Introduction to Taxonomies and Controlled Vocabularies

Heather Hedden
Senior Vocabulary Editor, Cengage Learning

SLA Annual Conference
Boston, Massachusetts, June 13, 2015
About Heather Hedden

- Continuing education online workshop instructor, Simmons College School of Library and Information Science
- Author of *The Accidental Taxonomist* (Information Today, Inc.)
- Part-time freelance back-of-the-book indexer
- SLA Taxonomy Division – previous mentoring committee chair
- American Society for Indexing – Board member, previous Taxonomies & Controlled Vocabularies Special Interest Group manager

Previously

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

1. Introduction, Definitions, and Types
2. Creating & Wording of Terms
3. Nonpreferred Terms
4. Creating Term Relationships
5. Structural Design: Hierarchies and Facets
6. Display Options
7. Software Tools
8. Sources for Terms
9. Project Process & Management
10. Resources
1. Introduction, Definitions, and Types
2. Creating & Wording of Terms
3. Nonpreferred Terms
4. Creating Term Relationships
5. Structural Design: Hierarchies and Facets
6. Display Options
7. Software Tools
8. Sources for Terms
9. Project Process & Management
10. Resources
Introduction

1. Applications and uses of taxonomies and other controlled vocabularies
2. Benefits to taxonomies and other controlled vocabularies
3. Definitions and types of controlled vocabularies
Applications and uses

1. Indexing support (and secondarily, retrieval)
2. Retrieval support
   a) Search
   b) Hierarchical browse
   c) Faceted search/browse
1. Indexing Support

- List of agreed-upon terms for indexing multiple documents and/or by multiple indexers to ensure consistency
- For periodical articles, multi-volume works, database records, digital asset collections, etc.

Examples:
- Library of Congress Subject Headings (LCSH) http://authorities.loc.gov
- Medical Subject Headings (MeSH) http://www.nlm.nih.gov/mesh/MBrowser.html
- Getty Art & Architecture Thesaurus http://www.getty.edu/research/tools/vocabularies/aat
Introduction:

Indexing Support

Examples (continued):

Cengage Learning Gale Subject Thesaurus: Internal indexer alphabetical browse view
2. Retrieval Support
   a. Search
      Controlled list of terms and their synonyms/equivalents to aid online retrieval

      - For website or intranet search engines, online databases, online directories, enterprise search
      - Might be displayed as type-ahead auto-suggest terms, or might not be displayed at all.
      - May be used for automated indexing / auto-tagging
Introduction: Applications and Uses

2. Retrieval Support

b. Hierarchical browse
   Categorization scheme for information organization, classification, guided search
   - For web site structural design, online information services, intranet content organization, corporate content management
   - Examples of web site navigational taxonomies:
     - Digital Web Magazine topics http://www.digital-web.com/topics
2. Retrieval Support

c. Faceted search/browse
Multiple term lists of different types, also called facets/filters/refinements, for the user to search on in combination

- For online databases, catalogs, e-commerce sites
- Usually does not have synonyms or term hierarchies
- Examples:
  - NCSU Libraries catalog (browse new titles)  
    www.lib.ncsu.edu/catalog
  - Internet Movie Database (advanced title search)  
    www.imdb.com/search/title
Benefits of taxonomies/controlled vocabularies

1. **Controlled vocabulary (CV):**
   - Brings together different wordings (synonyms) for the same concept
   - Helps people search for information by different names

2. **Classification and structure:**
   - Organizes information into a logical structure
   - Helps people browse or navigate for information
A controlled vocabulary gathers synonyms, acronyms, variant spellings, etc.

- Documents not missed due to use of different words (e.g. **Automobiles**, instead of **Cars**)
- User does not need to guess the spelling of unusual or foreign names (e.g. **Condoleezza Rice**)

A search restricted on the controlled vocabulary retrieves concepts not just words.

- Documents excluded for mere text-string matches (e.g. **monitors** for computers, not the verb “observes”)

**Introduction: Benefits**
Introduction: Benefits

Controlled Vocabulary contains all synonyms:

Users may enter:
- Oil industry
- Oil & gas industry
- Oil & gas industries
- Petroleum industry

Text may contain:
- Oil and gas industry
- Oil companies
- Big oil
- Oil producers
A hierarchical taxonomy provides guided search.

- Users can browse and locate narrower (more specific) subjects of interest.
- Users find out what is included and what is not.
- Users may find related subjects of interest.

Taxonomies can reflect natural categories.
Definitions and Types

Controlled vocabulary (CV)

- The most general, broadest concept for all applications

- An authoritative, restricted list of terms (words or phrases) mainly used for indexing/tagging content to support retrieval

- Controlled in who and when new terms can be added

- Usually makes use of equivalent non-preferred terms (synonyms, etc.) to point to the correct, preferred terms

- May or may not have structured relationships between terms
Introduction: Definitions & Types

Controlled Vocabulary types

- Simple term list/pick list
- Synonym ring (search-support “thesaurus”)
- Authority file (controlled list with variants; no hierarchy)
- Taxonomy (definition #1)
  - Hierarchical taxonomy
  - Faceted taxonomy
- Thesaurus
- Ontology
Term List

- A simple list of terms
- Usually alphabetical, but could be in other logical order
- Lacking synonyms, it is usually short enough for quick browsing
- Can appear in drop-down scroll boxes
- May be used for various metadata values
- Part of a larger set of controlled vocabularies
Synonym ring

- A controlled vocabulary with synonyms or near-synonyms for each concept
- No designated “preferred” term: All terms are equal and point to each other, as in a ring.
- Usually just called a “controlled vocabulary”
Introduction: Definitions & Types

Taxonomy

A controlled vocabulary with broader/narrower (parent/child) term relationships that include all terms to create a hierarchical structure

- With focus for categorizing and organization concepts
- May or may not have equivalent non-preferred terms (synonyms, etc.) to point to the correct, preferred terms
- May comprise several hierarchies or facets (A facet can be considered a hierarchy.)
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

Course
- Main Dishes (15504)
- Desserts (7530)
- Side Dishes/Vegetables (6182)

Convenience
- Entertaining (23804)
- Make-Ahead (13917)
- Quick/Easy (13186)

Cost Per Serving
- $1 and Under (388)
- $1.01 to $2 (394)
- $2.01 to $3 (250)
- $3.01 to $4 (94)
- $4.01 and Up (28)

Cuisine
- American (28614)
- Italian (3129)
- New American (2370)

Main Ingredient
- Vegetables (11246)
- Fruits (6297)
- Poultry (5287)

Dietary Consideration
- Meatless (11299)
- Low Cholesterol (7534)
- Low Saturated Fat (7444)
Thesaurus

- A controlled vocabulary that has standard structured relationships between terms
  - Hierarchical: broader term/narrower term (BT/NT)
  - Associative: related terms (RT)
  - Equivalence: preferred term (“use for” or “used for”) / non-preferred term (use) (USE/UF)
- Also supports notes, such as scope notes (SN), for terms, as needed.
- Created in accordance with standards, such as ANSI/NISO Z39.19 Guidelines for Construction, Format, and Management of Monolingual Controlled Vocabularies.
- “Thesaurus” is most often the kind of controlled vocabulary used in indexing periodical literature.
Introduction: Definitions & Types

Thesaurus entry examples

Government lending
  >BT Economic policy
  >>BT2 National policy
  <NT Veterans' loans
  RT Agricultural credit
  RT Federally-assisted loans
  RT Federally-guaranteed loans
  RT Government and business
  RT Government insurance
  RT Loans
  RT Student loan funds
  UF American domestic economic assistance
  UF Federal aid to depressed areas
  UF Federal credit programs
  UF Federal domestic assistance programs
  UF Government loans

materials acquisitions
  UF acquisitions (of materials)
  library acquisitions
  BT collection development
  NT accessions
    approval plans
    gifts and exchanges
    materials claims
    materials orders
    subscriptions
  RT book vendors
    jobbers
  subscription agencies
  subscription cancellations
Ontology

- A complex thesaurus-type of controlled vocabulary, in which terms have specified attributes and relationships are further specified.
- Relationships contain meaning, are “semantic”.
- A form of “knowledge representation”.
- An ontology is not just for indexing. It is often an end itself in representing a domain of information.
Ontology example
## Controlled Vocabularies - Complexity

<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) Hierarchical relationships</td>
<td>Ambiguity control Synonym control Hierarchical relationship Associative relationships</td>
<td>Ambiguity control (Synonym control) Semantic relationships Classes</td>
</tr>
</tbody>
</table>
Taxonomies

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

Thesauri

- All terms have relationships, but “hierarchies” can comprise 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.)
- Approach is term-centered and what terms are linked to/from it.
- Especially serving indexers/indexing.
## Introduction: Definitions & Types

### Taxonomies, for:
- Content/terms that naturally can be categorized
- A subject area with defined scope and limits
- Browsable hierarchical displays
- Non-expert users, who benefit from guidance of hierarchies
- Relatively small collections of terms
- Somewhat broad categories

### Thesauri, for:
- Content/terms that are *not* easily categorized
- Multiple, overlapping subject areas or domains with diverse content.
- Alphabetical displays (indexes), or type-ahead display excerpts
- Users who are subject-matter experts and will likely look for specific terms
- Large and/or constantly changing vocabulary
- Highly specific terms for detailed indexing
“Taxonomy”
Any kind of controlled vocabulary, in a/an…
- enterprise, corporate setting
- content management system
- website navigation (e.g. ecommerce site)

“Thesaurus”
Any kind of controlled vocabulary…
- for indexing articles / literature retrieval databases
- used by librarians, indexers, or other information professionals
- that includes nonpreferred terms (synonym rings)
1. Introduction, Definitions, and Types
2. Creating & Wording of Terms
3. Nonpreferred Terms
4. Creating Term Relationships
5. Structural Design: Hierarchies and Facets
6. Display Options
7. Software Tools
8. Sources for Terms
9. Project Process & Management
10. Resources
A taxonomy comprises…

What are variously called:

- **Terms**: generic, common designation
- **Nodes**: in hierarchical “trees”
- **Concepts**: term + synonyms, attributes, relationships, etc.
- **Objects**: same as concept, in object-oriented databases
- **Descriptors**: preferred term, excludes non-preferred
- **Preferred terms**: excludes non-preferred terms
- **Topics**: preferred term, not a named entity
Creating & Wording of Terms

Term creation issues

- Deciding whether to include a term
- Choosing a preferred term name
- Term format
- Distinguishing different terms with the same name
- Pre-coordination vs. post-coordinate of terms
Whether a concept should be included as a term

1. Is it within the subject-area scope of the CV?
2. Is it important, likely to be looked up?
3. Is there enough information on the subject? Current and *anticipated* documents on the topic
4. Do users want and expect it?
Choosing the preferred term wording

- You must always choose a “preferred term” (except in synonym ring vocabularies)
- Variants and synonyms are designated as “non-preferred terms”
- Wording is based on user expectations and needs, more than on content, which varies.
Choosing the preferred term wording

Choosing between two “synonyms”:

- **Doctors** vs. **Physicians**
- **Movies** vs. **Motion pictures**
- **Cars** vs. **Automobiles**

Consider:

1. Wording of terms most likely looked up by the intended users/audience, especially in browsed CV
2. Enforcing organizational/enterprise controlled vocabulary
3. Conforming to academic or professional standards
4. Consistency in style throughout the CV
5. Wording with in the documents/content indexed
Choosing the preferred term wording
Differentiate closely related terms, or use one as preferred (The other becomes a nonpreferred term.)
- Foreign policy vs. International relations
- Colleges & universities vs. Higher education

Differentiate topics from actions, or use one as preferred
- Contracts vs. Contracting
- Investments vs. Investing

Differentiate broader and more specific concept, or use one
- Electric power plants vs. Hydroelectric power plants
- Plants & factories vs. Factories

Consider likely occurrences of the more specific topic in the content.
Creating & Wording of Terms

Term format

- Lower case or initial caps; not title caps
- Single words or multi-word phrases
- Nouns or noun phrases
- Common nouns or proper nouns
- Adjectives alone can be terms in special circumstances and where noun is obvious from context.
- Countable nouns are usually plural.
- Parenthetical qualifiers may be used for disambiguation, not modification.
- Avoid term inversions.
Distinguishing different terms with the same name

Example of alternatives:

- **French (language)** and **French (people)**
- **French language** and **French people**
- **French** and **French people**
- **French** in the Language hierarchy/facet and **French** in the People hierarchy/facet
## Exercise: Which format for Preferred Term?

