Transportation Research Thesaurus

The Why, the What, the How

• • •

DAVID BATTY

Effective communication is always facilitated by the use of a common language that is understood by both parties. In human speech, common language may be ambiguous or misleading because of the way it is used, but exchange and feedback usually allow the parties to communicate. Recorded information does not offer the same opportunity for immediate interaction and feedback. Writers may use inconsistent terminology, so that a searcher who chooses to scan the full text of a document must include in a search statement all the synonyms, related terms, and levels of detail the author might have used. Therefore, we need a substitute for the result of spoken interchange—the "Ah, yes, I see. What you mean is...." That substitute is the common, controlled index language (thesaurus) used by the indexer to interpret and represent the themes, concepts, and language of the author, and used by the searcher to interpret and represent sometimes vague expressions of a need to know.

Unfortunately, the value of controlled index languages was ignored in the United States for many years. Early on-line database providers were often dissuaded from using a thesaurus for what seemed to be valid economic considerations: fear of the apparent expense of designing and using a thesaurus, and doubt that the size of the database would justify that expense. These apprehensions turned out to be ill founded. Databases grew to almost unmanageable size, and the cost of searching titles and abstracts far exceeded what would have been the investment in the development and use of a controlled index language.

With the increase in CD-ROM publication, we have at our fingertips packages that can contain 600 MB in a single unit—roughly the equivalent of 150,000 digitally stored single-spaced pages or 8,000 printed book pages stored as graphic images. But this accessibility is misleading, since we cannot flip through the CD-ROM anywhere near as easily as we can browse a printed book or document file. Full-text searching will always be valuable for browsing in any size of file, but in large files, controlled searching will be needed for efficient retrieval.

This point is one that even law firms, long convinced of the efficacy of full-text searching, are beginning to appreciate. Several studies of litigation support activities have shown that the use of full-text searching is not as effective as the use of even a simply designed controlled index language. Litigation support companies hired by law firms to gather discovery documents and have them ready for attorneys at the drop of a question have long used controlled languages, called taxonomies because they are used in hierarchical form. Lawyers have now found that a taxonomy and indexing of discovery documents require much less of an investment than the continuing cost of paralegals (or attorneys) to do repeated text searching—sometimes for the same documents.

Another common objection to controlled index languages is their apparent complexity. But people have learned to use and value other organized reference works, including encyclopedias, back-of-the-book indexes, and even the yellow pages of a telephone directory.

Perhaps it is the idea of an intermediary language that is unsettling: "Why can't I go straight to what I want?" Yet no one arriving in a strange city would expect to go directly to the local city hall (even knowing the address) without a map or a knowledgeable cab driver (reference librarian) who already knows the map.

The author is president, CDB Enterprises, Inc.

Need for a Transportation Research Thesaurus (TRT)

It is not uncommon today for a database to contain records collected and indexed during a span of 50 years by a variety of indexing languages—or indeed, none at all. Over time the working language of a discipline changes, and eventually reaches the point at which a new, consistent indexing language is needed.

Such was the case for the Transportation Research Information Services (TRIS) database, the on-line database operated and maintained by the Transportation Research Board. TRIS had grown greatly from the Highway Research Information Service, with 26,000 records not indexed by any controlled language, through the incorporation of many other indexed databases: the International Road Research Documentation (IRRD) database (see article page 13); databases related to specific areas of maritime and rail transport and urban transit; and databases of the major transportation libraries at Northwestern University and the University of California, Berkeley.

In December 1993 a contract was awarded to CDB Enterprises, Inc., for the development of a new language for transportation research. This new language would be used primarily with TRIS and secondarily as a general transportation thesaurus.

Before the Transportation Research Thesaurus (TRT) project began, the TRIS database contained 386,845 records, which were indexed by 70,416 postable terms. (Postable terms are terms intended for use in indexing or searching; nonpostable terms are usually synonyms, quasi-synonyms, or terms deemed too specific for inclusion as postable terms.) These 70,416 terms had been reduced by eliminating all terms used fewer than 10 times, by removing extraneous punctuation, and by normalizing plural forms to the extent that this could be done mechanically with a simple algorithm. The thesaurus project team used this reduced list as the basic vocabulary resource for the new thesaurus. The team drew many additional terms from other vocabulary sources, including the original list of TRIS index terms and other thesauri, glossaries, manuals, and dictionaries. By the end of the project, the thesaurus contained 8,518 postable terms; many additional, nonpostable terms (commonly called *USE references* or *lead-in vocabulary*) remain as an aid to users.

