Provenance, Journal of the Society of Georgia Archivists

Volume 8 | Number 1 Article 3

January 1990

Evolution of the Thesaurus of University Terms

Jill M. Tatem
Case Western Reserve University

Evolution of the *Thesaurus of University Terms*

Jill M. Tatem

Three years ago the Society of American Archivists published a modest pamphlet—*Thesaurus of University Terms* (TUT). This thesaurus was developed at Case Western Reserve University (CWRU) by Jeff Rollison and Jill Tatem, with the assistance of Ruth W. Helmuth, then university archivist, and their colleagues Fred Lautzenheiser and Bob Psuik.

In agreeing to publication of TUT it was hoped that the thesaurus might contribute to the discussion about the ways archivists analyze and describe college and university archival materials. A secondary goal was that other similar repositories might be able to use TUT as a starting point to develop or examine their own descriptive vocabularies. Almost as an afterthought it occurred to the compilers that other repositories might actually use TUT to describe their records.

PROVENANCE, Vol. VIII, No. 1, Spring 1990

Experiences during the intervening years have led to the conclusion that college and university archivists are either very kind or very desperate. The anticipated criticisms and suggestions were not forthcoming. The responses have been almost entirely of the "We've bought your thesaurus and we really like it, but we're not sure we're doing it right. How do you use it?" type.

The purpose of this article is two-fold: to complete an obligation to all those gentle or desperate college and university archivists who have invested seven dollars to purchase TUT and costly hours to figure out what to do with it. A selfish motive, and second purpose, is that, in explaining what CWRU was and is attempting to accomplish, someone will be prompted (perhaps through irritation at seeing the thing done badly) to suggest a better way.

TUT began life in 1983 as an experiment based on a notion of Jeff Rollison's. Specifically, he wanted to build a mechanism to describe CWRU's archival records based on the functions carried on in the university. It was to be simple to create, simple to use, and detailed. What the experiment became was a vocabulary used in two online files. One is a post-coordinate folder-level index to records. The other is a description of record-creating entities. It was hoped that this would become an important part of a total descriptive system.

Of course, the notion of explicit access to records based on function was not new to the archives. The classification system developed by Ruth Helmuth in the mid-1960s had served as the foundation of arrangement and, consequently, of access to archival records since the archives was

established. Briefly, this system classifies university offices by their functional responsibilities, not by administrative hierarchies. The notation, because it represents a given type of office such as a registrar the same way in each record group, links offices with similar responsibilities across record groups. Thus, the first step in retrieval, that of linking a topical request to the most likely relevant sources via knowledge of the primary function of a record creator is well supported on a macro level.

The classification system was supplemented by other more detailed finding aids, of course. Rarely could the archives not provide some information about a topic within its collection scope. But there was a growing unease on the part of the staff about its ability in, say five years, to continue to provide the level of service users had grown to expect without devoting every working hour to reference.

In 1983, the archives's last staff increase was seven years old. While the staff was not growing, the collection and the number of service requests were—at an alarming rate. As an institutional archives, the universe the archives documents is small and cohesive. An overwhelming majority of collection use is by the staff of the archives providing research services for university administrators who need detailed but comprehensive answers, not references to likely sources. Typically, these answers are needed yesterday. Very little document retrieval—what librarians refer to as "known-item" searches—occurs. And visitors who require only that they be shown possibly relevant records series and left to browse dozens of boxes of correspondence files are rare.

The immediate need was for a kind of information retrieval disaster prevention plan. The more long-range goal is to build a descriptive system that 1) actively helps users define (and continuously refine) their information needs, and 2) locates information or sources of information relevant to their needs. Ideally, this should be a progressive process, not a series of frustrating dead-ends and false starts. And this leads to the third goal: from the users' perspective the system must be consistent and predictable, that is, what is learned in one search should be useful in subsequent attempts. Under no circumstances should users have to "unlearn."

The compilers worked from several basic hypotheses (none of them new insights, but mentioned to explain the context in which TUT is used). First, different users have different perceptions of the nature of the collection. A corollary is that often the same user has different perceptions of the nature of the collection at different times. Second, users have widely differing precision and recall requirements. And, third, most of the time, in stating their information needs, users are trying to define the unknown.

