An Introduction to Taxonomies

Heather Hedden
Hedden Information Management

Society for Technical Communication - Boston Chapter November 13, 2008



Background

Heather Hedden's taxonomy development experience

- Developed controlled vocabularies for periodical index databases (Gale)
- Matched controlled vocabulary to keywords for consumer products/services directories (various "yellow pages" clients)
- Created enterprise taxonomies for corporate web sites and intranets (Earley & Associations)
- Developed base and custom taxonomies for integration within an enterprise search software product (Viziant)



Outline

- 1. Definitions
- 2. Applications of Taxonomies
- 3. Creating & Wording of Terms
- 4. Non-preferred Terms
- 5. Term Relationships
- 6. Software
- 7. Implementation & Maintenance
- 8. Resources



Outline

- 1. Definitions
- 2. Applications of Taxonomies
- 3. Creating & Wording of Terms
- 4. Non-preferred Terms
- 5. Term Relationships
- 6. Software
- 7. Implementation & Maintenance
- 8. Resources

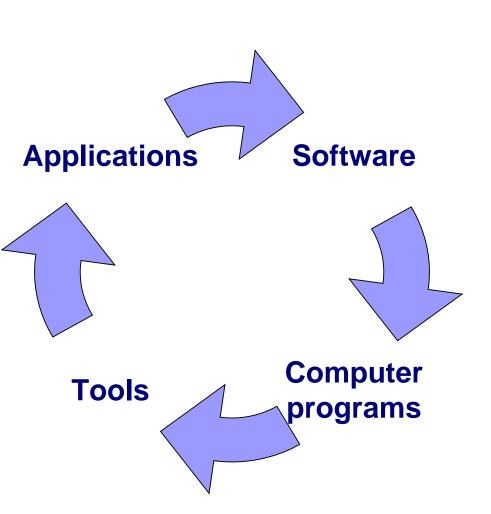


Controlled vocabulary

- 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 (broader/narrower)
- Also called "authority files," especially if containing just named entities
- "Controlled vocabulary" is the most general, widely used, broadest concept for all applications (CV for short)

Synonym ring

- A controlled vocabulary with synonyms or nearsynonyms 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"





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
- "Thesaurus" is most often used for controlled vocabularies used in indexing periodical literature

Thesaurus entry example

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



Thesaurus

Created according to standards:

- ANSI/NISO Z39.19 (2005) Guidelines for Construction, Format, and Management of Monolingual Controlled Vocabularies.
- ISO 2788 (1986) Guidelines for the Establishment and Development of Monolingual Thesauri
- ISO 5964 (1985) Guidelines for the Establishment and Development of Multilingual Thesauri



Taxonomy

- A controlled vocabulary with broader term/narrower term relationships that includes all terms to create a hierarchical structure
- An emphasis on categories and classification
- May or may not have associative term relationships or non-preferred terms
- "Taxonomy" is most often used for controlled vocabularies used in corporate/enterprise applications, where it's used for any controlled vocabulary.
- Sometimes used synonymously for any controlled vocabulary.

Taxonomy (Hierarchical) Example





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 defines a set of representational primitives with which to model a domain of knowledge or discourse... ontology can be viewed as a level of abstraction of data models, analogous to hierarchical and relational models" -- Tom Gruber

http://tomgruber.org/writing/ontology-definition-2007.htm



Ontology

Created according to W3C guidelines:

OWL Web Ontology Language Guide W3C Recommendation 10 February 2004

http://www.w3.org/TR/2004/REC-owl-guide-20040210/



Ontology relationship examples

- Broader/narrower types examples
 - Is located in/contains
 - Is comprised of parts/is a part of
- Associative types examples
 - Is purchased by/is sold to
 - Is a member of/has members of
- An ontology is not just for indexing. It is often an end itself in representing a domain of information.



