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 assent 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

Back-of-the-book index excerpt example

Locators (page numbers)

Single locators

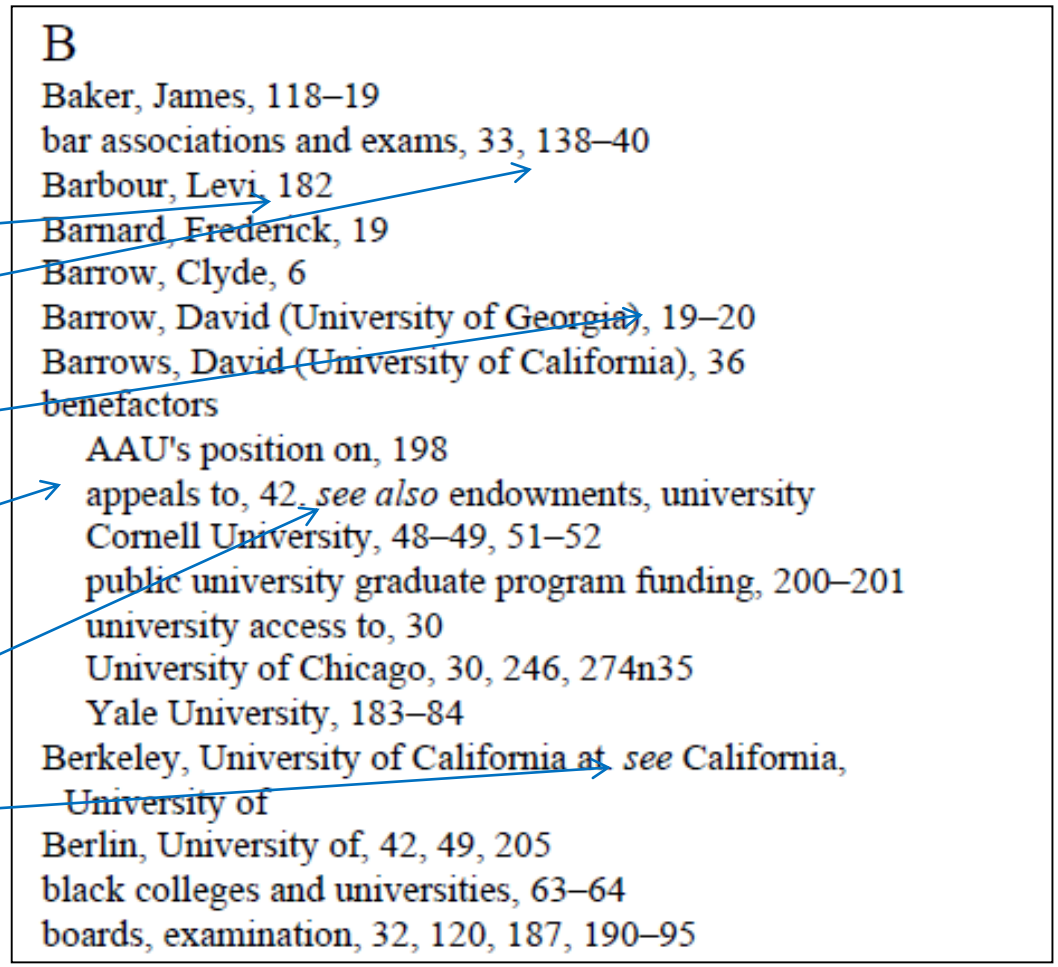
Multiple locators

Range locators

Indented subentries

See also cross-references

See cross-references



B
Baker, James, 118–19
bar associations and exams, 33, 138–40
Barbour, Levi, 182
Barnard, Frederick, 19
Barrow, Clyde, 6
Barrow, David (University of Georgia), 19–20
Barrows, David (University of California), 36
benefactors
AAU's position on, 198
appeals to, 42, *see also* endowments, university
Cornell University, 48–49, 51–52
public university graduate program funding, 200–201
university access to, 30
University of Chicago, 30, 246, 274n35
Yale University, 183–84
Berkeley, University of California at, *see* California, University of
Berlin, University of, 42, 49, 205
black colleges and universities, 63–64
boards, examination, 32, 120, 187, 190–95