One path through this maze of ambiguities and unknowns is to present multiple views of the collection. Variables determining these views include which portions of the collection are described (both as physical and intellectual entitites), by what criteria are those portions linked, how detailed/comprehensive is the description, and for what kinds of users is it meant.

In this context, the files the archives is building using TUT form two layers of a multi-tiered system of finding

aids. Too, among many of the biggest problems in this approach, are identifying useful perspectives for which and from which to create collection "views", and integrating or linking the different views so they form a coherent and navigable whole—not a mess of pieces and parts.

The classification system provides one useful tool as a skeleton which links offices horizontally through functional relationships. TUT could provide a way to put flesh on the skeleton both as a translation into English of the functional concepts embodied in the classification schedules and as a way of extending those linking concepts into more specific descriptions of detailed activities of which functions are composed.

The compilers tried, however, to be realistic about what they could achieve. For current users, finding aids are irrelevant. For the price of a phone call they are accustomed to receiving answers at the exact levels of precision and recall they require. Any descriptive system that did not either make users less dependent on the archives staff and more willing and able to conduct their own research, make the archives staff more efficient without sacrificing quality, or both, would be a wasted effort. As appealing as the first possibility was, the compilers knew it would have to be an awfully sexy system to lure people away from those phone calls. So they concentrated on working out some way to help themselves first, secure in the virtuous knowledge that, in helping themselves first, they would really be helping their university.

On this noble and altruistic note, they set about the task of deciding what was unpleasant and time-consuming about the way they currently worked. Surprisingly enough,

they managed to compress what started as a very large list into two problems:

1) Everyone hated scanning pages and pages of box lists to extract the few folders that looked promising;
2) A way was needed to break out of the cycle of starting most searches with the same record series, because those were the ones this staff knew best, even though there might be better sources. Of course, the more the best-known ones were used, the better-known they were, and the more they were used, the less the rest of the collection was exploited. And there was that awful dreaded wondering about what might have been missed.

After weeks of brainstorming, the first wheel had been reinvented. (There were to be more.) Anyone familiar with information retrieval theory will recognize that the first problem was a need to improve precision, that is, the number of relevant documents retrieved as a proportion of the total documents retrieved. The second problem was the need to improve recall, that is, the number of relevant documents retrieved as a proportion of the total relevant documents in the system.

Invigorated by the realization that their experience had validated thirty years of research in information science, the compilers forged ahead to determine how best to solve these two problems—problems that had stymied some of the best minds in the field. Unfortunately, the experts claimed that both these problems could not be solved at

once. It was possible to have better precision or better recall but not both—choose one.[1] Not liking the sound of this, the archives staff ignored it. (This was not to be the last good advice they ignored.)

Instead, they opted to turn the precision problem over to the computer. It should be noted that the archives had decided very early to build an online system. In 1983 microcomputers were quite expensive and turning one into a $7000 typewriter, instead of exploiting its powerful retrieval capabilities, appealed to no one. The computer was ideally suited to scanning pages of descriptions and would do it faster. The humans would then devote their energies to the recall problem, which sounded more interesting, as it would probably involve the rediscovery of forgotten treasures.

This is an oversimplification, of course. Because of the kind of information that was to be extracted from the collection, several decisions to aid precision were made. One of these was to focus on folder-level descriptions.

It would have been simple to have cleaned up the substance of the existing finding aids and left the basic structure alone. For example, storing accurate box lists in machine-readable form for online searching would certainly speed the process of scanning folder titles. Unfortunately, easy-to-use but sophisticated text retrieval software for microcomputers was not available in 1984. And the use of existing folder labels would not solve language problems.

---

[1] Elaine Svenonius, "Directions for Research in Indexing, Classification, and Cataloging," *Library Resources and Technical Services* 25 (January/March 1981).

While increasing the depth of indexing at the series level would certainly direct staff attention to less frequently used but possibly useful records, it was concluded that a great deal of work would produce very little advantage.