<table>
<thead>
<tr>
<th>Term</th>
<th>Preferred Format</th>
</tr>
</thead>
<tbody>
<tr>
<td>Printers (people)</td>
<td>Printers as people</td>
</tr>
<tr>
<td>Schools (elementary)</td>
<td>Elementary schools</td>
</tr>
<tr>
<td>Walnut (wood)</td>
<td>Walnut tree wood</td>
</tr>
<tr>
<td>Photography, digital</td>
<td>Digital photography</td>
</tr>
<tr>
<td>Prisoners of war</td>
<td>War prisoners</td>
</tr>
<tr>
<td>Deoxyribonucleic acid</td>
<td>DNA</td>
</tr>
<tr>
<td>Ears</td>
<td>Ear</td>
</tr>
<tr>
<td>Taxonomies</td>
<td>Taxonomy</td>
</tr>
</tbody>
</table>
Pre-coordination vs. Post-coordination

Pre-coordination: creation of relatively specific, compound modified terms
   Example: **Hispanic women writers**

Post-coordination: leaving terms as simpler and broader, assuming that they will be combined by the user in searching
   Example: **Hispanics** AND **Women writers**
   or: **Hispanic women** AND **Writers**
Creating & Wording of Terms

Pre-coordination vs. Post-coordination

Advantages to pre-coordinated terms

- Provide more precise retrieval results, if used correctly
- Better suited for specific, custom taxonomies
- Better suited for phrase search-string searching

Disadvantages to pre-coordinated terms

- Require creating a larger, more complex taxonomy
- Narrower terms might be overlooked by the user
- More complex to correctly index

Flexibility in degree of pre- or post-coordination is OK, but consistency of application makes the taxonomy more usable.
Pre-coordination vs. Post-coordination

When do you leave terms simple for post-coordination?

- If faceted search is supported
- If controlled vocabulary is to be kept small

When do you create pre-coordinated terms?

- If post-coordinated cannot reliably retrieve the desired content
- If subject area is focused and thus deep: a relatively specialized controlled vocabulary
- If a lot of content is expected for the pre-coordinated concept
- If indexed document-types are varied
Inadequacies of post-coordination

- **Hispanic women writers**
- **Hispanics AND Women writers**
  - could also be for “women writers, who write about Hispanics”
- **Russian foreign policy**
- **Russia AND Foreign policy**
  - could also be for “foreign policy toward Russia”
- **French embassies**
- **France AND Embassies**
  - could also be for “foreign embassies in France”
Exercise: Post-coordinate: How? When?

Federal aid to higher education

United States-Russian relations

African-American women writers

Plastic farm machinery parts
Term Notes

- Terms may have notes.
- If included, not all terms need notes.
- May have multiple types/purposes of notes
- For searcher, indexer, or both
- Basic standard note is: Scope Note (SN)
- Others: History Note
  - Indexer Note
  - Usage Note
Scope Notes

1. To restrict or expand the application of a term
2. To distinguish between terms of overlapping meaning (may have reciprocal notes)
3. To provide advice on term usage

Concise information, not a dictionary definition
Must be original, not copied copyright text
Scope Note Examples

Defining what is in scope:

**Inequality** (from ProQuest)

**SN:** Socioeconomic disparity stemming from racial, cultural, or social bias

Defining what is in scope and explaining what is not:

**Dedications** (from Gale, a part of Cengage Learning)

**Scope:** Inscriptions prefixed to literary, musical, or artistic works in tribute to a person or cause. For church dedications, use Consecration.

Limiting what is in scope and explaining what is not:

**Nonverbal Communication** (from Medical Subject Headings)

**Annotation:** human only; for animals use **ANIMAL COMMUNICATION** or **VOCALIZATION, ANIMAL**
Term Attributes

- A sophisticated thesaurus/taxonomy database can store additional data about a term.
- Additional data may be free text or controlled.
- Additional data may be a field to sort upon.

Examples:

- *Companies*: address, industry code, private/public status
- *Person names*: title/occupation, birth date, nationality
- *Products*: part number, price, market, intro date
- *Places*: latitude and longitude
1. Introduction, Definitions, and Types
2. Creating & Wording of Terms
3. Nonpreferred Terms
4. Creating Term Relationships
5. Structural Design: Hierarchies and Facets
6. Display Options
7. Software Tools
8. Sources for Terms
9. Project Process & Management
10. Resources
Nonpreferred Terms

Terms for which a designated preferred term should be used instead.

Have various designations:

- Variants
- Nonpreferred terms
- Non-postable terms
- NPT
- Nondescriptors
- Equivalent terms
- Equivalency relationship
- Alternate terms
- Entry terms
- Used for terms
- Use for terms
- See references
- Use references
- Cross-references
- Synonyms
- Aliases
- Keywords
Nonpreferred Term Types include:

- synonyms
- near-synonyms
- variant spellings
- lexical variants
- foreign language names
- acronyms/spelled out forms
- scientific/technical names
- phrase variations (in print)
- antonyms
- narrower terms and instances that are not preferred terms
Nonpreferred Term Types include:

- **synonyms**: 
  - Cars USE Automobiles
- **near-synonyms**: Junior high schools USE Middle schools
- **variant spellings**: Defence USE Defense
- **lexical variants**: Hair loss USE Baldness
- **foreign language names**: Luftwaffe USE German Air Force
- **acronyms/spelled out forms**: UN USE United Nations
- **scientific/technical names**: Neoplasms USE Cancer
- **phrase variations (in print)**: Buses, school USE School buses
- **antonyms**: Misbehavior USE Behavior
- **narrower terms and instances that are not preferred terms**: Power hand drills USE Power hand tools
Standard thesaurus notation: USE / UF

Examples:

- Public procurement USE Government purchasing
- Dress design USE Costume design
- Eskimos USE Inuit
- Inundations USE Floods

Sometimes instead: See

Also considered a kind of “Relationship”: Equivalency

Relationship is reciprocal: Use and Used for (Used from)

- Inundations USE Floods
- Floods UF Inundations
Depending on the search/retrieval system, often a non-preferred term may point to only one preferred term. No “multiple USE” references.

Many-to-one, not one-to-many.
How many non-preferred terms to create?

Are non-preferred terms for retrieval only, or also for search?

- If taxonomy is small and browsable, and especially in facets, non-preferred terms are not needed for the user search, only for the retrieval/indexing side (automated or human).

- If “search box” is present, numerous non-preferred terms are needed, to match to all user-entered text.

- If automated indexing, even more non-preferred terms are needed to match phrases in texts.
How many non-preferred terms to create?

If users can input text in search box:

- Include non-preferred terms that are alphabetically close (unlike in A-Z index)

  - Ethnic groups
  - UF Ethnic minorities
  - UF Ethnicities

If system supports “smart” search on words within terms:

- Do not include simple inversions.

  - Political fundraising
  - UF Fundraising, political
How many non-preferred terms to create?
Plural/singular?
Depends on whether the system supports automatic “stemming”

Stemming might exist for single words but not phrases.

**Stations** stems to **Station**

**Train stations** does not stem.

Need to add non-preferred term: **Train station**
Exercise: Create nonpreferred terms

Oil industry

Political campaign finance

Aviation
Multiple Use

Occasionally one-to-many is permitted.

- Not more than one-to-two
- For AND, not OR
- Both desired results, both preferred terms must be used in combination, in both indexing and in searching.

Example: **Folk drama** USE **Drama** AND **Folk culture**

- Used in thesauri, not simply taxonomies, browsable by both indexers and users
- Sometimes called: “Used for And,” “Use for Plus”
- Commercial thesaurus software may not support it.
Nonpreferred terms for named entities: organizations, companies, countries, etc.

Official long, official short, common name, acronym, partial abbreviation, and all combinations/permutations. Example:

- United States Department of State
- United States State Department
- U.S. Department of State
- U.S. State Department
- U.S. Dept. of State
- U.S. State. Dept.
- US State Dept.

Plus native language name(s) for foreign entities
Nonpreferred terms for named entities: person names

- Include/exclude middle name
- Include/exclude middle initial
- Diminutives and nicknames, e.g. Mike/Michael
- Variations used by the specific individual, not merely all possible – could refer to someone else
- In auto-indexing, surname only or title+surname are OK, if unique within content.

**President Obama** – OK, unambiguous
**President Bush** – problematic, ambiguous
Exercise: Create nonpreferred terms

United Kingdom

September 11 Attacks

Anglican Church of Canada

French Navy
1. Introduction, Definitions, and Types
2. Creating & Wording of Terms
3. Nonpreferred Terms
4. Creating Term Relationships
5. Structural Design: Hierarchies and Facets
6. Display Options
7. Software Tools
8. Sources for Terms
9. Project Process & Management
10. Resources
Relationships between terms

Between preferred and nonpreferred terms:
1. Equivalence: Use (USE) / Used for (UF)

Between preferred terms:
2. Hierarchical: Broader term (BT) / Narrower term (NT)
3. Associative: Related term (RT)
4. Customized relationships (more specific types of BT/NT or RT)

Relationships are reciprocal between terms.
Hierarchical Relationships

Reciprocal (bi-directional) relationships, but asymmetrical

Broader term (BT)

SOME  ALL

Narrower term (NT)

SOME  ALL

Fruits NT Oranges  Oranges BT Fruits

Three types:
1. Generic – Specific
2. Common noun – Proper noun
3. Whole – Part
1) Hierarchical - Generic/Specific:

Category or class
– members
– more specific types

Narrower term
“is a” or “are a kind of”
broader term

Plants
NT Trees

Financial services
NT Investment services

Romance languages
NT Italian
2) **Hierarchical - Instance:**

Common noun

– Proper noun

Narrower term

“is a” broader term

- **Smartphones**
  - NT **Samsung Galaxy**

- **Golfers**
  - NT **Woods, Tiger**

- **Holidays**
  - NT **Thanksgiving**
### 3) Hierarchical – Whole/Part:

<table>
<thead>
<tr>
<th>Concept or entity</th>
<th>United Nations</th>
</tr>
</thead>
<tbody>
<tr>
<td>– part</td>
<td>NT UNICEF</td>
</tr>
<tr>
<td>– subentity</td>
<td>Massachusetts</td>
</tr>
<tr>
<td></td>
<td>NT Boston</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Narrower term</th>
<th>Digestive system</th>
</tr>
</thead>
<tbody>
<tr>
<td>“is in” broader term</td>
<td>NT Stomach</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Must be an integral part that cannot be taken out</th>
<th>Engineering</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NT Electrical engineering</td>
</tr>
</tbody>
</table>
Associative Relationships

- Suggestions to the user of possible related terms of interest
- Not used in simple hierarchical taxonomies
- Required feature of standard thesauri
- Standard designation of RT
- Default is symmetrically bi-directional relationship
- Between terms within the same hierarchy or in different hierarchies

Called:

- Related term
- Associated term
- See also
Term Relationships

Associative Relationships
Between terms within the same hierarchy

1. Having a shared broader term (siblings) and overlapping meaning

Example:

- **Taxes**
  - **Local taxes** → **Property taxes**

Required RT links, according to the standards
2. Siblings *without* overlapping meaning (mutually exclusive terms)

Optional, and usually not done. Not incorrect, but better left out.

Example:

```
Consumer Electronics

Radios

TV sets
```
As with non-preferred terms, if a search box exists, don’t assume user relies on A-Z browse. Include alphabetically neighboring relationships.
Term Relationships

Associative Relationships Between terms in different hierarchies:

- Process and agent
- Process and instrument
- Process and counter-agent
- Action and property
- Action and product
- Action and target
- Cause and effect
- Object and property
- Object and origins
- Raw material and product
- Discipline and practitioner
- Discipline and object
Term Relationships

Associative Relationships Between terms in different hierarchies:

Process and agent: Skiing RT Skiers
Process and instrument: Ventilation RT Fans
Process and counter-agent: Infections RT Antibiotics
Action and property: Environmental cleanup RT Pollution
Action and product: Glassblowing RT Glassware
Action and target: Auto repair RT Automobiles
Cause and effect: Hurricanes RT Flooding
Object and property: Plastics RT Elasticity
Object and origins: Petroleum RT Oil wells
Raw material and product: Timber RT Wood products
Discipline and practitioner: Physics RT Physicists
Discipline and object: Literature RT Books
Exercise: Suggest one or more related terms for each

Newspapers
Fires
Germany
Athletics
U.S. Congress
Term Relationships

Specific/customized relationships

- Relationships containing meaning: “semantic”
- Variations on either BT/NT or RT
- Reciprocal but usually asymmetrical, or directional, not plain RT
- Specific enough to convey the necessary meaning, but not uniquely specific
- Taxonomist defines the relationships and their codes.
- A defining characteristic of ontologies.
Specific/customized relationships

Sample variations on BT/NT:

Located in (LOCIN) / Contains Location (LOCCAN)
Empire State Building LOCIN New York, NY
New York, NY LOCCAN Empire State Building

Has parent organization (PAR) / Has sub-organization (SUB)
Internal Revenue Service PAR Dept. of the Treasury
Dept. of the Treasury SUB Internal Revenue Service

Is part of a whole (WHOL) / Has parts (PART)
Automobile engines WHOL Automobiles
Automobiles PART Automobile engines
Specific/customized relationships

Sample variations on the associative relationship (RT):

Has produced the work (WRK) / Created by (CRE)
Twain, Mark WRK The Adventures of Tom Sawyer
The Adventures of Tom Sawyer CRE Twain, Mark

Produces the product (PRD) / Is manufacture by (MAN)
Apple Inc. PRD iPod
iPod MAN Apple Inc.

