Introduction to Indexes, Indexing, and Controlled Vocabularies

Heather Hedden
Senior Vocabulary Editor
Gale, A Cengage Learning Company
About Heather Hedden

- Senior vocabulary editor, Gale/Cengage, 1996-2004, 2014-present
- 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
- SLA Taxonomy Division past chair of Mentoring Committee and Membership Committee.
- NISO Bibliographic Roadmap working group member
- Previously: taxonomy consultant, indexer
About Cengage

- Publisher of educational (textbook) and reference content and software
- Divisions & brands: Gale, Questia, MindTap, Pathbrite, Learning Objects, WebAssign
- Headquartered in Boston. Offices in San Francisco, Detroit, Cincinnati, Albany, etc.
- Formerly Thomson Learning, spun off from Thomson and renamed in 2007

Gale, A Cengage Company

- Subscription databases to libraries: GVRL ebooks, In Context, Academic OneFile, Business Collection, Literature Resource Center, etc.
- Public web products: Questia, Books & Authors, HighBeam, Encyclopedia.com
- Gale Research reference books, directories, other book imprints (Greenhaven, Thorndike, St. James Press, etc.)
- Legacy companies: Information Access Company (InfoTrac), Predicasts, Primary Source Media (Artemis)
Introduction to Indexes, Indexing, and Controlled Vocabularies

- Indexes
  - Types
  - Examples
- Indexing
  - Types
  - Software
- Controlled Vocabularies
  - Uses & Benefits
  - Types
  - Software
Indexes
Indexes

To indicate, to point to

"A systematic guide designed to indicate topics or features of documents in order to facilitate retrieval of documents or parts of documents."

This definition covers:

- Books/single-document indexes and database/continuing indexes
- Displayed browsable indexes and non-displayed indexes that support search.
- Manually created indexes and automatically generated indexes based on algorithms or search engine indexes
Indexes

Indexes may be classified many ways. By…

- **Object referred to:** Subject index or Author index
- **Types of terms as headings:** Names (proper nouns), Subjects
- **Content upon which the index is based:** Full text, abstracts, citations, first lines
- **Arrangement of entries:** alphabetic, alphanumeric, classified, combination
- **Indexing method:** manual, automated, or a combination
- **Term selection:** from the text, from a controlled vocabulary, or a combination
- **Term coordination:** pre-coordinated or post-coordinated (Boolean or proximity)
- **Document type or genre:** books, periodicals, poetry, images, maps, videos
- **Medium:** print, microform, electronic/online
- **Periodicity:** one-time, closed index or continuing, open index
Indexes

One-time, Closed Indexes

- Back-of-the-book or other monograph index
- An index is created for the single work, then is done (closed).
- Index entries point to page numbers or section numbers.
- Index is fully displayed to end-users.
- Index entries are unique to that work and based on concepts in that work.
- May be for multi-volume works, if multiple volumes are due to the size of the work, and not based on periodicity/year.
- If there are subsequent editions, new editions have their own index, which may be based on/derived from a previous edition’s index, or may be completely original, but it is not the same index.
Indexes

Continuing, Open Indexes

- Periodical articles or other database content (records, reports, images, multimedia files, etc.)
- Comprise multiple documents by multiple creators
- Index entries point to complete documents, articles, image files, records
- Index may or may not be displayed to end-users
- Same, shared index entries are linked to multiple documents
- Utilizes a controlled vocabulary/thesaurus
- Continuing indexes can discontinue when periodical publication ceases or when a vast collection is completely indexed but it had been continually indexed over a long period (years).
Indexes

Back-of-the-book index excerpt example

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

Locators (page numbers)

Single locators

Multiple locators

Range locators

Subentries, indented

See also cross-references

See cross-references
Indexes

Print periodical index excerpt example

August 2005, Reader’s Guide to Periodical Literature

- Reference locators (as citations)
- Single reference
- Multiple references
- Subdivisions (like subentries)
- See also cross-references
- See cross-references
Indexes

Online periodical index excerpt example

Subdivisions

See also cross-references

See cross-references
Indexing

Indexing

- The creation of an index
- The linking of terms/headings/index entries to content/document units or to references to the content/document units.

- Based on analysis to determine meaningful concepts, not just words, with sufficient content
Indexing

One-time, Closed Indexes

- Typically a single indexer indexes a single work.
- The indexer:
  - Reads the work and identifies concepts for which there is sufficient information (such as a couple of sentences).
  - Decides how to name the concepts in the index
  - Creates variant entries for the same concept as double-posts or See references
  - Creates subentries to split/group subtopics of an entry that has many page locators
  - Decides where to enter See also cross-references

➢ In an integrated process, the indexer creates the index entries and associates them with page locators.
Indexing

Continuing, Open Indexes

- Typically involve multiple indexers, indexing in a multi-user indexing system.
- The indexers:
  - Skim the article or view the digital asset and identify key concepts, in accordance with policy guidelines on indexing specificity.
  - Browse or search the controlled vocabulary, displayed in the indexing system, to find the desired concepts.
  - Assign the controlled vocabulary terms to the document.
  - If terms cannot be found, suggesting terms or adding conditional/unapproved terms may be permitted.
  - May assign other metadata for the document.
  - Proceed to index the next document in the queue.

