Taxonomies for Search & Retrieval

Heather Hedden Hedden Information Management

January 5, 2009

Outline

- 1. Taxonomy definitions
- 2. Why have a taxonomy
- 3. Deciding on the taxonomy user interface
- 4. Deciding on the taxonomy scope
- 5. Choosing the taxonomy source
- 6. Taxonomy terms and relationships
- 7. Taxonomy governance and maintenance
- 8. Resources

Outline

- 1. Taxonomy definitions
- 2. Why have a taxonomy
- 3. Deciding on the taxonomy user interface
- 4. Deciding on the taxonomy scope
- 5. Choosing the taxonomy source
- 6. Taxonomy terms and relationships
- 7. Taxonomy governance and maintenance
- 8. Resources

Definitions

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
- Mainly used for indexing/tagging content to support retrieval
- 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, related)
- "Controlled vocabulary" is the most general, widely used, broadest concept for all applications (CV for short)

Definitions

Synonym ring

A type of 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.



Definitions

Taxonomy

- 1. A controlled vocabulary with broader/narrower (parent/child) term relationships that include all terms to create a hierarchical structure
 - For categorizing and organization things
 - May or may not have equivalent non-preferred terms (synonyms, etc.) to point to the correct, preferred terms
- 2. A controlled vocabulary used in corporate/enterprise applications
 - Sometimes used synonymously for any controlled vocabulary
 - To support information organization and/or information retrieval

Outline

- 1. Taxonomy definitions
- 2. Why have a taxonomy
- 3. Deciding on the taxonomy user interface
- 4. Deciding on the taxonomy scope
- 5. Choosing the taxonomy source
- 6. Taxonomy terms and relationships
- 7. Taxonomy governance and maintenance
- 8. Resources

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

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")

Users may enter:

Oil industry Oil & gas industry Oil & gas industries Petroleum industry CV contains all synonyms: Oil industry Oil & gas industry Oil and gas industry Oil & gas industries Oil and gas industries Petroleum industry Oil companies **Big oil Oil producers** Petroleum companies

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 reflect natural categories.



Outline

- 1. Taxonomy definitions
- 2. Why have a taxonomy
- 3. Deciding on the taxonomy user interface
- 4. Deciding on the taxonomy scope
- 5. Choosing the taxonomy source
- 6. Taxonomy terms and relationships
- 7. Taxonomy governance and maintenance
- 8. Resources

Deciding on the Taxonomy User Interface:

- Alphabetical browse
- Hierarchical tree
- Facets
- Search box
- Combination of one or more



Alphabetical browse

- Simple list
- Back-of-the-book style index with indented subentries (site A-Z index)

Ο

off-campus work, 78 office hours responsibility, 80 Office of Research Administration (ORA), 96 office space faculty, 109 librarian, 93 MSP/Lowell, 10 Official Designee of the Chancellor, 3 Official Personnel File, 65–66 orientation work of faculty, 81

Р

paid leaves of absence. *see* leaves of absence, paid parental/adoptive leave, 134 parking lots and spaces, 108–109 part-time faculty, 59 collective bargaining representative, 5 consecutive contracts, 60 full-time faculty retrenchment impact, 63 hiring process, 62 temporarily less than full-time, 122 temporary part-time (visiting), 60–61 union membership exclusion, 61–62 patent policy, 95–96 payroll deduction agency service fee, 148

Alphabetical (A-Z) browse is suitable for:

- Names (proper nouns), separated out or as a significant component
- Content not easily categorized
 - Content within a book and book-like narrative texts, including instructional texts
 - Unlimited-scope subject area content
- Content vast in volume and depth necessitating a large controlled vocabulary
- When there are resources for either index updating, or for frequent updates/additions of content

Hierarchical tree

1. Expandable, on same page

Online example: <u>Human body</u>



Hierarchical tree

 Separate page for each level of the hierarchy

> Arts & Entertainment > Automotive > Business & Professional Services > Clothing & Accessories > Community & Government > Computers & Electronics >

Find business listings for Arts & Entertainment.

Arcades & Amusements Amusement Equipment & Supplies, Video Arcades & Internet Gaming Centers,

<u>Bars</u>

Carnivals, Fairs, & Festivals Fairgrounds, Music Festivals, ...

Children's & Family Entertainment Miniature Golf Courses, Theme Parks, Zoos, ...

Concert Tickets

Cultural Attractions, Events, & Facilities Historical Places & Services, Libraries, Museums,

DVD Rentals

Entertainment Clubs & Nightlife Comedy Clubs, ...