Has member affiliation with (AFF) / Has members (MEM)
Saudi Arabia AFF OPEC
OPEC MEM Saudi Arabia

For treating (TRE) / Can be treated with the drug (DRUG)
ACE inhibitors TRE Hypertension
Hypertension DRUG ACE inhibitors
Specific/customized relationships

Sample semantic relationships (and codes) between the following terms (on next slide):

- Toyota Corolla
- Automobiles
- Motor vehicle industry
- Toyota Motor Corporation
- Akio Toyoda
**Term Relationships**

**Company name – Product name**
Toyota Motor Corp. – Toyota Corolla

**Produces (PRO) / Manufactured By (MAN)**
Toyota Motor Corporation PRO Toyota Corolla
Toyota Corolla MAN Toyota Motor Corporation

**Industry – Product type**
Motor vehicle industry – Automobiles

Motor vehicle industry PRO Automobiles
Automobiles MAN Motor Vehicle Industry

**Company name – Executive name**
Toyota Motor Corp. – Akio Toyoda

Is managed by person (MPN)/Manages company (MC)
Toyota Motor Corp. MPN Akio Toyoda
Akio Toyoda MC Toyota Motor Corporation

**Product type – Product name**
Automobiles – Toyota Corolla

Hierarchical relationship (instance)

Type of (TYP) / Includes the brand (BRD)
Toyota Corolla TYP Automobiles
Automobiles BRD Toyota Corolla

**Industry – Company name**
Motor vehicle industry – Toyota Motor Corp.
Associative relationship

Has members (MEM) / Is affiliation with (AFF)
Motor vehicle industry MEM Toyota Motor Corporation
Toyota Motor Corporation AFF motor vehicle industry
1. Introduction, Definitions, and Types
2. Creating & Wording of Terms
3. Nonpreferred Terms
4. Creating Term Relationships
5. Structural Design: Hierarchies and Facets
6. Display Options
7. Software Tools
8. Sources for Terms
9. Project Process & Management
10. Resources
Ways to structure controlled vocabularies

- One or more taxonomy/thesaurus
- Separate authority files
- Categories or classes
- Hierarchies
  - Polyhierarchies
- Facets
One or more taxonomy/thesaurus?

Thesaurus management software asks you to define a thesaurus/file/database.

A single taxonomy:

- Has standard relationships (BT/NT, RT, USE/UF) confined within it. Cross-taxonomy links, if any, are non-standard/semantic types.
- Has a defined, restricted subject scope.
- Has its own indexing/tagging policy.
- Could function in isolation, unlike a single facet (although may be supplemented by other controlled vocabularies).
- Has its own implementation, function, and purpose (although taxonomies can be reused and repurposed).

http://accidental-taxonomist.blogspot.com/2014/09/one-or-more-taxonomies.html
Separate authority files for different kinds of named entities

Examples:

- Person names
- Company names
- Product names

Each file may have its own:

- updating and maintenance policies and rights
- editorial style formats and policies
- term extended attributes
- stemming and search software rules
Classification within a single controlled vocabulary:

Categories or Classes

- To segment a master unified CV for different users, markets, audiences, end-use
- One or more category or class tags assigned to individual terms
- A feature (administrative) in most thesaurus software
- For taxonomy managers, not necessarily seen by indexers or searchers/users

Application Examples:

- for different internal departments: finance, R&D, customer service
- for different organizational uses: website, intranet, product database
- for different markets: healthcare, financial services, high-tech, etc.
- for the end-user to filter the taxonomy to contain only certain sub-sets of terms.
A single taxonomy may have one more top-term hierarchies.

Hierarchies:

- The extension of BT/NT to include all terms
- More important for taxonomies than thesauri
- Emphasize categorization, classification, sorting
- Involve working from the top down
- Also known as “tree” structures
Examples of hierarchies for classification:
Classifying of things – can only go in one place

- Linnaean taxonomy of classification of organisms (National Center for Biotechnology Information of the National Library of Medicine)

- Taxonomy of musical instruments
  http://www.ksanti.net/free-reed/description/taxonomy.html

- Dewey Decimal System
  http://www.oclc.org/dewey/resources/summaries/deweysummaries.pdf

- SIC or NAICS codes for industries

Classifying of concepts, electronically, allows terms in multiple places in the hierarchy.

- Categorized Web sites: dmoz directory http://www.dmoz.org
Linnaean taxonomy: National Center for Biotechnology Information, National Library of Medicine

Taxonomy Browser

- **Carnivora** (carnivores) Click on organism name to get more information.
  - **Caniformia**
    - **Ailuridae** (Red Panda family)
      - Ailurus
    - **Canidae** (dog, coyote, wolf, fox)
      - Atelocynus
      - Canis
      - Cynodictis
      - Chrysocyon
      - Cuon
      - Dusicyon
      - Lycaon
      - Nycterentes
      - Otocyon
      - Speothos
      - Urocyon
      - Vulpes
    - **Mephitidae** (skunks)
      - Conepatus
      - Mephitis
      - Mydaus
      - Spilogale
    - **Mustelidae** (weasel, mink, badger, martens and others)
      - Galictinae
      - Helictidinae
      - Lutrinae
      - Martinae
      - Melinae
      - Mellivorinae
      - Mustelinae
      - Taxidiinae
    - **Odobenidae** (walruses)
      - Odobenus
    - **Otaridae** (fur seals & sea lions)
      - Arctocephalus
      - Callorhinus
## Taxonomy of musical instruments

<table>
<thead>
<tr>
<th>Category</th>
<th>Bowed</th>
<th>Plucked</th>
<th>Struck</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Violin</td>
<td>Guitar</td>
<td>Whistle</td>
</tr>
<tr>
<td></td>
<td>Viola</td>
<td>Banjo</td>
<td>Recorder</td>
</tr>
<tr>
<td></td>
<td>Cello</td>
<td>Ukulele</td>
<td>Organ Flute Pipes</td>
</tr>
<tr>
<td></td>
<td>Contrabass</td>
<td>Harp</td>
<td>Flutes, mixtures, etc.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Harpsichord</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Harmeded Dulcimer</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Piano</td>
<td></td>
</tr>
</tbody>
</table>

### Aerophones

<table>
<thead>
<tr>
<th>Category</th>
<th>Pipe Aerophones</th>
<th>Reed Pipe Instruments</th>
<th>Brass Instruments</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>pitch determined by pipe</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>length</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Edge Instruments</td>
<td>Whistle Flutes</td>
<td>Single Reeds</td>
<td>Without Valves</td>
</tr>
<tr>
<td></td>
<td>Whistle</td>
<td>Saxophone</td>
<td>Conch shell</td>
</tr>
<tr>
<td></td>
<td>Recorder</td>
<td>Single Reed Bagpipe</td>
<td>Animal Horn</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Shofar</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Didgeridu</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Bugle</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Trombone</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Trumpet</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Cornet</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>French Horn</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Euphonium</td>
</tr>
</tbody>
</table>

### aerophones

- **Unframed Reed**
  - Wind Blown
  - Mouth Blown
  - Mouth Blown & Plucked

- **Mouth Blown**
  - *Sheng*
  - *Shō*
  - *Khaen*
  - Harmonica
  - Concertina
  - Bandonimé
<table>
<thead>
<tr>
<th>Dewey Decimal System 100s level</th>
</tr>
</thead>
<tbody>
<tr>
<td>000 Computer science, knowledge &amp; systems</td>
</tr>
<tr>
<td>010 Bibliographies</td>
</tr>
<tr>
<td>020 Library &amp; information sciences</td>
</tr>
<tr>
<td>030 Encyclopedias &amp; books of facts</td>
</tr>
<tr>
<td>040 [Unassigned]</td>
</tr>
<tr>
<td>050 Magazines, journals &amp; serials</td>
</tr>
<tr>
<td>060 Associations, organizations &amp; museums</td>
</tr>
<tr>
<td>070 News media, journalism &amp; publishing</td>
</tr>
<tr>
<td>080 Quotations</td>
</tr>
<tr>
<td>090 Manuscripts &amp; rare books</td>
</tr>
<tr>
<td>100 Philosophy</td>
</tr>
<tr>
<td>110 Metaphysics</td>
</tr>
<tr>
<td>120 Epistemology</td>
</tr>
<tr>
<td>130 Parapsychology &amp; occultism</td>
</tr>
<tr>
<td>140 Philosophical schools of thought</td>
</tr>
<tr>
<td>150 Psychology</td>
</tr>
<tr>
<td>160 Logic</td>
</tr>
<tr>
<td>170 Ethics</td>
</tr>
<tr>
<td>180 Ancient, medieval &amp; eastern philosophy/</td>
</tr>
<tr>
<td>190 Modern western philosophy</td>
</tr>
<tr>
<td>200 Religion</td>
</tr>
<tr>
<td>210 Philosophy &amp; theory of religion</td>
</tr>
<tr>
<td>220 The Bible</td>
</tr>
<tr>
<td>230 Christianity &amp; Christian theology</td>
</tr>
<tr>
<td>240 Christian practice &amp; observance</td>
</tr>
<tr>
<td>250 Christian pastoral practice &amp; religious orders</td>
</tr>
<tr>
<td>260 Christian organization, social work &amp; worship</td>
</tr>
<tr>
<td>270 History of Christianity</td>
</tr>
<tr>
<td>280 Christian denominations</td>
</tr>
<tr>
<td>290 Other religions</td>
</tr>
<tr>
<td>300 Social sciences, sociology &amp; anthropology</td>
</tr>
<tr>
<td>310 Statistics</td>
</tr>
<tr>
<td>320 Political science</td>
</tr>
<tr>
<td>330 Economics</td>
</tr>
<tr>
<td>340 Law</td>
</tr>
<tr>
<td>350 Public administration &amp; military science</td>
</tr>
<tr>
<td>360 Social problems &amp; social services</td>
</tr>
<tr>
<td>370 Education</td>
</tr>
<tr>
<td>380 Commerce, communications &amp; transportation</td>
</tr>
<tr>
<td>390 Customs, etiquette &amp; folklore</td>
</tr>
<tr>
<td>400 Language</td>
</tr>
<tr>
<td>410 Linguistics</td>
</tr>
<tr>
<td>420 English &amp; Old English languages</td>
</tr>
<tr>
<td>430 German &amp; related languages</td>
</tr>
<tr>
<td>440 French &amp; related languages</td>
</tr>
<tr>
<td>450 Italian, Romanian &amp; related languages</td>
</tr>
<tr>
<td>460 Spanish &amp; Portuguese languages</td>
</tr>
<tr>
<td>470 Latin &amp; Italic languages</td>
</tr>
<tr>
<td>480 Classical &amp; modern Greek languages</td>
</tr>
<tr>
<td>490 Other languages</td>
</tr>
<tr>
<td>500 Science</td>
</tr>
<tr>
<td>510 Mathematics</td>
</tr>
<tr>
<td>520 Astronomy</td>
</tr>
<tr>
<td>530 Physics</td>
</tr>
<tr>
<td>540 Chemistry</td>
</tr>
<tr>
<td>550 Earth sciences &amp; geology</td>
</tr>
<tr>
<td>560 Fossils &amp; prehistoric life</td>
</tr>
<tr>
<td>570 Life sciences: biology</td>
</tr>
<tr>
<td>580 Plants (Botany)</td>
</tr>
<tr>
<td>590 Animals (Zoology)</td>
</tr>
<tr>
<td>600 Technology</td>
</tr>
<tr>
<td>610 Medicine &amp; health</td>
</tr>
<tr>
<td>620 Engineering</td>
</tr>
<tr>
<td>630 Agriculture</td>
</tr>
<tr>
<td>640 Home &amp; family management</td>
</tr>
<tr>
<td>650 Management &amp; public relations</td>
</tr>
<tr>
<td>660 Chemical engineering</td>
</tr>
<tr>
<td>670 Manufacturing</td>
</tr>
<tr>
<td>680 Manufacture for specific uses</td>
</tr>
<tr>
<td>690 Building &amp; construction</td>
</tr>
<tr>
<td>700 Arts</td>
</tr>
<tr>
<td>710 Landscaping &amp; area planning</td>
</tr>
<tr>
<td>720 Architecture</td>
</tr>
<tr>
<td>730 Sculpture, ceramics &amp; metalwork</td>
</tr>
<tr>
<td>740 Drawing &amp; decorative arts</td>
</tr>
<tr>
<td>750 Painting</td>
</tr>
<tr>
<td>760 Graphic arts</td>
</tr>
<tr>
<td>770 Photography &amp; computer art</td>
</tr>
<tr>
<td>780 Music</td>
</tr>
<tr>
<td>790 Sports, games &amp; entertainment</td>
</tr>
<tr>
<td>800 Literature, rhetoric &amp; criticism</td>
</tr>
<tr>
<td>810 American literature in English</td>
</tr>
<tr>
<td>820 English &amp; Old English literatures</td>
</tr>
<tr>
<td>830 German &amp; related literatures</td>
</tr>
<tr>
<td>840 French &amp; related literatures</td>
</tr>
<tr>
<td>850 Italian, Romanian &amp; related literatures</td>
</tr>
<tr>
<td>860 Spanish &amp; Portuguese literatures</td>
</tr>
<tr>
<td>870 Latin &amp; Italic literatures</td>
</tr>
<tr>
<td>880 Classical &amp; modern Greek literatures</td>
</tr>
<tr>
<td>890 Other literatures</td>
</tr>
<tr>
<td>900 History</td>
</tr>
<tr>
<td>910 Geography &amp; travel</td>
</tr>
<tr>
<td>920 Biography &amp; genealogy</td>
</tr>
<tr>
<td>930 History of ancient world (to ca. 499)</td>
</tr>
<tr>
<td>940 History of Europe</td>
</tr>
<tr>
<td>950 History of Asia</td>
</tr>
<tr>
<td>960 History of Africa</td>
</tr>
<tr>
<td>970 History of North America</td>
</tr>
<tr>
<td>980 History of South America</td>
</tr>
<tr>
<td>990 History of other areas</td>
</tr>
</tbody>
</table>
The following table provides detailed information on the structure of NAICS. Also included, on this page, are downloadable links to the complete NAICS structure.

