

# The NISO Standard for Controlled Vocabularies: A Blueprint for Revision

by Bella Hass Weinberg

Bella Hass Weinberg is a professor in the Division of Library and Information Science, St. John's University. She consults on the design of thesauri, indexes and interfaces. She prefers voicemail to email: 718-990-1456.

I chaired the committee that developed the NISO *Guidelines for the Construction, Format, and Management of Monolingual Thesauri* [1], which was published in 1994 and reaffirmed in 1998. I participated in the NISO Workshop on Electronic Thesauri, held in 1999, at which it was decided to update the guidelines, but I was not part of the revision committee that produced the *Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies* [2], issued (only in electronic form) in 2005. That standard should soon be up for review and public comment; in this article I share my thoughts on the need for revision of the standard.

## Precoordination

At the initial meeting of the committee that developed the 1993 thesaurus standard, I recall asking, “How are we going to distinguish thesauri from subject heading lists?” The consensus was that thesauri are designed for postcoordination (combination of terms at the searching stage) and subject heading lists for precoordination (combination of terms at the indexing stage). Late in the comment stage, someone asked why the draft standard provided no guidance on subheadings to be used with descriptors, and I replied, “Because they are out of scope for a standard on thesauri.”

Given that the 2005 standard substituted the term *controlled vocabularies* for *thesauri* in the title, I expected to find guidance on subheadings in it. Therefore, I was stunned to read, “The rules for developing precoordinated indexing terms

are generally outside of the scope of this standard. These guidelines are available in sources such as the Library of Congress *Subject Cataloging Manual: Subject Headings . . .*” (p. 37). Nothing was thus accomplished by broadening the scope of the standard if it provides no guidance on subheadings.

My impression is that a global change was done to convert the term *thesaurus* to *controlled vocabulary* throughout the standard. This process resulted in the following ludicrous sentence in the definition of “subject heading”: “Precoordination of terms for multiple and related concepts is a characteristic of subject headings that distinguishes them from controlled vocabulary terms” (p. 9). (The term *controlled vocabularies* includes subject heading lists.) Another example: “Each term included in a controlled vocabulary should represent a single concept . . .” (p. 23). That is a requirement of thesauri, not subject heading lists, which often include conjoined terms, for example, **Banks and Banking**. As the previous quotation notes, subject headings precoordinate “multiple and related concepts.”

## Definitions

The 2005 *Guidelines* have eight pages of definitions close to the beginning of the standard (section 4) and a more extensive glossary in an (unnumbered) appendix. I much prefer the structure of the 1993 standard, which defined terms at the point of their first use in the text, and collected the definitions in a glossary that was treated as part of the text.

The 2005 *Guidelines* have inconsistencies between the definitions and the text. For example, *entry term* is defined as

“The non-preferred term in a cross reference . . .” (p. 5). The guideline on place names, however, says, “The form most familiar to the users of the controlled vocabulary should be designated as the entry term [i.e., the preferred term], and cross-references should be provided from the variants” (p. 33).

The replacement of *descriptor* (the word for the preferred term in a thesaurus) by term in this standard led to many awkward expressions, such as “entry terms and terms” (p. 63). The definitions section has a separate entry for *preferred term*, but that expression is not used consistently instead of *descriptor*. The definition of *indexing term* states, “Terms, subject headings, and heading-subheading combinations are examples . . .” and concludes, “Also called descriptor” (p. 6). The index says, “descriptors *see* terms” (p. 169).

In the definition of *controlled vocabulary*, the note “This rule does not apply to synonym rings” was erroneously placed after the rule for qualifiers of ambiguous terms rather than the rule for selecting a preferred term (p. 5). This observation is confirmed by the correct placement of this note under the rule for preferred terms in the body of the standard (p. 13). The exclusionary note for synonym rings is missing in the section on the purpose of controlled vocabularies, which specifies that they should “Provide consistent and clear hierarchies” (p. 11). Hierarchy is conspicuously absent from the abstract of the standard.

The 2005 *Guidelines* refer from *quasi-synonym* to *near-synonym* (p. 8). The definition previously given for quasi-synonym was assigned to *generic posting*, with the example “**furniture** UF beds” (p. 6). I like the expression quasi-synonym for this case, where the terms are treated as equivalent, although they are not even near-synonyms. *Longman Advanced American Dictionary* explains quasi- as “acting in some way like” [3, p. 1177].

The 2005 *Guidelines* treat *candidate term* and *provisional*

*term* as synonyms (p. 4), in contrast to the 1993 *Guidelines*, which had separate definitions for them. Yet chart B.1 in the 2005 edition (p. 135) has separate entries for “Candidate terms” and “Provisional terms.”

