THE INDEXER

Volume 23 Number 3, April 2003 **ISSN 0019-4131**

The International Journal of Indexing

OFFPRINT

129–133 Indexing and the 'organized' researcher

Hope A. Olson and Lisa M. Given

Indexing and the 'organized' researcher

Hope A. Olson and Lisa M. Given

This article proposes that indexing concepts relating to relevance, precision, recall, coextensiveness, exhaustivity, specificity and consistency offer a ready-made model that can be applied to the organization of research data. This knowledge organization model contributes significantly to the ability of researchers to collect and organize data in a manner most likely to shed light on the research problems they address.

The power of indexing for organizing and retrieving information is well established. Referring to indexers' autonomy, David Lee (2001: 191) wrote: 'What a powerful person the indexer is...', while making clear (describing how indexes are judged for the UK's prestigious Wheatley Medal) that indexers earn that autonomy through quality based in certain principles. These principles are sometimes enunciated and sometimes pragmatically intuitive. When enunciated they become transferable to other endeavours. Our purpose in this article is to illustrate how the power of indexing can be extended to the realm of original research. To do so we begin with an explanation of why research methodology needs the help of indexers. We then offer a brief overview of the formal concepts involved in indexing to create a model for general application. Finally, we describe such an application in an actual research study.

Why researchers need indexing

Across disciplines and research interests, researchers have one very important consideration in common: what to do with the reams of data collected during a research project? Regardless of the research method chosen (interviews, questionnaires, focus groups, etc.) or the orientation of the study (quantitative, qualitative, or textual), the primary goal of research is to make sense of the data collected. This process necessitates order; the researcher must be able to gather pieces of data together to answer the research questions at hand.

In quantitative research, concepts, variables and hypotheses are selected before the study begins 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 understand certain phenomena. In studies that use inductive, qualitative methodological approaches, thematic codes emerge from the information gathered during and following data collection. These codes, analysed in the context of the data gathered, aid in the development of patterns and theories to explain various phenomena. Textual research has elements of both quantitative and qualitative methods; texts are selected at the outset of the study based on the themes or discourses being considered, and manifestations of those discourses emerge at the point of data gathering through close reading of those texts.

Organizing data is particularly troublesome for researchers using qualitative methods, which generally result in rich (and lengthy) in-depth interview transcripts, personal diaries, video transcripts with multiple voices, and so on. Unfortunately, methods textbooks, previously published studies, and even university-level research methods courses rarely discuss (or tend to gloss over) issues related to the organization and management of data for analysis of the results. Researchers are often left to forge their own paths in order to organize the data they have collected in a way that facilitates analysis – to clearly identify and report on the distinct themes emerging from the dataset.

While some pieces of data (e.g. a participant's age or gender) are easily marked and labeled for analysis, others (e.g. participants' feelings of social isolation) are more complex and difficult to manage. In indexing, strategies for effective organization and retrieval of such different types of information have been honed by examining specificity, exhaustivity, relevance, and other elements central to the process of subject analysis. Linking the research process to indexing concepts and techniques may help researchers, educators, students and reviewers of research across academic disciplines to benefit from a new and valuable approach to data preparation – and professional indexers may find yet another avenue for marketing their skills. The model discussed in this article, therefore, offers one potential solution to the problem of organizing research data: the application of core concepts of knowledge organization to the data management process in which virtually all researchers engage.

The Knowledge Organization Model

The following sections develop indexing concepts into a simple model (Fig. 1) for applying to the organization of research data (for a more complete description of these concepts see Olson and Boll, 2001: ch. 5). The focus of the Knowledge Organization Model is to retrieve what is relevant from a mass of information. It is the same goal that Richard Raper is seeking when he asks 'Would I use this entry if I were searching for information in the book?' (in Booth, 1996: 91). Applying this model to research data, relevance is determined by whether or not the information contributes to answering the research questions or hypoth-