Entertainment Industry Entertainment Agencies & Bureaus, Motion Picture Producers & Visual Arts Studios, Music & Recording Industry, Art Supplies & A

Event Planning

Convention & Meeting Planning Services, Party Planners, ...

Movie Theaters

Movies, Videos, & DVDs Video Production Services, ...

Music Music Instruction, Music Stores, Musical Instruments Retail, ...

Night Clubs

Performing Arts Entertainers, Live Theater, Orchestras, Symphonies, & Bands, ...

Sports Tickets Concert Tickets

Tours & Charters Boat & Yacht Charters, Rental, & Leasing, Fishing Guides & Charters, ...

Video Game Rentals

Art Supplies & Art Supply Stores, Artists & Art Studios, Professional Photographers, ...

Hierarchical tree

Separate page for each level of the hierarchy Online examples:

- ebay <u>http://www.ebay.com</u>
- Amazon.com book subject categories <u>http://www.amazon.com/gp/homepage.html</u>
- Open Directory Project <u>http://www.dmoz.org</u>

Hierarchical categories are suitable for:

- Content that is naturally categorizable: products, industries, government agencies, academic disciplines, scientific things, technologies
- Taxonomies of any size, but especially smaller taxonomies
- Large volume and/pr changing content (e.g. news feeds), resulting in multiple records per term (category) and the frequent need for new subcategories
- Automatic indexing/ auto-categorization, which puts documents into approximate categories

Hierarchical categories are also used in site organization

- Simpler and smaller taxonomies
- No use of synonyms (non-preferred terms)

Examples of web site organizational taxonomies:

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

Facets

- For serving faceted classification, which allows the assignment of multiple classifications to an object
- A "dimension" of a query; a type of concept; domain of content
- Intended for searching with multiple terms in combination (postcoordination), one from each facet
- Can be for topics or for named entities, but generally not both



21

Facets for Web sites

Rosenfeld and Morville, *Information Architecture for the World Wide Web*

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

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)
- Example types of facets:
- Products (name, series number, category, size, color, price, etc.)
- People (name, job title, gender, birth year, location, department)
- Reports (title, author, subject, audience, document type)

Online examples of faceted taxonomies:

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

Search box



Ideally should:

- Support Boolean, truncation, wildcards, phrase search, and other "advanced" searching
- Search on the taxonomy, not merely on the full-text
- Be in addition, not instead of a taxonomy browse

Outline

- 1. Taxonomy definitions
- 2. Why have a taxonomy
- 3. Deciding on the taxonomy user interface
- 4. Deciding on the taxonomy scope
- 5. Choosing the taxonomy source
- 6. Taxonomy terms and relationships
- 7. Taxonomy governance and maintenance
- 8. Resources

Deciding the Taxonomy Scope

- Content scope
- Subject area scope
- Term types to include
- Usage
- Breadth and depth



Content Scope

Content to be covered by the taxonomy can be limited to some or all of:

Content types:

- Employee handbooks
- Manuals
- Policies
- Standards
- Product data
- Transaction records
- Reports & white papers
- Marketing literature
- External publications

Document/file formats:

- Word documents
- PDF documents
- Presentations
- Spreadsheets
- Web pages
- Image files
- Videos
- Database records
- Other published formats



Subject Area Scope

Example: Area to be covered by an enterprise taxonomy can be limited to some or all of:

- Products and services (names, descriptions, features, user issues)
- Technology and technical terminology (science, research & development oriented)
- People and their areas of expertise
- Human resources
- All information in an enterprise

Scope of Term Types

(could also be considered as metadata or "facets")

Term types to be included can be limited to some or all of:

- Subjects/Topics
- Person names
- Organization/company names
- Product names
- Locations
- Activities/Actions
- Job titles
- Price ranges

etc.



Breadth and depth

- How many terms (approximate)
- How many (if any) hierarchical levels deep
- How specific the terms get
 A specific term vs. combining two less specific terms

Drug trials	Product testing and Drugs		
CRM Software	Customer Relations Management and Software		

Outline

- 1. Taxonomy definitions
- 2. Why have a taxonomy
- 3. Deciding on the taxonomy user interface
- 4. Deciding on the taxonomy scope
- 5. Choosing the taxonomy source
- 6. Taxonomy terms and relationships
- 7. Taxonomy governance and maintenance
- 8. Resources

Choosing the taxonomy source

- 1. Buy
- 2. Build
- 3. Automate

- 1. Purchasing a Taxonomy
- List of those available: Taxonomy Warehouse <u>www.taxonomywarehouse.com</u>
- Vary to the extent of what's included: term relationships (broader/narrower, related), "used from" synonyms, scope note definitions, additional term attributes
- May or may not include implementation services
- Suitable for generic areas:
 - Industries, product types, geographic places, science & medicine, academic disciplines
- Not for unique organization needs
- Not for small web navigation taxonomies

