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)



Outline

Introduction to Indexes, Indexing, and Controlled Vocabularies

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





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."

-- Anderson, James D. 1997. NISO-TR02, Guidelines for indexes and related information retrieval devices. http://www.niso.org/publications/tr/tr02.pdf

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 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





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.



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).



Back-of-the-book index excerpt example

Locators (page numbers)

Single locators

Multiple locators

Range locators

Subentries, indented

See also cross-references

See cross-references

В

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



August 2005, Reader's Guide to Periodical Literature

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

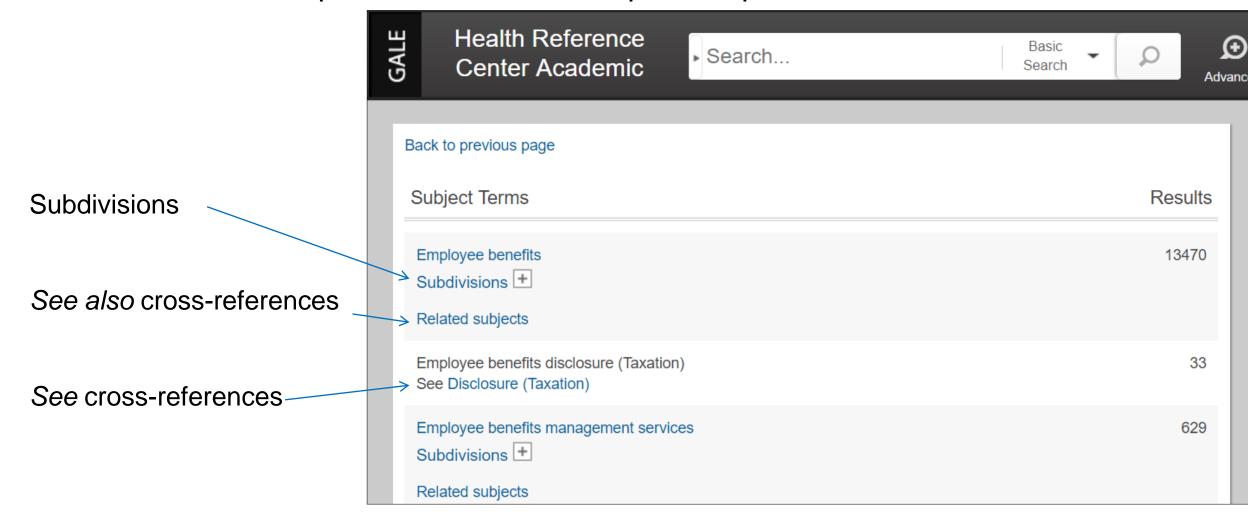
Costs

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



Online periodical index excerpt example









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



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.



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



Software for indexing

For one-time, closed indexes

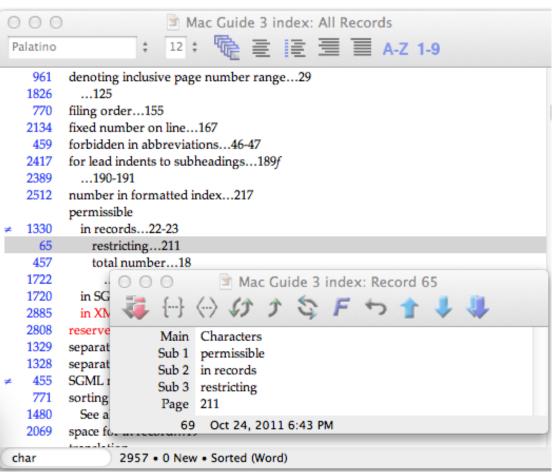
- Software to manage manual closed/book indexing
- Embedded indexing features of publishing software
- Tools to embed index tags in XML documents
- Software to semi-automate closed/book indexing