It is unclear at what point the project focused on vocabulary control as the most useful beginning or how seriously other possibilities were explored. Because discussions frequently returned to vocabulary problems, this was undoubtedly seized as the solution very early. It was necessary to circumvent problems created by using folder titles of originating offices and, frankly, some very eccentric processors. Some of the worst of these were extensive use of proper names without any context, changes in terminology both over time and across the university, and the ubiquitous non-descriptive horrors like "correspondence, 1954." The biggest language problems were the need for descriptive descriptions and generic posting.

In spite of the fact that experimental testing of information retrieval systems has been going on for thirty years, there is more information on what is not known than what is known about what factors make for good systems. While conclusions of many of these studies have limited generalizability or are simply not reliable because of flawed methodologies, they have produced a small body of conventional wisdom. Some of the pieces of wisdom are that complex descriptive structures do not work much better than simple ones and that artificial indexing

languages do not work much better than natural language.[2] Clearly, controlling the descriptive vocabulary was not a panacea. From the research findings reviewed (by no means an exhaustive review), the most useful conclusion found was that natural language and controlled vocabularies each aid precision and recall, but in different ways, and that many other system variables have at least as significant an effect on information retrieval performance as does the descriptive language. It is generally acknowledged that vocabulary control aids recall by controlling synonymy and relatedness, and that precision problems with controlled vocabularies stem from lack of currency and specificity.[3] The need for control of synonymy and relatedness were two of the most

---

[2] Bert R. Boyce and Donald H. Kraft, "Principles and Theories in Information Science," *Annual Review of Information Science and Technology* 20 (1985): 159-60. Several recent publications have reviewed the results of the last few decades of research. Among them are Karen Sparek Jones, ed., *Information Retrieval Experiment* (London: Butterworths, 1981), especially the author's own articles in this compilation, "The Cranfield Tests" and "Retrieval Test Systems." Also helpful are Pauline A. Cochrane, *Redesign of Catalogs and Indexes for Improved Online Subject Access* (Phoenix, AZ: Oryx Press, 1985) and *Subject Retrieval in the Seventies: New Directions* (Westport, CT: Greenwood Publishing Co., 1972).

[3] Elaine Svenonius, "Unanswered Questions in the Design of Controlled Vocabularies," *JASIS* 37 (1986): 331-340. Jean Aitchison and Alan Gilchrist, *Thesaurus Construction: A Practical Manual*, 2nd. ed. (London: ASLIB, 1987), 3-9.

troublesome problems, so this became a priority in spite of the discouraging research findings. The staff reassured themselves with the hope that between their ability to modify the thesaurus quickly and easily, reliance on folder level descriptions, and the relatively stable terminology, adequate precision levels could be maintained.

Having decided on a controlled vocabulary of some species, it was a relatively simple matter to decide on a thesaurus using minimal precoordination. It was important to keep the list of terms small. The compilers also wanted to avoid all the aggravation of striving to maintain consistency of word order that comes as a necessary consequence of high levels of precoordination. And since this was to be an online index, the combination of terms necessary to achieve desired levels of specificity would be handled at the time of searching.

Finding the words was easy. Putting them into some useful kind of order was not. The staff attempted to apply the principles and techniques of facet analysis to functional descriptors as a means of imposing order. The first difficulty was in defining a function. If it is simply a purposeful, authorized action, then the restricting vocabulary describes concepts like FUNDRAISING, AUDITING, ESTABLISHING, TERMINATING. Some of these are understandable on their own, but many do not really mean anything useful until the object of the activity is known. Programs, departments, employees (which is usually called firing, if its involuntary or resignation or retirement if it is not) can all be terminated. Students are terminated (usually by graduating or withdrawing), as are buildings (usually thought of as demolition). In order to

clarify these syntactic relationships, functions can be redefined as purposeful, authorized actions upon objects. In constructing a vocabulary, however, the result is a very long list of pre-coordinated descriptors. The staff then turned to facet analysis.

Facet analysis identifies the fundamental aspects of a subject and then organizes the subject's descriptive terminology into groups or facets. The trick is determining what aspects of a subject are fundamental. A number of criteria have been used over the years in developing different thesauri. They generally are variations on entities, processes, properties, space, and time.

