illinois.

Markey, K. (1984a). *Subject searching in library catalogs: Before and after the introduction of online catalogs (OCLC Library Information and Computer Science Series No. 4)*. Dublin, OH: OCLC.

Markey, K. (1984b). Interindexer consistency tests: A literature review and report of a test of consistency in indexing visual materials. *Library & Information Science Research*, *6*, 155–177.

Matthews, J. R., Lawrence, G. S., & Ferguson, D. K. (Eds.) (1983). *Using online catalogs: A nationwide survey.* New York: Neal-Schuman.

Milstead, J. L. (1984). *Subject access systems: Alternatives in design*. Orlando, FL: Academic Press.

Mizzaro, S. (1997). Relevance: The whole history. *Journal of the American Society for Information Science*, *48*, 810–832.

Modern Language Association. (1995). *MLA handbook for writers for research papers*. New York: MLA.

Olson, H. A. (1997). The feminist and the emperor's new clothes: Feminist deconstruction as a critical methodology for library and information studies. *Library & Information Science Research*, *19*, 181–198.

Olson, H. A. (1999). Exclusivity, teleology and hierarchy: Our Aristotelean legacy. *Knowledge Organization*, *26*, 65–73.

Olson, H. A. (2000). Reading "Primitive Classification" and misreading cultures: The metaphysics of social and logical classification. In C. Beghtol, L. C. Howarth, & N. J. Williamson (Eds.), *Dynamism and stability in knowledge organization: Proceedings of the Sixth International ISKO conference, 10–13 July 2000, Toronto, Canada* (pp. 3–9). Würzburg, Germany: Ergon.

Olson, H. A. (2001). Cultural discourses of classification: Indigenous alternatives to the tradition of Aristotle, Durkheim and Foucault. In H. Albrechtsen, & J. Mai (Eds.), *Proceedings of the 10th ASIS SIG/CR Classification Research Workshop, October 13, 1999: Advances in classification research v. 10*, (pp. 91–106). Medford, NJ: Information Today, for the American Society for Information Science and Technology.

Preschel, B. M. (1972). *Indexer consistency in perception of concept and in choice of terminology.* New York: Columbia University.

Schamber, L., Eisenberg, M. B., & Nilan, M. S. (1990). A re-examination of relevance: Toward a dynamic, situational definition. *Information Processing & Management*, *26*, 755–776.

Sparck Jones, K. (1973). Does indexing exhaustivity matter? *Journal of the American Society for Information Science*, *24*, 313–316.

Spink, A., Jansen, B. J., & Ozmultu, H. C. (2000). Use of query reformulation and relevance feedback by Excite users. *Internet Research*, *10*, 317–328.

Swanson, D. R. (1965). The evidence underlying the Cranfield results. *Library Quarterly*, *35*, 1–20.

Taylor, A. G. (1999). *The organization of information*. Englewood, CO: Libraries Unlimited.

Wellisch, H. (1991). *Indexing from A to Z*. Bronx, NY: H.W. Wilson.

Zyzanski, S. J., McWhinney, I. R., Blake Jr., R., Crabtree, B. F., & Miller, W. L. (1992). Qualitative research: Perspectives on the future. In B. F. Crabtree, & W. L. Miller (Eds.), *Doing qualitative research* (pp. 231–248). Newbury Park, CA: Sage.