A Tale of Two Worlds

Designing controlled vocabularies to meet specific business needs

Heather Hedden Senior Vocabulary Editor Gale, A Cengage Company

heather.hedden@cengage.com

Taxonomy Boot Camp London 17 October 2017





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," through Hedden Information Management
- American Society for Indexing board member, 2016 present
- Experienced taxonomy consultant and indexer



Agenda

- Introduction
 - Different business needs for Cengage controlled vocabularies
- 2. Comparisons of taxonomies and thesauri
 - Definitions, examples, and comparisons
 - Comparisons of suitability
 - When it's a compromise / in-between
- 3. Cengage's use of a thesaurus and of taxonomies
 - Views of the thesaurus and taxonomies
 - Why the difference between thesaurus and taxonomies
 - Extent of the difference between thesaurus and taxonomies / compromise
 - Meeting in the middle: links across vocabularies



Introduction



Various needs for taxonomies or thesauri:

- Internal documentation, including market and product research, policies and procedures.
- Digital assets for content managers to reuse in publishing content
- Product information (ecommerce, product reviews) for customers
- Curated, premium content for subscribers
- Informational or entertainment content for the public





Cengage

- Publishing company of educational and reference content
- Formerly Thomson Learning, spun off from Thomson and renamed in 2007
- Two different, but related businesses:
- 1. Gale, a Cengage Company reference/research content:
 - Library subscription databases of periodical articles, reference book chapters/essays, digital archives, and audio/video files
 - Public web products: Questia, Books & Authors, HighBeam Research, Encyclopedia.com
 - Gale Research reference books, directories, and other book imprints
- 2. The rest of Cengage educational/textbook/ebook and related software publishing:
 - Former Houghton-Mifflin College Division, Wadsworth, Delmar, Heinle, Schirmer, etc.
 - Imprint brands include: 4LTR Press, Milady, Chilton, etc.
 - Companion course software (MindTap, WebAssign, Aplia, CNow, SAM, etc.)













1. Controlled Vocabularies for Gale Digital Products

- Goal: Support efficient and effective search and retrieval of items from within content repositories
- Design Strategy: ANSI/NISO Z39.19 thesaurus-compliant vocabularies aimed at Boolean queries & browse navigation
- Content: articles from academic journals, magazines, newspapers, reference books; some multimedia
- Users: Library patrons, students, reference and research librarians,
- Purpose: retrieve sufficient relevant content on a subject
- Typically, relevancy & precision takes precedence over recall. (Retrieving accurate results, not necessary all results)



2. Controlled Vocabularies for Higher Education Digital Learning Products

- Goals: Wide variety of specifics, but all are focused on supporting software features and functions
- Design Strategy: Taxonomies with flexibility to meet specific needs of specific disciplines
- Content: text, multimedia, assessments, activities, flashcards, learning objectives/outcomes, etc.
- Users: Instructors, students, Cengage internal staff
- Purpose: various
 - Linking of content, activities, learning objectives, etc.
 - Building custom courses (by instructors)
 - Exploration of topics in more depth (by students)
- Full recall, with more limited content, is desired.







Summary of Controlled Vocabulary Types

Less

Controlled Vocabularies - Complexity

More

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

Definitions – according to ISO 25964-2: 2013 standards

Taxonomy

Scheme of categories and subcategories that can be used to sort and otherwise organize items of knowledge or information.

The typical taxonomy is presented as a hierarchical vocabulary, used for classifying or categorizing, organizing, browsing, navigating, searching and/or filtering.

Thesaurus

Controlled and structured vocabulary in which concepts are represented by terms, organized so that relationships between concepts are made explicit, and preferred terms are accompanied by lead-in entries for synonyms or quasi-synonyms.

Sections 3 Terms and Definitions, 9.9.1 Taxonomy



Definitions – according to ANSI/NISO Z39.19 standards

Taxonomy

A taxonomy is a controlled vocabulary consisting of preferred terms, all of which are connected in a hierarchy or polyhierarchy.

