Organizing Knowledge = Organizing Data: Applying Principles of Information Organization to the Research Process

Lisa M. Given
School of Library & Information Studies
University of Alberta
lisa.given@ualberta.ca

Hope A. Olson
School of Library & Information Studies
University of Alberta
hope.olson@ualberta.ca

ABSTRACT

Organizing research data for effective analysis has been insufficiently addressed in the methodological literature. This paper proposes that principles of knowledge organization relating to relevance, precision, recall, exhaustivity and specificity offer a ready-made model that can be applied to research data. The primary example is drawn from qualitative research to demonstrate how the model can be reinterpreted for that context. Reference is also made to the model's transferability to quantitative and textual research.

RÉSUMÉ

L'organisation des données de recherche pour l'analyse efficace a été insuffisamment abordée dans la littérature méthodologique. Cette étude propose que les principes de l'organisation de connaissance qui se rapportent à la pertinence, à la précision, au rappel, à l'exhaustivité et à la spécificité offrent un modèle tout fait qui peut être appliqué sur les données de recherche. L'exemple primaire est tiré de la recherche qualitative pour démontrer comment le modèle peut être réinterprété pour ce contexte-là. L'allusion est also faite à la transferabilité du modèle à la recherche quantitative et textuelle.

INTRODUCTION

Qualitative, quantitative and textual interpretation approaches to research cover a wide spectrum, but share one important feature - the organization of research data to enable analysis. From the definition of variables, to the coding of in-depth interviews, to the close reading of texts, researchers must develop techniques for creating appropriate categories for gathering and interpreting their data. While research methods texts offer some direction for classifying and coding data, and while software packages (e.g. *Ethnograph*; *SPSS*) have eased the process of data management, there are few precise approaches that explore the intellectual work involved in organizing research data. By linking the research process to the principles and techniques of knowledge organization, researchers across many academic disciplines may benefit from a new and valuable approach to meaningful data analysis.

In information science, questions of relevance, specificity and exhaustivity are central to the process of subject analysis, offering strategies for effective information organization and retrieval. While these principles and techniques have been well-documented and honed within the field for many decades, their utility for the research

process has yet to be explored. As quantitative, qualitative and textual research are all framed by formal or informal processes for organizing data, the principles of knowledge organization hold great promise for enhancing the rigour of nearly any kind of research, including projects in Library and Information Studies (LIS). This paper links an established core of principles from information science to the process of organizing data in four sections: first, definition of the problem of coding with examples from qualitative research; second, explanation of a model for useful principles of knowledge organization; third, application of this model in an extended example from qualitative research; and fourth, explanation of how the model can be applied to quantitative and textual research.

THE PROBLEM OF ORGANIZING DATA: A QUALITATIVE EXAMINATION

While librarians have been examining and implementing these core principles for millennia and researchers have been testing them for decades, no one has made what we believe to be the natural link between the principles of knowledge organization and the analysis process in which all researchers engage. Yet the need to organize data effectively is a constant refrain in methodology texts. For example, Benjamin F. Crabtree and William L. Miller's, *Doing qualitative research* (1992, 17-23) outlines four qualitative analysis styles and associated research traditions and techniques from basic content analysis through heuristic approaches. In each, the development of categories is central to the data analysis process.

In organizing data for analysis, the ideal is to turn the raw data into a beautiful data rainbow, where emergent themes will be as distinct and clearly identified as the colours of the spectrum, and will fit into an overall structure that makes sense given the research questions. However, few projects fit this ideal, and categories more commonly resemble a game of pick-up sticks. Here, themes are identified like the colours of the sticks, but need to be picked carefully from the pile during the analysis process. In a worst-case scenario, data are splattered all over like the colours in a Jackson Pollock painting. Potential themes may be identifiable, but overall, the data gives little direction for rigorous analysis. How then do we manage data so that the end result looks more like a rainbow and less like data splatter?

Texts and journal articles intended to guide qualitative researchers through the coding and analysis process display two general trends. First, these texts tend to gloss over the process of data coding in favour of discussions of data collection, data analysis and the writing process. Crabtree and Miller typify this approach. They note that the goal in applying grounded theory "is to develop classifications and theory grounded in the particular social scene investigated" (1992, 26-27), but they offer little guidance for the process of developing these classifications. Second, there is a clear focus on the benefits of computer software (such as *Ethnograph* and *Nudist*) for qualitative analysis; as Stephen J. Zyzanski et al. note,

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

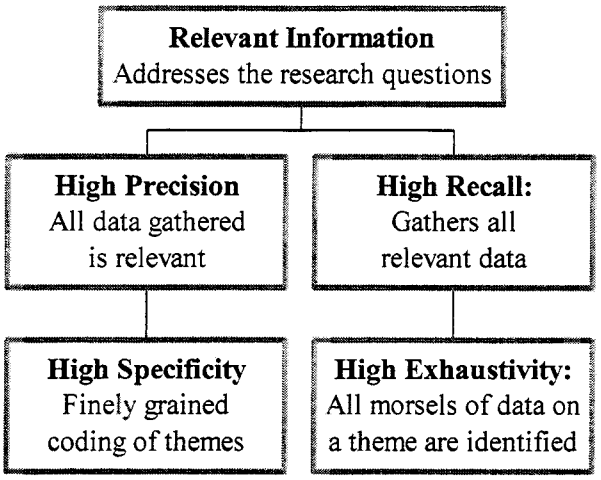
also commercially available for text retrieval, data base management, and a variety of analysis functions. These are essential programs since qualitative research often results in large volumes of verbal text that must be coded, sorted, interpreted, and summarized using text-based analysis techniques (1992, 235).