<table>
<thead>
<tr>
<th>Sector</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>Agriculture, Forestry, Fishing and Hunting</td>
</tr>
<tr>
<td>21</td>
<td>Mining, Quarrying, and Oil and Gas Extraction</td>
</tr>
<tr>
<td>22</td>
<td>Utilities</td>
</tr>
<tr>
<td>23</td>
<td>Construction</td>
</tr>
<tr>
<td>31-33</td>
<td>Manufacturing</td>
</tr>
<tr>
<td>42</td>
<td>Wholesale Trade</td>
</tr>
<tr>
<td>44-45</td>
<td>Retail Trade</td>
</tr>
<tr>
<td>48-49</td>
<td>Transportation and Warehousing</td>
</tr>
<tr>
<td>51</td>
<td>Information</td>
</tr>
<tr>
<td>52</td>
<td>Finance and Insurance</td>
</tr>
<tr>
<td>53</td>
<td>Real Estate and Rental and Leasing</td>
</tr>
<tr>
<td>54</td>
<td>Professional, Scientific, and Technical Services</td>
</tr>
<tr>
<td>55</td>
<td>Management of Companies and Enterprises</td>
</tr>
<tr>
<td>56</td>
<td>Administrative and Support and Waste Management and Remediation Services</td>
</tr>
<tr>
<td>61</td>
<td>Educational Services</td>
</tr>
<tr>
<td>62</td>
<td>Health Care and Social Assistance</td>
</tr>
<tr>
<td>71</td>
<td>Arts, Entertainment, and Recreation</td>
</tr>
<tr>
<td>72</td>
<td>Accommodation and Food Services</td>
</tr>
<tr>
<td>81</td>
<td>Other Services (except Public Administration)</td>
</tr>
<tr>
<td>92</td>
<td>Public Administration</td>
</tr>
<tr>
<td>Category</td>
<td>Subcategories</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------------------------------------------------------------------------</td>
</tr>
<tr>
<td><strong>Arts</strong></td>
<td>Movies, Television, Music</td>
</tr>
<tr>
<td><strong>Games</strong></td>
<td>Video Games, RPGs, Gambling</td>
</tr>
<tr>
<td><strong>Kids and Teens</strong></td>
<td>Arts, School Time, Teen Life</td>
</tr>
<tr>
<td><strong>Reference</strong></td>
<td>Maps, Education, Libraries</td>
</tr>
<tr>
<td><strong>Shopping</strong></td>
<td>Clothing, Food, Gifts</td>
</tr>
<tr>
<td><strong>World</strong></td>
<td>Català, Česky, Dansk, Deutsch, Español, Esperanto, Français, Galego, Hrvatski, Italiano, Lietuvių, Magyar, Nederlands, Norsk, Polski, Português, Română, Slovensky, Suomi, Svenska, Türkçe, Български, Ellinikó, Русский, Українська, ייִדיש, العربية, ไทย, 日本語, 简体中文, 繁體中文</td>
</tr>
<tr>
<td><strong>Business</strong></td>
<td>Jobs, Real Estate, Investing</td>
</tr>
<tr>
<td><strong>Health</strong></td>
<td>Fitness, Medicine, Alternative</td>
</tr>
<tr>
<td><strong>News</strong></td>
<td>Media, Newspapers, Weather</td>
</tr>
<tr>
<td><strong>Regional</strong></td>
<td>US, Canada, UK, Europe</td>
</tr>
<tr>
<td><strong>Society</strong></td>
<td>People, Religion, Issues</td>
</tr>
<tr>
<td><strong>Computers</strong></td>
<td>Internet, Software, Hardware</td>
</tr>
<tr>
<td><strong>Home</strong></td>
<td>Family, Consumers, Cooking</td>
</tr>
<tr>
<td><strong>Recreation</strong></td>
<td>Travel, Food, Outdoors, Humor</td>
</tr>
<tr>
<td><strong>Science</strong></td>
<td>Biology, Psychology, Physics</td>
</tr>
<tr>
<td><strong>Sports</strong></td>
<td>Baseball, Soccer, Basketball</td>
</tr>
</tbody>
</table>
Hierarchy purposes

1. Serving users who are browsing, exploring, discovering, not searching, to whom the hierarchy is displayed.
   - Don’t even have to know the first word or few letters, as in alphabetical browsing

2. Instructing users on classification

3. Enabling “recursive”/“rolled up” retrieval results
   (A term retrieves what is indexed to it and what is indexed to each one of its narrower terms, all together.)
Hierarchical taxonomy is suitable for:

- Content that is naturally categorizable: products, industries, government agencies, academic disciplines, scientific things, technologies
- Small and medium-sized taxonomies. Very large taxonomies are not so suitable.
- End-user browse displays.
- Certain kinds of auto-categorization, which puts documents into approximate categories.
Issues in creating hierarchies

- Number of levels and the number of terms per level
- Arrangement of the term hierarchy
- Named entities within subject hierarchies
- Numeric code usage
- Term names – In context of broader term or unambiguous
- Top term usage – As indexing terms or just labels
- Recursive retrieval – Broader term additionally retrieves content indexed with all narrower terms or not
Deep hierarchy: Many levels

Geographies
- North America
  -- United States
  --- New England
  ---- Massachusetts
  ----- Boston
  ------ North End
  -------- Old North Church

- South America
- Europe
- Asia
  -- Central Asia
  -- Middle East

- Africa
- Oceania
  -- Southeast Asia

Broad hierarchy: Many terms per level

Geographies
- U.S. cities
  -- Albuquerque
- U.S. States
  -- Alabama
- Countries
- World cities
- Continents
- Landmarks
Deep: Many levels (NAICS style with 10-20 upper level terms)

Industries
- Transportation services
  -- Air transportation
  --- Schedule air transportation services
  ---- Scheduled air freight transportation services

Broad: Many terms per level (job search sites, 50 - 80 terms per level)

Industries
- Accounting/Auditing
- Administrative Support Services
- Advertising/Marketing/Public Relations
- Aerospace/Aviation/Defense
- Agriculture, Forestry, & Fishing
- Airlines

Second levels at select terms only: Healthcare, Sales
Deep vs. broad decision factors

- Display interface/horizontal and vertical space
- Speed of displaying deeper levels
- User needs, and expectations
  - Industry experts, internal employees, general public, students, etc.

Need to balance how much can be easily skimmed in one view vs. how many levels down the user has patience to click down through.

More levels lead to less consistency across levels.
Arrangement of term hierarchy

Decision: What’s the best method to handle different means of classification within the same hierarchy?

- Industries by traditional SIC/NAICS classification or by vertical market
- Products by manufacturing technology or by end-use
- Places by physical geographic location or by type
- Organizations by goals/objectives or by political/religious affiliation
- Government agencies by type or by country/state of affiliation

Even within facets, there often are hierarchies.

Even if polyhierarchies are allowed, a top-level classification is still needed. Too many polyhierarchies can be confusing.
1. Governmental bodies & agencies
   - U.S. governmental bodies & agencies
     -- U.S. Courts
     -- U.S. executive branch agencies
     -- U.S. legislative branch
     -- State bodies & agencies
   - Foreign governmental bodies & agencies
     -- Foreign courts
     -- Foreign legislatures
     -- Foreign national agencies
     -- Foreign state & provincial government agencies

2. Governmental bodies & agencies
   -- Foreign legislatures (+ instances)
   -- U.S. legislatures (+ US federal and state instances)

3. Governmental bodies & agencies
   - Legislative bodies
     -- National legislatures (+ instances, both foreign and US)
     -- State & provincial legislatures (+ all instances alphabetical for US and foreign)

4. Governmental bodies & agencies
   - Legislative bodies (+ all instances, US and foreign, in one alphabetical list)
Hierarchy arrangement decision factors

- User audience needs and expectations
  - How the users classify the subject matter
  - Whether users will browse the hierarchy or rather use a search box to search on a given topic

- Support for polyhierarchies

- Permissibility of terms as category labels, not linked to content, at various intermediate levels within the hierarchy
  - e.g. Foreign legislatures

Need to consider

- Whether to create terms that are difficult to distinguish in indexing
  - e.g. both Legislative bodies and National legislatures
Named entities in hierarchies
As separate hierarchies or integrated into the appropriate subjects.
If integrated, linked out at what level, lowest, or at a higher level?

Heads of state & government ← All heads of state here, or
- Presidents ← names of presidents here and
- Prime ministers ← names of prime ministers here
Named entities separate out in hierarchies:
Numeric codes in hierarchies

Advantages:
- Supports more precise indexing and retrieval
- Enables faster/more efficient indexing with memorized codes
- Preference of technical subject matter experts (e.g. engineers)

Disadvantages:
- Limited ability for revising
- More complex to implement multiple-field terms
- Inflexibility
Term names in context or unambiguous

In context:
French cinema
  Comedy films
  Documentary films
  Historical films
  Suspense films

Unambiguous:
French cinema
  French comedy films
  French documentary films
  French historical films
  French suspense films

Easier for user browsing
Easier to manage in thesaurus management software
Top term usage for indexing or just a label

**Fictional work**
- NT Folktale
- NT Novel
- NT Novella
- NT Play
- NT Poem
- NT Short story

**Information technology occupations**
- NT Computer consultant
- NT Computer engineer
- NT Computer imaging specialist
- NT Computer scientist
- NT Information technology manager
- NT Information technology specialist
- NT Software developer
- NT Systems analyst

**Health care**
- NT Anatomy and physiology
- NT Clinical procedures
- NT Diagnosis
- NT Diseases and disorders
- NT Drugs
- NT Health care profession
- NT Health care services
- NT Medical equipment and supplies
- NT Medical specialties
- NT Patients
- NT Personal health
- NT Treatments

Examples of Broad Level Access Terms not used for indexing, from Cengage Learning vocabularies
Hierarchies can use “facet indicators” (node labels) for categorizing terms at the same hierarchical level.

Similar, but not exactly the same as “facets” for faceted classification.

Categorizing by subject, then by facet-aspect.

Example:

Automobiles
  by size
  Compact automobiles
  Mid-size automobiles
  Full-size automobiles
    by body type
    Coupes
    Sedans
    Station wagons
    Minivans
  Sport-utility vehicles
    by engine type
    Diesel engine automobiles
    Electric automobiles
    Natural gas engine automobiles
    Alcohol engine automobiles
Recursive or “rolled up” retrieval

Does a term for which there are narrower terms retrieve the sum or all of what is indexed with all of the narrower terms?

- If not, then its meaning is the broader topic “in general.”
- If yes, then an additional term for “in general” might need to be created.
Polyhierarchies
Sometimes a term can have two or more broader terms.

- Polyhierarchy is permitted if the hierarchical relationship is valid in both/all cases
- Remember “All-and-Some” test for each generic hierarchical relationship
- Systems may or may not support it.
Polyhierarchy

Hierarchy → Polyhierarchy
Polyhierarchy based on generic relationship

Structural Design: Hierarchies

Professions
- Musicians
- Educators
  - Music Teachers

Motor vehicles
- Cars
- Trucks
  - Light trucks
Polyhierarchy based on different kinds of hierarchical relationships/different means of categorizing (in thesauri, but less common in taxonomies)

- Bodies of Water
  - Lakes
    - Great Salt Lake
  - United States
    - Utah
Pluses

Polyhierarchy is useful when...

- It is obviously logical for select terms (cross-overs/hybrids, e.g. Music teachers or Light Trucks)
- It is indicated by different stakeholder views
- Indexers/taggers browse the taxonomy
- End-user testing/input (e.g. card-sorting) indicates users are split as to where in the hierarchy an item belongs
- It’s a thesaurus (focused on term relationships)
Retail website polyhierarchy example:

Health & Fitness
  › Portable Fitness Electronics
    › Fitness GPS Watches

Car, Marine & GPS
  › GPS Navigation
    › Handheld GPS
      › Fitness GPS Watches

Sports taxonomy polyhierarchy example:

Back exercises
  › Dead lifts

Hamstring exercises
  › Dead lifts
Minuses

Polyhierarchy is *not* so good when…

- It violates hierarchical relationship standards
- It becomes excessive, perhaps more common than monohierarchies
- It is the result of different kinds of a categorization, and the presence of different kinds of categorization is confusing
- It is a small taxonomy and the user doesn’t need or expect polyhierarchy
Problems with excessive polyhierarchies

- Familiar tree structure is lost.
- Users cannot see the logical hierarchy.
- Users spend too much time clicking through categories.
Logical polyhierarchies, if done consistently, could become too extensive in a hierarchical taxonomy (but OK in a thesaurus).