Summary of definitions

Synonym Ring

- + preferred terms
- =Authority file (or controlled vocabulary)
 - +broader/narrower terms
- =Taxonomy
 - +related terms
- = Thesaurus
- -- Fred Leise

м

Outline

- 1. Applications of Taxonomies
- 2. Definitions
- 3. Creating & Wording of Terms
- 4. Non-preferred Terms
- 5. Term Relationships
- 6. Software
- 7. Implementation & Maintenance
- 8. Resources



Applications of Taxonomies & Thesauri

- Indexing support (and secondarily, retrieval)
- 2. Retrieval support
 - Retrieval support with facets
- 3. Organization and navigation support



Applications: Indexing support

List of agreed-upon terms as index entries for indexing multiple documents and/or by multiple indexers to ensure consistency

- For: periodical articles, database records, multivolume works, encyclopedias, web pages
- Called: controlled vocabulary, thesaurus, authority file
- Format: print or electronic



Applications: Indexing support

Examples database index thesauri

- Library of Congress Subject Headings (LCSH) http://authorities.loc.gov
- Library of Congress Graphic Materials (images) http://lcweb2.loc.gov/pp/tgmiquery.html
- Medical Subject Headings (MeSH)
 http://www.nlm.nih.gov/mesh/MBrowser.html
- Maternal and Child Healthcare Thesaurus http://www.mchthesaurus.info/thesaurus/home.htm



Applications: Retrieval support

Mapping-table of terms and their synonyms or equivalents to aid online retrieval

- For: on-site search engines, online databases, online yellow pages, enterprise search
- Called: controlled vocabulary, synonym ring
- Format: electronic
- Usually for automatic "indexing"



Applications: Retrieval support

Examples of online search taxonomies

- Verizon Superpages
 http://www.superpages.com/yellowpages
- Amazon.com book subject categories http://www.amazon.com/gp/homepage.html



Applications: Retrieval support: Facets

- Multiple term lists of different types, also called facets (i.e. people, places, events, products, laws, etc.), for the user to search on in combination
 - For: online databases, e-commerce sites
 - Called: faceted taxonomy, faceted browse
 - Format: electronic
 - Depending on length, may or may not make use of synonyms; may or may not have hierarchies



Applications: Retrieval support: Facets

Examples of faceted taxonomies

- Microbial Life Educational Resources http://serc.carleton.edu/microbelife/resources/index.html
- Shoebuy.com advanced search http://www.shoebuy.com/s.jsp/r_as



Applications: Retrieval support: Facets

Facets for enterprise taxonomies:

Patrick Lambe,

Organising Knowledge

- People and organizations
- Things and parts of things
- Activity cycles
- Locations

Facets for Web sites: Rosenfeld and Morville, Information Architecture

- Topic
- Product
- Document type
- Audience
- Geography
- Price



Applications: Organization/navigation support

Hierarchical categorization/classification systems for information organization

- For: web site structural design, online information services, intranet content organization, corporate content management
- Called: web site taxonomy, enterprise taxonomy
- Format: electronic
- Emphasis on classification and guided user search;
 hierarchical rather than A-Z



Applications: Organization/navigation support

Examples of web site taxonomies

- Digital Web Magazine topics
 http://www.digital-web.com/topics
- Information Architecture Institute http://iainstitute.org/en/site-map.php
- MyFlorida.com State of Florida site map http://www.myflorida.com/taxonomy

Outline

- 1. Applications of Taxonomies
- 2. Definitions
- 3. Creating & Wording of Terms
- 4. Non-preferred Terms
- 5. Term Relationships
- 6. Software
- 7. Implementation & Maintenance
- 8. Resources



Variously Called:

- Terms
- Nodes
- Concepts
- Objects
- Descriptors
- Preferred terms
- Topics

- generic, common designation
- > in hierarchical "trees"
- term + synonyms, attributes, relationships, etc.
- same as concept, in objectoriented databases
- preferred term, excludes nonpreferred
- excludes non-preferred terms
- preferred term, not a named entity



Deciding whether a concept should be a term

- 1. Is it within the subject-area scope of the taxonomy?
- 2. Is it important and likely to be looked up?
- Is there enough information on the topic?
 (Current documents and anticipated documents or articles on the topic)
- 4. Do users want and expect it?



Choosing the "preferred" term

- Choosing between two synonyms:
 - Doctors vs. Physicians
 - Movies vs. Motion pictures
 - Cars vs. Automobiles

Consider:

- Wording of terms most likely looked up by the intended users/audience, especially in browsed taxonomy
- 2. Enforcing organizational/enterprise controlled vocabulary
- 3. Conforming to academic or educational standards
- 4. Consistency in style throughout the taxonomy
- 5. Wording with in the documents/content indexed

м

Creating & Wording of Terms

Term format issues

- 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.



