Indexing languages

6.2.4. Thesaurus development

Overview

The developers of subject authority files are greatly concerned with the completeness and accuracy of controlled vocabulary. They make careful decisions about content coverage (domain and scope), indexing criteria (specificity and exhaustivity), and other factors. This module explains the major decisions. It is geared toward thesaurus development for purposes of the Information Organization Project, but the same principles apply to subject headings lists and to controlled vocabularies intended for physical description fields.

Why develop a thesaurus?

If you love to play with words, thesaurus development is really fun! Professionally speaking, a thesaurus makes it possible to . . .

- adequately describe intellectual content of documents at an appropriate conceptual level
- provide vocabulary control; lead indexers and searchers to authorized terms
- enhance search capabilities and improve retrieval effectiveness (precision, recall)

Thesaurus developers must consider:

- Whether to modify an existing thesaurus or develop a new thesaurus
- Literary warrant (concepts/terms in documents) and user warrant (concepts/terms in users’ queries, based on their domain knowledge, language skills, etc.)
- Content coverage: domain and scope
- Indexing criteria: specificity and exhaustivity
- Thesaurus maintenance

Development of a good thesaurus requires a major intellectual effort as well as clerical operations like data entry and production of sorted lists. The initial decision is whether to modify an existing thesaurus or develop a new thesaurus.
It is better to plan for a larger and more comprehensive controlled-vocabulary system than a smaller system that rapidly becomes inadequate as collection grows. The scope and complexity of the subject should help predict the potential scope and complexity of the thesaurus.

Thesaurus maintenance requires continuing monitoring of and adaptation to changes in language based on literary warrant and user warrant. The vocabulary system should allow for additions, deletions, and other changes in content.

---

**Content coverage: domain and scope**

Domain and scope are decisions that define and limit the content of the thesaurus. Thesaurus terms are assigned to just one field in the record structure.

**Domain**

Domain is the overall subject, topic, discipline, or theme represented in the controlled vocabulary. Domain addresses the question:

- What is the full range of concepts to be represented in this field?

The domain of the thesaurus is the same as the domain of the collection. Hint: The title of your collection should indicate to you your domain.

**Scope**

Scope is the extent or limitations of the domain. Scope addresses the question:

- What kinds of concepts or terms are excluded from the vocabulary? Is the Thesaurus just the broadest terms in the domain?

Often there are no special constraints, so the domain and scope are the same. Types of limitations vary.

**Examples**

- Domain is middle-school subjects such as math.
  Scope is subjects designated by the school district.

- Domain is medical information for families.
  Scope is common, not medical, terminology.
Indexing criteria: specificity and exhaustivity

Specificity and exhaustivity are decisions that determine which terms are assigned to the record field. Decisions to make specificity and exhaustivity high or low are based on realistic understandings of the nature of the concepts and terms; they are not judgments of thesaurus quality.

Specificity

Specificity is the extent to which index terms precisely or exactly represent the subject of a document or query. Specificity depends in part on the concreteness and complexity of terms in the subject area.

Specificity is a criterion for indexing language: the level of specificity is determined at the time the controlled vocabulary is developed. Specificity addresses these questions:

- To what extent are the terms capable of precisely representing subject concepts?
  
  If the terms represent subjects accurately and reliably (are not ambiguous), specificity is high. Specificity tends to be high for concrete nouns (e.g., "school") and for subjects that have few synonyms. Specificity tends to be highest in disciplines with distinct vocabularies that are not based on common words (e.g., law, medicine—not LIS!).

  If it is difficult to identify terms that exactly, accurately, and consistently represent the subjects, specificity is low. Specificity tends to be low for subjects that are abstract (e.g., "love"), difficult to identify and define, or have many equivalent and related terms.

- Do terms in the indexing language match the precise concepts needing representation?
  
  If the terms represent a great deal of detail, specificity is high. This may be the case for technical or professional content.

  If the terms are mostly general or broad, specificity is low. This level may be appropriate for content intended for nonprofessionals and children.

Here's a subtle but important point: indexing specificity pertains to the ability of terms to precisely represent the concepts. The terms themselves can be broad or narrow; that is, the level of specificity may or may not be related to term hierarchies in the thesaurus. In other words, terms can be precise (high specificity) or imprecise (low specificity) at any level of a hierarchy. Although narrower terms may be added to a hierarchy to increase specificity, high specificity does not require extensive hierarchical relationships in a thesaurus.

Exhaustivity

Exhaustivity is the extent to which indexing represents all subjects in a document. It is a decision that determines how many terms an indexer assigns to represent a document.

Exhaustivity is a criterion for indexing practice: the degree of exhaustivity is determined at the time the index terms are entered in the record.
Exhaustivity involves degrees of coverage on a continuum with these extremes:

- **Depth indexing** covers all the main topics plus subtopics.
- **Summarization** covers only the dominant topic of a document.

Exhaustivity addresses these questions:

- To what extent should the range of subject(s) be represented by the indexer?
- If a document has many subjects, how many should be represented? Should broad terms, narrow terms, or both be entered?
- Is the goal to retrieve the information (depth indexing) or to retrieve documents that can provide the information (summarization)?

If the choice is **depth indexing**, the level of exhaustivity is high. This usually means that many terms are assigned to each record; the data input rules may say to enter all the terms that apply. In this case, both broad and narrow terms in a hierarchy may be assigned to provide a range of possible access points to match users’ query terms. The goal of searching is to retrieve the information.

If indexing is geared toward **summarization**, the level of exhaustivity is low. This usually means that one or few terms are assigned to each record; the data input rules may limit the number of terms to enter. The goal of searching is to retrieve the document, based on the assumption that a few terms adequately describe the content or that the document provides its own access points, as in a back-of-the-book index.

Another subtle but important point: Summarization does not mean that a term is chosen only from the top level of a hierarchy. Rather, the term should be chosen from an appropriate level. A narrow term may best summarize the subject of a whole document. The idea of assigning the most specific term available in the vocabulary is known as the **principle of specific entry**. It is often invoked in subject cataloging, where subject headings describe the topic of the whole document, not subtopics.

**Examples**

Specificity is high if the vocabulary includes detailed math topics (e.g., set operations). Exhaustivity is high if all math operations in a particular textbook are indexed.

Specificity is low if the vocabulary contains terms for abstract, hard-to-identify concepts such as love and kinds of love (romantic, parental, spiritual). Exhaustivity is low if only one term, such as "parental love," is assigned.

**Project Alert!**

- You must state the domain and scope of your thesaurus and describe your decisions on specificity and exhaustivity.
- **Specificity**: Remember that the terms can be precise representations (high specificity) even if they are broad concepts in term hierarchies.
- **Exhaustivity**: If you state that the approach is depth indexing (high exhaustivity), your instructor understands that you cannot demonstrate this in your records because of the limited number of terms in the sample thesaurus.
Final thoughts

Content coverage and indexing criteria decisions for thesaurus development are critical to the success of searching and information retrieval. Think about these factors before moving on to the thesaurus tutorial, where you actually create the semantic and syndetic structure of the sample thesaurus.

Small changes in indexing can have surprisingly large effects on IR performance. See the module on how indexing languages affect retrieval, which explains the relationship of indexing specificity and exhaustivity to the IR performance measures precision and recall.