Thesaurus Creation and Indexing Compared

Indexing Society of Canada Annual Conference
Montreal, June 3, 2017

Presented by
Heather Hedden
About Heather Hedden

- Senior vocabulary editor, Gale/Cengage, 1996-2004, 2014-
- Author of *The Accidental Taxonomist* (2010, 2016)
- Online course instructor “Taxonomies & Controlled Vocabularies” (formerly through Simmons College School of Library and Information Science)
- American Society for Indexing board member, 2016 – present
- American Society for Indexing Taxonomies & Controlled Vocabularies SIG founder, past manager
- SLA Taxonomy Division past chair of Mentoring Committee and Membership Committee.
- NISO Bibliographic Roadmap working group member
- Previously: indexer, taxonomy consultant
Outline

- Introduction
  - Book indexing vs. Database indexing
  - Thesauri
- Book Index vs. Thesaurus Design
  - Terminology Comparison
  - Index and Thesaurus Points of Comparison
  - Activity Comparison
Three related functional/skill areas:

1. **Back-of-the-book indexing**
   - Identifying the concepts and names mentioned in the book and organizing them into an index

2. **Periodical/database indexing**
   - Identifying the main ideas of an article or content item and assigning the most appropriate index terms available from a controlled vocabulary

3. **Controlled vocabulary (thesaurus) creation**
   - Creating and editing a structured list of terms used for database indexing and for supporting end-user retrieval
Back-of-the-book indexing

- Also called “closed indexing”: the index is created for the single work, for contained (closed) content, then is done (closed)

- Indexing subsequent editions may involve referring to previous edition’s index, but usually are indexed from scratch again

- Embedded indexing (linking to text location in the electronic file) may enable index reuse and revision in subsequent editions (although this usually is not the case)
Database indexing

- Also called “open indexing”: indexing is an ongoing process as additional periodical issues or content is added, and the index is used yet never “finished” (open)

- A controlled vocabulary is necessary to provide consistent indexing to the same concepts from different sources indexed by different indexers over time.

- Originally was mostly for periodical articles. Now for any content in a content management system of digital asset management system: HTML files, PDFs, PPTs, brochures and ads, test questions and learning activities, images, audio, video, etc.
Introduction: Two Kinds of Indexing

1. Similarities between the two kinds of indexing
   - Read/examine and analyze content for what the main concepts are
   - Consider different ways the concepts might be named
   - Consider to how much detail to index

2. Differences between the two kinds of indexing
   - Tasks:
     - Back-of-the-book indexing requires the indexer to additionally come up with (invent) all of the index terms and their variants and arrange them into an index
     - Database indexers merely utilize the existing controlled vocabulary (and may suggest terms subject to approval)
   - Indexers:
     - A book is indexed by a single indexer
     - Multiple indexers share working on a database indexing product
Introduction: Book Indexes

Back-of-the-book index excerpt example

Locators (page numbers)
- Single locators
- Multiple locators
- Range locators

Indented subentries

See also cross-references

See cross-references
Introduction: Periodical Indexes

Print periodical index excerpt example

Reference locators (as citations)

Single reference

Multiple references

Subdivisions (like subentries)

See also cross-references

See cross-references
Online periodical index excerpt example

Subdivisions

See also cross-references

See cross-references
A controlled vocabulary is:

- An authoritative, restricted list of terms (words or phrases) mainly used for indexing/tagging content to support content management and retrieval.
- Controlled in who, when, and how new terms may be added.
- Each term stands for an unambiguous concept.

Supports consistent indexing

- When there are multiple indexers
- When there are multiple documents to be indexed over time.

