Anderson, J., & Rowley, F. (1991). Building End-User Thesauri from Full-Text. 2nd ASIS SIG/CR Classification Research Workshop, 1-10. doi:10.7152/acro.v2i1.12543

Building End-User Thesauri from Full-Text

James D. Anderson
School of Communication, Information, and Library Studies
Rutgers, The State University of New Jersey

Frederick A. Rowley
ARIS: Anderson Rowley Information Systems
Riverdale New York

0. OVERVIEW

We are interested in the possible contribution of end-user thesauri to the improvement of information retrieval by end- users. Thesauri (from the Greek for treasure or treasury) in information retrieval attempt to record and display relations among concepts and terms -- to be treasuries of concepts and the terms that represent them. End-user thesauri are designed to guide and facilitate end-user searching of textual databases (both full-text databases and reference databases that contain only surrogates of full-texts, such as abstracts). End-user thesauri link the vocabulary of the searcher and the vocabulary of the database, functioning as part of the user-database interface. End-user thesauri are not designed to guide indexing, although they can be used to suggest terms, much like writers have used Roget's thesaurus for centuries.

We have been working on software to facilitate the creation and use of end-user thesauri. We believe the best sources of terms are user queries; the second best sources are texts containing potential answers to queries; the third best source is expert human indexing; and the fourth best source are existing linguistic reference tools, such as other thesauri, dictionaries, glossaries, handbooks, etc. Our software is designed to facilitate the large-scale gathering, structuring, and displaying of vocabularies gathered from these sources. Its three main components are a textual database management program; a thesaurus construction program; and an end-user search interface program, all running on MS-DOS microcomputers.

So far, we have created two small prototype thesauri, one based on minutes of meetings of the Board of Governors at Rutgers University and one based on abstracts of articles on various aspects of knowledge representation for information retrieval.

Once terms are gathered from searchers, texts, indexers, and reference sources, thesaurus editors face the same kinds of challenges as those creating more traditional indexing-oriented thesauri; however, the optimum solutions may not be the same. These challenges include: sorting and categorizing terms to create useful hierarchies and displays for end-user searching; use of bound terms vs. elemental descriptors; optimum term relationships; inclusion of variant forms and equivalent terms; homographs; designing helpful displays; and testing and validation.

1. TRACKING AND MANAGING vs. CONTROLLING VOCABULARY

Traditionally, thesauri have been used in information retrieval to guide the indexer rather than the searcher. Indeed, many database producers with indexer thesauri do not even publish or make them available for use in searching. Indexer thesauri tend to be smaller than end-user thesauri, since they

ISSN: 2324-9773

Anderson, J., & Rowley, F. (1991). Building End-User Thesauri from Full-Text. 2nd ASIS SIG/CR Classification Research Workshop, 1-10. doi:10.7152/acro.v2i1.12543

record only authorized terms and a relatively small number of synonyms or equivalent terms. End-user thesauri, in contrast, attempt to record the entire vocabulary of a discipline or client population.

End-user thesauri are descriptive rather than prescriptive. They strive to record the actual language used, rather than language that should be used. Relations among terms are, likewise, more suggestive than prescriptive, reflecting the dynamics of natural language usage more than the controlled structure of a planned, artificial indexing language.

Marcia Bates is one of the principal proponents of end-user thesauri as search tools [Bates 1986].

2. PULLING IN TEXT

Our software is designed to pull in free-text terms from unstructured ASCII files or from free-text database record fields, as well as delimited database descriptors. For best results, multi-word free-text terms (e.g., information science, birth control, etc.) should be linked before term acquisition. We are doing this "manually" rather than algorithmically; it's an easy and fast process and well worth the time in terms of results. Work by others to identify multi-word phrases through natural language processing may be able to take over this job in the future.

3. CATEGORIZING

Hundreds of words and phrases can be pulled in from even small texts or a small number of database records. On the first round of acquisition, these include all of the most commonly used non- substantive words that are usually found in stop-lists. The first job is to sort these terms into three groups: stop list terms; terms to discard; and substantive terms to be added to the thesaurus. The stop list is treated as a thesaurus hierarchy. Once a term is added to it, all future occurrences will match the original term and it will not again appear on a list of new terms to be considered for integration into the thesaurus. In contrast, if a term is discarded, it will reappear the next time it is encountered in a text. Generally speaking, if in doubt, a term should be added to the stop list. It can always be removed from the stop list for integration into substantive hierarchies later.