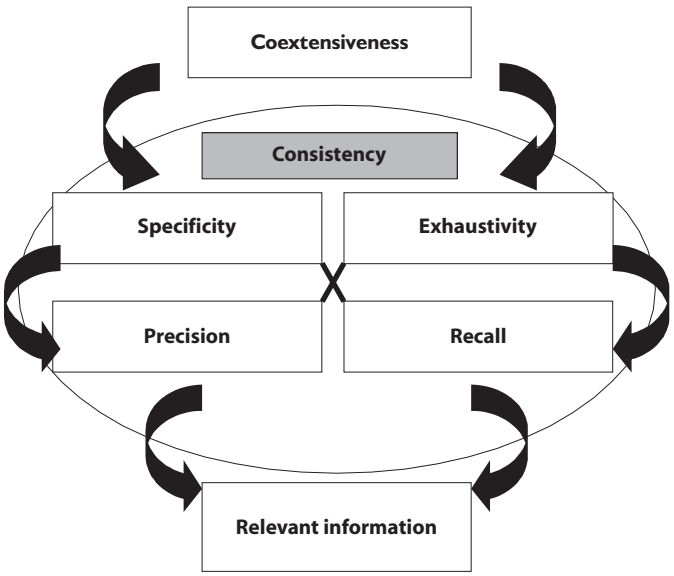

Figure 1. Knowledge organization model

eses under study. The Knowledge Organization Model, when used for indexing, governs the development of vocabulary terms, their application to texts or entries, and the creation of a logically organized index ready for searching. When used for organizing research data, the model governs the development of variables or themes to be coded, their application to data, and the creation of an organized body of data suitable for analysis.

Coextensiveness

The first element is choosing what will be covered and how it will be divided into categories. *Coextensiveness* is what Pat Booth is seeking when she scans a book 'to get an idea of the overall content, the structure of the text, and the principal themes and subthemes, important names, etc.' (Booth, 1996: 92). For example, categories of animals might follow the traditional zoological taxonomy based on absence or presence of a backbone, diet, and means of reproduction. In this taxonomy cats are categorized as felines, felines as carnivores, carnivores as mammals, and mammals as vertebrates. These categories are useful if searchers have questions about carnivores or felines. However, if searchers want to know about pets, these categories are useless. Here, categories based on animals' relations to humans are more appropriate, such as 'domesticated' and 'wild'. Each of these can then be further divided, so that the category 'domesticated' would include pets, livestock and working animals. 'Pets' might then be subdivided to include cats, dogs, parrots, goldfish, and so on – with 'cats' further developed into Siamese, Burmese, Manx, Himalayan, and so forth.

Similarly, the definition of variables and thematic codes used in research data 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 but a social agency in another. Depending on the goals of the study, the categories selected may differ substantially from one study to another, to be coextensive with the research problem at hand.

Specificity and precision

Specificity is the relative detail within the vocabulary – the number of hierarchical levels defined. In organizing animals, a scheme that stops at differentiating only domesticated from wild animals would have very low specificity. Specificity increases with each level as the hierarchy becomes deeper, extending to a level that differentiates particular breeds of particular animals (such as 'sealpoint Siamese cats'). As with coextensiveness, the level of specificity should serve the needs of users of that organizational scheme; an index in veterinary science will probably need much higher specificity in regard to animals than an index used by elementary schoolteachers. It will also be driven by the material itself, as long strings of page numbers call out to an indexer to devise subheadings. 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, which is, of course, standard indexing practice. Each item must be coded at the most specific level available. This application of specificity would dictate that Siamese cats be indexed as 'Siamese' and not as 'cats' or 'pets' or 'domesticated animals'.

Precision is one standard way of measuring how effectively a system retrieves relevant information. It refers to how much of the body of data gathered is relevant compared to how much is irrelevant. If precision is high, then all the information retrieved is relevant and little or no irrelevant information is retrieved. For research data this would mean that all the data gathered for a particular variable or coded as having a particular attribute do actually have that attribute. To approach this ideal it is important to remember that precision is enhanced by high specificity, meaning that data are indexed at a very precise level – the selected terms are finely grained, using very detailed levels of categorization.