Structure of TRT

The TRT is based on a set of clusters of associated terms called facets (or hierarchies). Each facet has a Top Term, for example, TESTING. For ease of reference, each facet is tagged with a capital letter of the alphabet:

A TRANSPORTATION
B TRANSPORTATION OPERATIONS
C MANAGEMENT AND ORGANIZATION
D COMMUNICATION AND CONTROL
E PLANNING AND DESIGN
F CONSTRUCTION AND MAINTENANCE
G TESTING
H SAFETY AND SECURITY
J ENVIRONMENT
K ECONOMIC AND SOCIAL FACTORS
M PERSONS AND PERSONAL CHARACTERISTICS
N ORGANIZATIONS
P FACILITIES
Q VEHICLES AND EQUIPMENT
R MATERIAL
S PHYSICAL PHENOMENA
T DISCIPLINES
U MATHEMATICS
V AREAS AND REGIONS
W TIME
X INFORMATION ORGANIZATION

Levels within a hierarchy are tagged with additional lowercase letters of the alphabet. A part of the hierarchy for T: DISCIPLINES is shown below. There are gaps in the notational codes to allow for the addition of other terms in the future.

T	DISCIPLINES
Tp	Science
Tpd	Biology
Tpdb	Biophysics
Tpdbb	Bioacoustics
Tpdbd	Biodynamics
Tpdbk	Biokinematics

Terms that combine two or more postable terms (e.g., CULVERT DESIGN, from CULVERTS and DESIGN) are called precoordinated terms. The TRT includes only a limited number of frequently used precoordinated terms (e.g., HIGHWAY DESIGN) in order to reduce the size of the vocabulary and facilitate the work of indexers. In fact, in the TRT the concept of CULVERT DESIGN would be represented only by the two terms CULVERTS and DESIGN.

One of the important functions of a thesaurus is to provide notes on the use of terms and references to related terms. The TRT therefore includes Scope Notes (SN), giving brief descriptions when necessary of the terms' scope; Use For (UF) indicators,

for synonyms or other nonpostable terms; Broader Terms (BT), indicating more general terms that are one level higher in the hierarchy; Narrower Terms (NT), consisting of all postable terms that are one level lower in the hierarchy; and Related Terms (RT), comprising other postable related terms.

TRT Display Formats

The machine-readable form of the TRT offers displays in four formats: (1) hierarchical, (2) rotated (in this case a keyword out of context [KWOC] index), (3) alphabetical, and (4) the full display traditionally associated with a thesaurus. Each of these formats conforms basically to the current American National Standards Institute/National Information Standards Organization Guidelines for Monolingual Thesauri. However, it should be noted that both the hierarchical and rotated-term displays of the TRT include enhancements that allow these two shorter displays to be used effectively as a complete thesaurus without the need to consult the much lengthier full display.

Hierarchical Display

This format displays a taxonomy or hierarchy of all postable terms in the order of their notational codes, and thus shows all family relationships. It also provides SNs and UF and RT references to terms in other facets. For example:

```
Pmf    Air transportation facilities
Pmfc       Airports
Pmfk       Airstrips
         SN        Runways without airport or
                     airbase facilities
         UF        Landing strips
         RT        Airport runways (Pmfcgmr)
Pmfke          Emergency airstrips
Pmfm       Airways
```

KWOC Index

The KWOC index is a vital cross-referencing and searching tool. It displays in alphabetical order every significant word of every term or phrase in the TRT, regardless of the word's position in the term or phrase or its context in the TRT. For example:

```
AIRPORTS
   Access to AIRPORTS        <use>  Bmkha
   AIRPORTS                  Pmfc
AIRSTRIPS
   AIRSTRIPS                 Pmfk
   Emergency AIRSTRIPS       Pmfke
```

The notational code of each term is included in the KWOC index to allow quick and easy cross-referencing to the appropriate facet in the hierarchical display. Note that nonpostable terms are included in the KWOC display, with a "use" cross-reference to the notational code of the postable term; thus every significant word of both postable and nonpostable terms is accessible through the KWOC index.