If an acceptable existing stop list is available, it can be loaded into the thesaurus in advance, thereby blocking the appearance of all of its stop-list terms.

After stop-list and discard terms are marked, sorting and categorization of the remaining terms must begin.

If this is the beginning of a new thesaurus, it is generally best to sort all terms into broad categories or facets which reflect the principal interests of the potential users of the database. In the words of Dagobert Soergel, "A properly designed facet frame captures the essential conceptual structure of a field and is instrumental in eliciting the concepts to be included in the index language, in assisting in the analysis of a search topic, and in the analysis of an entity in indexing" (p. 397).

The Indian librarian and philosopher Ranganathan has identified the most basic or fundamental categories or facets that generally apply to all fields: personality (usually called "entities" in The

ISSN: 2324-9773

Anderson, J., & Rowley, F. (1991). Building End-User Thesauri from Full-Text. 2nd ASIS SIG/CR Classification Research Workshop, 1-10. doi:10.7152/acro.v2i1.12543

West); material (meaning raw or constituent material); energy (referring to operations, processes and events); space; and time. Fans of Ranganathan often abbreviate these five fundamental categories as "PMEST". Western theorists in the Classification Research Group (Great Britain) expanded and revised these basic categories to include the following (Vickery, p.46-47):

- Things, entities, including Naturally occurring (e.g., Minerals, Animals, Plants, Soils), Products (e.g., Bridges, Engines, Fibre), Mental constructs (e.g., Equations, Rectangles, Formulae).
- Parts, components, structure (e.g., Beam, Wheel, Wing), including Organs (e.g., Heart, Seed).
- Materials, constituents (e.g., Metal, Glass, Nitrogen).
- Attributes, including Qualities, properties (e.g., Cohesion, Color, Solubility); Processes, behavior (e.g., Vibration, Inflammation).
- Operations, including Experimental (Cutting, Breeding); Mental (Calculation, Reasoning).
- Operating agents (any thing or entity can act as an agent)
- Place, condition

Every field, discipline or subject area has its own particular set of facets, or important categories. Vickery cites an example from Soil Science (p. 47):

- Soil, the entity of central interest.
- Parts of soil, e.g. Gravel.
- Constituents of soil, e.g., Nitrogen.
- Structure of soil, e.g., Profile.
- Layers of soil, e.g. Horizon.
- Organisms in soil, e.g., Bacteria.
- Parent materials of soil, e.g., Muck.
- Processes in soil, e.g., Mineralization.
- Properties of soil, e.g., Cohesion.
- Measures of property, e.g., Sticky point.
- Operations on soil, e.g., Amendment.
- Equipment for operations, e.g., Plough.

For literature, the Modern Language Association has determined the principal aspects or facets to be the following:

- specific literatures (by nationality or culture)
- performance media
- languages

ISSN: 2324-9773

Anderson, J., & Rowley, F. (1991). Building End-User Thesauri from Full-Text. 2nd ASIS SIG/CR Classification Research Workshop, 1-10. doi:10.7152/acro.v2i1.12543

- periods
- individuals (e.g., authors)
- groups/movements
- genres
- works
- features
- literary techniques
- themes/motifs/figures/characters
- influences (recipients)
- sources
- processes
- types of scholarship
- methodological approaches
- theories
- devices/tools
- disciplines
- scholars
- document types

Our thesaurus on knowledge representation for information retrieval includes the following primary facets:

- abstract entities, e.g., ideas, rules, theory.
- attributes, e.g., degree of order, specificity, effectiveness.
- concrete entitles, e.g., documents, buildings, symbols.
- materials, e.g., ceramics.
- operations, e.g., filing, arrangement, indexing, cognition.
- persons, e.g., authors, experts, indexers, librarians.
- places, e.g., environment, Asia, addresses (locations).
- tools, e.g., computers, thesauri, technology.