A problem arises, however, because this practice presumes that the chosen categories can be mutually exclusive. If particular breeds of cat 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 they were the majority in the dataset. Achieving precision may not be quite as obvious a task as it seems at first glance.

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. At the level of the vocabulary, exhaustivity is concerned with the different aspects or facets included. For example, the taxonomies of animals described above might be combined with facets to represent functions (digestion, reproduction, etc.), environment (aquarium, barn, house, etc.) and type of contact with humans (food, companionship, etc.). Each facet added to the scheme raises the exhaustivity of the vocabulary. Exhaustivity leads to the important criterion of book indexing: 'comprehensiveness; that is, that everything

in the book is covered' (Lee, 2001: 192). In a given research project, the nature of the research questions and/or hypotheses must govern the choice of facets or topics included.

The application of the concept of exhaustivity is related to the level of indexable matter: how much of a particular topic must be covered by a book before it is represented in the index? Does it need to be particularly insightful or will a passing mention suffice? How is it related to the core themes of the book? Choices need to be made regarding the level of representation appropriate to the research problem under study. A related question is how many concepts will be represented? Here, exhaustivity meets specificity. If a discussion on pets includes cats and dogs, it seems reasonable to be both specific and exhaustive in categorizing those concepts. However, if the discussion includes cats, dogs, fish, birds, rodents, rabbits and llamas 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.

Exhaustivity is closely related to *recall* – how much of the available relevant information or data is actually retrieved. 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 but no irrelevant information. That is why precision and recall tend to be inversely correlated and this will have an impact on the construction of data categories and codes for analysis. One way to enhance recall is to seek high exhaustivity. If exhaustivity is high more codes are used, which will allow more data to be retrieved and analysed. 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.

Complications in the Knowledge Organization Model

It would seem that the notions of relevance, precision, recall, specificity and exhaustivity could produce perfect categories and codes, but as indexers know, there are several potential problems in applying these principles. Two problems particularly complicate the pursuit of ideal data organization.

Precision versus recall

The first complication is the inverse relationship between precision and recall (as denoted by the **X** in the center of Fig. 1). High exhaustivity tends to lower precision: adding more and more codes results in the retrieval of irrelevant data alongside the relevant. Conversely, high specificity tends to lower recall. Since high specificity uses narrower categories it will produce fewer data in each category than will 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.

Consistency

Interindexer consistency, the second complication in the Knowledge Organization Model, refers to solving problems of inconsistency in the application of terms and concepts. If an indexer uses headings differently while indexing a particular book, the end result will be a frustrating experience for readers. For example, entries for 'cats, development,' 'cats, reproduction,' and 'kittens' will overlap, even though they are obviously distinguishable conceptual terms. Effective indexing will require their consistent usage (in addition to cross-references). Consistency is a basic attribute of a quality index (Lee, 2001: 192). The same applies to categorizers or coders of research data; inconsistencies in term usage will produce potentially misleading results at the point of data analysis. Unfortunately, consistency is extremely difficult to achieve. The three factors that most often increase consistency are:

1. documentation to aid application of vocabularies or encoding schemes as stressed in indexing standards (Calvert, 1996: 74);
2. low specificity;
3. low exhaustivity.

Obviously, the solution of consistency problems may create other inadequacies; using low specificity and low exhaustivity to achieve consistency will lead to low precision and low recall.

An application of the Knowledge Organization Model: qualitative data analysis

This section presents one example of how the Knowledge Organization Model may be applied to qualitative data, using an example drawn from the interview phase of Lisa M. Given's (2000) dissertation research. While the examples provided here are drawn from the field of library and information studies and use a qualitative approach, the Knowledge Organization Model may be generalized to apply to data collected in other areas of social sciences and humanities research.

The interviews in this study examined the personal contexts, perceptions and information-seeking behaviours of 25 mature undergraduates at one Canadian university. The data analysis followed a grounded theory approach, where themes (e.g. the role of the library in facilitating success, mature students' social isolation from younger peers) were coded as they emerged from the data in an ongoing and iterative process. The study included the research question: 'What are the academic information behaviours in which mature university students engage?' Information behaviour refers 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 illustrates how indexing principles can be applied to real qualitative data.