Not all technical terms have definitions in the standard. An example is “descriptive metadata” (p. 19). Librarians distinguish “descriptive cataloging” from “subject cataloging,” but “descriptive metadata” includes subject headings. Typographical errors in definitions are discussed below in the section entitled “Editorial Flaws.”

### Illustrations

The 1993 standard placed illustrations from published thesauri in an appendix because some formatting guidelines of the standard may not have been followed in these thesauri. The 2005 *Guidelines* embed small excerpts from published thesauri into the text. Many of these illustrations violate the rule that terms – except for proper names – should be lowercased. Even the original illustration provided to illustrate ambiguity (p. 13) capitalizes “Mercury (metal),” although lowercasing serves to distinguish common nouns from proper names. The metal is lowercased on page 21. The parenthetical qualifiers for the four meanings of *mercury* are not in alphabetical order and are in different sequences on pages 1 and 13.

The illustration for the section of the standard that defines *thesaurus* (5.4.4) is a single descriptor record from the DTIC [Defense Technical Information Center] *Thesaurus*, with terms in all uppercase and no RT (related term) reference in the sample record.

Some of the annotations to illustrations are incorrect. For example, the explanation of Example 185, the record for “Ferula” in *Medical Subject Headings* (MeSH), says, “This history notes (*sic*) indicates that from 1986-1990 the term was not authorized . . .” (p. 98). That is not true. In MeSH, the note

“was see under PLANTS, TOXIC 1986-90” refers to a period when there were two categories of subject headings: those authorized for use in the print index as well as online and those authorized only for use in the online database. “Ferula” was in the latter category, and users of the print index were referred to a broader term. The distinction was abolished in 1991.

The illustration of a pick list includes the term “Miscellaneous” (p. 78). This selection is a poor model for the design of a category structure. If it can be justified at all, the heading Miscellaneous can be assigned only after all other headings have been examined, but the screen shot shows only a partial array – from Manuscripts to Vessels.

Another poor illustration is given for Top Term Structure (p. 71): a top term with only a single level of narrower terms, all aligned. The essence of a top term structure is that it shows multiple levels of hierarchy.

Most serious is the misidentification of the illustration for the multilevel display. The *Guidelines* include an example produced with Synaptica software (p. 69), which can generate a top term structure, but not a multilevel display. The latter gives all levels of broader and narrower terms for each descriptor, often coded BT1, BT2; NT1, NT2. There are small excerpts from multilevel thesauri in the 2005 *Guidelines*, for example, on p. 19, an “Online thesaurus entry” from AGROVOC, and on p. 85, an excerpt from the *UNESCO Thesaurus* used to illustrate a multilingual vocabulary. Both of these illustrations feature only two levels of broader terms, and no narrower terms.

### Examples

The preceding section focuses on examples excerpted from published controlled vocabularies. In this section, I discuss examples provided by the standard’s authors, without attribution to a published vocabulary.

The standard includes poor examples of precoordinated terms, for example, United States – History – Civil War, 1861-1865 (p. 37). Jessica Milstead’s dissertation pointed out the sorting problem caused by such headings [4, pp. 174-175], which are intended to be filed on the date; without complex programming, however, computers process the words for the historical event.

Another poor example used to illustrate precoordination is “Searching, Bibliographic” as this flouts the guideline to enter terms in natural language order (section 7.8). The justification for the inverted order – “Precoordination is often used to ensure logical sorting of related expressions” (p. 8) – brings to mind Charles Ammi Cutter’s murky exception to the rule for direct order: “inverting the phrase only when some other word is decidedly more significant . . .” [5, (rule) 76, p. 42]. This exception led to numerous classified arrays in Library of Congress Subject Headings. Little by little, these inverted terms are being flipped to direct order.

The first example of node labels has incorrect indention (p. 61); the subsequent page gets it right.

### Legalese

Throughout the standard, the words *may*, *should* and *must* appear in boldface italic type, which I find jarring. This boldfacing recurs in the Summary of Standard Requirements, in Appendix A. (I’m not sure a 31-page section – more than a third of the body of the work – merits the title “Summary,” but this appendix is useful.)

The legalese is inconsistent. In some places, the standard says that semantic relationships of thesauri *should* be reciprocated, for example, on pp. 9 and 166; in others, for example, in section 8.1.1 and the definition of *reciprocity* (p. 8 and p. 164), that they *must* be. I favor the latter requirement.