Thesaurus

A thesaurus is a controlled vocabulary arranged in a known order and structured so that the various relationships among terms are displayed clearly and identified by standardised relationship indicators.

Its purpose is to promote consistency in the indexing of content objects...

Sections 4.1 Definitions, 5.4.3 Taxonomy, and 5.4.4 Thesaurus



Taxonomies

- All terms belong to a limited number of major hierarchies (or facets)
- May bend standard (ISO or ANSI/NISO) hierarchical rules.
- Supports classification, categorisation, concept organisation. (Like Linnaean taxonomy.)
- > Approach is a top-down navigation.
- Especially serving end-users when browsing.

Thesauri

- ➤ All terms have relationships, but "hierarchies" may be as few as 2 terms.
- Standard (ISO or ANSI/NISO) rules are strictly followed.
- Supports concept scoping, disambiguation, and relationships with similar concepts. (Like Roget's.)
- Approach is term-centered and what terms are linked to/from it.
- Especially serving indexers/ indexing.



Comparison of Taxonomies and Thesauri: Suitability

Taxonomies are for:

- Content and terms that naturally can be categorised
- > A subject area with defined scope and limits
- Browse navigations
- guidance of hierarchies
- Relatively small collections of terms

Thesauri are for:

- > Terms that are not easily categorised into hierarchies or facets
- Multiple, overlapping subject areas or domains with diverse content.
- Highly specific terms for detailed indexing
- Non-expert users, who benefit from > Users who are subject-matter experts and will likely look for specific terms
 - Vocabulary that is large and/or constantly growing



Taxonomy/Thesaurus Compromise

- You want features of both taxonomies and thesauri.
 Such as structured as hierarchies or facets and also associative (related term) relationships.
- 2. Owners or users prefer to call it one thing, even though the other is needed.

 Such as users referring to a "taxonomy" for research content that needs a thesaurus: varied and broad in scope and manually indexed.

And "thesaurus management" software is for both/either.



Cengage's use of a Thesaurus and Taxonomies



Cengage Controlled Vocabularies

"Controlled vocabularies"

For research/library products (Gale), comprising:

- A single Subject thesaurus, ANSI/NISO standards-compliant.
 Around 60,000 preferred terms and always growing
- Multiple (11) named entity authority files (People, Companies, Organizations, Geos, Works, etc.)
- Other classification metadata (Dates, Languages, Information Types, Genres, NAICS, etc.)

2. Multiple (42) "Taxonomies" For educational/curriculum produces

For educational/curriculum products:

- One for each academic discipline
- Ranging from 400 4500 preferred terms in each











Academic Journals (201)

Search Terms: Subject Guide Search (Hijab (Custom))

Sort by: Newest

Full-text: ARTICLE

Army now allows soldiers to wear turbans, beards, and headscarves

Adelle M. Banks.

The Christian Century.

134.4 (Feb. 15, 2017) p12. Word Count: 451. Reading Level (Lexile): 1440.

New army regulations will allow Sikh and Muslim soldiers to wear turbans, beards, and hijabs--the headscarves worn by some Muslim women--under most circumstances. "Based on the successful examples of soldiers currently...

Link to CDL Materials Catalog ILL My Library OpenURL Link



Full-text: ARTICLE

Chancellor Angela Merkel of Germany calls for a ban of burqas as election nears

David Iaconangelo.

The Christian Century.

134.1 (Jan. 4, 2017) p15. Word Count: 558. Reading Level (Lexile): 1320.

In accepting her party's nomination for a fourth term in office, German chancellor Angela Merkel delivered a speech that took aim at conservative critics of her liberal asylum policies, yet also embraced a proposal to...