Example: creating polyhierarchies for products based on different classifications
Do not create a polyhierarchy to both a “parent” \textit{and a “grandparent.”}

\begin{itemize}
  \item \textbf{C}ameras
  \item \textbf{D}igital \textbf{C}ameras
  \item \textbf{D}igital \textbf{S}LR \textbf{C}ameras
\end{itemize}

\textit{Grandparent} of Digital SLR Cameras

\textit{Parent} of Digital SLR Cameras
Might be better not to have polyhierarchies when the taxonomy is small and the number of top-level categories are few.

Example: Client management documents of a financial services company has 114 topical terms categorized with just five broader terms:

- Account Information
- Client Information
- Client Status
- Disclosures & Notifications
- Approvals/Guidance

Decided against polyhierarchies.

Reason: Repeat users can memorize the small hierarchy. They don’t expect polyhierarchy here.
Polyhierarchy Summary:

- Some is good. More isn’t necessarily better.
- Polyhierarchies are best for isolated terms that can fall into two categories.
- Polyhierarchies can become too many in cases of overlays of two different categorization methods for numerous terms. (Facets may be better.)
- Polyhierarchies are useful, no matter how extensive, in term-focused thesauri.
- Polyhierarchies should be more limited in fully displayed hierarchical taxonomies.
Exercise: Polyhierarchies

Suggest two (or more) broader terms for each term within the same taxonomy:

- Hotel managers
- Printers
- Fish
- Egypt
Facets

- For serving faceted classification, which allows the assignment of multiple classifications to an object
- A “dimension” of a query; a type of concept
- Intended for searching with multiple terms in combination (post-coordination), one from each facet
- A refinement, filter, limit by, narrow by
- Can be for topics or for named entities, but generally not both
- Reflect the domain of content
- Similar, but not identical to an authority file
Example types of facets:

- For Products: *name*, *series number*, *category*, *size*, *color*, *price*
- For People: *name*, *job title*, *gender*, *birth year*, *location*, *dept.*
- For Reports: *title*, *author*, *subject*, *audience*, *document type*

Facets are suitable for:

- Structured data with discernable metadata fields or database records
- Homogeneous data with similar types of characteristics (e.g. products in an e-commerce site)
Additional examples

For enterprise taxonomies:
Patrick Lambe, *Organising Knowledge*
- People and organizations
- Things and parts of things
- Activity cycles
- Locations

For Web sites:
Rosenfeld and Morville, *Information Architecture*
- Topic
- Product
- Document type
- Audience
- Geography
- Price
Faceted Classification History

Mathematician/librarian S.R. Ranganathan (1920s) developed as an alternative to the Dewey Decimal System for books: “Colon Classification”

1. *Personality* – topic or orientation
2. *Matter* – things or materials
3. *Energy* – actions
4. *Space* – places or locations
5. *Time* – times or time periods
Ecommerce Facets

Narrow Selection By:
- Size Range
  - Show All
  - (Reg, Plus, Slim, Big & Tall, etc.)
- Specific Size
  - Show All
- Color
  - Show All
- Sleeve Length
  - Show All
- Fabric
  - Show All
- Style
  - Show All
  - All Others (56)
  - Casual (51)
  - Tailored (11)
  - Buttondown (6)
  - Split Neck (3)
Advantages

- Supports more complex search queries by users
- Allows users to control the search refinement, narrowing or broadening in any manner or order
- Familiar to novice users; suitable for expert users

Disadvantages

- Only suitable for somewhat structured, unified type of content that all share the same multiple facets
- May not support “advanced search” of multiple terms selected at once (“or”) from the same facet
- Requires investment of thorough indexing/tagging
Placement of certain terms may not be obvious

- Institutions could be Places or Organizations: Educational institutions, Museums, Libraries, Places of worship
- Business activities could be Actions or Topics: Acquisitions, Contracts, Joint ventures, Sales

Deciding in which facet to put these terms

- Whether two (parenthetically modified) terms for the concept should be created, one for each facet, e.g. **Hotels (buildings)** and **Hotels (companies)**
- Whether a term can be in more than one facet
  - Can systems and search implementations distinguish between two?
  - Automated indexing may not distinguish between different facet meanings of a term: e.g. action or topic.
Facet Design Tips

- Number of facets: 3-8, with 5-6 as ideal
- Facets listed in logical, not alphabetical order
- Number of terms per facet: 2-25
  - Ideally not much more than can be viewed in a scroll box
  - If the list is obvious (US states), then more is OK.
  - Exception can be made for hierarchical “Topics” facet
- If <12 terms, then a logical display order
  If >12 terms, then alphabetical order
- A two-level hierarchy (indented) within a facet is possible
Combined hierarchies and facets

When facets are useful, but topics are too many and are best displayed in a hierarchy

1. Start with facets, and then have hierarchies within facets
2. Start with hierarchical categories, then refine further with facets
1. Hierarchies within facets
   Hierarchies indented with the facet
   World Bank documents advanced search
2. Start with hierarchical categories, then limit with facets

Target [www.target.com](http://www.target.com)
Buzzillions product reviews [www.buzzillions.com](http://www.buzzillions.com)
Start with hierarchical categories and have facets available from the top level Amazon.com (shop by Department) www.amazon.com/gp/site-directory/ref=nav_shopall_btn
Exercise: Create Facets

Designate a set of 4-7 facets for a tour operator web site selling vacation packages.
Outline

1. Introduction, Definitions, and Types
2. Creating & Wording of Terms
3. Nonpreferred Terms
4. Creating Term Relationships
5. Structural Design: Hierarchies and Facets
6. Display Options
7. Software Tools
8. Sources for Terms
9. Project Process & Management
10. Resources
There are multiple ways to display a controlled vocabulary.

1. Thesaurus display options
2. Sorting display options
3. Hierarchy end-user display options
4. Facet end-user display options

- A thesaurus, with non-preferred terms and term details, may have different display options: alphabetical, hierarchical, and variations of each.
- A hierarchical taxonomy (without non-preferred terms) is better left hierarchical and not sorted alphabetically.
Display Options: Thesaurus Displays

Standard thesaurus A-Z display options:

- Simple list with or without non-preferred terms
- With all immediate relationships (flat format) listed at each term: UF, BT, NT, RT
- Full term hierarchy (multilevel hierarchy) displayed at each term: BT1, BT2; NT1, NT2
- Top Term – alphabetical listing of highest level terms with hierarchies under each
- Permuted/Rotated - lists each term multiple times for each word within the term
Flat format – alphabetical list of terms, listing all immediate relationships at each term (most common display)

**Preferred term name**
- **SN**: Scope note text
- **UF**: Non-preferred term
- **BT**: Broader term
- **NT**: A narrower term
- **NT**: Another narrower term
- **RT**: A related term
- **RT**: Another related term
Full term Hierarchy Displays

Multi-Level Hierarchy

**Recreation facilities**
- UF: Recreation centers
- UF: Recreational facilities
- NT1: Amusement parks
- NT1: Athletic facilities
  - NT2: Fitness centers & gyms
  - NT2: Skating rinks
  - NT2: Swimming pools
  - NT2: Tennis courts
- NT1: Bowling alleys
- NT1: Golf courses
- NT1: Skiing facilities
- NT1: Stadiums & arenas
- BT1: Public buildings & facilities
  - BT2: Facilities & Infrastructure

Two-Way Hierarchy

: Facilities & Infrastructure
  : Public buildings & facilities
**Recreation facilities**
  . Amusement parks
  . Athletic facilities
    . Fitness centers & gyms
    . Skating rinks
    . Swimming pools
    . Tennis courts
  . Bowling alleys
  . Golf courses
  . Skiing facilities
  . Stadiums & arenas
Display Options: Thesaurus Displays

Top Term – alphabetical listing of highest level terms with hierarchies under each

In a hierarchical taxonomy: only one top-term and one complete hierarchy

Politics & Government
  . Domestic policy & programs
  .. Agricultural policy
  .. Economic policy
  ... Monetary policy
  ... Fiscal policy
  .. Energy policy
  .. Health policy & programs
  .. Social policy
  . Foreign policy
  .. Appeasement
  .. Bilateralism
  .. Foreign assistance
  ... Foreign military assistance
  .. Foreign intervention
  .. Unilateralism

In a flatter thesaurus: multiple top terms, with multiple small hierarchies

Political actions
  . Campaigning
  . Lobbying
  . Political protests

Political ideologies
  . Conservatism
  . Environmentalism
  . Liberalism
  .. Socialism

Politicians
  . Political office holders
  . Political candidates
Display Options: Thesaurus Displays

Permuted/Rotated Index
Alphabetical list of (key) words within terms.
Lists each term multiple times for each word within the term
1. Keyword in context (KWIC)
2. Keyword out of context (KWOC)

- Useful for printed thesauri.
- Unneeded for searchable thesauri in electronic format.
KWIC:
NASA Thesaurus
(Volume 2: Rotated Display)
[Link to PDF file]
**KWOC:**
Maternal and Child Healthcare Thesaurus
(Rotated Term List)
www.mchthesaurus.info/Rotated/Acontent.html
Display Options: Term Sorting Displays

Display of terms

Alphabetical sort

Personnel management
- Hiring
- Letting go of employees
- Promotions
- Recruiting
- Training

Logical sort

Personnel management
- Recruiting
- Hiring
- Training
- Promotions
- Letting go of employees
Hierarchy end-user display options

One level per web page

- Dex Media Superpages
- Amazon.com (shop by Department)

Expandable tree

Fly-out subcategories
One level per web page

Advantages:
- Large number of terms can display at each level
- Good for large volumes of content
- Works well with polyhierarchies
- Technically easy to implement

Disadvantages:
- Display may take entire screen width, compromising content display
- Users see only one level at a time
- Less appropriate for taxonomies with varied/inconsistent levels or levels containing just one or a few terms
- May lack method to indicate whether a level is final or contains more lower levels
Expandable tree

USA Today

http://content.usatoday.com/community/tags/topic-index.aspx
Expandable tree

Advantages:
- Allows the user to explore/expand multiple subcategories simultaneously
- Presence of plus signs indicates presence of deeper levels
- Accommodates inconsistent numbers of terms per level
- Display takes up only part of screen
- Good for hierarchically deep taxonomies

Disadvantages:
- Insufficient for displaying very large taxonomies or large numbers of terms at the same level
- May have system constraints, such as single-word labels
- More complex to develop and may take more time to load display. Not good for the web. Usually for internal systems.
Fly-out subcategories:
Lynda.com video courses [www.lynda.com](http://www.lynda.com)
Books & Authors http://books.wiseto.com

Know What You're Going to Read Next?
Looking to curl up with a good book, but finding yourself stuck on the question of "what do I read next"? Let Books and Authors help with our expert title recommendations, reading lists, and more. Find out more in our MONTHLY HIGHLIGHTS.

Seasonal Suggestions
Libraries unite this year at the ALA Midwinter Meeting

Spotlight On
Check out these inspirational tales for Black History Month

Best Sellers
U.S. and Canadian Bestsellers based on domestic sales.

Award Winners
The 2015 Newbery Medal went to Kwame Alexander.
Display Options: Hierarchy Displays

*PLOS ONE* scholarly online journal [www.plosone.org](http://www.plosone.org)
Fly-out subcategories

Advantages:
- Users can visualize the hierarchy (see both/multiple levels) at the same time, while also seeing additional page content

Disadvantages:
- Not so suitable for multiple levels or large taxonomies
Display Options: Facets

NARROW SELECTION BY:
- Size Range
  - Show All
- Specific Size
  - Show All
- Color
  - Show All
- Sleeve Length
  - Show All
- Fabric
  - Show All
- Style
  - Show All
- Featured Collection
  - Show All

narrow by:
- Category
  - Banquet Table (3)
  - Bistro Table (9)
  - Counter-height Table (7)
  - Dinette Table (2)
  - Dining Table (66)
  - Kitchen Table (4)
  - Nook Table (2)
  - Pub Table (31)
- Shape
  - Novelty (1)
  - Oval (2)
  - Rectangle (47)
  - Round (53)
  - Square (21)
- Finish
  - Chrome (10)
  - Distressed (7)
  - Ebony (6)
  - Espresso (20)
  - Lacquered (9)

Refine By
- Price Range
  - $9 to $20
- Characters
  - All Characters
- Clothing Sizes
  - All Clothing Sizes
- Sleep Body Style
  - All Sleep Body Style
- Colors

Narrow Your Results
- Price
  - $200 - $249.99 (2)
  - $250 - $299.99 (8)
  - $300 - $349.99 (19)
  - $350 - $399.99 (17)
  - $400 - $449.99 (1)
  - $450 - $499.99 (2)
- Screen Size
  - 17" and Up (1)
  - 15" - 18" (9)
  - 14" and Under (37)
- Processor Brand
  - AMD (4)
  - Intel® (45)
- System Memory RAM
  - 1GB (1)
  - 2GB (4)
  - 4GB (31)
  - 6GB (8)
  - 8GB (4)
  - 16GB (1)
- Laptop Features
  - Touchscreen (4)
  - Backlit Keyboard (13)
  - Built-in Webcam (50)
  - Blu-ray (2)
  - HDMI Output (44)
  - Wireless Capability (5)
  - Wireless Display (15)
  - Ultraportable (40)
  - ENERGY STAR Certified (41)
  - 3D-Ready (1)
- Operating Platform
  - Windows (40)
Shoebuy retail site: [http://www.shoebuy.com](http://www.shoebuy.com)
This report explores the connection between geomicrobiology and astrobiology, explaining NASA’s interest in the natural history of Yellowstone National Park. Multi-institution research teams are...

