

Shareable and machine-processable vocabulary structures: report on a NISO thesaurus workshop



earch engines and natural language processors are increasingly powerful retrieval tools.

We still need controlled vocabularies, however, as some experiences with retrieval at Cisco reveal. We also need vocabulary structures that are machine-processable so that we can manage vocabularies and share them with others. Recently, the National Information Standards Organization (NISO) held a planning workshop to consider a new standard for sharing vocabulary in a networked environment.

Probably all of us who work with specialized vocabularies have retrieval horror stories. At Cisco recently, I tested a search engine that offered advanced natural language processing (NLP) technology. I typed the term *route processor* (a specific product at Cisco) into the search engine's query box. Without consulting me, the search engine matched the root of just one of my terms (*route*) against synonyms (both nouns and verbs) found in general dictionaries and thesauri.

When I examined the results, I discovered the search engine had given me documents containing the following matches for *route*: *dispatch*, *forward* and *forwarded*, *overcome* and *overcomes*, *pass*, *path*, *send* and *sends*, *ship* and *shipped*, *transmit* and *transmits* and *transmitting*. My search for a very specific product returned all 2000 documents in the test database.

By Ruthanne Lowe, Cisco Systems, Inc.

I'm not surprised that general dictionaries, thesauri, and the automatic features of search engines and natural language processors don't work for Cisco's specialized needs. I am surprised, however, at the amount of work and expertise that are needed to build a shareable, machine-processable vocabulary from scratch. I was, therefore, very pleased to participate in the process to consider a NISO standard for electronic thesauri.

NISO Workshop on electronic thesauri

Barbara E. Cohen, an STC Indexing SIG member, and I were invited to participate in the NISO workshop on electronic thesauri along with approximately 60 other participants representing publishers, universities, libraries, information services, professional associations, government agencies, vendors, and corporations.

We met near Washington, D.C., on November 4th and 5th, 1999, to discuss the need for a new standard. The group was convened on the recommendation of another group that had met earlier to review and reaffirm the *Guidelines for the*

Construction, Format, and Management of Monolingual Thesauri (ANSI/NISO Z39.19-1993, revised 1998). The workshop was sponsored by the National Information Standards Organization (NISO), the American Psychological Association (APA), the American Society of Indexers (ASI), and the Association for Library Collections and Technical Services (ALCTS, a division of the American Library Association).

Issues and recommendations

We were asked to consider whether or not there was a need for a new standard for generating thesauri by electronic means. In addition, we were asked to consider the need for tools that aid in showing semantic relationships among terms; the need for building structures that support a variety of thesaurus displays; and the need for developing interoperability protocols, structures, and semantics.

By the end of the second day, we agreed that NISO should sponsor the writing of a standard for sharing vocabulary information in a networked environment. We also

Probably all of us who work with specialized vocabularies have retrieval horror stories.

agreed that the new standard should be wider in scope than the current standard. In addition to thesauri, it should potentially include other knowledge organization structures such as classifications, ontologies, taxonomies, semantic networks, and subject

(Continued on Page 4)

heading lists. As such, the proposed standard would supplement, not supersede, the current NISO Z39.19 standard.

Due to our diverse backgrounds and viewpoints, we recognized that key terms would have to be defined in the standard. What exactly were, for example, *metadata*, *ontologies*, *taxonomies*, *semantic networks*, *concepts*, *terms*, and even *indexes*? (The term *index*, for example, was used in group discussions to describe back-of-the-book or open-end retrieval devices, computer-generated lists running behind search engines, and structures on Web pages.) Was the *electronic* in *electronic thesauri* necessary? What did *electronic* mean in this context: did it mean generated electronically or available electronically?

We spent some time discussing thesaurus displays, but finally agreed that while displays were important and should inform the content of the term record, displays themselves should be the responsibility of individual corporations, vendors, or institutions.

The idea of a registry for relationship types was enthusiastically supported. Several participants supported the potential for embedding graphic images in the term record. While we had lively discussions regarding communications formats, MARC (Machine-readable Cataloging) formats, the Dublin Core, Web encoding and exchange specifications, and the role of preferred terms in an electronic knowledge structure, we didn't have unanimity on these topics. But then, for the most part, we were not seeking consensus. The keynote speaker, James Ander-

son of Rutgers, suggested that we be primarily concerned with getting all of our issues on the table.

Vocabulary and retrieval at Cisco

Since 1995, Cisco has been publishing product literature and technical documentation on the Web. As Cisco employees continually try to improve information retrieval on the Web site, they are evaluating the roles that controlled vocabularies, search engines, and NLP technologies might play in effective information retrieval.

Many editors, writers, and others at Cisco have made significant efforts to record, if not control, vocabulary usage, including the use of multilingual terms. As part of those efforts, Cisco publishes a glossary called *Internetworking Terms and Acronyms* each year (available on our Web site) and maintains word lists in the *Cisco Systems Documentation Style Guide*. A major metadata framework project is ongoing.