Pre-coordination vs. Post-coordination

 Pre-coordination: creation of relatively specific, compound modified terms

Example: Hispanic writers

 Post-coordination: leaving terms as simpler and broader, assuming that they will be combined by the user in searching

Example: Hispanics AND Writers

1

Outline

- 1. Definitions
- 2. Applications of Taxonomies
- 3. Creating & Wording of Terms
- 4. Non-preferred Terms
- 5. Term Relationships
- 6. Software
- 7. Implementation & Maintenance
- 8. Resources



Non-preferred Terms

- Terms for which a designated preferred term should be used instead.
- Have various designations:

- Variants
- Non-preferred terms
- Non-postable terms
- NPT
- Nondescriptors
- Equivalent terms
- Equivalency relationship
- Used for terms
- Use for terms
- See references
- Use references
- Cross-references
- Synonyms
- Aliases
- Keywords



Non-preferred Terms

Types include:

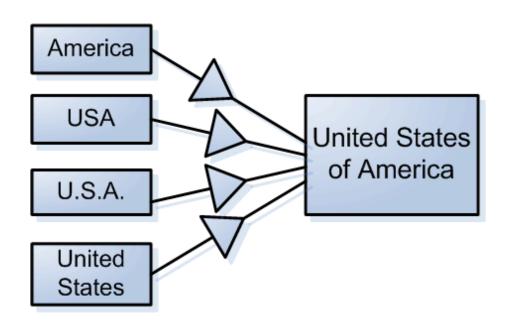
- synonyms: Cars USE Automobiles
- near-synonyms: Junior high USE Middle school
- variant spellings: Defence USE Defense
- lexical variants: Hair loss USE Baldness
- foreign language terms: Luftwaffe USE German Air Force
- acronyms/spelled out forms: UN USE United Nations
- scientific/technical names: 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 plants



Non-preferred Terms

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



w

Outline

- 1. Definitions
- 2. Applications of Taxonomies
- 3. Creating & Wording of Terms
- 4. Non-preferred Terms
- 5. Term Relationships
- 6. Software
- 7. Implementation & Maintenance
- 8. Resources



Creating Term Relationships

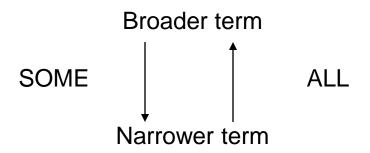
- Broader term (BT) / Narrower term (NT)
- 2. Related term (RT)
- More specific types of BT/NT or RT (in ontologies)

Relationships are reciprocal between terms.



Broader/Narrower Relationships

Asymmetrical reciprocal relationships



Fruits NT Apples Apples BT Fruits

Three types:

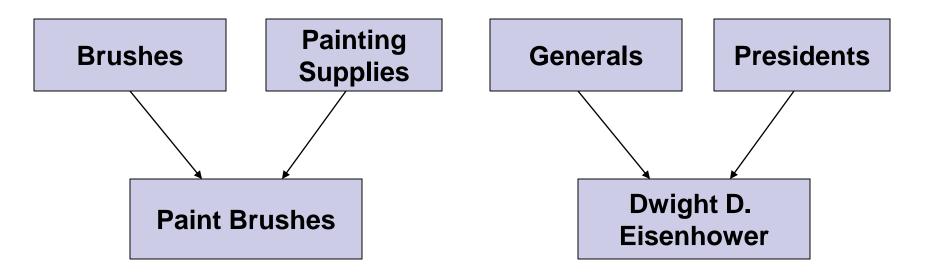
- Generic Specific
- 2. Common noun Proper noun (instance)
- 3. Whole Part



Broader/Narrower Relationships

Polyhierarchies

Sometimes a term can have two or more broader terms.

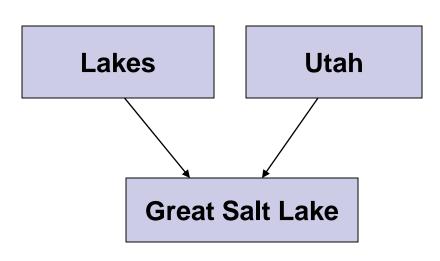


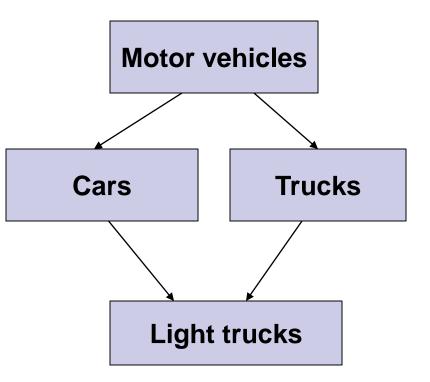


Polyhierarchies

Based on generic relationship:

Based on different means of categorizing:





м

Associated Term 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 terms
 - Associated terms
 - See also



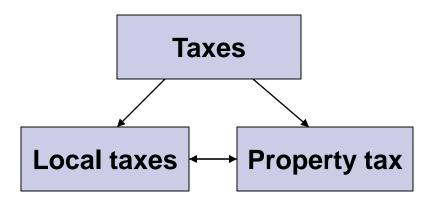
Associated Term Relationships

Between terms within the same hierarchy

 Having a shared broader term (siblings) and overlapping meaning

Required RT links, according to the standards

Example:





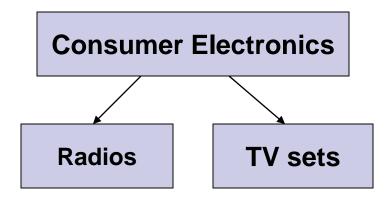
Associated Term Relationships

Between terms within the same hierarchy

Siblings without
 overlapping meaning
 (mutually exclusive terms)

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

Example:



r.

Outline

- 1. Definitions
- 2. Applications of Taxonomies
- 3. Creating & Wording of Terms
- 4. Non-preferred Terms
- 5. Term Relationships
- 6. Software
- 7. Implementation & Maintenance
- 8. Resources



Software tools for creating and editing taxonomies and thesauri

- Taxonomies, thesaurus, or controlled vocabulary: the same software is used
- 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.
- Originally in-house developed mainframe computer programs



- Spreadsheet software (Excel)
- Thesaurus software: Single-user Windows PC desktop software
- Thesaurus software: Larger-scale, multi-user clientserver or Web-based systems
- Taxonomy creation & editing module of a search/indexing system
- "Home-grown" programs developed in-house in large organizations



Thesaurus software should:

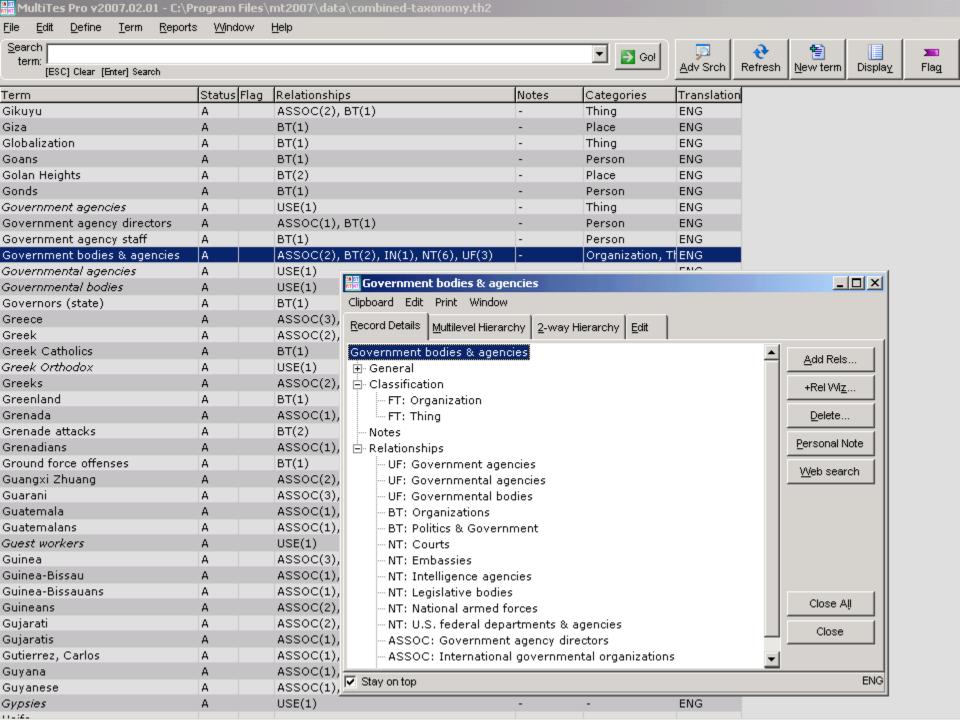
- Maintain terms and their relationships (UF, BT/NT,RT)
 - As reciprocals
 - When renaming, merging, subsuming, or deleting terms
- Disallow invalid relationships (according to standards)
- Create scope notes and attributes for terms
- Support candidate and approved terms; include term creation and update dates
- Generate reports in various thesaurus display formats
- Export data in format for importing into a content indexing/search/retrieval system