Alphabetical Display

This display provides an alphabetical listing of all postable and nonpostable (lead-in) terms, also showing the notational code for each term. This listing allows easy scanning of the descriptors for experienced users who know the terminology.

Full Display

The full display for each postable term shows all relationships. The following example is based on one of the postable terms in the hierarchical example given above:

```
Airstrips          Pmfk
   SN    Runways without airport or airbase
           facilities
   UF    Landing strips
   BT    Air transportation facilities
   NT    Emergency airstrips
   RT    Airport runways (Pmfcgmr)
         Airways
```

Each nonpostable term is also listed alphabetically in the full display, together with the notational code of the preferred term that should be used:

```
Landing strips        <use>  Pmfk
```

Development of TRT

The TRT was developed using the now well-recognized process of facet analysis. A facet is a group of terms that share a single, principal characteristic. The terms may share multiple characteristics, but only one is chosen as the principal characteristic.

For example, Wood, Nylon, Steel, Copper, and Wool all share the characteristic of being MATERIALS, regardless of any other characteristics some of them may share, such as Combustibility. These terms thus constitute the beginnings of a MATERIALS facet. Sometimes the terms in a facet can be divided into subfacets based on secondary characteristics. For example, within the MATERIALS facet, Wood and Wool are

ORGANIC MATERIALS, Steel and Copper are METALS, and Nylon is a PLASTIC. Facets and subfacets are then arranged as simple hierarchies of terms, from general to specific.

Any hierarchy of terms requires a mechanism to preserve the order and level of subordination of its contents, and to represent the terms themselves by a set of unique codes that can link the hierarchical order to alphabetical orders. In the TRT, the notational codes use the alphabet: a capital letter to indicate the facet or Top Term, and lowercase letters for the detail of the hierarchy.

Notational codes can be used effectively throughout thesaurus development—to allocate terms to facets and indicate their current location in the facet hierarchy; to allow efficient manipulation of large numbers of terms; and to ensure that references reciprocate, for example, that every BT reference has a corresponding NT reference. The codes also facilitate the manipulation of terms to produce listings of the whole thesaurus in a variety of orders and formats, and to produce the final thesaurus showing all the relationships.

The use of facets has many advantages. With the terms organized into smaller, related groups, each group of terms can be examined more easily and efficiently for consistency, order, hierarchical relationships, relationships to other groups, and acceptability of the language used in the terms.

The facet approach is also useful for its flexibility in dealing with the addition of new terms and relationships. Because each facet can stand alone, changes in a facet can usually be made easily at any time without disturbing the rest of the thesaurus.

The facet approach combined with the use of notational codes is especially amenable to software applications. The TRT project team used software developed by CDB Enterprises, Inc., customized for thesaurus construction. This software was used to ensure the integrity of the structure of hierarchies and references between them, to generate printed and on-screen displays of different thesaurus formats, and even to display the thesaurus for point-and-shoot allocation of terms in on-line indexing. Nonetheless, the process of facet analysis, and the construction and maintenance of a thesaurus, are and must always remain essentially an intellectual endeavor.

A further and final benefit of the facet approach becomes apparent in the use of the thesaurus: an indexer or searcher finds it easy to understand a set of hierarchically organized facets as a conceptual map showing the precise level and set of associations of a term. Categories of related terms are easier to negotiate than a long list of alphabetized terms.

TRT Production

During the project it was agreed that a machine-readable format for thesaurus publication and use would be acceptable. Accordingly, the project team developed software to display the thesaurus. This program (called The Thesaurus Viewer or simply The Viewer) allows the user not only to display the thesaurus in alphabetical, hierarchical, and rotated formats, but also to navigate the displayed vocabulary. For example, when a term is highlighted in the KWOC index display, a click of the mouse on the HIERARCHY button at the top of the screen will produce a display of the term in its hierarchical position, a click on the ALPHABETICAL button will show the term's position in an alphabetized list of terms, and a double click on the term in any display will show the full display of the term with all SNs and relationships.

The TRT in its electronic form contains all the thesaurus files in every display format; The Viewer; and ASCII files suitable for local printing of the full TRT User Guide and Maintenance Manual, an outline of the TRT, a file of the TRT in its hierarchical format, and a file of all descriptors in alphabetical order (with notation) to serve as an authority file for those who need it for their own software applications.