Different types of controlled vocabularies with different features

- Variants/synonyms that redirect to the preferred term name
- Relationships between terms
- Notes, definitions, attributes attached to individual terms
A thesaurus is a kind of controlled vocabulary or taxonomy. Thesauri have certain inter-term relationship types:

1. Equivalence (use/used from nonpreferred terms or synonyms; USE/UF)
2. Hierarchical (broader term/narrower term; BT/NT)
3. Associative (related terms; RT)

Thesauri are described in:

- ANSI/NISO Z.39.19 guidelines
- ISO 25964 Part 1
Introduction: Thesauri

Thesaurus excerpt example

Alphabetical browse:

- Corporate trust services (Subjects)
- Corporate turnarounds (Subjects) (NPT)
- Corporate videos (Subjects) (NPT)
- Corporate welfare (Subjects)
- Corporate wellness programs (Subjects) (NPT)
- Corporation directors (Subjects) (NPT)
- Corporation executives (Subjects) (NPT)
- Corporation law (Subjects)
- Corporation reports (Subjects) (NPT)
- Corporation secretaries (Subjects)
- Corporations (Subjects)
- Corporatism (Subjects) (NPT)
- Corporative state (Subjects) (NPT)
- Corporativism (Subjects) (NPT)

Selected term details:

- Descriptor: Corporation law
  - Relationships:
    - UF Company law (Subjects)
    - UF Corporate law (Subjects)
    - BT Commercial law (Subjects)
    - NT Antitrust law (Subjects)
    - NT Business judgment rule (Subjects)
    - NT Disregarding corporate entity (Subjects)
    - NT Incorporation (Subjects)
    - NT Railroad law (Subjects)
    - RT Articles of incorporation (Subjects)
    - RT Business enterprises (Subjects)
    - RT Business trusts (Law) (Subjects)
    - RT Bylaws (Subjects)
    - RT Corporate counsel (Subjects)
    - RT Corporate domicile (Subjects)
Introduction: Thesauri

Thesaurus excerpt example

Hierarchical view excerpt
Introduction: Book Indexes and Thesauri

Three related functional/skill areas

- Indexing
- Database Indexing
- Thesaurus Creation

Shared activity of term creation and organization
Background: Terminology Comparison

Concepts

- Book index: entries (main entries and subentries)
- Thesaurus: terms

Connections between concepts (entries or terms)

- Book index: cross-references
- Thesaurus: relationships

Connection/link to content:

- Book index: locators (page numbers)
- Thesaurus: references or links
Points of Comparison

1. Concept style
2. Hierarchical structure
3. Multiple points of entry
4. Indication of related concepts
Concept Style

Similarities: Book index main entries and thesaurus terms
- Nouns or noun phrases
- Names or generic concepts
- Countable nouns in the plural
- Concise (for easy scanning), yet clear and unambiguous
- Capitalization style varies, set by the publisher

Differences: Book index *subentries* and thesaurus terms
- Subentries can additionally be prepositional phrases, adjectives, etc.
- Subentry meaning is always with respect to main entry and can be ambiguous in the index as a whole.
- Subentries are usually lower case.
Hierarchical Structure

Same goal:
- To guide the users to more precise topics

Same approach:
- If a term has (or is likely to have) too many locators/references, it needs to be broken out by creating multiple corresponding subordinate entries/terms
- Locators/linked content at subentries/narrower terms only, or at both the subentries/narrower terms and at the corresponding main entry/broader term, depending on the overall index/thesaurus editorial policy.
Hierarchical Structure Comparison: Differences

<table>
<thead>
<tr>
<th>Book Indexes: Subentries</th>
<th>Thesauri: Narrower Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subdivisions</td>
<td></td>
</tr>
<tr>
<td>1. Specific aspects of the main entry</td>
<td>1. Specific kinds or members of a class</td>
</tr>
<tr>
<td>2. Any additional concept in combination with the main entry</td>
<td>2. Named instances of a generic term</td>
</tr>
<tr>
<td></td>
<td>3. Parts of a whole</td>
</tr>
<tr>
<td>Must be related to main entry</td>
<td>Can and should stand on their own as terms</td>
</tr>
<tr>
<td>Can be prepositional phrases, gerunds, adjectives, etc.</td>
<td>Must be nouns or noun-phrases, just like main heading terms</td>
</tr>
<tr>
<td>“Flips” of main entry/subentry may have same meaning</td>
<td>Broader terms and narrower terms cannot be “flipped”</td>
</tr>
<tr>
<td>Hierarchy usually 2 levels, sometimes 3</td>
<td>Hierarchy is usually 3-4 levels, often more</td>
</tr>
<tr>
<td>Indicated by indentation or run-in following colon and semicolons</td>
<td>Indicated by reciprocal hierarchical relationships of broader term/narrower term (BT/NT); often displayed by indentation</td>
</tr>
<tr>
<td>Narrower concepts may be subentries or other main entries. No hierarchy among main entries.</td>
<td>Narrower concepts <em>must</em> be assigned NT relationships.</td>
</tr>
</tbody>
</table>
Hierarchical Structure Comparison: Examples