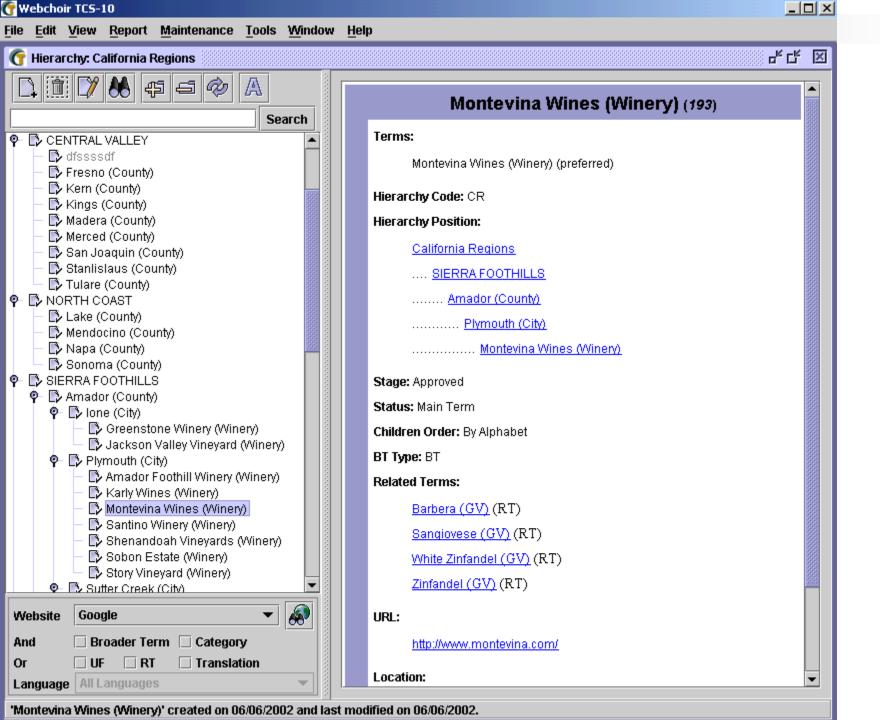


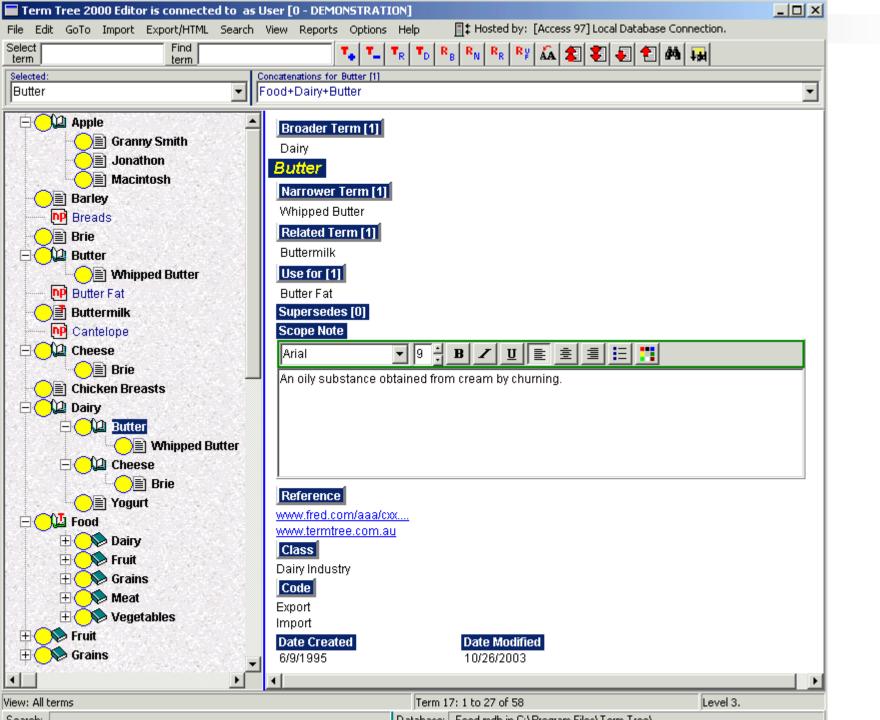
Single-user Windows desktop software (may also have multi-user versions):