Introduction: Periodical Indexes

Print periodical index excerpt example

Reference locators (as citations)

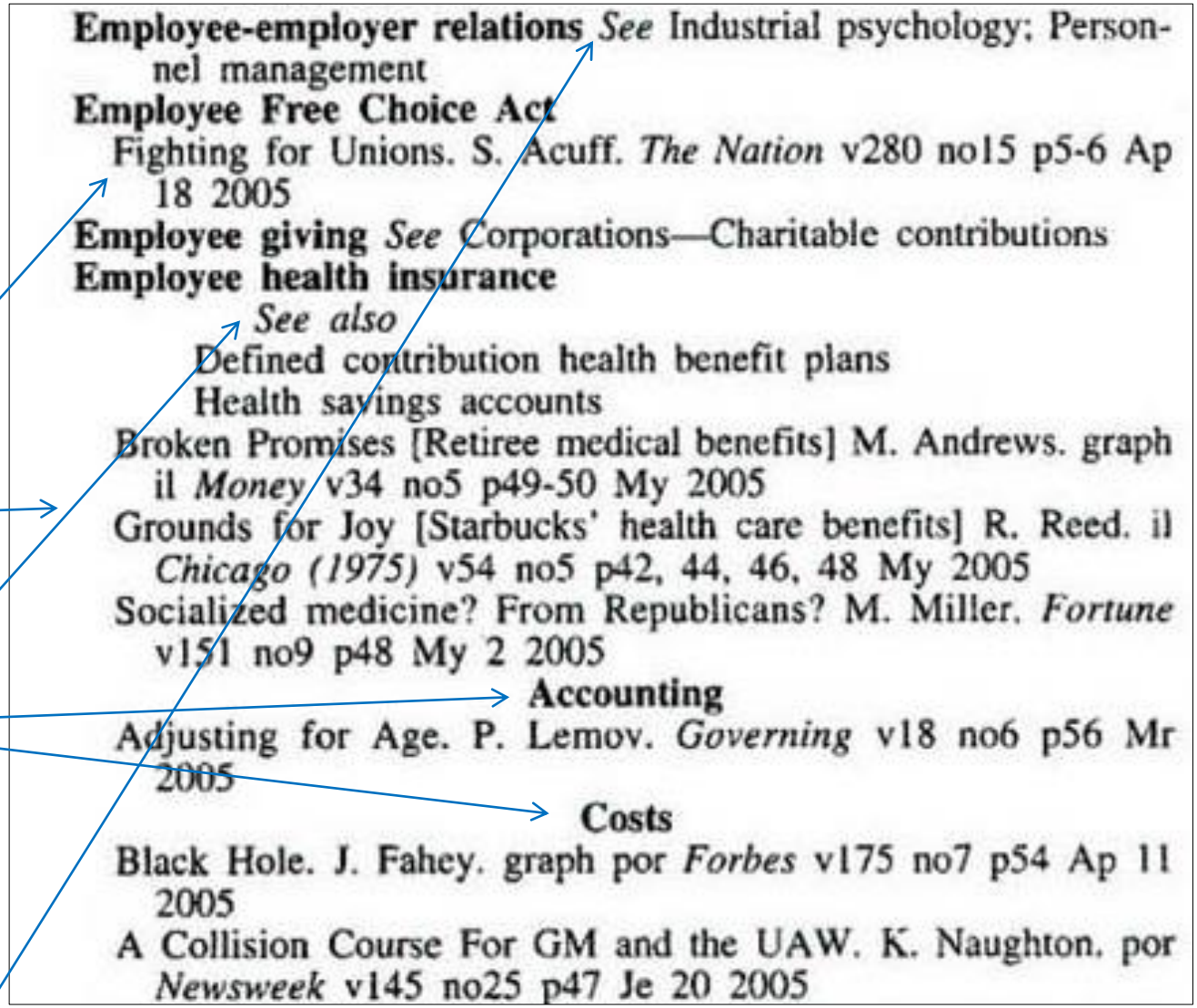
Single reference

Multiple references

Subdivisions (like subentries)