Link to CDL Materials Catalog ILL My Library OpenURL Link



Abstract: REPORT

On: The couch and the chador

Content Types

Academic Journals (201)

Magazines (109)

Books (1)

News (2,067)

Images (10)

Videos (4)

Search within results

Search within results O

Limit Search by:

- Full Text
- Peer Reviewed Journals
- Contains Images

Publication Dates:

Past Week

Past Month

Past Year

Custom Date Range

Subjects

Hijab (Custom) (201)

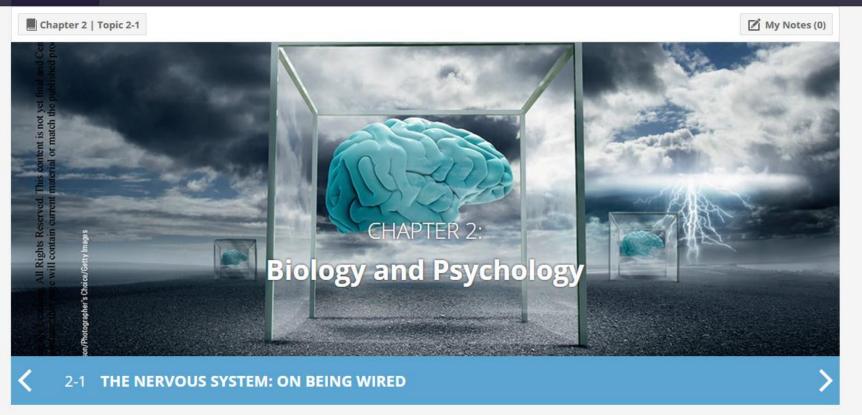
+

Islamic clothing (54)

Muslim women (50)







The nervous system is a system of nerves involved in thought processes, heartbeat, visual-motor coordination, and so on. The nervous system consists of the brain, the spinal cord, and other parts that make it possible for us to receive information from the world outside and to act on the world. It is composed of cells, most of which are *neurons*. Here we begin our study of the nervous system.