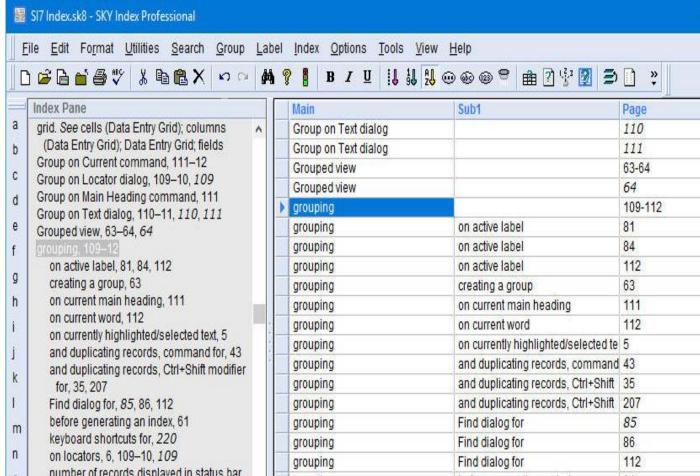
For continuing, open indexing

- 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



Software to manage manual closed/book indexing







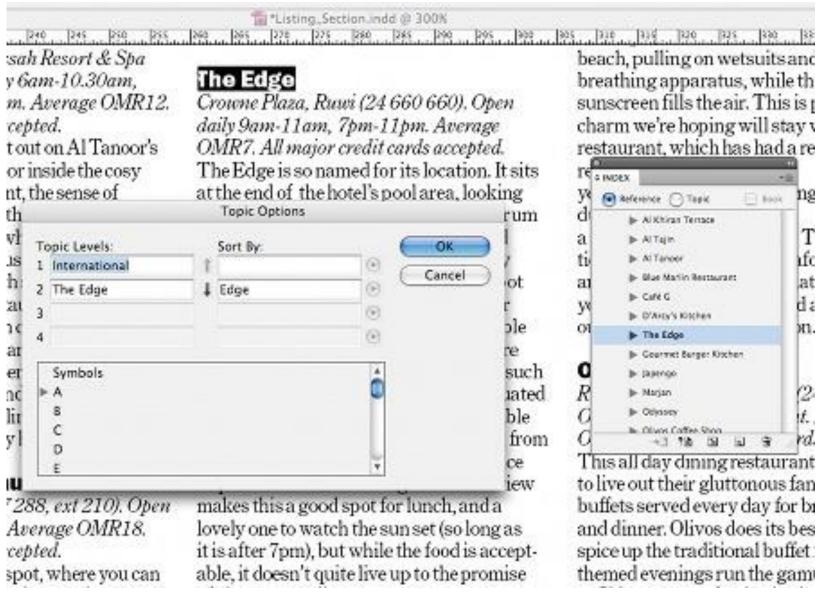






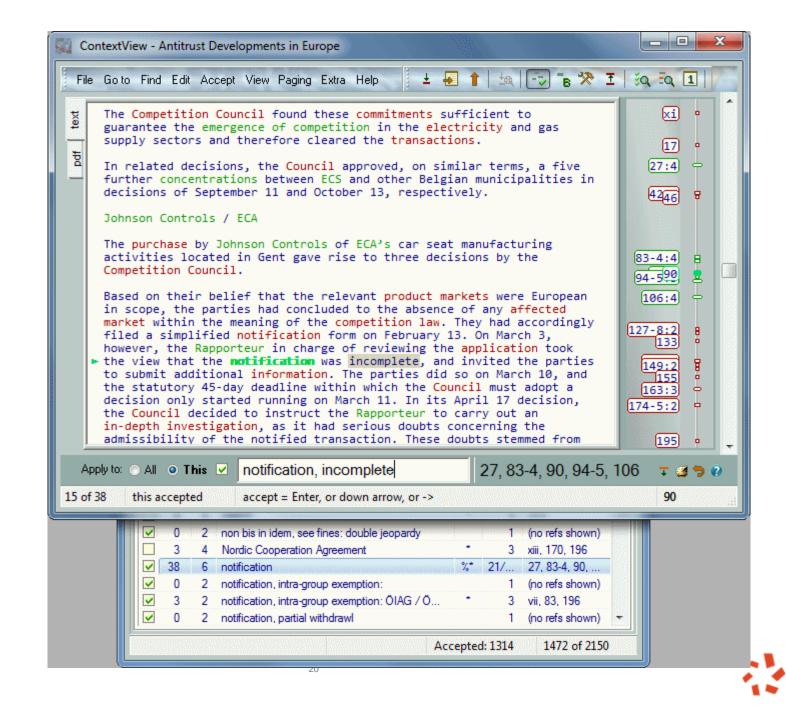
Embedded indexing feature of InDesign

Embeds index tags within the text





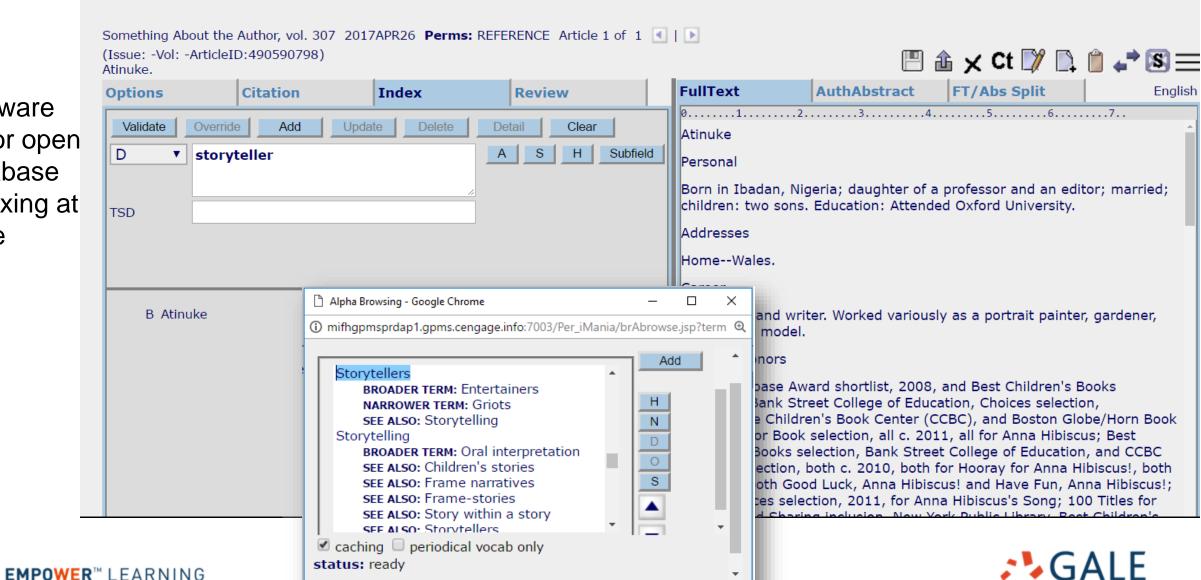
Software to semi-automate closed/book indexing:



Software UI for open database indexing at Gale

Index

Mercurv



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



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.







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 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 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"



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