- MultiTes Pro www.multites.com
- WebChoir TCS-10 Personal www.webchoir.com
- Term Tree 2000 www.termtree.com.au
- TheW32 (freeware) http://publish.uwo.ca/~craven/freeware.htm

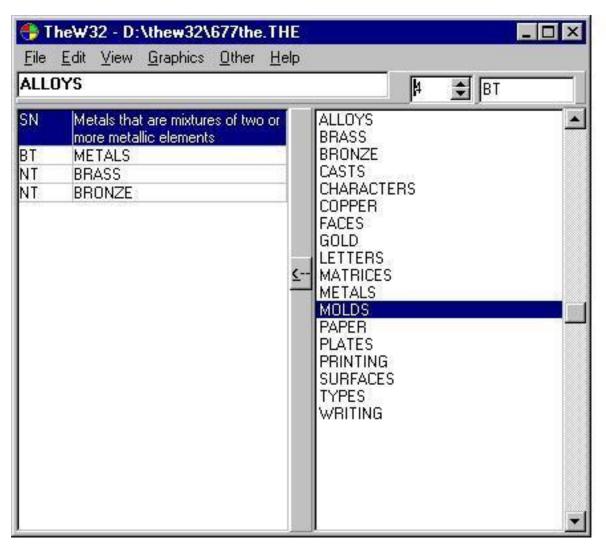
Other tools have come and gone: http://www.willpowerinfo.co.uk/thessoft.htm







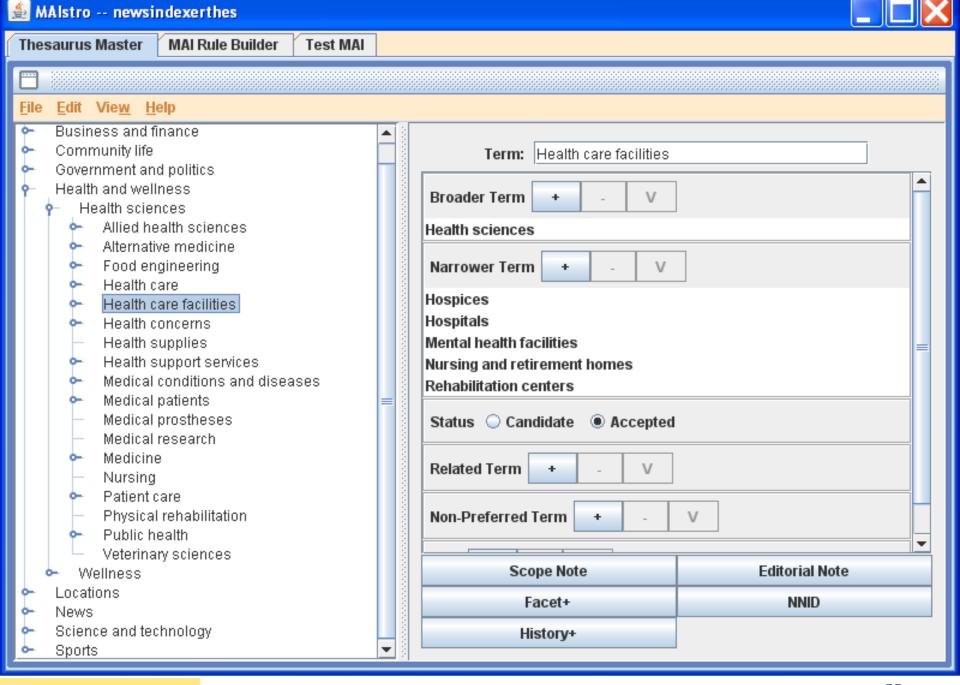
TheW32 interface





Large-scale multi-user client-server systems (Windows or Web):

- Data Harmony (Access Innovations) www.dataharmony.com
- Synaptica (Dow Jones) www.synaptica.com
- Wordmap (Wordmap Inc.) www.wordmap.com
- ConceptChoir (WebChoir Inc.) www.webchoir.com
- STAR/Thesaurus (Cuadra Associates) www.cuadra.com
- a.k.a. (Synercon Management Consultants) www.a-k-a.com.au



.