As soon as a term is added to the thesaurus as a candidate term, our software creates a thesaurus record for it, with the following standard fields (additional special fields can be added as needed):

rec\ Record number
des\ Descriptor or term
dbf\ Database field or source of term
scn\ Scope note
eqv\ Equivalent terms
err\ Errors or variant forms
nrt\ Narrower terms
brt\ Broader terms
rlt\ Related terms

The process of initial concept sorting is performed by simply assigning one (or more) of the primary facet terms as a "broader term" in each a term's thesaurus record. This process is analogous

ISSN: 2324-9773

Anderson, J., & Rowley, F. (1991). Building End-User Thesauri from Full-Text. 2nd ASIS SIG/CR Classification Research Workshop, 1-10. doi:10.7152/acro.v2i1.12543

to the physical sorting of concept cards into conceptual piles, as was done formerly in the "manual" creation of thesauri.

After initial categorization of terms into broad primary facets, these large categories must be further broken down into sub- categories for the purpose of creating useful displays of terms for visual scanning by end-users. Criteria for what constitutes useful displays (and therefore, useful categories) must be based on assessment of utility for users. These are qualitative, subjective criteria, not open to rigorous judgments of right, wrong, or correct. For any particular vocabulary, there are an infinite number of possible displays. The objective is to create displays that help the user get an overview of the field and to navigate effectively from one conceptual area to another.

In our knowledge representation thesaurus, for example, we have broken up the large "attribute" category in terms of the usual focus of an attribute, e.g.,

- classificatory attributes, e.g., degree of order; fuzziness; inclusion.
- database attributes, e.g., data structures; domain; exhaustivity; representational predictability.
- reliability attributes, e.g., accuracy, authority, fidelity.

Similarly, operations are categorized into sub-facets such as:

- bibliographic operations, e.g., bibliographic control, bibliography (operation), citation (operation). The qualifier "(operation)" distinguishes these operations from "bibliography" as a list of citations and "citation" as a bibliographic description, respectively.
- computer operations, e.g., computer programming, text processing, scanning.
- intellectual operations, e.g., cognition, communication, reading.

The process of subdivision should continue so that a set of closely related terms, numbering no more than 10 or 20 at the most, remain in a single category.

The actual process of subdivision, or sorting into sub- categories, is carried out by replacing a broader "broader term" by a narrower "broader term" in the thesaurus records. An initial hierarchical display will list, for example, all the terms initially classed as "operations" as a single undifferentiated array in this hierarchy. Using this list, the thesaurus editor can open the record for each term and replace the broader term "operations" with narrower terms, such as "bibliographic operations," "computer operations," "intellectual operations," etc. These new gathering terms may be integrated into the overall thesaurus structure by listing them as narrower terms in the record for the former broader term, in this case "operations".

ISSN: 2324-9773

Anderson, J., & Rowley, F. (1991). Building End-User Thesauri from Full-Text. 2nd ASIS SIG/CR Classification Research Workshop, 1-10. doi:10.7152/acro.v2i1.12543

4. BOUND TERMS vs. ELEMENTAL DESCRIPTORS

What constitutes a "term" in an end-user thesaurus? Thesaurus experts have long debated the appropriate nature of terms in a thesaurus, some holding up a goal of listing only "single concept, elemental terms." The international standard for monolingual thesauri says, "it is a general rule that terms in a thesaurus should represent simple or unitary concepts as far as possible, and compound terms should be factored (i.e. split) into simple elements, except when this is likely to affect the users' understanding" (quoted in Aitchison and Gilchrist, p.24). The 1990 draft for the "Proposed American National Standard Guidelines for Thesaurus Construction, Structure and Use" acknowledges that "the establishment of procedures for dealing consistently with compound terms is one of the most difficult areas in the fields of thesaurus construction and indexing" (ANSI/NISO Z39.19, p. 29).

Few would object to the division of "workload of dentists in Scotland" (example in Aitchison and Gilchrist, p. 24) into "workload," "dentists," and "Scotland." but what about "philosophy of education," "history of science," "library science," "information science," "library school," "school library," or "birth control"? Thesaurus standards and textbooks provide guidelines for making these decisions. In an end-user thesaurus, the emphasis should be on user warrant -- actual linguistic behavior. If multi-word phrases are treated as terms by users -- that is, if they consistently or very frequently appear in the same form, they should be listed as terms in the thesaurus. If "information science" is always (or almost always) "information science," then it should be considered a single term. However, if it is just as often expressed as the "science of information," then it may not be an established term warranted by usage and can be split into two terms "information" and "science."