2-1a. Neurons: Into the Fabulous Forest

Within our brains lies a fabulous forest of nerve cells, or neurons. Neurons

NOTES		
MY UN	DERSTANI	DING
MY UN	DERSTANI FAIR	DING STRONG



Work Type

Get Work

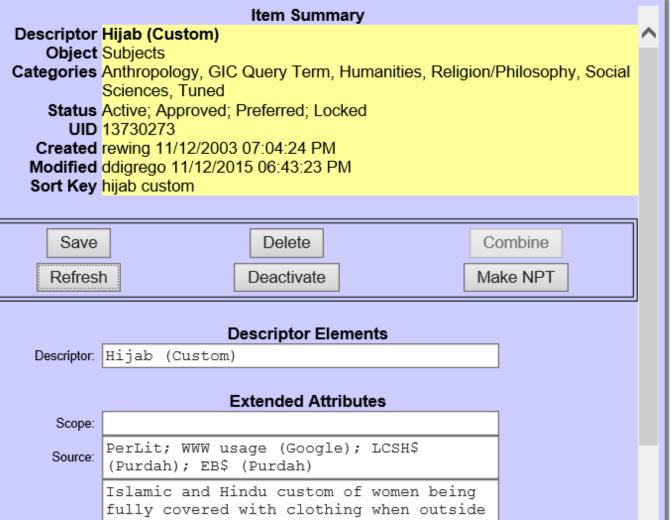
My Work

Preferences

Logoff

Something About the Author, vol. 307 2017APR26 Perms: REFERENCE Article 1 of 1 (Issue: -Vol: -ArticleID:490590798) Atinuke. **FullText** AuthAbstract FT/Abs Split English Citation Index Review **Options** 0.....1....2....3.....4....5.....6.....7.. Validate Delete Override Add Update Clear Detail Atinuke storyteller Subfield D Personal Born in Ibadan, Nigeria; daughter of a professor and an editor; married; children: two sons. Education: Attended Oxford University. TSD Alpha Browsing - Google Chrome ddresses imifhgpmsprdap1.gpms.cengage.info:7003/Per iMania/brAbrowse.jsp?term ome--Wales. areer Add Storvtellers B Atinuke toryteller and writer. Worked variously as a portrait painter, gardener, **BROADER TERM:** Entertainers nd artist's model. **NARROWER TERM:** Griots SEE ALSO: Storytelling N wards, Honors Storytelling **BROADER TERM:** Oral interpretation ranford Boase Award shortlist, 2008, and Best Children's Books 0 **SEE ALSO:** Children's stories election, Bank Street College of Education, Choices selection, SEE ALSO: Frame narratives poperative Children's Book Center (CCBC), and Boston Globe/Horn Book SEE ALSO: Frame-stories ward Honor Book selection, all c. 2011, all for Anna Hibiscus; Best SEE ALSO: Story within a story hildren's Books selection, Bank Street College of Education, and CCBC SEE ALSO: Storytellers hoices selection, both c. 2010, both for Hooray for Anna Hibiscus!, both caching periodical vocab only 011, for both Good Luck, Anna Hibiscus! and Have Fun, Anna Hibiscus!; status: ready CBC Choices selection, 2011, for Anna Hibiscus's Song; 100 Titles for

Cengage Controlled Vocabularies



[Add Relationship]

Relationships

- UF Purdah (Custom) (Subjects)
- BT Manners and customs (Subjects)
 - RT Dress codes (Subjects)
 - RT Hindu women (Subjects)
 - RT Islamic clothing (Subjects)
 - RT Islamic culture (Subjects)
 - RT Muslim women (Subjects)
 - RT Veils (Subjects)
 - **EQTAXON** Islamic clothing (Taxon Religion)
 - EQD Hijab (Custom) (FCDTerms)
 - IDX ai156vi0012 The French Headscarf Ban Is Oppressive (Doc)
 - IDX ai156vi0013 Fabric of Society: Banning Headscarves Is Right (Doc)
 - IDX GIC2008ov 000068 Headscarves: Religious and Political Issues (Doc)
 - IDX GIC2010stat 001701 G1278xb02 International Views of Bans on Full Islamic Veils: 2010 (IM)
 - IDX GICTPC1278 Headscarves: Religious And Political Issues (Portals)
 - IDX HID0416 Hijab and Islam (Doc)



1. Why different controlled vocabularies for research products and learning/educational products?

2. Why taxonomies vs. a thesaurus?



Why one subject thesaurus for research products:

- Multiple, overlapping subject areas.
- Some terms cannot be easily categorised.
- Both Hierarchical (NT, BT) and Related Term (RT) relationships important for communicating meaning and disambiguation.
- Terms can be broad because "sub-headings" are applied to focus meaning.
- Developed for use by trained indexers, with plenty of RTs and crossreferences.
- Vocabulary is very large and growing.



Why new, distinct taxonomies for academic disciplines:

- Can be more customised around a discipline.
- May contain more, specific terms.
- Instructional vs. research perspective often calls for different terms.
- Eliminates the need to disambiguate same terms with different meanings.
- Single discipline can allow for smaller, more efficient vocabulary.
- More flexibility for uses in new products not yet identified.
- Have complex terms and no need for a set of second-level "sub-headings" vocabulary.



Additional benefits of hierarchical structure taxonomies for learning disciplines:

- Facilitates limiting the taxonomy to a subject discipline scope.
- Easier to understand and review by non-taxonomist subject-matter experts.
- Categories with narrower terms are more typical in instructional material.

Such as learning objectives for "describe three types of X", with X being the broader term/category for three narrower terms.



Cengage Controlled Vocabularies: Extent of the Difference

Taxonomies for educational content

- Separate taxonomies for each discipline (subject area).
- Term choice/form reflects the discipline. No need for parenthetical qualifiers
- Each relatively small (typically 500 3,000 preferred terms).
- A single hierarchy for each discipline.
- Once built and tested, vocabularies are slow-growing.
- Easier use for non-indexers: subject-matter experts & product developers.
- Could be implemented in an end-user browse interface.