Outline

- 1. Definitions
- 2. Applications of Taxonomies
- 3. Creating & Wording of Terms
- 4. Non-preferred Terms
- 5. Term Relationships
- 6. Software
- 7. Implementation & Maintenance
- 8. Resources



Implementation & Maintenance

Implementation of a Taxonomy

- 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-use search system.
- The taxonomy is likely to need to be ported (for indexing and/or end-use searching). Consider:
 - Export format
 - Schedule and method for updates

•

Implementation & Maintenance

Indexing Support

- Need communication with indexers on new terms
 - Method for indexers nominate new CV terms
 - Method for taxonomist to announce new terms
- Need to maintain indexing quality and consistency
 - Conduct random indexing checking/editing
 - Are terms being overlooked (not used)?
 - Create more non-preferred terms or related-terms.
 - Are terms being misused?
 - Re-word terms or add scope notes.



Implementation & Maintenance

Taxonomy maintenance, because change happens:

- Terminology changes
- New content brings up new concepts
- Some content might get dropped
- New requirements, users, needs, trends, markets, etc. arise
- User feedback suggests improvements



Implementation & Maintenance

For maintenance, review the following:

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

w

Outline

- 1. Definitions
- 2. Applications of Taxonomies
- 3. Creating & Wording of Terms
- 4. Non-preferred Terms
- 5. Term Relationships
- 6. Software
- 7. Implementation & Maintenance
- 8. Resources



Resources: Books

- Aitchison, J., Gilchrist, A. & Bawden, D. (2000). Thesaurus construction and use: a practical manual (4th ed.). Chicago, IL: Fitzroy Dearborn.
- ANSI/NISO Z39.19 (2005) Guidelines for Construction, Format, and Management of Monolingual Controlled Vocabularies. Bethesda, MD: NISO Press.
- Lambe, Patrick. (2007). Organising Knowledge: Taxonomies, Knowledge and Organisational Effectiveness. Oxford, England: Chandos Publishing.
- Stewart, Darin L. (2008) Building Enterprise Taxonomies. Portland, OR: Mokita Press.



Resources: Organizations

- American Society for Indexing: Taxonomies and Controlled Vocabularies Special Interest Group http://www.taxonomies-sig.org
- Information Architecture Institute http://iainstitute.org
- Special Libraries Association (SLA) http://www.sla.org
- American Society of Information Science & Technology http://www.asis.org



Resources: Discussion Groups

- Taxonomy Community of Practice http://finance.groups.yahoo.com/group/TaxoCoP
- Taxonomies & Controlled Vocabularies SIG, ASI http://finance.groups.yahoo.com/group/taxonomies
- Metadatalibrarians http://metadatalibrarians.monarchos.com

.

Resources: Workshops, Seminars

- Taxonomy Community of Practice Webinar phone calls
 - \$50 each. Occasionally free vendor-sponsored calls.
 - Usually first Wednesday of the month, 1:00-2:00 pm EST
 - www.earley.com/TaxoCoP.asp
- "Taxonomies and Controlled Vocabularies"
 Simmons College Graduate School of Library and Information Science Continuing Education Program
 - Online: 5 weeks, \$250, next session: February 2009
 - Onsite: Boston, full day, \$220, next session: Saturday, March 28, 2009
 - www.simmons.edu/gslis/continuinged/workshops
- Taxonomy Boot Camp conference
 - Information Today Inc.
 - Next: November 19-20, 2009, San Jose, CA
 - www.taxonomybootcamp.com/2009



Resources: Web Sites

- Taxonomy Community of Practice Wikispace, http://taxocop.wikispaces.com
- Taxonomy Guide, Faculty of Information Studies, University of Toronto http://plc.fis.utoronto.ca/tgdemo/default.asp
- Construction of Controlled Vocabularies: A Primer http://www.slis.kent.edu/%7Emzeng/Z3919/index.htm
- Thesaurus Construction tutorial by Tim Craven http://publish.uwo.ca/~craven/677/thesaur/main00.htm
- Willpower Information: Publications on thesaurus construction and use http://www.willpowerinfo.co.uk/thesbibl.htm
- Taxonomy Watch Blog by Linda Farmer, http://taxonomy2watch.blogspot.com
- Earley & Associates www.earley.com
- Taxonomy Strategies www.taxonomystrategies.com



Questions

Heather Hedden
Hedden Information Management
98 East Riding Drive
Carlisle, MA 01741
www.hedden-information.com
heather@hedden.net
978-467-5195