This site describes how Yellowstone National Park is a focal point for cutting-edge microbiology research and how it provides a valuable setting for outreach education. Topics include questions for...

This article is a compilation of information about free-living eukaryotes in extreme environments. Written in summary form, it includes anaerobes, thermophiles, psychrophiles, acidophiles, ...

This is a survey reporting the phylogenetic diversity of Mono Lake bacteria as conducted by the Mono Lake Microbial Observatory. Samples were collected from different layers of the lake and analyzed...

This article describes a permafrost subglacial lake discovered beneath Antarctica. The lake offers scientists a chance to test their sterile drilling techniques before exploring elsewhere in search...
Sea Vent Viewer part of SERC Web Resource Collection
http://www.nsf.gov/news/overview/earth-environ/interact...
This website serves as an educational overview of National Science Foundation (NSF) earth and environmental science research focusing on hydrothermal vent systems. It features an interactive viewer ...

We're in Hot Water Now: Hydrothermal Vents part of SERC Web Resource Collection
http://www.nationalgeographic.com/xpeditions/lessons/07/...
In this National Geographic lesson, students will use National Geographic's Yellowstone internet module to learn about the processes that drive geysers. The activity involves learning about ...

Narrow the View ▼
Subject: Biology Show all Subject: Biology
Biology > Ecology > Habitats
Marine 7 matches
Benthic 7 matches

Resource Type Show all Resource Type
Audio/Visual > Images/Illustrations
7 matches

Extreme Environments: High Pressure, Extremely Hot
Ocean Environments: Deep Sea Floor/Abyssal: Hydrothermal Systems
Grade Level: Primary (K-2), Intermediate (3-5), College Lower (13-14), High School (9-12), Middle (6-8), Graduate/Professional, College Upper (15-16), General Public, Informal

Extreme Environments: Extremely Hot, High Pressure
Ocean Environments: Deep Sea Floor/Abyssal: Hydrothermal Systems
Grade Level: Middle (6-8)
Facet design options and issues

- Facets as post-search filters, as initial browse, or both
- Display of partial/start of list of values within a facet
- Ability to select multiple values from within the same facet at once (with check boxes)
- Inclusion of other metadata (not “taxonomy”) in the same set of facets (date, author, document type, etc.)
- Have all generic facets or also category-specific facets
- Designating kinds of topics into different facets
Outline

1. Introduction, Definitions, and Types
2. Creating & Wording of Terms
3. Nonpreferred Terms
4. Creating Term Relationships
5. Structural Design: Hierarchies and Facets
6. Display Options
7. Software Tools
8. Sources for Terms
9. Project Process & Management
10. Resources
Software tools for creating and editing thesauri and taxonomies

- Taxonomies, thesaurus, or controlled vocabulary: the same software is used; choose the types of relationships to implement.
- Simple controlled vocabulary (synonym ring) or simple hierarchical taxonomy could be created on paper (or Word or Excel), but combined features of UF, RT, BT/NT, and scope notes, attributes etc., require special software to maintain.
- They typically enforce standards, but not all the same standards: ANSI/NISO Z39.19, or SKOS/RDF, or records management.
Software feature of enforcing standards

- Software support of ANSI/NISO Z39.19 includes:
  - Preferred terms must be unique (no duplicate terms)
  - A nonpreferred term (synonym) can point to only one preferred term
  - A pair of terms can be either hierarchically (BT/NT) associatively (RT) related to each other but not both.
  - Hierarchical relationships extend:
    Term A is narrower to Term B, and Term B is narrower to Term C, Term C cannot be narrower to Term A.
Software Types

- Spreadsheet software (Excel)
- Thesaurus software: Single-user Windows PC desktop software
- Thesaurus software: Larger-scale, multi-user client-server or Web-based systems
- Taxonomy creation & editing module of a content management, document management, digital asset management, collaborative software (SharePoint)
- Vertical market software for creating classification structures
- Proprietary “home-grown” programs developed in-house in large organizations
Using spreadsheet software (Excel) for a simple hierarchy

- For BT/NT relationships only (not RTs)
- For very limited nonpreferred term ideas, not complete
- Not more than 4-5 levels deep
- For small Web site navigation taxonomies or for starter top levels of larger taxonomies
- Individual worksheets can be saved and imported into other systems, including thesaurus software
### Software

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Level 1</td>
<td>Level 2</td>
<td>Level 3</td>
<td>Level 4</td>
<td>Level 5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Diseases and disorders</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Age related diseases</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Alzheimer’s disease</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Cataracts</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Gout</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Osteoarthritis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Osteoporosis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Sarcopenia</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Senile dementia</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Dehydration</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Hyponatremia</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Metabolic syndrome X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Water intoxication</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Cardiovascular diseases and disorders</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Aneurysm</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>Arteriosclerosis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>Atherosclerosis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>Coronary artery disease</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>Heart diseases and disorders</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>Angina pectoris</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>Arrhythmias</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>Atrial arrhythmias</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>Heart block</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>Ventricular arrhythmias</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>Athlete heart syndrome</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>Cardiomyopathy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>Carditis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>coronary artery disease</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Thesaurus software should:

- Maintain terms and their relationships (UF, BT/NT, RT)
  - As reciprocals
  - When renaming, merging, subsuming, or deleting terms
- Disallow invalid relationships (according to standards)
- Create scope notes and attributes for terms
- Support candidate and approved terms; include term creation and update dates
- Generate reports in various thesaurus display formats
- Export data in format for importing into a content indexing/search/retrieval system: CSV, Excel, HTML, XML (including schemas of ZThes, RDF, SKOS, and OWL)
Thesaurus software points of comparisons:

- interface design and ease of use
- multiple taxonomy display options
- term searching
- spell-checking
- speed (limited mouse clicks) for repeated term and relationship additions
- single-step new term & relationship creation
- single-step branch (term and narrower terms) moving
- drag & drop relationship adding
- user-defined (customizable) relationships
- user-defined term notes and term attributes
- bilingual or multilingual taxonomy support
- importing and exporting formats
- connectors to enterprise search systems
MultiTes Pro
Multisystems (Miami, FL)

www.multites.com

- Single product independent vendor since 1983.
- $295 single user; $1295 for 5 users
  $2495 for 10 users; $3950 enterprise deployment
- Web/cloud-based option: $4950/year per thesaurus for 20 accounts
- Add-on products: web development kit, enterprise development kit
- Free 30-day downloadable trial and online tutorial
  www.multites.com/download.htm
<table>
<thead>
<tr>
<th>Term</th>
<th>Status</th>
<th>Flag</th>
<th>Relationships</th>
<th>Notes</th>
<th>Categories</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Account Management</td>
<td>A</td>
<td>USE(1)</td>
<td>USE(1)</td>
<td>-</td>
<td>-</td>
<td>ENG</td>
</tr>
<tr>
<td>Accountability</td>
<td>A</td>
<td>USE(1)</td>
<td>USE(1)</td>
<td>-</td>
<td>-</td>
<td>ENG</td>
</tr>
<tr>
<td>Acquisitions</td>
<td>A</td>
<td>USE(1)</td>
<td>USE(1)</td>
<td>-</td>
<td>-</td>
<td>ENG</td>
</tr>
<tr>
<td>Actions</td>
<td>A</td>
<td>USE(1)</td>
<td>USE(1)</td>
<td>-</td>
<td>-</td>
<td>ENG</td>
</tr>
<tr>
<td>Adaptable</td>
<td>A</td>
<td>USE(1)</td>
<td>USE(1)</td>
<td>-</td>
<td>-</td>
<td>ENG</td>
</tr>
<tr>
<td>Advancing</td>
<td>A</td>
<td>USE(1)</td>
<td>USE(1)</td>
<td>-</td>
<td>-</td>
<td>ENG</td>
</tr>
<tr>
<td>Agility</td>
<td>A</td>
<td>USE(1)</td>
<td>USE(1)</td>
<td>-</td>
<td>-</td>
<td>ENG</td>
</tr>
<tr>
<td>Alliances</td>
<td>A</td>
<td>USE(1)</td>
<td>USE(1)</td>
<td>-</td>
<td>-</td>
<td>ENG</td>
</tr>
<tr>
<td>Anticipating Change</td>
<td>A</td>
<td>USE(1)</td>
<td>USE(1)</td>
<td>-</td>
<td>-</td>
<td>ENG</td>
</tr>
<tr>
<td>Art of the Turnaround</td>
<td>A</td>
<td>USE(1)</td>
<td>USE(1)</td>
<td>-</td>
<td>-</td>
<td>ENG</td>
</tr>
<tr>
<td>Articulating Your Message</td>
<td>A</td>
<td>USE(1)</td>
<td>USE(1)</td>
<td>-</td>
<td>-</td>
<td>ENG</td>
</tr>
<tr>
<td>Assessing Performance</td>
<td>A</td>
<td>USE(1)</td>
<td>USE(1)</td>
<td>-</td>
<td>-</td>
<td>ENG</td>
</tr>
<tr>
<td>Automation</td>
<td>A</td>
<td>USE(1)</td>
<td>USE(1)</td>
<td>-</td>
<td>-</td>
<td>ENG</td>
</tr>
<tr>
<td>Behavior of Leaders</td>
<td>A</td>
<td>USE(1)</td>
<td>USE(1)</td>
<td>-</td>
<td>-</td>
<td>ENG</td>
</tr>
<tr>
<td>Being Accountable</td>
<td>A</td>
<td>USE(1)</td>
<td>USE(1)</td>
<td>-</td>
<td>-</td>
<td>ENG</td>
</tr>
<tr>
<td>Borrowing</td>
<td>A</td>
<td>USE(1)</td>
<td>USE(1)</td>
<td>-</td>
<td>-</td>
<td>ENG</td>
</tr>
<tr>
<td>Brand</td>
<td>A</td>
<td>USE(1)</td>
<td>USE(1)</td>
<td>-</td>
<td>-</td>
<td>ENG</td>
</tr>
<tr>
<td>Brand Management</td>
<td>A</td>
<td>USE(1)</td>
<td>USE(1)</td>
<td>-</td>
<td>-</td>
<td>ENG</td>
</tr>
<tr>
<td>Brand strategy</td>
<td>A</td>
<td>USE(1)</td>
<td>USE(1)</td>
<td>-</td>
<td>-</td>
<td>ENG</td>
</tr>
<tr>
<td>Branding</td>
<td>A</td>
<td>USE(1)</td>
<td>USE(1)</td>
<td>-</td>
<td>-</td>
<td>ENG</td>
</tr>
<tr>
<td>Brands</td>
<td>A</td>
<td>USE(1)</td>
<td>USE(1)</td>
<td>-</td>
<td>-</td>
<td>ENG</td>
</tr>
<tr>
<td>Budget</td>
<td>A</td>
<td>USE(1)</td>
<td>USE(1)</td>
<td>-</td>
<td>-</td>
<td>ENG</td>
</tr>
<tr>
<td>Budget Management</td>
<td>A</td>
<td>USE(1)</td>
<td>USE(1)</td>
<td>-</td>
<td>-</td>
<td>ENG</td>
</tr>
<tr>
<td>Budget Planning</td>
<td>A</td>
<td>USE(1)</td>
<td>USE(1)</td>
<td>-</td>
<td>-</td>
<td>ENG</td>
</tr>
<tr>
<td>Budget Strategy</td>
<td>A</td>
<td>USE(1)</td>
<td>USE(1)</td>
<td>-</td>
<td>-</td>
<td>ENG</td>
</tr>
<tr>
<td>Budgeting</td>
<td>A</td>
<td>USE(1)</td>
<td>USE(1)</td>
<td>-</td>
<td>-</td>
<td>ENG</td>
</tr>
<tr>
<td>Budgets</td>
<td>A</td>
<td>USE(1)</td>
<td>USE(1)</td>
<td>-</td>
<td>-</td>
<td>ENG</td>
</tr>
<tr>
<td>Budgeting &amp; Costs</td>
<td>A</td>
<td>USE(1)</td>
<td>USE(1)</td>
<td>-</td>
<td>-</td>
<td>ENG</td>
</tr>
<tr>
<td>Budgeting &amp; Costs</td>
<td>A</td>
<td>USE(1)</td>
<td>USE(1)</td>
<td>-</td>
<td>-</td>
<td>ENG</td>
</tr>
</tbody>
</table>