See also cross-references

See cross-references



Employee-employer relations *See* Industrial psychology; Personnel management
Employee Free Choice Act
Fighting for Unions. S. Acuff. *The Nation* v280 no15 p5-6 Ap 18 2005
Employee giving *See* Corporations—Charitable contributions
Employee health insurance
See also
Defined contribution health benefit plans
Health savings accounts
Broken Promises [Retiree medical benefits] M. Andrews. graph il *Money* v34 no5 p49-50 My 2005
Grounds for Joy [Starbucks' health care benefits] R. Reed. il *Chicago (1975)* v54 no5 p42, 44, 46, 48 My 2005
Socialized medicine? From Republicans? M. Miller. *Fortune* v151 no9 p48 My 2 2005
Accounting
Adjusting for Age. P. Lemov. *Governing* v18 no6 p56 Mr 2005
Costs
Black Hole. J. Fahey. graph por *Forbes* v175 no7 p54 Ap 11 2005
A Collision Course For GM and the UAW. K. Naughton. por *Newsweek* v145 no25 p47 Je 20 2005

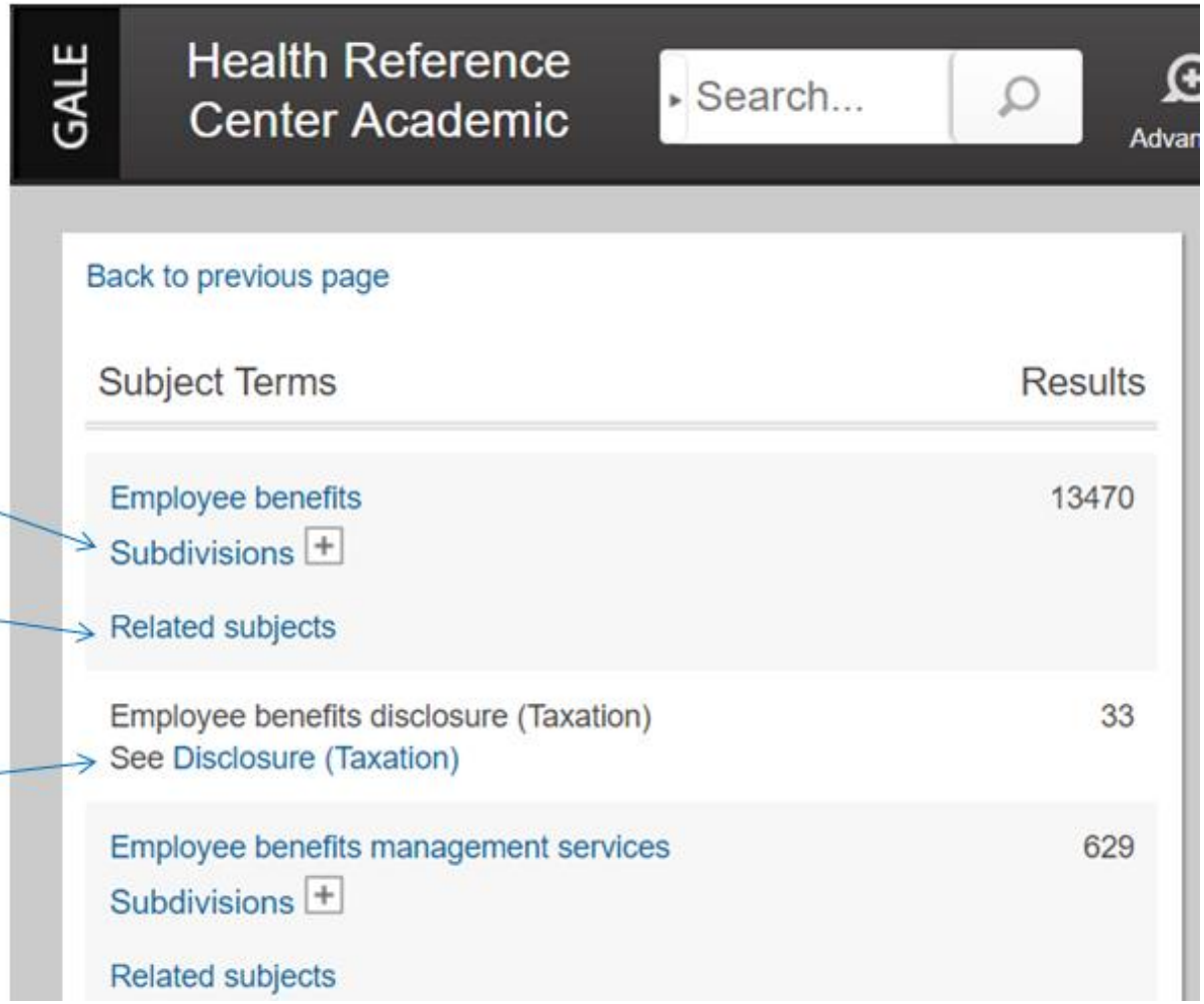
Introduction: Periodical Indexes

Online periodical index excerpt example

Subdivisions

See also cross-references

See cross-references



Introduction: Thesauri

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 te

Introduction: Thesauri

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
http://www.niso.org/standards/resources/Z39-19.html

