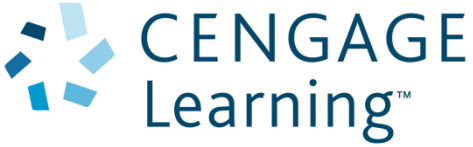
Thesaurus Creation and Indexing Compared

Heather Hedden
Senior Vocabulary Editor, Cengage Learning

American Society for Indexing Annual Conference
Seattle, Washington, April 30, 2015



About Heather Hedden

- Controlled vocabulary editor at a library database vendor, Gale/Cengage Learning, 1996 – 2004, 2014 – present
- Continuing education online workshop instructor, Simmons College Graduate School of Library and Information Science
- Author of *The Accidental Taxonomist* (Information Today, Inc.)
- Part-time freelance back-of-the-book indexer

Previously

- Taxonomy consultant
- Periodical article/reference database indexer (Information Access Company/Predicasts/Gale)

Outline

Introduction

- Book indexing and periodical/database indexing compared
- Introduction to back-of-the-book indexes and thesauri

Book indexing and thesauri creation comparison

1. Style of entries/terms
2. Hierarchical structure
3. Multiple points of entry
4. Indication of related concepts

Work comparison

- Activity and skills
- Work life

Further information on thesauri

Introduction

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*, which results in creating an index

3. Controlled vocabulary (thesaurus) creation

- Creating and editing a structured list of terms used for database indexing (and sometimes multi-volume book indexing) for supporting end-user retrieval

Back-of-the-book indexing vs. periodical/database indexing

1. Back-of-the-book indexing

- Also called "closed indexing": the index is created for the single work, 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) enables index reuse and revision in subsequent editions

2. 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

A. 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 utilize the existing controlled vocabulary (and may suggest terms subject to approval)

2. Differences between the two kinds of indexing (continued)

B. Differences in the resulting indexes

- Back-of-the-book indexing results in a fully displayed browsable alphabetical index.
- Database indexes may or may not be displayed to end-users. Maybe just portions (such as terms in a type-ahead scrollbox)

C. Differences in the indexers

- A book is indexed by a single indexer.
- Database indexing projects are shared by multiple indexers.

Introduction: Book Indexes

Back-of-the-book index excerpt example

Locators (page numbers)