Cengage Controlled Vocabularies: Extent of the Difference

Cengage learning discipline "taxonomies" are a hybrid:

Thesaurus-like:

- > Have all standard thesaurus relationships
 - Broader term/narrower term
 - Related terms (although not as extensive)
 - Nonpreferred terms
 - Scope notes

Taxonomy-like:

- Each taxonomy has a single top term and limited number of 2nd level terms.
- > Sometimes bends standard hierarchical relationship rules at the top levels.



Cengage Controlled Vocabularies: Extent of the Difference

Health Insurance

term in the Subject thesaurus

- NT Capitated payment systems (Medical care) (Subjects)
- NT Group health insurance (Subjects)
- NT Hospitalization insurance (Subjects)
- NT Long term care insurance (Subjects)
- NT Managed care plans (Medical care) (Subjects)
- □ NT Medicare (Subjects)
 - NT Medicare Part D (Subjects)
 - NT Medigap (Subjects)
 - NT National health insurance (Subjects)
 - NT Point of service plans (Subjects)
 - NT Preferred provider organizations (Medical care) (Subjects)
- □ NT Prescription drug plans (Subjects)
 - NT Medicare Part D (Subjects)
- ☐ NT Prospective payment systems (Medical care) (Subjects)
 - NT Diagnosis related groups (Subjects)
 - NT Relative value scale payment systems (Medical care) (Subjects)
 - NT Single payer system (Health care) (Subjects)
 - RT Accident insurance (Subjects)
 - RT Altmark Trans GmbH v. Nahverkehrsgesellschaft Altmark GmbH 2003 C.M.L.R. 12 (E.C.J.)
 - RT American Ass'n of Retired Persons v. EEOC 489 F.3d 558 (3d Cir. 2007) (Court Cases)
 - RT Coinsurance (Subjects)
 - RT Commission of the European Communities v. Ireland 2012 C.E.C. 229 (C.J.E.U.) (Court Cases)
 - RT Cooperative insurance (Subjects)
 - RT Cost shifting (Medical care) (Subjects)
 - RT Dental insurance (Subjects)
 - RT Diagnosis related groups (Subjects)
 - RT Disability insurance (Subjects)
 - RT Flexible spending accounts (Subjects)
 - RT Fraternal organizations (Subjects)
 - RT Germany, Sickness Insurance Act 1883 (Statutes)
 - RT Health care costs (Subjects)
 - RT Health care industry (Subjects)
 - RT Health care reform (Subjects)
 - RT Health insurance claims processing software (Subjects)
 - RT Health insurance exchanges (Subjects)
 - RT Health insurance industry (Subjects)
 - RT Health insurance tax credits (Subjects)
 - RT Health maintenance organizations (Subjects)

4 UFs 14 direct NTs (16 NTs total) up to 2 levels deep 36 RTs

Health Insurance term in the Health discipline taxonomy

3 UFs

5 RTs

13 direct NTs

(58 NTs total)