### Thesaural Relationships

The 2005 standard has less information on thesaural relationships than did its predecessor. The current *Guidelines* mention the possibility of developing more refined codes for the associative (RT) relationship (section 8.4.4), but omit the illustration of such codes that was appended to the prior edition (Figure A22). Given that the 2005 standard discusses semantic networks (section 10.9.3), which have *more* refined relationships than thesauri, this omission is hard to understand. Moreover, after the 1993 standard was issued, most thesaurus software vendors developed the capability to accommodate special relationship indicators.

### Vocabulary Mapping

After the 1993 *Guidelines* were published, I was invited to participate in a panel about the NISO standard, held at a conference of the American Library Association. Another panelist criticized the standard for not helping him with his problem of integrating two disparate controlled vocabularies. I replied that Z39.19-1993 was a standard for developing a monolingual thesaurus, not for merging incompatible subject heading lists.

The current *Guidelines* have sections on merging vocabularies (10.6-10.8) that essentially enumerate the problems inherent in this process. I'm not convinced that these sections belong in a standard. The appendix on interoperability is useful, except that it is replete with vogue words, for example, *satellite vocabularies* (p. 143), for the well-established *microthesauri*. My *Bulletin* article on vogue words [6] could use a sequel.

### Index

The index is not comprehensive. For example, there is one locator for the heading "spelling," but the important point in

section 11.4.2 about non-displayed spelling variations and typographical errors is not indexed. The same unindexed point occurs on page 93.

The summary is not indexed, except for its range of pages under the heading-subheading combination "standard: summary of" (p. 171). (Would you look this up?)

The index lacks continuation headings, even on left-hand pages. For example, p. 170 begins with the subheading "and metadata." One must flip back to the prior page to find the heading "interoperability." Having participated in several juries of the American Society of Indexers, I know that the lack of continuation headings in a book index causes it to be placed on the reject pile immediately, even before the structure of the index is examined.

The index has an incorrect locator for "vocabulary switching" – "44-145." That type of error can be caught by a simple reading of a draft index.

### Information Design

The only running head in the document is the standard number. The 1993 edition had running heads for chapter titles and appendix numbers, which made the standard browsable and facilitated following references such as "see Appendix D." Given the focus of the 2005 edition on web display, in which orienting the reader is so important, more attention should have been given to the information design of the standard.

Although the standard provides a guideline on the use of running heads to identify various sequences in a printed vocabulary (section 9.4.1.3), in the glossary, *running head* is defined incorrectly as "A page heading indicating the first and last entries that appear on that page" (p. 165). The correct term for that feature, which is often found in dictionaries and encyclopedias, is *guideword* [7, p. 498].

### Editorial Flaws

The sequence of bulleted items in several sections is not logical and would benefit from rearrangement. The hyperlinking of the document is incomplete; cross references to related sections and definitions are often missing. For example, there is no link between the definitions of *term* and *preferred term* even though these expressions are used interchangeably. Thorough hyperlinking would have compensated, in part, for the incomplete index.

There is an error in the summary's reference to section 9.2.5 (p. 131) for guidelines on sorting. That section deals with capitalization. Conversely, there is a correct number in the reference to section 11.3 (p. 141), but the reference is inappropriate, as it follows the statement that relationships in graphic “displays are often built dynamically (in real-time)” [by computer], while section 11.3 deals with human maintenance of vocabularies.

In most cases, the summary replicates the rules from the body of the standard and omits the examples. In some cases, the summary provides only rubrics, resulting in misleading information. For example, the summary of section 8.4.1 (p. 115) lists among the uses of Related Term references “Mutually exclusive sibling terms.” The body of the standard, however, explains that RT references should *not* be provided between such terms (section 8.4.1.2).

This American standard uses the British spelling “Acknowledgements” (p. viii). Within a single paragraph (sect. 2.2), the spellings “multi-lingual” and “multilingual” co-occur. The heading of section 6.4.2.3 hyphenates “Cross-references,” while the text does not. A gross spelling error occurs on p. 14: “rather then.” The headword “graphics display” in the glossary (p. 160) should read “graphic.”

The definitions section does not consistently italicize terms within definitions that are themselves headwords, for

example, *keywords* in the definition of *natural language* (p. 7). Quite a few terms are not typographically distinguished from the surrounding text, for example, “if the editor changes passenger cars to automobiles . . .” (p. 101). The font size of Example 131 (p. 56) is much larger than that of other examples on the same page. The sample Candidate Term Form from DTIC (p. 147) is very fuzzy; it seems to have been faxed.

### Bibliographic References

Appendix F of the *Guidelines* contains the references, arranged numerically in two sections: (1) Controlled Vocabularies Used as Examples and (2) Documents Referenced in the Standard. The opposite sequence would have been more logical. In any case, the references in the two sections are numbered consecutively, and these numbers are used in the body of the work.