Single locators

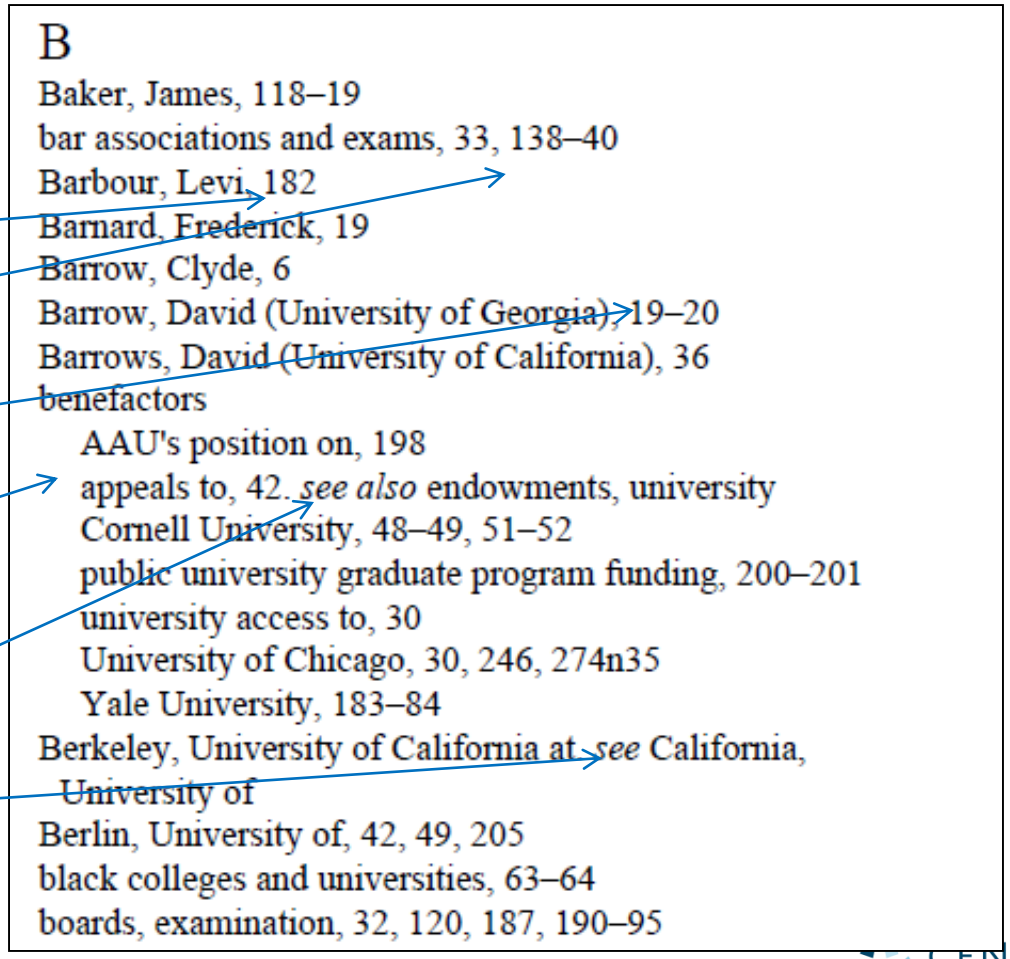
Multiple locators

Range locators

Indented subentries

See also cross-references

See cross-references



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: Thesauri

A thesaurus is a kind of controlled vocabulary or taxonomy

- Each term stands for an unambiguous concept
- There is control over the addition of terms to the vocabulary

that has the full set of 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)

As described in ANSI/NISO Z.39.19-2005 guidelines

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

Thesauri compared with “taxonomies”

Thesauri

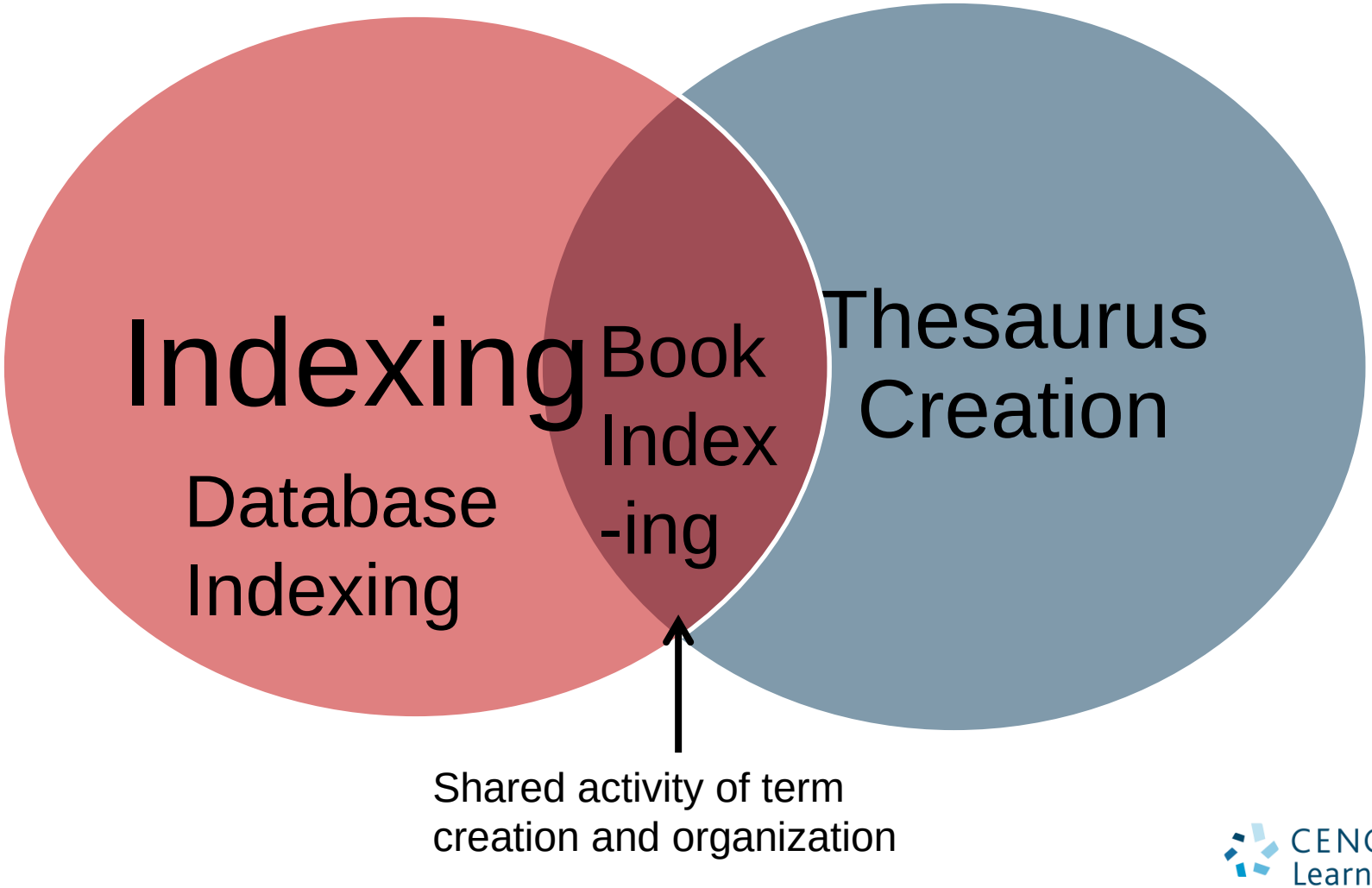
- Include all relationship types (equivalence, hierarchical, and associative/related)
- All terms have relationships, but hierarchies can have as few as 2 terms.
- ANSI/NISO rules are strictly followed.
- Supports concept scoping, disambiguation, and relationships with similar concepts.
  (Like looking up in Roget’s thesaurus.)
- Approach is term-centered and what terms are linked to/from it.
- Especially serving indexers/ indexing.