up to 4 levels deep

- NT Health care reimbursement systems (Taxon Health)
 - NT CMS reimbursement (Taxon Health)
 - NT Health insurance coverage (Taxon Health)
 - NT Health insurance history (Taxon Health)
- ☐ NT Health insurance laws (Taxon Health)
 - NT ENTITY Health Care and Education Affordability Reconciliation Act o
 - NT ENTITY Patient Protection and Affordable Care Act (Statutes)
- NT Health insurance terminology (Taxon Health)
- ☐ NT Insurance claims process (Taxon Health)
 - NT Accept assignment (Taxon Health)
 - NT Assignment of benefits (Taxon Health)
 - NT Claims adjudication (Taxon Health)
 - NT Claims tracking (Taxon Health)
 - NT Electronic data interchange (EDI) (Taxon Health)
- NT Filing insurance claims (Taxon Health)
 - NT CMS-1500 claim form completion (Taxon Health)
 - NT National provider identifier (NPI) (Taxon Health)
 - NT Payer responsibility determination (Taxon Health)
- NT Insurance claim files maintaining (Taxon Health)
- ☐ NT Insurance claim forms (Taxon Health)
- □ NT CMS-1500 claim forms (Taxon Health)
 - .. More levels exist.
 - NT UB-04 claim forms (Taxon Health)
- □ NT Insurance forms (Taxon Health)
 - NT First report of injury form (Taxon Health)
- ☐ NT Insurance claim forms (Taxon Health)
- MT CMS-1500 claim forms (Taxon Health)
 - .. More levels exist.
 - NT UB-04 claim forms (Taxon Health)
- NT Insurance payments (Taxon Health)
- NT Insurance plan types (Taxon Health)
- NT Commercial insurance plans (Taxon Health)
 - NT Automobile, disability, and liability insurance (Taxon Health)
 - NT Medigap (Taxon Health)
- NT Consumer-driven health plans (Taxon Health)
- NT Fee-for-service insurance (Taxon Health)
- NT Government health plans (Taxon Health)
- ☐ NT Medicaid (Taxon Health)
- . More levels exist.



Cengage Controlled Vocabularies: Meeting in the Middle

Discipline taxonomy terms are linked to equivalent Subject Thesaurus terms.

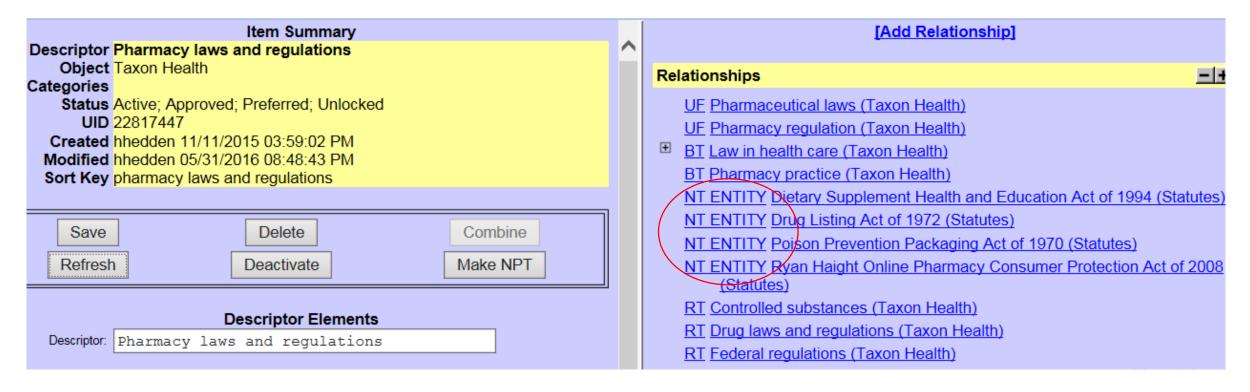
- Vocabulary management system "Atlas" supports customised crossvocabulary equivalency links (crosswalks).
- When implemented in products, students can jump from an instructional topic to relevant articles in the Gale reference databases, indexed with the equivalent term.



Cengage Controlled Vocabularies: Meeting in the Middle

Discipline taxonomy terms are linked to selected narrower named entities in the Gale name authority files.

- "Atlas" supports customised cross-vocabulary hierarchical relationships.
- Eliminates the need for duplicate maintenance.



Summary

- 1. For multiple purposes (different content and different users) a single controlled vocabulary may not suffice, even within the same organisation.
- For different purposes sometimes a thesaurus is needed, and sometimes a taxonomy is needed.
- 3. The distinction between a thesaurus and a taxonomy can be blurred.
- 4. Multiple controlled vocabularies can be linked term-by-term.



Questions/Contact

Heather Hedden

Senior Vocabulary Editor

Gale, A Cengage Company

20 Channel Center Street

Boston, MA 02210 U.S.A.

heather.hedden@cengage.com heather@hedden.net

www.cengage.com, www.gale.com
www.cengage.co.uk, www.gale.cengage.co.uk