Our preference is to err on the side of preference for bound terms. Our knowledge representation thesaurus has a term "prototype fuzzy query processors." This of course could have been broken down into "prototypes," "fuzzy queries," and "processors," (and "fuzzy queries" could be factored into "fuzziness" and "queries") but by keeping it together, as it appeared in the text from which it was drawn, it can be listed in displays of narrower terms under "prototypes" (giving users an array of specific prototypes to choose from); under "fuzziness," illustrating the application of this attribute; under "queries" (leading users to procedures applied to queries); and under "processors" (illustrating a wide range of processing applications).

Actually, the question is not whether to factor a term that represents a complex or compound concept, but whether to list compound/complex terms as well as their more elemental parts. In the "prototype fuzzy query processors" example, the complex term is listed under each of its more elemental factors.

5. TERM RELATIONSHIPS

Display of relations among concepts and the terms that represent them is a hallmark of the modern information retrieval thesaurus. The role of such relations, and which ones are useful, has long been a topic of debate. Several decades ago, Jason Faradane developed his "relational indexing" (summarized in Faradane, 1980), which incorporated what he claimed to be the fundamental

ISSN: 2324-9773

Anderson, J., & Rowley, F. (1991). Building End-User Thesauri from Full-Text. 2nd ASIS SIG/CR Classification Research Workshop, 1-10. doi:10.7152/acro.v2i1.12543

relationships used in the human mind to organize concepts: concurrence, equivalence, distinctness, self-activity, dimensional, action, association, appurtenance, and functional dependence. Based on Faradane's work, as well as earlier theories of Hume and the Classification Research Group (U.K.), Richard Diener, while a doctoral student at Rutgers, suggested the following relations as key: concurrence, hierarchy (broader), hierarchy (narrower), equivalence, distinctness, action (transitive), action (intransitive), causation, time-space, property (constant), property (variable), material, and quality/quantity. Fans of Derek Austin's PRECIS (Preserved Context Indexing System) will recognize the accommodation of many of these relations in that system. More recently, Wang, Vandendorpe, and Evens have tested the efficacy of an even larger list of 43 "lexical-semantic relations," including part-whole, head-organization, personnel- object, count-mass, set-element, substance, provenance, typical agent, typical object, typical result, typical instrument, habitat-object, offspring-parent, cause-action effected, taxonomy, synonymy, complementarity, antonymy, converseness, and reciprocal kinship.

These relatively large numbers of relations have not been widely used, because, I suspect, most of them are not clearly apparent and are therefore difficult to identify and use. The consensus among thesaurus makers and users, as reflected in national and international standards, is that all relations should be subsumed into just three broad types: equivalence, hierarchical ("broader" and "narrower"), and other (for "related" or "associated" terms).

But even the distinction between hierarchical and non- hierarchical (i.e., merely related) concepts, although perhaps clear in theory, is not always easy (or useful) in practice. Our "Prototype fuzzy query processors" can be used to illustrate this issue.

While "prototype fuzzy query processors" are clearly "prototypes" as well as "processors," they are not "fuzziness," nor are they "queries," and purists would properly insist that the relationship between "prototype fuzzy query processors" and "fuzziness" or "queries" be listed as "related" rather than hierarchical or "narrower/broader." The problem with this approach is that related terms are generally absent from hierarchical displays, yet hierarchical displays are one of the most useful working displays for thesaurus editors as they structure vocabularies and for end users as they explore terms and concepts.

At least in the initial compilation of a thesaurus, it is useful to treat all term relationships as hierarchical (broader/narrower). After each pass at sorting or categorizing terms, a hierarchical display can be generated which will include all terms sorted into a particular category. The terms in each category can then be sorted, or categorized into yet another level of subordinate categories. If terms were tagged as related, rather than broader or narrower terms, at this early stage of classification, they would be lost from the hierarchical display.

Once sorting and categorization has taken place, term relations which are not strictly hierarchical may be converted to associative or related.