However, there is little guidance in terms of the intellectual work that is needed in order 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.

THE KNOWLEDGE ORGANIZATION MODEL

Our view is that the principles of knowledge organization offer a model for the intellectual work involved in preparing data for analysis. This section of our paper develops these principles into a simple model (see figure 1) for easy application in organizing research data. The focus of the knowledge organization model is to retrieve, from a mass of information, that which is relevant. Relevance is determined by whether or not the information contributes to answering the question at hand. In applying this model to research data analysis, relevance is determined in relation to the research questions. This criterion resembles the concept of relevance as applied to search queries or users' needs but, in this case, specifically addresses research questions.

Figure 1: Knowledge Organization Model



Precision is one standard way of measuring how effectively a system retrieves relevant information. It refers to how much of the data gathered is relevant compared to how much is irrelevant. If precision is high, then all of the information retrieved is relevant and no irrelevant information is retrieved. For research data, this would mean that all of the data 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 – the higher the specificity, the higher the level of precision.

High specificity means that data is indexed (or coded) at a very precise level – the categories are finely grained, using very detailed levels of categorization.

For example, if we are coding data about dogs we might code just that it is about 'dogs' or we might code individual breeds such as 'terriers' or 'poodles', or more specifically 'Yorkshire terriers', 'Boston terriers', 'miniature poodles' or 'toy poodles'. A problem arises because we are presuming that our categories can be mutually exclusive. If we code particular breeds each dog has to fit a breed. Dogs that are crosses between breeds tend to fall between categories and these dogs may well be the majority of dogs in any random data set. So achieving precision may not be quite as obvious a task as it at first seems.

The other common way of assessing relevance is recall, which is concerned with retrieving all relevant information or data. Maximum recall means retrieving every last instance of a theme or variable. However, in achieving high recall 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. One way to enhance recall, however, is with high exhaustivity. Exhaustivity refers to the breadth of coding – the number of factors coded for any particular unit of data. If exhaustivity is high that means that more codes are used, which will allow more data to be retrieved and analysed. Every last theme will be identified and coded. So a particular dog may be coded as a poodle who likes Beethoven, eats watermelon and often sleeps next to a cat. If all of these factors are coded it will be possible to identify poodles with a penchant for watermelon if categories are combined, but this poodle will also be retrieved in a search on cats and Beethoven. 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.

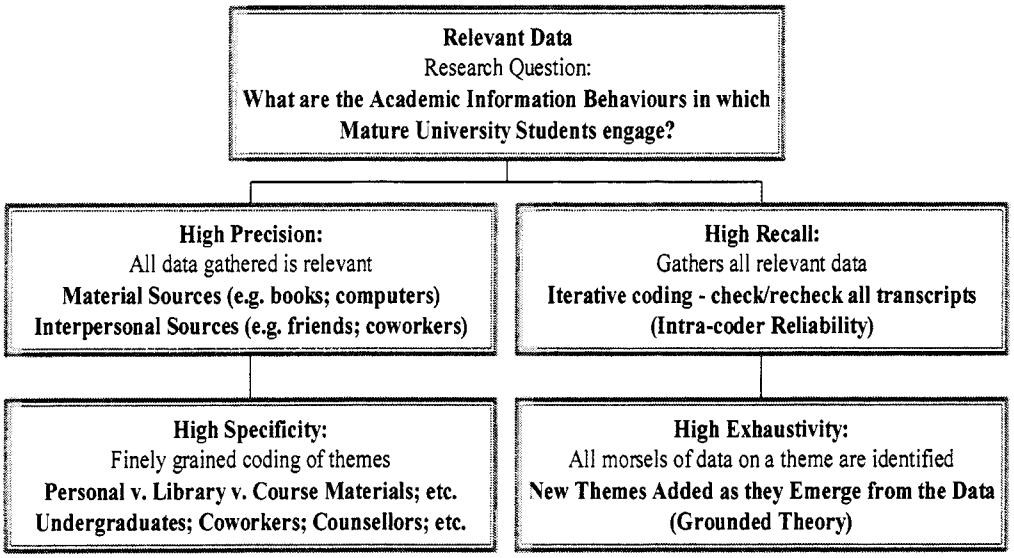
A QUALITATIVE APPLICATION

This section of the paper will use an example from Lisa Given's (2000) research that examines the ways that mature undergraduates come to terms with being socially identified as 'students', and the impact of this identification on the ways that these students locate and use information to support their studies. 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 behaviours in which mature university students engage?" Information behaviours 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. While this research question was only one of many addressed in the study, it offers an illustration of the ways that the principles of knowledge organization can apply to real qualitative data. The coding process for this question involved several considerations. First, the recognition of an information behaviour 'theme' in a transcript (*e.g.* reading a book) and coding that theme (*e.g.* 'reading' – for the act of reading; and, 'book' – for the item

itself). Then, in examining additional pages of the transcript (or reviewing the next transcript) the question arose as to whether or not 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 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. However, where the process of reading a bedtime story brought to mind a potential research topic for a paper for class, 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 just not practical as it would take a very 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) in order to code all instances. The more exhaustive the coding, the more iteration that is required.