- 2. Building a Taxonomy
- Decide who does it:
 - Contracted consultancy (large project)
 - Taxonomist temp/freelancer (small project)
 - Existing information architect or staff librarian (ongoing or recurring small tasks)
 - Taxonomist new hire (ongoing or recurring large tasks)

Taxonomists – usually have education in library & information science and experience in information architecture

3. Automated Taxonomy Creation

Can serve in:

- Term extraction
- Synonym ("used for") term association
- Related term suggestion

Useful for large search taxonomies, not for hierarchy design

Technologies:

- Entity extraction
- Text mining
- Text analytics

"Auto-categorization" or "auto-classification" is not the same as automated taxonomy creation. They makes use of an existing controlled vocabulary with extracted terms.

Outline

- 1. Taxonomy definitions
- 2. Why have a taxonomy
- 3. Deciding on the taxonomy user interface
- 4. Deciding on the taxonomy scope
- 5. Choosing the taxonomy source
- 6. Taxonomy terms and relationships
- 7. Taxonomy governance and maintenance
- 8. Resources

- 1. Choosing the preferred term **Doctors** vs. **Physicians Movies** vs. **Motion pictures Cars** vs. **Automobiles**
- Creating synonyms (for search, not for navigation)
 Cars
 - UF Automobiles
 - **UF** Autos
- Structuring the hierarchy or relationships (broader/narrower or related)

Motor vehicles See also Drivers

- Cars
- -- Sports cars

1. Choosing the Preferred Term

Sources for taxonomy terms include:

- People
 - > Stakeholders, users, subject matter experts
 - > Via interviews, card-sort exercises, surveys, search logs
- Content
 - Web pages
 - Documents (Word, PDF, Excel, PPT, in print)
 - Images, video
 - External reference sources
 - Glossaries
 - Industry standards
 - Regulatory agencies
 - Other taxonomies

2. **Creating Synonyms** (Non-preferred, variants, usedfrom terms, cross-references, See references, keywords)

Types include:

- synonyms: Automobiles USE Cars
- near-synonyms: Weather USE Climate
- variant spellings: Defense USE Defence
- Iexical variants: Hair loss USE Baldness
- foreign language terms: Luftwaffe USE German Air Force
- acronyms/spelled out forms: CD USE Compact discs
- scientific/technical names: Neoplasms USE Cancer
- antonyms (for characteristics): Softness USE Hardness
- narrower terms and instances that are not preferred terms: Hydroelectric power plants USE Power plants

3. Structuring the hierarchy

Reciprocal (bi-directional) relationships, but asymmetrical



Fruits NT Oranges Oranges BT Fruits

Three types:

- 1. Generic Specific (Fruits Oranges) "NT is a type of BT"
- 2. Common noun Proper noun (Cities Cairo) "NT is an instance of BT"
- 3. Whole Part (Schools Classrooms) "NT is a part of BT" or "NT is within BT" © 2008 Hedden Information Management 41

3. Structuring the other relationships

Associative relationships (Related terms, "See Also" relationships) Reciprocal (bi-directional) relationships that are symmetrical

- Suggestions to the user of possible related terms of interest
- Not used in simple hierarchical taxonomies
- Between terms in the same hierarchy with overlapping meaning
 - Boats Ships
 - Taxonomists Information architects
 - Middle East North Africa
 - Telecommunications industry Media industry
 - Between terms in different hierarchies
 - Food Cooking
 - Infections Antibiotics
 - Environmental protection Pollution
- Auto repair Automobiles
- Engineers Engineering
 - Literature Books

Outline

- 1. Taxonomy definitions
- 2. Why have a taxonomy
- 3. Deciding on the taxonomy user interface
- 4. Deciding on the taxonomy scope
- 5. Choosing the taxonomy source
- 6. Taxonomy terms and relationships
- 7. Taxonomy governance and maintenance
- 8. Resources

For large taxonomies

Taxonomy "governance" encompasses:

- Maintenance
- Policy/rules creation and enforcement
- Metadata specification

Concerns the taxonomy and associated materials:

- Editorial style guide
- Taxonomy training materials
- Metadata standard
- Team rules and procedures (subject to executive review)

Need to 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

Maintenance deals with changes:

- Terminology changes over time.
- New content (documents, texts, images) brings up new concepts.
- Some content might get dropped.
- New requirements, users, needs, trends, markets, etc. arise.
- User feedback suggests improvements.