The first reference from the text is “[29]” (p. 2). It would have been useful to explain the reference apparatus at the beginning of the standard. Given the lack of running heads in the document, it is not easy to locate the reference list. The second reference (*ibid.*), to the NISO technical report on indexes, follows the statement, “For indexing procedures and practices see. . . .” That report [8] deals only with indexing *principles*, not procedures and practices.

A useful bibliography follows the references. The entries are grouped in ten categories whose headings are not arranged alphabetically. Every classification needs an outline, but no list of the section headings is given at the head of the bibliography or in the table of contents.

Clearly, the bibliography will require updating – at least to cite new editions of books and standards, including RDA [9] instead of *Anglo-American Cataloguing Rules*, and the new *IFLA Guidelines for Multilingual Thesauri* [10].

In several places in the body of the standard, the reader is

told to “Check the Internet,” for example, “for current software tools” (p. 99). References to specific websites that list such tools would have been more helpful.

### Conclusion

The 2005 *Guidelines* definitely require revision – at the very least to correct errors and inconsistencies and to update the bibliography.

I think it is useful to place thesauri within the spectrum of controlled vocabularies, but nothing was gained by broadening the scope of Z39.19 from its 1993 version to the 2005 edition. The latter provides no guidelines on the development of lists of subheadings, and it provides less information on thesauri than did its predecessor.

In my view, the reason the draft revision of the (now withdrawn) NISO standard *Basic Criteria for Indexes* failed to be approved (and hence was issued as a technical report) is that it had too broad a scope – book indexes, journal indexes, database indexes and automatic indexes. The Talmudic proverb that comes to mind is “Tafasta merubah, lo tafasta,” which means – If you try to grab too much, you grab nothing at all.

I recommend that the next edition of Z39.19 reinstate the word *thesauri* in the title, while explaining how thesauri differ from other controlled vocabularies. One might say that the NISO standard for controlled vocabularies should be withdrawn and the new British Standard [11] adopted instead. That standard is very expensive, however. I was told that the (reasonably priced) 1993 NISO standard for thesauri was a best-seller, and I’m sure that the 2005 edition is often downloaded.

I perceive that interest in thesauri is increasing; therefore, we need an accurate set of guidelines for this important tool for indexing and searching. ■

### Resources Cited in the Article

- [1] Natural Information Standards Organization. (1994). *Guidelines for the construction, format, and management of monolingual thesauri* (ANSI/ NISO Z39.19-1993). Bethesda, MD: NISO Press.
- [2] National Information Standards Organization. (2005). *Guidelines for the construction, format, and management of monolingual controlled vocabularies* (ANSI/ NISO Z39.19-2005). Bethesda, MD: NISO Press. Available as a free download: [www.niso.org](http://www.niso.org)
- [3] *Longman advanced American dictionary*. (2001). Essex, England: Pearson Education Limited.
- [4] Harris, J. L. (1970). *Subject analysis: Computer implications of rigorous definition*. Metuchen, NJ: Scarecrow Press.
- [5] Cutter, C.A. (1876). *Rules for a printed dictionary catalogue*. In *Public libraries in the United States of America: Their history, condition, and management, Part 2*. Washington: Government Printing Office.
- [6] Weinberg, B. H. (April/May 1990). Vogue words in information science. *Bulletin of the American Society for Information Science*, 16(4), 15.
- [7] Wellisch, H.H. (1996). *Indexing from A to Z* (2nd ed.). New York: H.W. Wilson.
- [8] Anderson, J. D. (1997) *Guidelines for indexes and related information retrieval devices* (NISO TR02-1997). Bethesda, MD: NISO Press. Available as a free download: [www.niso.org](http://www.niso.org)
- [9] Joint Steering Committee for Development of RDA. (2009). *RDA: Resource description and access: Prospectus*. Retrieved August 23, 2009, from [www.collectionscanada.ca/jsc/rdaprospectus.html](http://www.collectionscanada.ca/jsc/rdaprospectus.html)
- [10] IFLA Classification and Indexing Section. Working Group on Guidelines for Multilingual Thesauri. (2009). *Guidelines for multilingual thesauri* (IFLA Professional Reports, no. 115). The Hague: International Federation of Library Associations and Institutions. Available as a free download: <http://archive.ifla.org/VII/s29/pubs/Profrep115.pdf>
- [11] British Standards Institution. (2007). *Structured vocabularies for information retrieval* (BS 8723, parts 1-5). London: BSI.