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/nonpreferred 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 Vocabulary Types

Less

Controlled Vocabularies - Complexity

More

Pick List	Synonym Ring	Authority File	Taxonomy	Thesaurus	Ontology
Ambiguity control	Synonym control	Ambiguity control Synonym control	Ambiguity control (Synonym control) Hierarchical relationships	Ambiguity control Synonym control Hierarchical relationship Associative relationships	Ambiguity control (Synonym control) Semantic relationships Classes

Software for Taxonomy/Thesaurus Management

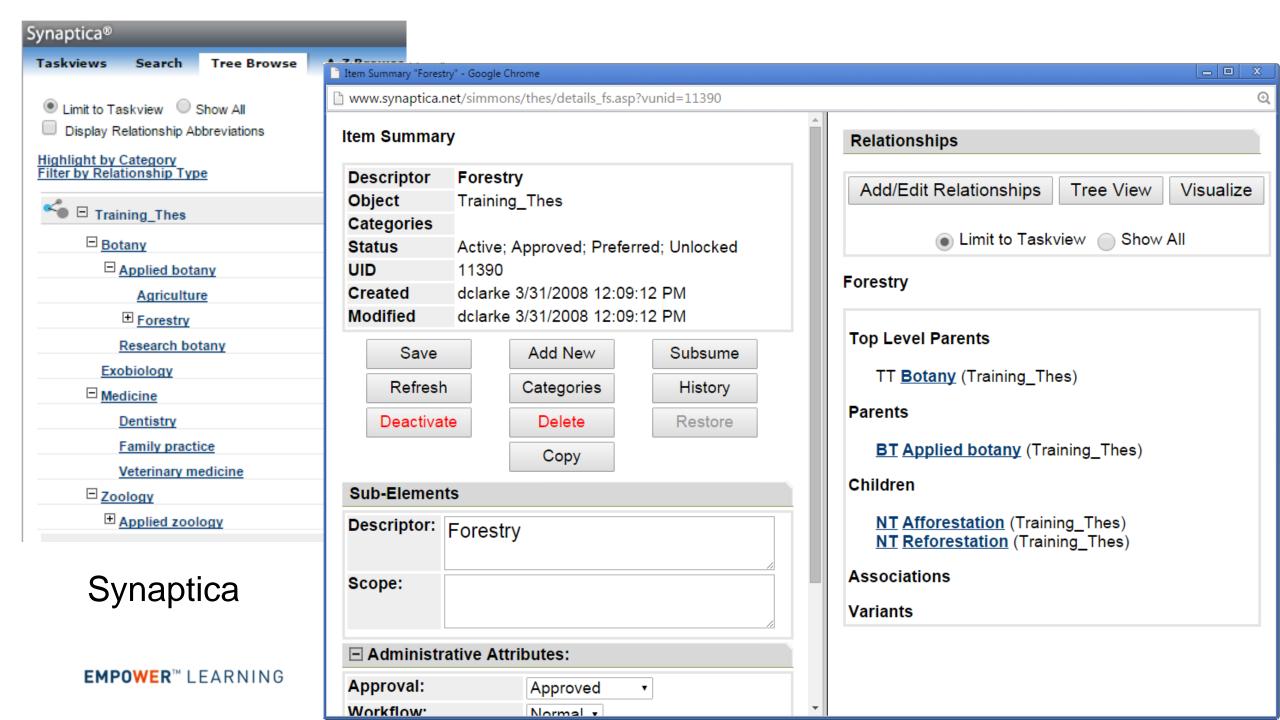
- Dedicated thesaurus management software
- Taxonomy creation & editing module of a content management, document management, digital asset management, collaborative software (SharePoint), or automated indexing software
- Vertical market software for creating classification structures
- Proprietary thesaurus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