**Budgeting & Costs**

- **General**
- **Classification**
  - FAC: Theme Theme
- **Notes**
  - SN: Managing expenses, costs, and budgets
- **Relationships**
  - UF: Budget
  - UF: Budget Management
  - UF: Budget Planning
  - UF: Budget Strategy
  - UF: Budgeting
  - UF: Budgets
  - UF: Cost Cutting
  - UF: Costs
  - UF: Expenses
  - BT: Finance

**Add relationships**

- **Budgeting & Costs**
  - SC
  - SN
  - UF
  - USE
  - BT
  - NT
  - RT

**Clipboard**

- **Edit**
- **Print**
- **Window**

**Record Details**

- **Multi-level Hierarchy**
- **2-way Hierarchy**

**Add Rel...**

**+Rel Wiz...**

**Delete...**

**Personal Note**

**Web search**

**Save**

**Close**
Synaptica
Synaptica Software LLC (Franktown, CO)
http://synapticasoftware.com
- Web browser-based, inside the firewall or hosted.
- Priced per user, per year, per vocabulary.
- Features 12 graduations of permission levels
- Related products: indexing module IMS (Indexing Management System), Synaptica SharePoint connector, Ontology Publishing Suite.
Software
Semaphore Ontology Manager
Smartlogic Semaphore Ltd. (London, UK)

www.smartlogic.com

- Although called “ontology software” it supports ISO 2788 thesaurus standard (international equivalent of ANSI/NISO Z39.19)
- Imports/export in CSV, XML, Zthes, SQL databases, and MultiTes files
- Related products:
  - Classification Server for automated classification
  - Ontology Service for a navigation system

Download free 30-day trial:
http://www.smartlogic.com/home/knowledge-zone/semaphore-trial-software
PoolParty
Semantic Web Company (Vienna, Austria)

http://www.poolparty.biz

- Software product offered since 2009.
- Thesaurus tool built on W3C Semantic Web Standards: SKOS, RDF, OWL, SPARQL
- Installed server or web-hosted options
- Can link domain-specific thesauri to Linked Open Data
- Integrations with SharePoint, WordPress, and Drupal.
- Imports ZThes XML, CSV
- Integrated text extraction and semi-automatic tagging to enable semantic search
- Download free 30-day trial: http://www.poolparty.biz/test-demo/thesaurus-server-entity-extractor
TopBraid Enterprise Vocabulary Net (EVN)
TopQuadrant Inc. (Raleigh, NC)

www.topquadrant.com

Added product since 2010, in addition to TopBraid Composer ontology software

- Web-based access to a Linux server installation
- Pricing based on number of users
- Free 30-day trial account
- Video demos: www.topquadrant.com/knowledge-assets/videos
Software
Data Harmony Thesaurus Master
Access Innovations (Albuquerque, NM)
www.dataharmony.com

Indexing services since 1978, commercial software (originally used in-house) offered since 1998.

- Multi-platform java-based (used on Windows, Mac, Solaris, Linux). Client software allows remote access.
- Also a web-hosted version.
- Separately or combined with M.A.I. (Machine Aided Indexer) as MAIstro. Other software extensions available. DH also offers taxonomy creation services.
Term: Health care facilities

Broader Term

Health sciences

Narrower Term

Hospices
Hospitals
Mental health facilities
Nursing and retirement homes
Rehabilitation centers

Status: Candidate

Related Term

Non-Preferred Term

Scope Note

Editorial Note

Facet+

History+
Software Demos
1. Introduction, Definitions, and Types
2. Creating & Wording of Terms
3. Nonpreferred Terms
4. Creating Term Relationships
5. Structural Design: Hierarchies and Facets
6. Display Options
7. Software Tools
8. Sources for Terms
9. Project Process & Management
10. Resources
Sources for terms when building new controlled vocabularies from scratch:

1. People as sources
2. Content/material as sources
3. External reference sources

- Sources for concepts
- Sources for preferred wordings and non-preferred terms
People as sources for terms:

- Owner/manager of the controlled vocabulary
- Taxonomist
- Subject matter experts
- Users
Owner/manager of the controlled vocabulary determines:

- Some or all of top-level terms or facets
- Some sample terms
- More for concepts than for preferred wording

More often the case for commercial, publicly used databases and search services and products

- Based on strategic/business need
- Database product/service design
- Perceived customer needs/ market research
Taxonomist determines:

- In hierarchical or faceted taxonomy, upper level terms (2-3 levels deep)
- More for concepts than for preferred wording

Based on:

- “General knowledge” of knowing what’s important
- Past similar taxonomy projects worked on
- Familiarity with standard classifications (SIC/NAICS industries, UNPSC products, academic disciplines, Dewey decimal system, etc.)
- What one may have learned in library school
Subject matter experts determine:

- Specific terms and their relationships (BT/NT, RT)
- For both concepts and preferred wording

Subject matter experts (SMEs):

- Involved for highly specialized/technical CVs
- For internally used CVs, usually internal employees, perhaps borrowed from other departments (e.g. engineers)
- My actually develop the draft taxonomy, which then would need improving
Users determine:

- What the needs are, scope
- What term concepts need to be included
  - For both concepts and term preferred wording

User input can be gathered:

- From internal users for “enterprise” taxonomy projects from the beginning
- From external, commercial, or public users for taxonomies and thesauri after some content/service is made available
Sources for Terms: People

Gathering information from internal users:
- User interviews (in-person or phone)
  - For concepts and scope
- User card-sorting exercises
  - For hierarchical relationships and concepts

Gathering information from internal or external users:
- Query logs/search logs (search engine use)
  - For both concepts and preferred wording
- Surveys/questionnaires, feedback forms
  - For both concepts and preferred wording
Sources for Terms: Content

Content/material as sources for terms

Content to be indexed:
- Web pages
- Intranet pages
- Images
- Video
- Word documents
- Excel documents
- PDF documents
- Printed material

Look for concepts within:
- Article titles and sub-article headings
- Web site navigation menu labels, site maps, Web page titles
- Image captions
- Listed products, services, goals, people-types, issues
- Existing meta data (keywords, titles, short description)
Perform a content “audit” or “inventory” with term extraction. Enter terms and supporting information into a spreadsheet. Include:

- Terms
- Source (URL, location, citation)
- File format type (html, pdf, doc, etc.)
- Content/document type (article, policy, form, manual, navigation page, etc.)
- Audience (manager, customer support personnel, customers, etc.)
- Notes
<table>
<thead>
<tr>
<th>Link (URL)/Location</th>
<th>Content Type</th>
<th>Terms</th>
<th>Audience</th>
</tr>
</thead>
<tbody>
<tr>
<td>BB_BioForm.rtf</td>
<td>blank form</td>
<td>Bio form, employee bio form, employee skills for HR personnel?</td>
<td></td>
</tr>
<tr>
<td>BB_Certification_Application_Form.x</td>
<td>form template with instructions</td>
<td>Black Belt certification application instructions</td>
<td>certification applicants</td>
</tr>
<tr>
<td>black belt project descriptions.htm</td>
<td>table of projects, leaders, and black belt projects</td>
<td>Black Belt projects</td>
<td>all employees</td>
</tr>
<tr>
<td>BULLeaders.htm</td>
<td>table of business leaders and leaders</td>
<td>business unit leaders, business unit contacts</td>
<td>all employees</td>
</tr>
<tr>
<td>Business Assessment Process.ppt</td>
<td>training/information</td>
<td>business assessment, DMAIC roadmap, value</td>
<td>Six Sigma participants</td>
</tr>
<tr>
<td>Control Charts.htm</td>
<td>navigation(?) To charts</td>
<td>control charts</td>
<td>Six Sigma participants</td>
</tr>
<tr>
<td>DMAIC_curriculum.xls</td>
<td>table of values</td>
<td>DMAIC curriculum</td>
<td>Six Sigma participants</td>
</tr>
<tr>
<td>DOEFORM.xls</td>
<td>blank form</td>
<td>DOE (Design of Experiments) approval form</td>
<td>Six Sigma participants</td>
</tr>
<tr>
<td>GB_faq.htm</td>
<td>FAQ</td>
<td>certification requirements, Green Belt program</td>
<td>Six Sigma participants</td>
</tr>
<tr>
<td>Information Sharing Meetings.doc</td>
<td>guidelines</td>
<td>meeting guidelines, information sharing meetings</td>
<td>Six Sigma participants</td>
</tr>
<tr>
<td>Leadership Training.htm</td>
<td>presentation</td>
<td>training, presentation, leadership training, presentation</td>
<td>Six Sigma participants</td>
</tr>
<tr>
<td>lean tools.jsp.html</td>
<td>resource information</td>
<td>lean thinking, lean tools, waste elimination, value</td>
<td>Six Sigma participants</td>
</tr>
<tr>
<td>LEAN.htm</td>
<td>program summary</td>
<td>lean activities, lean programs</td>
<td>Six Sigma participants</td>
</tr>
<tr>
<td>Link_Spring_2002.pdf</td>
<td>newsletter</td>
<td>The Link, Transformation, Full Supply Chain, Culture, Organization</td>
<td></td>
</tr>
<tr>
<td>MBB_Certification_Process.ppt</td>
<td>presentation</td>
<td>Master Black Belt Candidate Selection, Master Black Belt Certification</td>
<td></td>
</tr>
<tr>
<td>mission.htm</td>
<td>mission statement</td>
<td>mission statement</td>
<td></td>
</tr>
<tr>
<td>New Product &amp; Service Introduction</td>
<td>navigation</td>
<td>New Product &amp; Service Introduction; Design for Six Sigma (DFSS)</td>
<td></td>
</tr>
<tr>
<td>People.jsp.html</td>
<td>resource information</td>
<td>people, key players, key roles - master black belts, black belts, green belts</td>
<td></td>
</tr>
<tr>
<td>Problem Solving Techniques.htm</td>
<td>navigation to resource information</td>
<td>Problem Solving Techniques (Brainstorming, Pareto Analysis, Voting)</td>
<td></td>
</tr>
<tr>
<td>Program Life-Cycle Process.doc</td>
<td>policy</td>
<td>Program Life-Cycle Process, program management policy</td>
<td></td>
</tr>
<tr>
<td>Program Life-Cycle Process</td>
<td>letter-style statement</td>
<td>Lean</td>
<td></td>
</tr>
<tr>
<td>Reduce Variability - DMAIC.jsp.html</td>
<td>resource information</td>
<td>DMAIC, Reduce Variability - Define Phase, Measure Phase, Analyze Phase</td>
<td></td>
</tr>
<tr>
<td>Six Sigma Acronyms.htm</td>
<td>resource information, glossary</td>
<td>acronyms, glossary</td>
<td></td>
</tr>
<tr>
<td>Six Sigma in Government.pdf</td>
<td>article</td>
<td>Six Sigma in federal government case studies</td>
<td>all employees</td>
</tr>
<tr>
<td>Software Downloads.jsp.html</td>
<td>software download</td>
<td>software downloads, simulation software - Minitab, Process Simulator</td>
<td></td>
</tr>
<tr>
<td>Solution Trees.htm</td>
<td>resource information, glossary</td>
<td>solution trees, diagramming solutions</td>
<td></td>
</tr>
</tbody>
</table>
External reference sources

For preferred format rather than for concepts

- Library of Congress Subject Headings (LCSH) and Name Authorities: [http://authorities.loc.gov](http://authorities.loc.gov)
- Google: [http://www.google.com](http://www.google.com)
- Specialized glossaries on the Web – Some listed at Taxonomy Warehouse: [http://www.taxonomywarehouse.com](http://www.taxonomywarehouse.com)
- Other published thesauri – Also for relationships [http://www.asindexing.org/site/thesonet.shtml](http://www.asindexing.org/site/thesonet.shtml)

*Be careful, don’t copy the thesaurus!*
External reference sources (continued)

- Industry standards
- Regulatory agencies
- Other reference materials

Dictionary-type thesaurus, such as Roget’s?

- Limited use: only generic topics, only words not phrases
- Different purpose: different kinds of variants
- Use with caution; don’t import such a thesaurus
Existing vocabularies for license:

Taxonomy Warehouse directory [www.taxonomywarehouse.com](http://www.taxonomywarehouse.com)

- Make sure the licensing agreement does not preclude modifying/editing the vocabulary.