Continued on page 41

Future Work on Transportation Research Thesaurus

A continuation phase of the Transportation Research Thesaurus (TRT) began in December 1997. The purpose of this phase is to (1) design and conduct training sessions on using the TRT, especially with reference to TRIS; (2) revise the TRT in light of practical applications and review comments from the transportation community; and (3) revise TRIS indexing to bring it in line with the TRT, so that the whole TRIS database will be searchable by the same language back to its beginnings.

Ways of implementing and expanding the utilization of the TRT will also be explored. Consideration will be given to the development of mini- and micro-thesauri in specialized areas, such as intelligent transportation systems and transportation engineering. These will be self-sufficient thesauri, but compatible with the TRT so that indexers will be able to call on other parts of the TRT to augment indexing terms taken from a mini-thesaurus.

The TRT project team has also been approached with questions about foreign-language translations of the TRT. Any such considerations will almost certainly involve cooperation with the Organisation for Economic Co-operation and Development-supported preparation of a multilingual lexicon for users of the TRANSPORT CD-ROM.

Technology Transfer on the World Wide Web

Although the goal of applied research is to develop new products or techniques that will be used by practitioners, research results often take many years to be incorporated into common practice. A practitioner facing a problem frequently is unaware of relevant research and does not have ready access to research reports-the usual mode of technology transfer.

During the past 2 years, the Cooperative Research Programs (National Cooperative Highway Research Program and Transit Cooperative Research Program), administered by the Transportation Research Board, have been steadily increasing their use of the World Wide Web to assist in technology transfer, as well as other aspects of their business. Today, using the Web, a practitioner can quickly find out whether an NCHRP or TCRP project is relevant to his or her needs.

The foundation of the Cooperative Research Programs Web site (http://www2.nas.edu/trbcrp/) is a searchable database of status reports on all research projects since 1988. This database includes information on more than 480 projects covering all aspects of highway and transit research. The write-up for each project includes information on its status, objective, and principal investigator. The search engine is easy to use and identifies all write-ups containing the search terms. One can also find a project write-up by looking up the project number. The write-ups contain more up-to-date information on the status and work being performed on NCHRP and TCRP projects than is found in TRB's Transportation Research Information Services (TRIS) Research in Progress database (http://www3.nas.edu/rips/). As described below, some write-ups also provide additional value.

NCHRP and TCRP have begun publishing some reports and report appendices on line. In most cases, these reports would not be published in print because they have a limited audience that would not justify the printing costs. And the on-line report is much more accessible to users than the microfiche alternative. A future print-on-demand capability is envisioned that will make hard copies of these reports easier to obtain. Write-ups on the following projects include links to on-line reports. A current list of on-line reports is available at the Cooperative Research Programs Web site.

- NCHRP Project 1-31, Smoothness Specifications for Pavements
- NCHRP Project 3-46, Capacity and Level of Service at Unsignalized Intersections
- NCHRP Project 3-51, Communications Mediums for Signal, IVHS, and Freeway Surveillance Systems
- NCHRP Project 8-32(1), Innovative Practices for Multimodal Transportation Planning for Freight and Passengers
- NCHRP Project 8-32(5), Multimodal Transportation Planning Data
- NCHRP Project 14-12, Highway Maintenance Quality Assurance
- TCRP Project F-4, Bus Operator Workstation Evaluation and Design Guidelines

In addition, the contractors for some NCHRP projects have established their own Web sites. These sites typically include interim material that would normally not be available until project completion. Write-ups on the following projects include links to contractor Web sites:

- NCHRP Project 24-8, Scour at Bridge Foundations: Research Needs
- NCHRP Project 8-33, Quantifying Air-Quality and Other Benefits and Costs of Transportation Control Measures
- NCHRP Project 12-42, LRFD Bridge Design Specifications Support
- NCHRP Project 20-7/TASK 80, Assessment of AASHTO Needs for an Information Clearinghouse on the Principles and Programs of Continuous Quality Improvement

Some projects have developed executable programs and other types of files that are being distributed through the Cooperative Research Programs Web site. Write-ups on the following projects include attached files:

- NCHRP Project 3-51, Communications Mediums for Signal, IVHS, and Freeway Surveillance Systems
- NCHRP Project 4-18, Design and Evaluation of Large Stone Mixtures
- NCHRP Project 4-22, Pavement-Marking Materials: Health, Environmental, and Performance Assessment
- NCHRP Project 24-5, Downdrag on Bitumen-Coated Piles
- NCHRP Project 24-6, Expert System for Stream Stability and Scour Evaluation

The Cooperative Research Programs continue to look for ways to better serve practitioners. Please send any comments or questions to the staff member mentioned in the project listing or to Robert J. Reilly, director, Cooperative Research Programs, at the Transportation Research Board, 2101 Constitution Ave., N.W., Washington D.C. 20418. E-mail correspondence is also welcome; staff addresses are provided at the Cooperative Research Programs Web site.

Information provided by Ray Derr, Senior Program Officer, Transportation Research Board.

continued from page 6

It captures international materials for use by U.S. researchers and firms. International documents represent a wealth of untapped and often undiscovered resources. Underutilization of foreign resources results in duplication of costly research and waste of the valuable time of researchers. However, acquisition of foreign documents is often a difficult and lengthy endeavor. Problems associated with use of foreign documents include a lack of document ordering information, inadequate bibliographic citations, a lack of abstracts in English that can be used to determine a document's relevance, delays in obtaining a hard copy, the high cost of translation services and of foreign document acquisition, a lack of knowledge of what is accessible or available, and a lack of familiarity with foreign government organizations.

It applies rational and common information policies. To be effective, the NTL will have to be flexible. It would be counterproductive to replace fragmentation with rigid controls or arbitrary conformity requirements. U.S. DOT, as a leader in the development of the NTL network, will have to remain open and responsive to a wide range of information producers and users. The agency will have to work with the transportation community to identify the most pressing needs for better information and address those needs first. It will have to adapt its priorities as the needs of the community change. The policies and standards developed will have to be based on community-wide consensus.

It serves as an information advocate. As part of the NTL pilot, BTS has held a number of discussions with transportation librarians during the past year. A theme that keeps emerging is that the transportation research community needs strong advocates—one or more organizations that can sound the call for measures such as those discussed here. The advocates' ongoing goals should include identifying areas in which information requirements remain unmet and highlighting promising collections that are not generally accessible.

The rest of the pieces are more difficult to figure out, and it will not be possible to assemble the whole puzzle at once. But U.S. DOT and professionals around the country have taken two important steps. First, they have started to imagine or reimagine the possibilities of a full national transportation library system. Second, they have started to build such a system, using modern networking tools that may provide the efficiency and ease of construction that will finally make the NTL a reality.

continued from page 20

Using TRT

In indexing a document, the indexer uses as many postable terms from as many facets as needed to describe the document fully—typically from 6 to 10 terms per document. For example:

Title: Measuring traffic congestion and delay on an urban freeway following a ramp accident

Postable Term	Notation
Traffic measurement	Bte
Traffic congestion	Bthfc
Traffic delay	Bthd
Freeways	Pmrccdf
Ramps (Interchange)	Pmrcpjsrr
Traffic accidents	Hbbgt

In searching, the user constructs a search statement consisting of postable terms that together represent the request. The KWOC index provides the most comprehensive access since it shows all contexts of any word, even if buried within a phrase. The KWOC index also shows use references and the notations for all terms, thus allowing easy cross-reference to the hierarchy so the user can consult all the family relationships. The user may also double-click on a term in any display to see its full display with all notes and relationships.

If the search statement does not retrieve a satisfactory set of citations, it must be modified. The user refines the search to get fewer documents by adding terms with the AND operator or by using more specific terms within a hierarchy. The user broadens the search by dropping an ANDed term from the search statement or by adding more terms with the OR operator.

Availability of TRT

The latest version of the TRT will be published on CD-ROM by the Bureau of Transportation Statistics and will be available at no charge in the spring of 1998. It will also be available on the BTS Web site. Meanwhile, CDB Enterprises is issuing it at no charge on a set of three diskettes with a printed manual. CDB Enterprises is also prepared to make available a printed version of the hierarchical and KWOC listings at a cost of $50, plus shipping, anywhere in the world. For information, contact the author (telephone: 301-593-8901; fax: 301-593-1867; e-mail: davidbatty@aol.com).