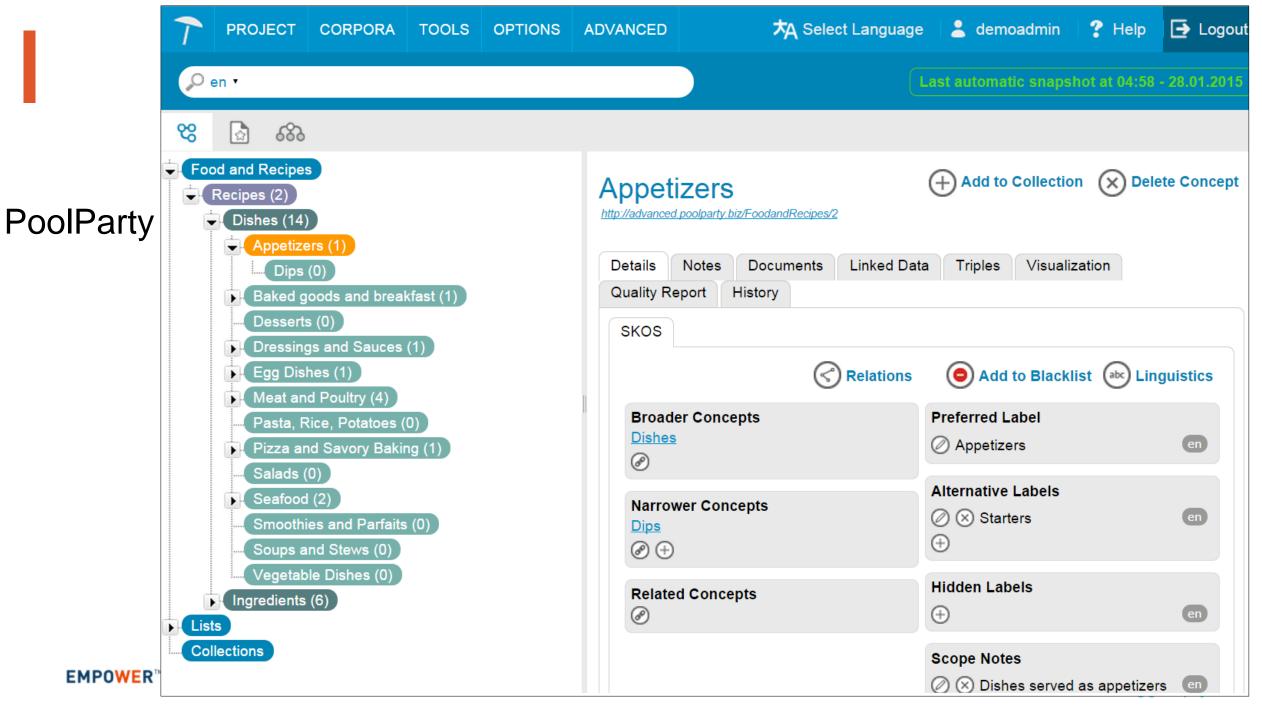
> Sreenshots from four dedicated thesaurus management software products