The coding process for this question involved several considerations. First, the recognition in a transcript of an

information behaviour theme (e.g. reading a book) and the selection of terms to represent that theme (e.g. 'reading' for the act of reading and 'book' for the item itself). Then, in examining additional pages of one student's interview transcript (or reviewing the next student's 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 in code assignment. For example, instances of reading a book that were not related to the student's academic life (e.g. 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 – it reflected an information behaviour related to the student's academic work. Whether or not this single instance was sufficiently relevant to be assigned a separate code is a question of both specificity and exhaustivity. Is this instance indexable matter?

The other half of the model relates to exhaustivity: how many themes are needed to address the research questions? A completely exhaustive codebook (i.e. a list of all relevant thematic terms) is just not practical as it would take too long to develop all the themes and search for them 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)? Secondly, to achieve optimal recall for each new theme or code that is assigned, the researcher must go through each transcript (often, many times – in an iterative process – much like the way an indexer reviews a text before completing an index) in order to code all instances. The more exhaustive the coding, the more iteration is required.

It is important to remember, however, that over-coding leads to extreme 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, including interesting themes that are simply not relevant – a challenge that all researchers must address in managing data. This problem can be difficult to avoid in a grounded theory approach, as the data captured in qualitative research are extremely rich and filled with engaging details; elements of individuals' life stories often arise in these in-depth interviews and researchers must refrain from labeling the entire transcript as relevant to one degree or another. On the other hand, they must not restrict their coding to the point that they miss 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 – and balance this with appropriate levels of exhaustivity and specificity, in order to facilitate the analysis of themes that simply cannot be missed.