Cisco's extensive vocabulary has some interesting and unusual characteristics. First, Cisco has made up new vocabularies for commands, products, and technologies, and these terms do not necessarily follow standard English-language rules. Second, Cisco uses the same common words over and over again, and all that differentiates them is a singular or plural form, a prefix or a suffix, an uppercase or lowercase letter, or an adjacent word in a compound term. Third, Cisco's vocabulary grows rapidly as new companies and new technologies are acquired, and it changes as quickly as all vocabulary in the industry

changes. These characteristics make it difficult for standard search engine software or NLP technologies alone to help us with information retrieval.

For instance, search engines can be programmed to search for singular and regular plural forms and can be configured to be case insensitive, but while these features might be desirable at times (for example, to search for *router* or *routers*, *interface* or *interfaces*, *VLAN* or *VLANs*, *boot* or *BOOT*), we must also be able to search for a singular or a plural form or for an uppercase or lowercase letter when it represents a different topic or concept (for example, *show interface command* or *show interfaces command*, *apart* or *APaRT*, *LECs* or *LECS*).

NLP technologies are promising, but I'm not sure they will always help us. We will not improve precision if we allow automatic truncation of suffixes on many of our common words. In Cisco's intranet, for example, *routed* protocols are not the same as *routing* protocols, and each must be searchable. In the Cisco IOS® command language, *ipx route*, *ipx router*, and *ipx routing* are three different commands, and each must be searchable.

For all of these reasons, Cisco anticipates building its own thesaurus-type structure with term records that include synonyms, variants, user warrant terms (the terms our users type into the query box on the search page), relationships, scope notes, and usage information. Such a structure must contain all of the information needed for vocabulary control and management and must, at the same

(Continued on Page 5)

time, be able to run behind our search engine, be displayed for consultation by users at the point of search, be made available to our writers, editors, and indexers, and, of course, be shared with interested and willing parties.

Conclusion

Despite advances in natural language retrieval by search engines and NLP technologies, many companies and organizations who want to improve retrieval will need to develop and manage controlled vocabularies. Having terms and relationships in machine-processable formats based on a standard will facilitate the sharing of those controlled vocabularies and other vocabulary structures. Developing the standard won't be easy because the issues are so complex. For one thing, a term record has two components: a machine-processable structure (fields and subfields) and the contents of the term record itself (terms, definitions, notes, and relationships). Also, the process of developing the standard will require the input of diverse groups of subject experts, and experts may have different and perhaps conflicting interests.

The composition of the committee that will go forward to write the new standard is suggested by Jessica Milstead in her final report on the workshop: "The standard committee should include not just thesaurus builders, but retrieval system vendors, software designers (vendors), and a knowledgeable user." Many STC members are, in fact, highly knowledgeable users of complex vocabulary structures;

their expertise could be invaluable on the standards committee.

STC members who wish to participate may play one of several roles. First, NISO has organizational members, such as STC, not individual members; therefore, STC as a whole can vote. A vote on draft standards is cast by STC Standards Council Manager and NISO Representative (currently Annette D. Reilly) who seeks input from STC experts as necessary before approving, disapproving, or abstaining with or without comments. Second, if STC members are experts in areas where NISO standards are being developed, they may be invited to serve on working groups.

STC members who wish to play a role may get in touch with me or with Annette D. Reilly (annette.d.reilly@lmco.com), our NISO representative.

Resources

Information about the NISO workshop, including a list of participants, issues, and background reading, is available at <http://niso.org/thesau99.html>.

Jessica Milstead was instrumental in planning the workshop. She also wrote the final report in which she included information about the workshop's background and planning, key issues, discussions, themes, issues presented by major speakers, notes from breakout sessions, and additional resources. Her full report is available at <http://www.niso.org/thes99rprt.html>.

Cisco's *Internetworking Terms and Acronyms* is available on Cisco's Web site at <http://www.cisco.com>.

Ruthanne Lowe is a technical editor for indexes and retrieval in the Enterprise Line of Business at Cisco Systems. She received her MLS from UCLA where she studied all aspects of information retrieval including indexing (she took a seminar in indexing from Robert Collison, former president of the Society of Indexers and author of Indexes and Indexing). Before coming to Cisco, she worked as a technical services librarian, cataloging instructor, and indexer for many years. She can be reached at rulowe@cisco.com, or at (408) 527-6276.

New book

Beyond Book Indexing: How to Get Started in Web Indexing, Embedded Indexing, and Other Computer-Based Media, edited by Marilyn Rowland and Diane Brenner is ready for shipping.

In this book, prominent indexing professionals provide an in-depth look at current and emerging computer-based technologies and offer suggestions for obtaining work in these specialties.

ASI members \$28.00
non-ASI members \$35.00

Shipping charges are \$3.95 for the first book, \$1.00 for additional copies.

Order by calling 1-800-300-9868 or 609-654-6266.

For more information e-mail custserv@infotoday.com or visit the web site at <http://www.infotoday.com>