6. VARIANT FORMS AND EQUIVALENT TERMS

In the process of sorting or categorizing terms, the thesaurus editor will come across different terms that have essentially the same meaning (e.g., lawyer and attorney) and variant forms of the same

ISSN: 2324-9773

Anderson, J., & Rowley, F. (1991). Building End-User Thesauri from Full-Text. 2nd ASIS SIG/CR Classification Research Workshop, 1-10. doi:10.7152/acro.v2i1.12543

term (e.g., index, indexes, indices). At some point, such equivalent and variant forms should be combined into a single record. For this purpose, one term is chosen to be the "preferred" term to represent the concept; equivalent terms are placed in the "equivalent terms" field; and variant forms, both for the preferred term and for equivalent terms, are placed in the variant terms field. The operational difference between the "equivalent" and "variant" fields is that terms in the equivalent field get lead term entries in alphabetical displays, while variant terms do not.

In some thesauri, the equivalent term field is called the "used for" field, but this is inappropriate for an end-user thesauri, since all terms are usable and, in fact, used. The terminology "used for" and "use" should be limited to indexer thesauri, rather than end-user thesauri. (Actually, I'd like to see it disappear altogether. My students get thoroughly confused about which term is used for which.)

The choice of a preferred term, e.g., "attorney" versus "lawyer" is, in the final analysis, somewhat arbitrary. The choice should go to the most-used term, if that is possible to determine. But the only function of a "preferred" term in an end-user thesaurus is to serve as a "hitching post" around which to gather equivalent, variant, broader, narrower, and other related terms. Access to the thesaurus and through it to databases will be available through all terms, regardless of which term is selected to serve as a preferred, or posting, term for a particular concept.

7. HOMOGRAPHS

As terms are sorted and categorized, it will also become apparent that some terms stand for two or more very different concepts, e.g., "mercury" can refer to a Roman god, a metallic element, a planet, or an automobile; "term" can refer to a word or phrase used in indexing and also to a period of time, as in a term of office or prison sentence; "sentence" can refer to a verbal phrase or to a criminal penalty, "bibliography" is both a process and a result of that process (as are many "-tion" words - - e.g., translation, citation), and so on. When such instances occur, separate records should be established for each concept, and the preferred term should be qualified, e.g., mercury (god); mercury (element); mercury (planet). In each such record, the unqualified term should be listed as an equivalent term, so that any reference to "mercury" will lead to the equivalent preferred terms "mercury (god)," "mercury (element)," "Mercury (automobile)," and/or "mercury (planet)," providing an opportunity to choose among these alternatives.

8. END-USER DISPLAYS

The chief end of end-user thesauri is to present to end-users useful displays of terms from which they can select terms for searches. Thesaurus developers have experimented with a wide range of textual and graphic displays, but two basic approaches predominate: alphabetical and relational. Relational displays are usually limited to hierarchical relations. The move of thesauri from print to computer-mediated displays has created a whole new set of constraints and opportunities. We have been exploring possible computer mediated displays for both hierarchical and alphabetical access to thesauri. Our current hierarchical display begins with tops terms (see figures 1-3). By clicking on a top term, the user can open a window with the next level of terms in that hierarchy, and continue down the hierarchy by picking a particular term at each step. We find this "single strand"

ISSN: 2324-9773

Anderson, J., & Rowley, F. (1991). Building End-User Thesauri from Full-Text. 2nd ASIS SIG/CR Classification Research Workshop, 1-10. doi:10.7152/acro.v2i1.12543

approach easier to follow than the display of complete hierarchies, as is often done in print displays. On a small video screen, one simply gets lost in too much detail.

Our alphabetical display invites keyword searches for embedded strings; once an interesting term is found, a click brings up its complete record, with all equivalent, broader, narrower, and related terms, plus scope note, source, etc., if included (see figures 4-7). A click on any term in this record brings up the record for that term, so that in effect, you have a giant hypertext in which every term is linked to every other term in a hierarchy in accordance with the relationships recorded in each thesaurus record. At our present stage of development, users may tag terms that they wish to use to search a database as they browse through the thesaurus.

9. TESTING, VALIDATION

There is a great need for testing and validating the impact of end-user thesauri, or for that matter, thesauri in general, on the effectiveness of information retrieval. Thesauri are expensive to create and maintain, and we should seek solid, rigorous evidence that the benefits they provide are commensurate to their costs. So far, there has been very little research to demonstrate their effectiveness.