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

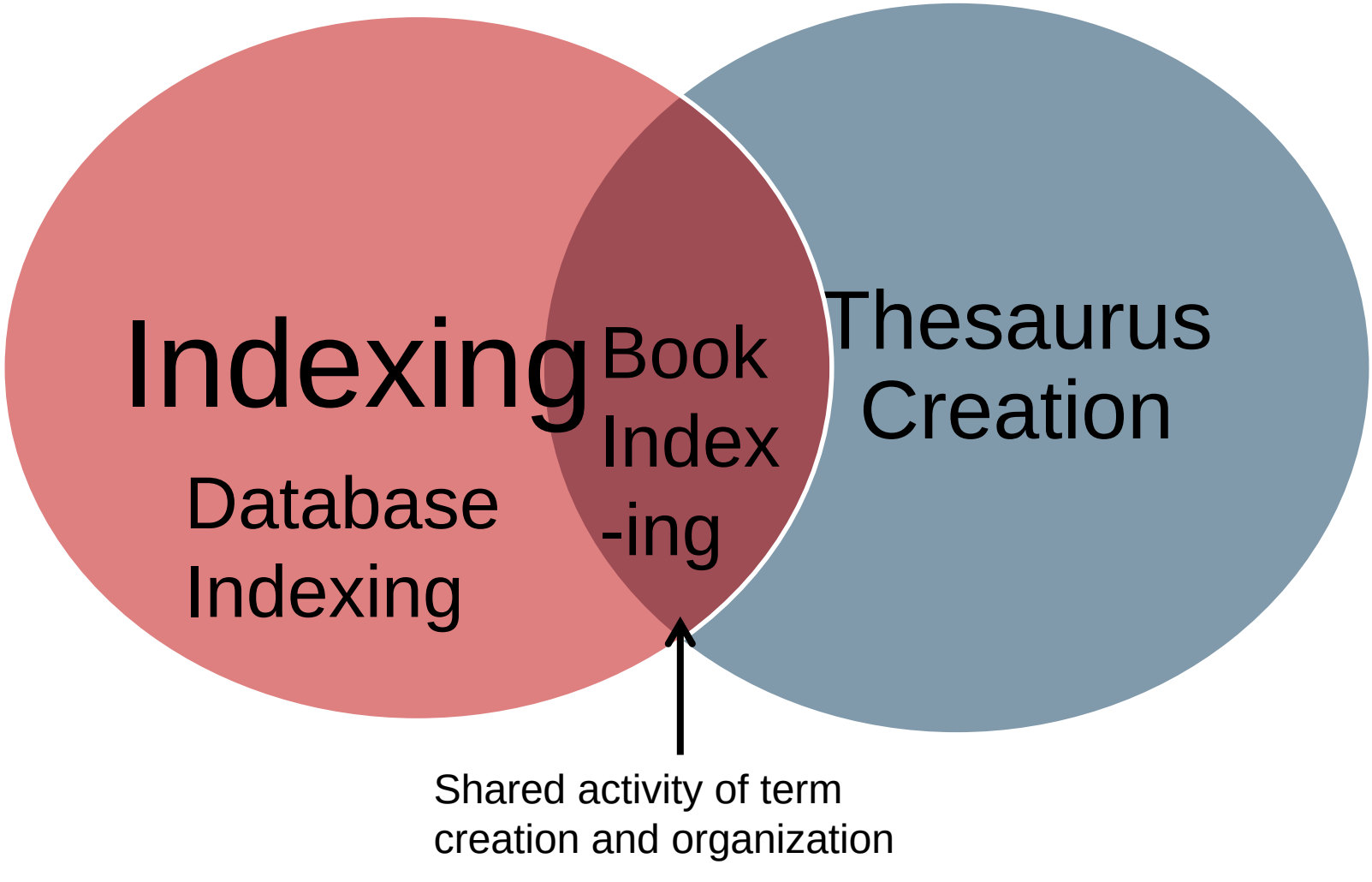
Thesaurus excerpt example

Hierarchical view excerpt

(NT1) Commercial law
- (NT2) Accounting law
- (NT2) Banking law
  - (NT3) Banking Act of 1935
  - (NT3) Disclosure (Banking law)
  - (NT3) Fair Credit Reporting Act
  - (NT3) Glass-Steagall Act
- (NT2) Bankruptcy law
- (NT2) Collection law
- (NT2) Construction law
  - (NT3) Building codes
- (NT2) Corporation law
  - (NT3) Antitrust law
    - (NT4) Antitrust law (International law)
    - (NT4) Rule of reason (Antitrust law)
    - (NT4) State action (Antitrust law)
  - (NT3) Business judgment rule
  - (NT3) Disregarding corporate entity
  - (NT3) Incorporation
    - (NT4) Articles of incorporation
  - (NT3) Railroad law
- (NT2) Economic loss doctrine
- (NT2) Food law
  - (NT3) Dairy laws
  - (NT3) Sugar laws
- (NT2) Insurance law

Introduction: Book Indexes and Thesauri

Three related functional/skill areas



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

Book Index vs. Thesaurus Design

Points of Comparison

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

Concept (Entry or Term) Style Comparison

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 Comparison: Similarities

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

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

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

Hierarchical Structure Comparison: Examples (continued)

Book Index

Flipping of main entry and subentry

light, 111, 114
  colors of, 62

color, 58–63
  of light, 62

Thesaurus

[Not done in thesauri]

Multiple Points of Entry Comparison: Similarities

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.

Multiple Points of Entry Comparison: Differences

Book Indexes	**Thesauri**