Book Index

Egypt
- Arab League and, 101
- Gaza Strip rule, 86
- Mamluk rule, 78
- peace with Israel, 100
- politics, 86
- Six Day War, 89–92
- Suez Crisis, 88

Thesaurus

Egypt
- NT: Alexandria
- NT: Cairo

Alexandria
- BT: Egypt
Hierarchical Structure Comparison: Examples (continued)

Book Index

Islam

- holidays in, 61, 63–64
- jihad, 51–52
- Muhammad and spread of, 46–47
- on nonbelievers, 39–40
- origins of, 43–46
- overview, 41–42
- principals, 53–54

Thesaurus

Islam

- NT: Shiite Islam
- NT: Sunni Islam

Shiite Islam

- BT: Islam

Sunni Islam

- BT: Islam
<table>
<thead>
<tr>
<th>Book Index</th>
<th>Thesaurus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flipping of main entry and subentry</td>
<td>[Not done in thesauri]</td>
</tr>
<tr>
<td>light, 111, 114</td>
<td></td>
</tr>
<tr>
<td>colors of, 62</td>
<td></td>
</tr>
<tr>
<td>color, 58–63</td>
<td></td>
</tr>
<tr>
<td>of light, 62</td>
<td></td>
</tr>
</tbody>
</table>
Multiple Points of Entry

Same goal:
- To direct various users, who use various terms that mean the same thing, to the same content location

Same approach:
- Utilizes synonyms, near synonyms, sometimes antonyms (e.g. behavior/misbehavior), slang or jargon, abbreviations or acronyms and spelled out forms, former and current names, pseudonyms, phrase variations and inversions, etc.
<table>
<thead>
<tr>
<th><strong>Book Indexes</strong></th>
<th><strong>Thesauri</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Two different methods:</td>
<td>One method only:</td>
</tr>
<tr>
<td>1. Double-posts</td>
<td>(Nothing like double-posts)</td>
</tr>
<tr>
<td>Both or all of equivalent-meaning entry terms have equal standing</td>
<td>Nonpreferred terms / Equivalency relationship: Use</td>
</tr>
<tr>
<td>2. <em>See</em> references</td>
<td>- Points the user from an entry term <em>not</em> used in the index to one that <em>is</em> used in the index</td>
</tr>
<tr>
<td>- Points the user from an entry term <em>not</em> used in the index to one that <em>is</em> used in the index</td>
<td>Thesaurus editor decisions:</td>
</tr>
<tr>
<td>Indexer decisions:</td>
<td>- In all cases, what the preferred term will be</td>
</tr>
<tr>
<td>- When to create double-posts versus <em>See</em> references (usually based on presence of subentries)</td>
<td>Equivalency relationships are bi-directional and reciprocal: Use and Used from (USE/UF)</td>
</tr>
<tr>
<td>- If using a <em>See</em> reference, then what the preferred term will be</td>
<td></td>
</tr>
</tbody>
</table>

*See* reference are one-directional: *See* (no corresponding “*Seen from*”)
**Book Index**

With double posts:

- computers in typography, 99–100, 145–146, 181
- digital typography, 99–100, 145–146, 181
- typography, digital, 99–100, 145–146, 181