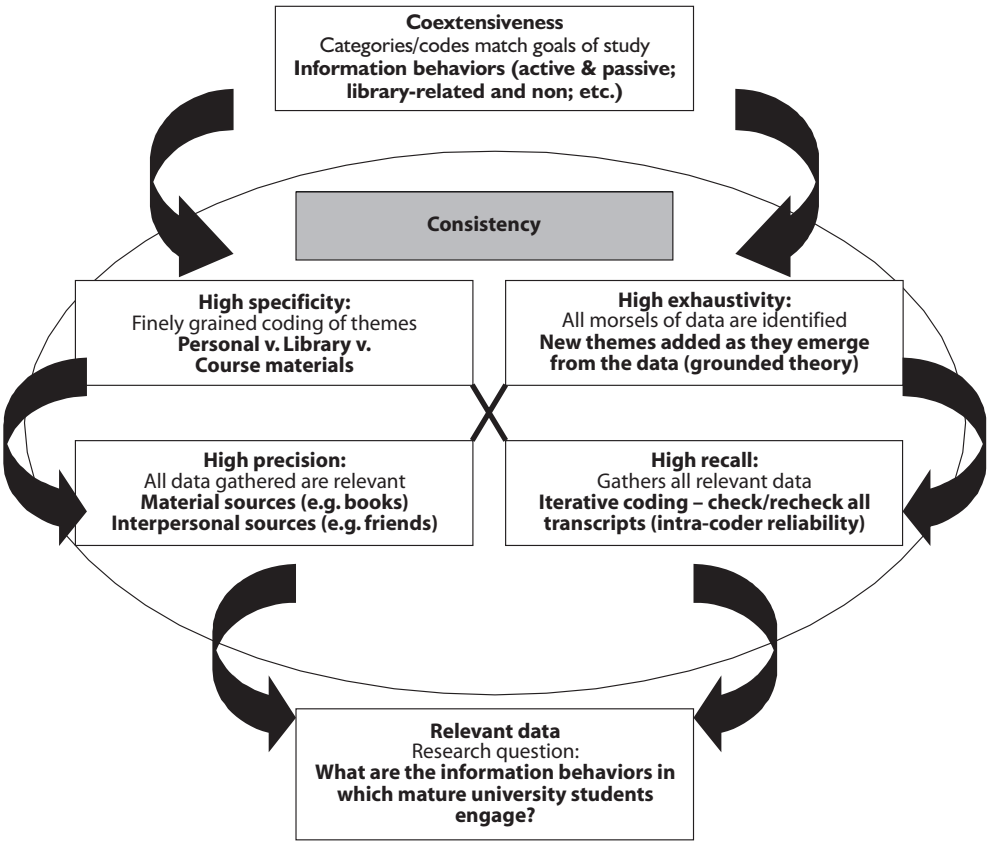

Figure 2. Qualitative data analysis model

There is yet another wrinkle in the process: it is vital that when researchers add new (or more specific) codes to the manual, they realize that they must thoroughly review all previously coded transcripts to reveal 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 suggest 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. 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 – much as it does in documents being indexed. Researchers must therefore approach consistency as closely as possible and bear inconsistencies in mind when drawing conclusions from the data. Many qualitative research methods texts refer to processes for testing both inter- and intra-coder reliability, which can enhance the level of consistency in the assigned codes.

The result is the Qualitative Data Analysis Model (Fig. 2), in which new themes are coded as they emerge from the data (exhaustivity), and the data are 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.

Conclusion

The Knowledge Organization Model is a robust one, growing from concepts rooted in centuries of indexing practice and research. This article has demonstrated its applicability to original research. Ultimately, concepts of coextensiveness, specificity and exhaustivity may be used consciously 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 provides access to information in a document. In the end, researchers can use the Knowledge Organization Model to add a new level of scholarly rigour to their work – to apply the power of indexing in order to organize knowledge in a new and creative way.

References

Booth, Pat F. (1996) How we index: six ways to work. *The Indexer* **20**(2), 90–3.

Calvert, Drusilla (1996) Deconstructing indexing standards. *The Indexer* **20**(2), 74–7.

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

Lee, David (2001) Judging indexes: the criteria for a good index. *The Indexer* **22**(4), 191–4.

Olson, Hope A. and Boll, John J. (2001) *Subject analysis in online catalogs,* 2nd edn. Englewood, CO: Libraries Unlimited.

Hope A. Olson is a professor in the School of Library and Information Studies at the University of Alberta where she teaches and researches in the area of organization of information. Her book, The power to name: locating the limits of subject representation in libraries*, was recently published by Kluwer Academic. She is also editor-in-chief of* Knowledge Organization*, the international journal on classification, thesauri, concept theory and other related topics. Email: hope.olson@ualberta.ca*

Lisa M. Given is an assistant professor in the School of Library and Information Studies at the University of Alberta. She teaches advanced topics in the organization of knowledge (e.g. indexing, abstracting, thesaurus construction), as well as courses in information architecture, instructional strategies for information professionals, and research methods. She conducts research into individuals' information behaviours, with a focus on higher education. Email: lisa.given@ualberta.ca

---

Poison pellets

In his diary column [*The Guardian*, 29 June 2002 – see Indexes Reviewed, Obiter Dicta, *The Indexer* **23**(2), p. 104] Simon Hoggart reported that he had inserted 'a little insult, a little poison pellet' into several entries in the index to his book of sketches, *Playing to the Gallery* (Atlantic Books, 2002). Here are a few of these entries (quoted in *The Bookseller*, 5 July 2002):

Blair, Cherie: smiles with her husband as if attending Moonie wedding, 29
Kaufman, Gerald: wears unfeasible suit, 227–9
Lilley, Peter: notable resemblance to Niles Crane in hit TV sitcom 'Frasier', 74–5
Longford, Lord: insists that book on humility should be placed in Hatchards window, 246
Prescott, John: takes credit for rain, 179–80; blames Tories for rain, 188–90; defeats hapless stenographer, 259–60
Thatcher, (Baroness) Margaret: unveils statue of self while making mad remarks, 269–71

The Bookseller adds: 'Recently, Julian Barnes took a similar approach to the indexing of his collection of *New Yorker* columns, *Letters from London*. But Mr Hoggart may have produced the first index with a running joke: it concerns the number of people who have been the subject of unfunny jokes by John Redwood.'