All members of each group (called a focus) of terms under the main facets share a single explicit characteristic. For example, entities might be grouped into abstract concepts, inanimate objects, etc.[4] Accordingly, the first-level division of TUT into four sections was made without much difficulty: form of record, places, individual record-creating entities, and everything else. The first three are straightforward alphabetical arrangements with related and preferred term cross-references. Since the last section is the heart of the thesaurus it was here that organizing terms was most important.

The difficulty was in identifying criteria for division that were sufficiently detailed to create cohesive groups,

---

[4] Phyllis A. Richmond, *Introduction to PRECIS for North American Usage* (Littleton, CO: Libraries Unlimited, 1981), 27-32; Aitchison and Gilchrist, 50-52; Lois Mai Chan, Phyllis A. Richmond, Elaine Svenonius, eds., *Theory of Subject Analysis: A Sourcebook* (Littleton, CO: Libraries Unlimited, 1985).

without being so detailed as to render the concept too specific. This is basically a problem of perspective. For example, DORMITORIES are both a type of building and a type of student service. Many thesauri solve this difficulty with polyhierarchies. The term appears in both foci, the notation identifying their different meanings. This approach was rejected in order to keep TUT small. Another concern was that this would require either greater precoordination or reliance on the notation to preserve the meanings of terms in use. Each term needed to be understandable out of context, and it was important to have minimal precoordination and a high degree of specificity. These are not complementary goals. A compromise was struck by reducing the clarity of distinctions among facets and foci. The result is that the characteristics by which terms are grouped are neither intuitively obvious nor made explicit.

This is TUT's most serious flaw. It not only limits ease of use of the existing vocabulary, but it will create obstacles to future modifications. In all fairness, however, neither of these problems has surfaced yet. TUT has been used, with some degree of success, for five years. (To what degree of success is not yet certain because controlled experiments on retrieval effectiveness have not been completed.) Nine new staff and eight students (one of whose primary language was not English) have been taught to use it without difficulty, and descriptors have been added successfully and easily.

Other problems which are being addressed include changing the display to improve ease of use. Since TUT's publication, efforts have been made to add scope notes and

cross references and to expand the entry vocabulary. It was clear four years ago TUT was lacking in these areas, but the primary concern was to get a working version ready for use and not to develop a definitive vocabulary.

TUT was an attempt to relate activities to the functions they support isolated from administrative structures, in such a way that each activity fit under one and only one function. This was probably an attempt to impose a two-dimensional model on a multi-dimensional world. What was achieved was a set of terms that describes activities and topics commonly found in the administrative records of colleges and universities. And TUT does that fairly well, because it is easy to use and fairly flexible. What TUT does not do is to aid retrieval by using the structure of a vocabulary to build paths through the mass of documentation that, because they are based on links that are inherent to the record and concepts that are part of the every-day work life of the intended users, are easy to follow.

Anyone contemplating a similar endeavor would do well to reflect on the croquet game Lewis Carroll's Alice played with the Queen of Hearts. It should be remembered that the croquet balls were live hedgehogs, the mallets live flamingoes, and the arches, soldiers doubled-over. As Carroll explained the procedure: "The chief difficulty Alice found at first was in managing her flamingo: she succeeded in getting its body tucked away, comfortably enough, under her arm, with its legs hanging down, but generally, just as she had got its neck nicely straightened out, and was going to give the hedgehop a blow with its head, it would twist itself round and look up into her face, with such a puzzled

expression that she could not help bursting out laughing: and when she had got its head down, and was going to begin again, it was very provoking to find that the hedgehog had unrolled itself, and was in the act of crawling away: besides all this, there was generally a ridge or a furrow in the way wherever she wanted to send the hedgehog to, and, as the doubled up soldiers were always getting up and walking off to other parts of the ground, Alice soon came to the conclusion that it was a very difficult game indeed."[5]

Jill M. Tatem is assistant university archivist, Case Western Reserve University. This article was originally presented at the 1989 Society of American Archivists annual meeting in St. Louis.

---

[5] Lewis Carroll, *Alice's Adventures in Wonderland*, reprinted edition (New York: Avenel Books), 121-22.