Pre-built taxonomies may be suitable for:

- Literature retrieval thesauri (of multiple sources)
- Very standard classifications (industries, product types, academic disciplines)
- Named entities (named persons, geographic places, companies, organizations)

Pre-built taxonomies are *not* suitable for:

- Internal/enterprise taxonomies
- Product categories of a specific brand/manufacturer
- Any unique content set
Outline

1. Introduction, Definitions, and Types
2. Creating & Wording of Terms
3. Nonpreferred Terms
4. Creating Term Relationships
5. Structural Design: Hierarchies and Facets
6. Display Options
7. Software Tools
8. Sources for Terms
9. Project Process & Management
10. Resources
Processes vary greatly as projects vary greatly. Depends on:

- Type of controlled vocabulary and purpose
  - Especially hierarchical vs. thesaurus focus
- Size and scope
- Whether technically specialized or not
- Whether work done in-house or by a consultant/contractor

Project Process & Management

- Project planning
- Literature retrieval thesaurus project process
- Enterprise taxonomy project process
- Implementation of a controlled vocabulary
- Maintenance of a controlled vocabulary
Project Planning Considerations

- What the taxonomy/thesaurus will be used for?
- Who are the users of the CV and the end-users?
- What is the topic area?
- What is team involved? Who are all the stakeholders?
- What are the project resources and constraints?
  - Skills and expertise
  - Funding
  - Technical support
  - Timing
  - Stakeholder expectations and involvement
  - Technology
- Approval and support
- CV policies & procedures needed (Documentation)
Planning stage

Determine the scope
- scope of content and subject area(s)

Determine the usage method
- search, browse, or navigation
- categorizing/classification vs. indexing
- human vs. automated indexing

Determine the technical/systems capabilities
- types of electronic content that can be included
- queried database or search engine
- support for combined term search/faceted browse
- user interface display
Single taxonomist or a team?

- Depends on project size
- Depends on project schedule/time frame (weeks, months, years)
- Multiple taxonomy editors could be working only part time on the project (in-house or freelance)
- Could divide work different ways:
  - By taxonomy hierarchy, facet, authority file, subject area
  - By task/function: interviewing, content audit, “straw man” taxonomy structure, researching terms and relationships, editing
For literature retrieval thesaurus projects

1. Gather term concepts from the literature being indexed
2. Start initial term lists in Excel, then import into thesaurus software
3. Consult reference sources (dictionaries, glossaries, other thesauri) and subject matter experts as needed
4. Build thesaurus of terms and relationships
5. Test with sample indexing and revise
6. Add apparently “missing” terms and relationships

Approach can be either top-down or bottom-up.
For literature retrieval thesaurus projects

4. Build thesaurus of terms and relationships, by either method:
   a) Complete each term’s set of non-preferred terms, full relationships, notes, and all attributes, slowly adding more terms – more appropriate for named entities
   b) Add many terms quickly and go back and add details – more appropriate for terms in hierarchies – more appropriate for technical terms requiring research
   c) Create preferred term, non-preferred terms and scope note at the same time. Create broader and narrower relationships before creating associated term relationships.

For (b) or (c), go back and add any additional non-preferred terms
For enterprise taxonomies

1. Interview users & stakeholders to find out search needs
2. If available, gather data from search logs/query logs
3. Gather term concepts from content to be covered: web/intranet pages, document management systems, shared files on servers, hard drives
4. Look for organizing principles, problems, and tasks
5. Create sample top level hierarchies, aka “straw man” taxonomy (in Excel or hierarchy-supporting thesaurus software) and check back with stakeholders.
6. Revise hierarchies as needed and build the taxonomies deeper, from top down
For all controlled vocabularies

- Review, test, use, and maintain. And edit at each stage.
- Proofread
- Verify Spelling
- Check for balance of coverage
- Analyze gaps
- Remove unnecessary duplication
- Use difference views
Implementation of controlled vocabularies

- The taxonomy/thesaurus creation tool is usually separate from the indexing program.
- Might have a combined thesaurus creation/indexing system, but usually separate from the end-user search system.
- The CV is likely to need to be ported (for indexing and/or end-use searching).

Consider:
- Export format
- Schedule and method for updates
XML as export format

XML generation alone may not be sufficient.

- May need to edit file format
- May need to write a schema
- May need to check

<CONCEPT>
  <DESCRIPTOR>Construction industry</DESCRIPTOR>
  <F-TYPE>Thing</F-TYPE>
  <TAXONOMY>Business & Finance</TAXONOMY>
  <UF>construction industry</UF>
  <UF>construction contractors</UF>
  <UF>building industry</UF>
  <UF>building contractors</UF>
  <NT>Commercial construction</NT>
  <BT>Industries</BT>
  <NT>Infrastructure construction</NT>
  <NT>Residential construction</NT>
  <NOTE>The range of businesses involved in the building of new structures, as well as additions, modifications, maintenance, repair, and improvements to existing ones.</NOTE>
</CONCEPT>
<term name="Child protection" status="Approved" id="57" type="preferred">
<relationships>
  <relationship type="hierarchical" name="Broader Term" termId="163">Care</relationship>
  <relationship type="hierarchical" name="Narrower Term" termId="1554">Children at risk</relationship>
  <relationship type="hierarchical" name="Narrower Term" termId="1555">Children in need</relationship>
  <relationship type="equivalence" name="Use For" termId="5534">protecting children</relationship>
  <relationship type="associative" name="Related To" termId="650">Child abuse</relationship>
  <relationship type="associative" name="Related To" termId="2805">Child safety</relationship>
  <relationship type="associative" name="Related To" termId="382">Domestic violence</relationship>
  <relationship type="associative" name="Related To" termId="2478">Sales to children</relationship>
  <relationship type="associative" name="Related To" termId="387">Sex offences</relationship>
  <relationship type="associative" name="Related To" termId="51">Child care</relationship>
</relationships>

<notes>
  <note name="Scope Note">Safeguarding children from neglect or physical, emotional or sexual abuse</note>
  <note name="Added In Version">1.00</note>
  <note name="Last Updated In Version">2.00</note>
</notes>

<attributes>
  <attribute name="Use for classifying content" />
  <attribute name="Use for concept mapping" />
  <attribute name="A-Z Entry" />
</attributes>
</term>
Implementation of controlled vocabularies

Consider where it will go:

- Indexing system and interface
- User browse interface
- Web site navigation and site map
- Web page HTML metadata tags
- Enterprise content management system or document management system, drop-down fields
- Third party clients, vendors, or partners
- Any combination of the above

It could end up being implemented in more places than originally intended.
Governance comprises:

- Maintenance
- Policy creation, documentation, and enforcement
- Metadata specification

Concerns the controlled vocabulary and associated materials:

- Editorial style guide
- Indexing policies and rules
- Taxonomy training materials
- Metadata standard
- Team rules and procedures (subject to executive review)
For governance, determine:

- Kinds of changes and their costs
- Kinds of info needed to determine the changes
- What group should maintain the taxonomy
- What kinds of rules should be followed.
- What the group does beyond maintaining the taxonomy
- The processes for changes
  - Comment-handling, appeals, issue logs, announcements, update schedules, etc.

Kinds of changes impacting vocabularies:

- Terminology changes
- New content, bringing up new concepts
- Content that gets dropped
- New requirements, users, needs, trends, markets, etc.
- User feedback suggesting improvements
For maintenance, review:

- Newly added content sources or latest issues of periodicals
- Search logs, click-trail reports
- Sections of the CV covering high-change topics (technology, current events, etc.)
- Indexing statistics (human or auto) to find:
  - high-use terms needing further differentiation
  - low/no-use terms that should be merged
  - identifiable indexing errors

If the taxonomy/thesaurus creation tool and the indexing system are integrated, no problem

If not, need to schedule maintenance updates
  - Daily, weekly, monthly?
  - Schedule downtime when indexers aren’t working
1. Introduction, Definitions, and Types
2. Creating & Wording of Terms
3. Nonpreferred Terms
4. Creating Term Relationships
5. Structural Design: Hierarchies and Facets
6. Display Options
7. Software Tools
8. Sources for Terms
9. Project Process & Management
10. Resources


“Taxonomies and Controlled Vocabularies” 5-week online course
Simmons College - School of Library and Information Science
Continuing Education Program. Next offered in September 2015
http://alanis.simmons.edu/ceweb

“Practical Taxonomy Creation” 3-part webinar course recording
http://www.asindexing.org/online-learning/taxonomy-hedden

Taxonomy Boot Camp conference
Next: October 16-17, 2012, Washington, DC
www.taxonomybootcamp.com

SLA Taxonomy Division Continuing Education Webinars
http://taxonomy.sla.org/category/ce/
Resources: Websites

Accidental Taxonomy book websites
www.hedden-information.com/Accidental-Taxonomist-Websites.htm

Taxonomy Warehouse
www.taxonomywarehouse.com

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

Thesauri Online, American Society for Indexing
www.asindexing.org/about-indexing/thesauri/online-thesauri-and-authority-files

The Accidental Taxonomist Blog
http://accidental-taxonomist.blogspot.com

Hedden Information Management past presentations
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
Appendix Topics

Related Topics

- Manual indexing & controlled vocabularies
- Automated indexing & controlled vocabularies
- Metadata & controlled vocabularies
- Social tagging/Folksonomies
Appendix Topics: Manual Indexing Support

Manual indexers (taggers, catalogers, content curators, database editors, digital asset managers, records managers, etc.)

Need support with:

- Alerts about new terms
- Ensuring indexing quality and consistency
- Indexing/tagging/cataloging software UI ease-of-use

Need communication/feedback loop on new terms:

- Method for indexers nominate new CV terms
- Filling out note/message field in indexing system
- Submission of candidate terms
- Creating and using unapproved terms (named entities)
Appendix Topics: Manual Indexing Support

Need to maintain indexing quality and consistency
Conduct random indexing checking/editing
Are terms being overlooked (not used)?
  - Create more non-preferred terms
  - Create more related-term links
Are terms being misused?
  - Re-word terms
  - Add scope notes
Is the detail of indexing inconsistent?
  - Provide more specific policies and documentation/training
Technologies

- Search engine combined with a search “thesaurus,” a synonym-type controlled vocabulary
- Entity extraction software to identify named entities (may be without a controlled vocabulary)
- Text analytics and text mining to identify topics (may be without a controlled vocabulary)
- Auto-categorization or auto-classification, making use of a controlled vocabulary/taxonomy with extracted terms.
  - Machine-learning and “training” documents
  - Rules-based categorization

Not as accurate as trained manual indexers, but much faster for large volumes of content
Machine-learning based auto-categorization

- Complex mathematical algorithms are created
- Taxonomist must then provide several (at least 10) representative sample documents for each CV term to “train” the automated indexing system.
- Like reverse periodical indexing
- If only using 10 documents, then profile/overview, encyclopedic articles are best.
- If pre-indexed records exist (i.e. converting from human to automated indexing), then hundreds of varied documents can be used for each term.
Rules-based auto-categorization

- Taxonomist must write rules for each CV term
- Like advanced Boolean searching

```
Bush
IF (INITIAL CAPS AND (MENTIONS "president**" OR WITH administration**" OR AROUND "white house" OR NEAR "george"))
USE
  U.S. President
ELSE
  USE Shrubs
ENDIF
```

*Data Harmony*
Appendix Topics: Automated Indexing Support

Automated indexing impact on controlled vocabulary

- Updated work is needed when new terms are created:
  - New training documents need to be found.
  - New rules need to be written for terms.
- Need more, varied nonpreferred terms.
- Need non-preferred terms of different parts of speech.
- Cannot have subtle differences between preferred terms.
- Avoid creating many action-terms.
- CV needs to be more content-tailored, content-based.
### Non-preferred terms, differences:

<table>
<thead>
<tr>
<th><strong>For human-indexing</strong></th>
<th><strong>For auto-indexing</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Presidential candidates</td>
<td>Presidential candidate</td>
</tr>
<tr>
<td>Candidates, presidential</td>
<td>Presidential candidacy</td>
</tr>
<tr>
<td></td>
<td>Candidate for president</td>
</tr>
<tr>
<td></td>
<td>Candidacy for president</td>
</tr>
<tr>
<td></td>
<td>Presidential hopeful</td>
</tr>
<tr>
<td></td>
<td>Running for president</td>
</tr>
<tr>
<td></td>
<td>Campaigning for president</td>
</tr>
<tr>
<td></td>
<td>Presidential nominee</td>
</tr>
</tbody>
</table>
Metadata

- Data about data (like MARC records, bibliographic data)
- May include: subject(s), author, title, date, source, document type, format, language, location, approval status, ownership, audience, purpose, brief description, etc.
- Controlled vocabularies can/should be used for most of these (subject, author, document type, language, location, source)

Metadata standards:
- Dublin Core Metadata Initiative - general
- NewsML (News Markup Language) - for global news exchange
- Metadata Object Description Schema (MODS) - bibliographic element schema set for various purposes, esp. library applications.
- Industry-specific XML schema standards

For content that is shared, distributed, public, etc., consider following a standard. For internal-only content, this is not necessary.
Social Tagging

- Also called collaborative tagging, social classification, social indexing
- Use of user-created keywords, shared non-controlled vocabularies, or only semi-controlled vocabularies
- Indexing (called tagging) done by content creators and by the users/consumers, not by indexers
- Folksonomy: resulting list of terms tagged
- Dynamic
- Popular on the Web since 2004: Flickr, del.icio.us, LibraryThing
- Now moving into enterprises
Advantages
- Reflects trends, up-to-date, can monitor change and popularity
- Cheaper and quicker than building and maintaining a taxonomy
- Facilitate workplace democracy and the distribution of management tasks
- Responsive to user needs

Disadvantages
- Inconsistent – precision & recall deficiencies
- Biased
- Requires critical mass of involvement to be useful

Solutions/trends:
- Some degree of vocabulary control
- Applicable to certain areas of content, not all
Further 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