**Thesaurus**

- Computers in typography
  - USE Digital typography
- Digital typography
  - UF Computers in typography
  - UF Typography, digital
- Typography, digital
  - USE digital typography
Multiple Points of Entry Comparison: Examples (continued)

Book Index

With See references:

AIGA. see American Institute of Graphic Arts

American Institute of Graphic Arts
awards, 6, 55–56, 63, 96, 100
founding of, 38
Nash, Ray, involvement in, 96
publications, 56
SP meetings with, 8

Thesaurus

AIGA
USE American Institute of Graphic Arts

American Institute of Graphic Arts
UF AIGA
Related Concepts

Same goal:
- To make the users aware of related topics of possible interest

Same approach:
- Related terms may be indicated anywhere within the index or thesaurus.
- It is somewhat subjective and takes experience to know when best to create them.
- Should be created consistently (not randomly, sporadically), but not excessively.
- Multiple See also or Related Terms at the same entry or term are OK.
## Book Indexes: *See also*

<table>
<thead>
<tr>
<th>See also is often two-way, indicated at both pairs of terms, but not necessarily always</th>
<th>Thesauri: Related Term (RT)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RT is always bi-directional reciprocal, indicated at both pairs of terms</td>
<td></td>
</tr>
</tbody>
</table>

| Not needed between entries that lie next to or near each other alphabetically, e.g. Engineers and Engineering. | Do not assume an alphabetical view is used. So, should be considered between terms that lie next to each other alphabetically |
|-------------------------------------------------------------------------------------------------------------------|

| If pointing to a subentry, the corresponding main entry needs to be named. *See also under* [main entry] | May point to terms at any level in the hierarchy without distinction |
|-------------------------------------------------------------------------------------------------------------------|

| May refer to a group of terms at once: *See also specific*… [class of terms] | Must refer to an individual term only |
|-------------------------------------------------------------------------------------------------------------------|
## Related Concepts Comparison: Examples

<table>
<thead>
<tr>
<th><strong>Book Index</strong></th>
<th><strong>Thesaurus</strong></th>
</tr>
</thead>
</table>
| legendary figures, 12–15. *see also* tall tales | Legendary figures  
RT: Tall tales |
| tall tales, 15–17. *see also* legendary figures | Tall tales  
RT: Legendary figures |

**Can be uni-directional in an index:**

Church of Jesus Christ of Latter-day Saints,  
93–95. *see also* Mormons

Mormons, 51, 64, 86, 93–95

**Always bi-directional in a thesaurus:**

Church of Jesus Christ of Latter-day Saints  
RT: Mormons

Mormons  
RT: Church of Jesus Christ of Latter-day Saints
### Book Index

- **Multiple OK (separated by semicolon):**
  - medications. *see also* drug therapy; side effects
    - combinations of, 10, 18
    - developments in, 196–199
    - targeted therapies, 38, 196–198, 201

### Thesaurus

- **Multiple OK:**
  - Medications
    - RT: Drug therapy
    - RT: Side effects

### See also for any term:

- Louisiana, 94. *see also* New Orleans

### Treated as narrower, not related terms:

- Louisiana
  - NT: New Orleans
Activity of Indexing vs. Thesaurus Creation

1. Similarities
   - Both do not require subject expertise, even less so for book indexing, except in technical subject areas

2. Differences
   - Indexing involves:
     - Greater specific content analysis
   - Thesaurus creation involves:
     - More broad-based analysis
     - More consideration of audience/users
     - Researching additional outside sources
Indexing may be either process:

1. Read and index page-by-page from the beginning

2. First skim the book and write down common themes and names, as likely index terms, then go back and begin indexing.

Thesaurus creation is more like the latter, without the second, indexing phase.
Heather Hedden
Senior Vocabulary Editor
Gale, A Cengage Learning Company
20 Channel Center St., Boston, MA 02210 USA
Heather.Hedden@cengage.com

heather@hedden.net
978-467-5195 (mobile)
www.hedden-information.com
http://accidental-taxonomist.blogspot.com