Such research will have to be based on real people doing real searches for answers to real questions. Earlier research comparing "natural language indexing" to "controlled vocabulary indexing" has been largely worthless for this purpose because it was based on artificial queries, small artificial databases, and artificial relevance judgments. The whole point of controlling, tracking or managing vocabularies is to improve the interface between real users and databases; if the users are removed, there is little point to vocabulary management.

The chief argument for vocabulary tracking and management is the tremendous variability in the human use of language. This variability has been verified over and over again in psychological and information science research. An excellent summary is provided by Furnas, Landauer, Gomez and Dumais, based on their own extensive research at Bell Communications Research (Bellcore). To quote: "The fundamental observation is that people use a surprisingly great variety of words to refer to the same thing. In fact, the data show that no single access word, however well chosen, can be expected to cover more than a small proportion of users' attempts. Designers have almost always underestimated the problem, and, by assigning far too few alternative entries to databases or services, created an unnecessary barrier to effective use" (p. 964). "Even with fifteen aliases [alternative terms], only 60-80 percent of 'attempts' will be satisfied. Clearly the only hope for untutored vocabulary driven access is to provide many, many alternative entry terms. Thus aliases are, indeed, the answer, but only if used on a much larger scale than usually considered" (p. 968).

This entire paper has focused on the creation of thesauri based on human conceptual analysis. We need to add into this mix the application of automatically generated thesauri in which term relations are based on co-occurrence. Thirteen years ago, Tamas Doszkocs demonstrated the possible benefits of such thesauri at the ASIS annual meeting (Doszkocs, 1978). For a search on prenatal toxicity, his "Associate Interactive Dictionary" provided the following list, ranked in order of frequency of co- occurrence: postnatal, gestational, fetus, gestations, teratogenicity, embryocidal, perinatal, placental. . . . I have always believed that this demonstration proves the useful potential

ISSN: 2324-9773

Anderson, J., & Rowley, F. (1991). Building End-User Thesauri from Full-Text. 2nd ASIS SIG/CR Classification Research Workshop, 1-10. doi:10.7152/acro.v2i1.12543

of this approach. We need to include such co- occurrence thesauri, as well as conceptual thesauri, in our continuing research and development.

10. REFERENCES

Aitchison, Jean; Gilchrist, Alan. Thesaurus Construction: A Practical Manual. 2d ed. London: Aslib, c1987. 173 p.

ANSI/NISO Z39.19. 1990. National Information Standards Organization. Proposed American National Standard Guidelines for Thesaurus Construction, Structure and Use. Gaithersburg, MD: National Institute of Standards and Technology, 1990. 113 p.

Bates, Marcia J. 1986. "Subject Access in Online Catalogs: A Design Model." Journal of the American Society for Information Science. 37(6): 357-376.

Doszkocs, Tamas D. "An Associative Interactive Dictionary (AID) for Online Bibliographic Searching." American Society for Information Science, Annual Meeting Proceedings. 1978: 100- 109.

Farradane, Jason. "Relational Indexing." Journal of Information Science 1(5): 267-276; January 1980.

Furnas, G. W.; Landauer, T. K.; Gomez, L. M.; Dumais, S. T. "The Vocabulary Problem in Human-System Communication." Communications of the ACM 30(11): 964-971; November 1987.

International Organization for Standardization. ISO 2788: Guidelines for the establishment and development of monolingual thesauri. 2d ed. Geneva: ISO, 1986.

Ranganathan, S. R. The Colon Classification. Rutgers Series on Systems for the Intellectual Organization of Information, v. 4. New Brunswick: Graduate School of Library Service, Rutgers, the State University, 1965. 298 p.

Soergel, Dagobert. Organizing Information: Principles of Data Base and Retrieval Systems. Orlando, FL: Academic Press, 1985, p. 450.

Vickery, B. C. Faceted Classification Schemes. Rutgers Series on Systems for the Intellectual Organization of Information, v. 5. New Brunswick, NJ: Graduate School of Library Service, Rutgers, the State University, 1966. 108 p.

Wang, Yih-Chen; Vandendorpe, James; Evens, Martha. "Relational Thesauri in Information Retrieval." Journal of the American Society for Information Science. 36(1): 15-27; January 1985.

ISSN: 2324-9773