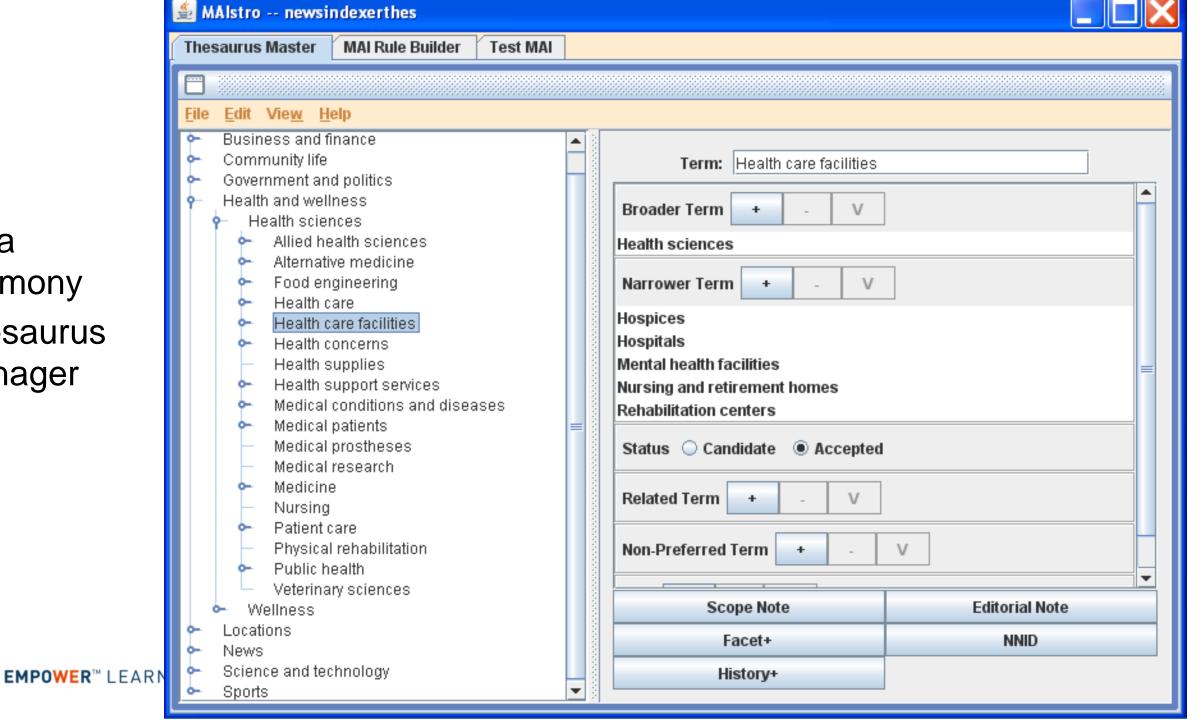
MultiTes Pro v2017.04.20 - C:\mtPro\data\Imported thesaurus.th2 File Edit Define Term Reports Window Help Search 4 ♦ Display Flag Advanced Manage Quick Refresh New Rel [ESC] Clear [Enter] Search Entry Search Categories List Term Term Wizard Imported thesaurus Relationships Status Notes Categories Translations Term 3-in-1 pictures USE ENG A la poupée prints BT, RT(2), UF SN **ENG** A trois crayons drawings SN **ENG** BT, UF(2) Äac prints USE **ENG** Α Abstract drawings BT(2) Aerial views Х Abstract paintings BT(2) Clipboard Edit Print Window Abstract photographs BT(2) Record Details Multilevel Hierarchy 2-way Hierarchy Edit Abstract prints BT(2) Abstract works BT, NT(4) Aerial views Add Rels... Accessories (Clothing & dress) BT, NT General Classification Acetate negatives BT, UF(5) +Rel Wiz... ■ Notes Acrylic paintings BT SN: Views from a high vantage point. Action comics USE Delete... □ Relationships Ads USE(2) UF: Air views Adventure comics BT, RT, UF Personal Note UF: Balloon views Rdd relationships \times BT, NT(4), RT(7) Advertisements UF: Views, Aerial UF Advertising Aerial views BT: Views BT(2), NT(3), R1 Advertising cards NT: Aerial photographs NT Advertising mail USE RT: Bird's-eye views SC Aerial photographs BT(2), RT RT: Panoramic views SN Browse... <u>S</u>ave UF Aerial views BT, NT, RT(2), USE Agfacolor transparencies BT USE ✓ Stay on top Close ВТ Agricultural equipment RT **EMPOWER**[™] Air conditioning, heating, & ventilating draw A USE Automatically add new terms if not found in thesaurus. Full a Page mode heather ENG 1580 terms (paged:120)

MultiTes





Data Harmony Thesaurus Manager



Questions/Contact

Heather Hedden

Senior Vocabulary Editor
Gale, A Cengage Company
20 Channel Center Street
Boston, MA 02210
heather.hedden@cengage.com

www.gale.com www.cengage.com