Taxonomies

- Have hierarchical relationships, but usually not related, and sometimes not even equivalence.
- All terms belong to a limited number of major hierarchies (or facets)
- May bend ANSI/NISO rules.
- Supports classification, categorization, and concept organization.
  (Like Linnaean taxonomy.)
- Approach is a top-down navigation.
- Especially serving end-users when browsing.

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. Style of entries/terms
2. Hierarchical structure
3. Multiple points of entry
4. Indication of related concepts

Entry or Term Style Comparison

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.

Entry or Term Style Comparison: Nuanced Differences

Differences: Book index main entries and thesaurus terms

Book Index

Concise entries (if not proper nouns) for easy browsing and minimal wrapping to the next line within a narrow column.

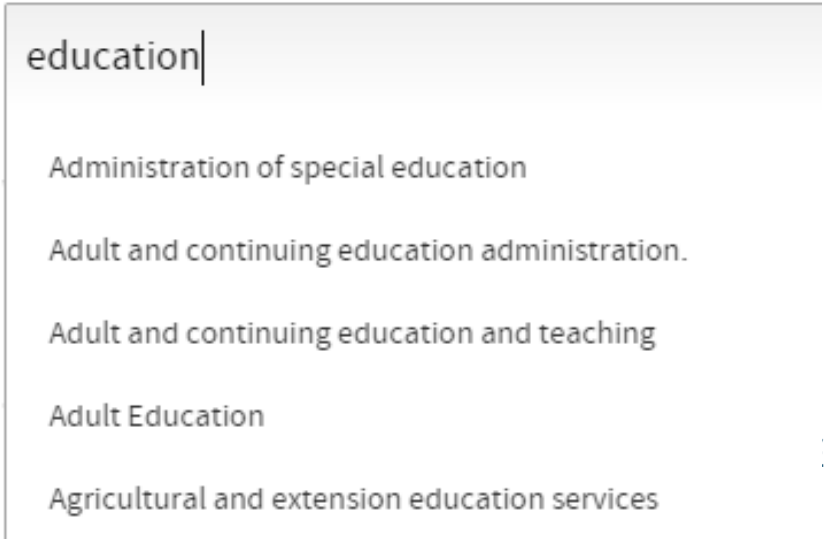
If main entry has subentries, these “refinements” enable having general one-word main entries

education
- administration
- adult and continuing
- agricultural extension

Thesaurus

Need not always be as concise (could be wide half-screen width scroll-boxes)

Without subentry “aspects” more complex, precoordinated terms are likely created



Hierarchical Structure Comparison: Similarities

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

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 (or triple) 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

Multiple Points of Entry Comparison: Nuanced Differences

Book Index

The user will skimming the printed index.