Maintenance requires review of:

- Newly added content sources or latest issues of periodicals
- Sections of the CV 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
 - Iow/no-use terms that should be merged
 - identifiable indexing errors
- Maintenance may require scheduled updates, if the taxonomy creation/maintenance tool and the indexing system are not integrated.

Outline

- 1. Taxonomy definitions
- 2. Why have a taxonomy
- 3. Deciding on the taxonomy user interface
- 4. Deciding on the taxonomy scope
- 5. Choosing the taxonomy source
- 6. Taxonomy terms and relationships
- 7. Taxonomy governance and maintenance
- 8. Resources

- Software
- Books
- Organizations
- Discussion groups
- Workshops and seminars
- Web Sites

Software

Can export terms, synonyms, term relationships, and other information as XML or CSV (comma delimited) files

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

Software

Large-scale multi-user client-server systems (Windows or via the 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

🔡 MultiTes Pro v2007.02.01 - C:\Pi	ogram Files\r	nt2007\data\combined-taxonomy.th2							
<u>File Edit D</u> efine <u>T</u> erm <u>R</u> eports	<u>W</u> indow j	<u>H</u> elp							
Search term: [ESC] Clear [Enter] Search				Go!	,⊡ <u>A</u> dv Srch	Refresh	1 New term	Display	>>> Fla <u>q</u>
Term	Status Flag	Relationships	Notes	Categories	Translation				
Gikuyu	A	ASSOC(2), BT(1)	-	Thing	ENG				
Giza	A	BT(1)	-	Place	ENG				
Globalization	A	BT(1)	-	Thing	ENG				
Goans	A	BT(1)	-	Person	ENG				
Golan Heights	A	BT(2)	-	Place	ENG				
Gonds	A	BT(1)	-	Person	ENG				
Government agencies	A	USE(1)	-	Thing	ENG				

Government agency directors	А	ASSOC(1),	BT(1) -	Person	ENG	
Government agency staff	A	BT(1)	-	Person	ENG	
Government bodies & agencies	A	ASSOC(2),	BT(2), IN(1), NT(6), UF(3) -	Organization, T	1 ENG	
Governmental agencies	A	USE(1)		Î.	ENC	
Governmental bodies	А	USE(1)	🚟 Government bodies & agencie	25		
Governors (state)	A	BT(1)	Clipboard Edit Print Window			
Greece	A	ASSOC(3),	Record Details			
Greek	A	ASSOC(2),				
Greek Catholics	А	BT(1)	Government bodies & agencies			A Add Date
Greek Orthodox	A	USE(1)		,		
Greeks	A	ASSOC(2),	⊡. Classification			+Rel Wiz
Greenland	A	BT(1)	- FT: Organization			
Grenada	A	ASSOC(1),	FT: Thing			Delete
Grenade attacks	A	BT(2)	- Notes			
Grenadians	А	ASSOC(1),	E. Relationships			Personal Note
Ground force offenses	A	BT(1)	- UF: Government agencie	s		V0/ab accerate
Guangxi Zhuang	А	ASSOC(2),	UF: Governmental agenc	ies		<u>eb search</u>
Guarani	A	ASSOC(3),	UF: Governmental bodies	5		
Guatemala	А	ASSOC(1),	- BT: Organizations			
Guatemalans	A	ASSOC(1),	- BT: Politics & Governmer	nt		
Guest workers	А	USE(1)	NT: Courts			
Guinea	A	ASSOC(3),	- NT: Embassies			
Guinea-Bissau	А	ASSOC(1),	- NT: Intelligence agencies	5		
Guinea-Bissauans	A	ASSOC(1),	- NT: Legislative bodies			
Guineans	А	ASSOC(2),	- NT: National armed force	es		Close All
Gujarati	A	ASSOC(2),	- NT: U.S. federal departm	ients & agencies		Class
Gujaratis	А	ASSOC(1),	- ASSOC: Government ag	ency directors		
Gutierrez, Carlos	A	ASSOC(1),	- ASSOC: International go	vernmental organization:	5	V
Guyana	A	ASSOC(1),				
Guyanese	A	ASSOC(1),	Stay on top			ENG
Gypsies	A	USE(1)	-	-	ENG	
11-16-						



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.



Organizations

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



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

Workshops and 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 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

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 http://www.earley.com
- Taxonomy Strategies http:// www.taxonomystrategies.com

Questions

Heather Hedden Hedden Information Management 98 East Riding Drive Carlisle, MA 01741 USA www.hedden-information.com heather@hedden.net 1-978-371-0822 1-978-467-5195 (mobile)