Two different methods: 1. Double-posts Both or all of equivalent-meaning entry terms have equal standing 2. *See* references - Points the user from an entry term *not* used in the index to one that *is* used in the index	One method only: (Nothing like double-posts) Nonpreferred terms / Equivalency relationship: Use - Points the user from an entry term *not* used in the thesaurus to one that *is* used in the thesaurus
Indexer decisions: - When to create double-posts versus *See* references (usually based on presence of subentries) - If using a *See* reference, then what the preferred term will be	Thesaurus editor decisions: - In all cases, what the preferred term will be
See reference are one-directional: *See* (no corresponding “Seen from”)	Equivalency relationships are bi-directional and reciprocal: Use and Used from (USE/UF)

Multiple Points of Entry Comparison: Examples

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 Comparison: Similarities

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.

Related Concepts Comparison: Differences

Book Indexes: *See also*	**Thesauri: Related Term (RT)**
See also is often two-way, indicated at both pairs of terms, but not necessarily always	RT is always bi-directional reciprocal, indicated at both pairs of terms
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

Book Index

legendary figures, 12–15. *see also* tall tales

tall tales, 15–17. *see also* legendary figures

Thesaurus

Legendary figures
RT: Tall tales

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

Related Concepts Comparison: Examples (continued)

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

Book Indexing vs. Thesaurus Creation: Activity Comparison

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.

Questions/Contact

Heather Hedden
Senior Vocabulary Editor
Gale, A Cengage Learning Company
20 Channel Center St., Boston, MA 02210 USA
www.cengage.com, www.gale.com
Heather.Hedden@cengage.com

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