Figure 2: Qualitative Data Analysis Model



The result is the Qualitative Data Analysis Model (figure 2) in which new themes are coded as they emerge from the data (exhaustivity), and the data is checked and re-checked in an iterative fashion in order to apply these new codes to all instances of the

relevant themes (high recall). All themes that are 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.

COMPLICATIONS IN THE MODEL

The notions of relevance, precision, recall, specificity and exhaustivity seem like they could produce perfect coding – that elusive data rainbow – but as indexers know, there are several potential problems. Two problems are particularly important in relation to organizing data. The first is the inverse relationship between precision and recall. High exhaustivity tends to lower precision, as the addition of more and more codes results in the retrieval of irrelevant data alongside the relevant. Conversely, high specificity will result in low recall. Since high specificity uses narrower categories it will produce fewer data in each category than low specificity. Theoretically it is possible to have an ideal level of both precision and recall, but in practice this does not occur. When developing themes and codes a researcher must decide which tendency is most important to the data analysis process.

Interindexer consistency is more commonly known to researchers as intercoder reliability, but both refer to solving problems of inconsistency in the application of terms and concepts. If different coders (or the same coders at different times) use different levels of specificity or exhaustivity during the coding process the analysis will produce potentially misleading results. Both recall and precision will be affected because both depend on accuracy of coding whether the emphasis is on specificity or exhaustivity. Inconsistency introduces noise into the coding process and has the potential to yield irrelevant results. Unfortunately, consistency is extremely difficult to achieve. The indexing literature is replete with studies that demonstrate considerable inconsistency even among experienced professionals using familiar, well-documented systems. The answer for qualitative coders is to approach as closely as possible to consistency and to bear in mind when drawing conclusions from research data that inconsistency is likely to be lurking there – like a flea on a dog. Many qualitative texts also refer to processes for testing both inter- and intra-coder reliability, which can enhance the level of consistency in the assigned codes.

Two ways for researchers to minimize these common problems relate to both the conceptual work involved with specificity and exhaustivity the problem of over-coding and their application the pre-iteration problem. Over-coding leads to too high levels of exhaustivity and specificity and thus to low precision and low recall. The problem of over-coding 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, as the data captured in qualitative research is 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. The pre-iteration problem occurs when new themes or more 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 it has been coded exhaustively. Similarly, if 'books' were originally coded, but the data suggests that more specific codes for 'textbooks' and 'library books' are required, recoding of earlier mentions of 'books' will be necessary. In either case, the researcher must ensure thoroughness in coding or it will be impossible to effectively retrieve relevant data for analysis.

APPLICATION OF THE MODEL TO QUANTITATIVE AND TEXTUAL RESEARCH

The Knowledge Organization Model (see figure 1) can also be applied to both quantitative research and textual interpretation. In a quantitative approach, decisions about specificity and exhaustivity come at the point at which 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 is needed to satisfy the research questions is really a matter of exhaustivity. Variables should be chosen to gather data relevant to the 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 questions and hypotheses will make collection of relevant data more likely to have an optimal level of precision, setting a framework for meaningful analysis.

The kinds of textual interpretation that LIS is increasingly drawing from the humanities are often considered to be at the other end of the research spectrum from quantitative research. However, once again the principles of knowledge organization apply. For example, a deconstruction is based on the notion of binary 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 logic/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 or related binaries examined 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 away from its focus. While deconstruction shows that our realities (in the form of binary oppositions) are constructed, discourse analysis reveals the factors that have constructed a particular reality. Again, the 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 Text Encoding Initiative 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.

Ultimately, concepts of specificity and exhaustivity may be consciously used in any type of research to expand or focus the data and its analysis. This tailoring of the data through its gathering and encoding provides relevant information for addressing research questions just as high quality indexing or classification offers a means for obtaining relevant information from a knowledge organization system.

REFERENCES:

Crabtree, Benjamin F., and William L. Miller. 1992. Primary care research: A multimethod typology and qualitative road map. In *Doing qualitative research*, edited by B.F. Crabtree and W.L. Miller,3-28. Newbury Park, CA: Sage.

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

Olson, Hope 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, no.2: 181-198.

Zyzanski, Stephen J., et al. 1992. Qualitative research: Perspectives on the future. In *Doing qualitative research*, edited by B.F. Crabtree and W.L. Miller,231-248. Newbury Park, CA: Sage.