Don't create cross-references that fall close to each other alphabetically (starting with the same word or with the same first 3-4 letters).

Do *not* create:

> biological sciences. *See* biology

Create adjective-noun inversions, as double-posts or cross-references to provide a different word to start on:

> business zoning
> zoning, business

Thesaurus

The user might *search* the thesaurus instead of browsing it.

Do create nonpreferred terms that would fall close to each other (starting with the same word or with the same first 3-4 letters).

Do create:

> Biological sciences
> *Use* Biology

If the thesaurus can be searched, do *not* create inverted nonpreferred terms. Use natural language only.

> Business zoning

Related Concepts Comparison: Similarities

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

Book Indexing vs. Thesaurus Creation: Activity Comparison

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.

Book Indexing vs. Thesaurus Creation: Skills Comparison

1. Similarities

- Analytical skills
- Organization/categorization skills
- Language skills
- Attention to detail
- Attention to user needs
- Ability to work independently

2. Differences

Thesaurus construction also needs:

- Understanding of thesaurus principles and standards
- Search skills
- Stronger communication skills
- Ability to work with diverse people

Book Indexing vs. Thesaurus Creation: Work Life Comparisons

Freelance book indexing	Freelance taxonomy work
Don't need to meet the client Always work from home	Need to meet and talk with people Work partially from home, partially onsite
Purchase and use your own software	Software provided by client
Assignments are usually clearly defined	Assignments are not usually clearly defined
Usually solo work	May work as part of a team
Usually submitted only when complete	Involves review/feedback; iterative
Projects are stand-alone	Often part of a larger project

Book Indexing vs. Thesaurus Creation: Work Life Comparisons

Freelance book indexing	Freelance taxonomy work
Clients: publishers, packagers, authors; sometimes nonprofits or government (occasionally subcontracting)	Clients: large enterprises, consultancies, information science and technology staffing/recruiting firms
Often have repeat clients	Usually one-time projects
Usually pays per page	Pays per hour (as subcontractor) or per project (as direct consultant)
Often predictable type of work	Rarely predictable type of work
Always freelance	As contractor or temporary employee. Can lead to permanent employee status.
Projects last one or a couple of weeks	Projects last several months.

Further information

Book indexing

- American Society for Indexing http://www.asindexing.org/about-indexing

Thesaurus creation

- American Society for Indexing http://www.asindexing.org/about-indexing/thesauri
- ANSI/NISO Z39.19-2005 (R2010) Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies http://www.niso.org/apps/group_public/download.php/12591/z39-19-2005r2010.pdf
- "Practical Taxonomy Creation" ASI Online Learning 3-part webinar course http://www.asindexing.org/online-learning/taxonomy-hedden
- Taxonomies and Controlled Vocabularies, 5-week online course, Simmons College, School of Library and Information Science http://alanis.simmons.edu/ceweb
- Construction of Controlled Vocabularies: A Primer http://marciazeng.slis.kent.edu/Z3919/index.htm
- Thesaurus Construction tutorial by Tim Craven http://publish.uwo.ca/~craven/677/thesaur/main00.htm
- Hedden Information Management http://www.hedden-information.com/presentations.htm

Questions?

Heather Hedden
Senior Vocabulary Editor
Cengage Learning
20 Channel Center St., Boston, MA 02210
(o) 617-757-8211 | (m) 978-467-5195
(e) Heather.Hedden@cengage.com | Heather@Hedden.net
www.cengage.com

www.hedden-information.com
http://accidental-taxonomist.blogspot.com

CENGAGE Learning