➢ Must follow indexing policy set for the database/product
Indexing

Software for indexing

For one-time, closed indexes

- Software to manage manual closed/book indexing
- Embedded indexing features of publishing software
- Tools to embed index tags in XML documents
- Software to semi-automate closed/book indexing

For continuing, open indexing

- Software intended for closed indexing, used on small-scale open indexing
- Add-on to controlled vocabulary management software
- Proprietary program, based on a database management system
- Automated or semi-automated indexing (auto-categorization) software
Indexing

Software to manage manual closed/book indexing

Cindex

SKY Index
Indexing

Embedded indexing feature of InDesign

Embeds index tags within the text
Software to semi-automate closed/book indexing: Textract
Software UI for open database indexing at Gale
Indexing

Automated approaches for continuing/open indexing

- **Auto-Indexing** – prominent terms extracted to create an index
  - Text analytics and text mining, based on natural language processing (NLP)
  - Information extraction, especially named entity extraction

- **Auto-Categorization/Classification** – documents assigned to categories
  - Main methods: machine-learning or rules-based (see next slide)
  - May also leverage results from text analytics, information extraction, text mining

- **Auto-Tagging** – terms assigned to documents
  - Same technologies as auto-categorization
  - May have more specific/granular taxonomy terms and additional automatically assigned metadata
Indexing

Auto-Categorization/Auto-Classification/Auto-Tagging Technologies

- **Machine-Learning/Statistical**
  - Automatically categorizes/tags based on previous examples.
  - System has complex mathematical algorithms.
  - Must supply the system with multiple representative sample documents for each taxonomy term to “train” the system. Results are reviewed and training sets are “tuned.”

- **Rules-Based**
  - Indexing rules, based on synonyms and other conditions, are created for each taxonomy term.
  - Some systems feature auto-generated suggested rules for each term/synonym which can be manually edited, in addition to writing rules from scratch.
  - Some systems feature more sophisticated rule-writing, like advanced Boolean searching (in reverse) and proximity operators or regular expressions.
Controlled Vocabularies
Controlled Vocabularies

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.
- The set of allowed values/terms for a designated descriptive metadata element/field.
- Controlled in who, when, and how new terms may be added.

Optional features:

- Variants/synonyms that redirect to the preferred term name
- Relationships between terms
- Notes, definitions, attributes attached to individual terms.
Controlled Vocabularies

Controlled Vocabulary Uses & Benefits

- Supporting consistent/accurate indexing
  - With multiple indexers and documents from multiple authors
- Supporting retrieval
  - Supporting greater precision and recall
    - Resolving synonymy - different words for the same concept, For better recall
    - Resolving polysemy - same word with different meanings, For better precision
  - Supporting guided topic browsing through classification structure for discovery
Controlled Vocabularies

Controlled Vocabulary Uses & Benefits

- Resolving *synonymy* - different words for the same concept
  - A controlled vocabulary gathers synonyms, acronyms, variant spellings, etc. – brings together that which is the same.
  - Documents not missed due to use of different words e.g. *Automobiles* vs. *Cars*

- Resolving *polysemy* - same word with different meanings
  - A search restricted on the controlled vocabulary retrieves concepts not just words – distinguishing that which is different.
  - Documents excluded for mere text-string matches e.g. *monitors* for computers, not the verb “observes”
Controlled Vocabularies

Taxonomy

- A controlled vocabulary with broader/narrower (parent/child) term relationships that include all terms to create a hierarchical structure
- Has focus on categorizing and organizing concepts
- May or may not have “synonyms” to point to the correct, preferred terms
- May comprise several hierarchies or facets

Leisure and culture

- Arts and entertainment venues
  - Museums and galleries
  - Children's activities
  - Culture and creativity
    - Architecture
    - Crafts
    - Heritage
    - Literature
    - Music
    - Performing arts
    - Visual arts
    - Entertainment and events
    - Gambling and lotteries
    - Hobbies and interests
    - Parks and gardens
    - Sports and recreation
      - Team sports
        - Cricket
        - Football
        - Rugby
        - Water sports
        - Winter sports
      - Sports and recreation facilities
      - Tourism
      - Passports and visas
      - Young people's activities
Controlled Vocabularies

Thesaurus

- A controlled vocabulary with standard structured relationships between terms
  - Hierarchical: broader term/narrower term (BT/NT)
  - Associative: related terms (RT)
  - Equivalence (synonyms): preferred term/non-preferred term (USE/UF)
- Not necessarily in neat, top-down navigation hierarchies
- Especially helpful for manual indexing (tagging)
- Created in accordance with ANSI/NISO Z30.19 or ISO 25964 standards.

Folk music
  UF Traditional music
  BT Music
  NT American folk music
  NT Canadian folk music
  NT Folk dance music
  NT Folk-rock music
  NT Folk songs
  NT Irish folk music
  NT Mariachi music
  NT Mexican son (Music)
  RT Folk festivals
  RT Folk singers
  RT Folklore
  RT Gypsy music
  RT Nueva trova (Music)
  RT Popular music
## Controlled Vocabularies

### Controlled Vocabulary Types

<table>
<thead>
<tr>
<th>Pick List</th>
<th>Synonym Ring</th>
<th>Authority File</th>
<th>Taxonomy</th>
<th>Thesaurus</th>
<th>Ontology</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Ambiguity control</td>
<td>- Synonym control</td>
<td>- Ambiguity control (Synonym control)</td>
<td>- Ambiguity control (Synonym control)</td>
<td>- Ambiguity control (Synonym control)</td>
<td>- Ambiguity control (Synonym control)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Synonym control</td>
<td>- Hierarchical relationships</td>
<td>- Hierarchical relationship</td>
<td>- Semantic relationships</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>- Associative relationships</td>
<td>- Classes</td>
</tr>
</tbody>
</table>
Controlled Vocabularies

Software for Taxonomy/Thesaurus Management

- Dedicated thesaurus management software
- Taxonomy creation & editing module of a content management, document management, digital asset management, collaborative software (SharePoint), or automated indexing software
- Vertical market software for creating classification structures
- Proprietary thesaurus management software developed in-house

➢ Screenshots from four dedicated thesaurus management software products
MultiTes