Background

• Viziant Corporation
  – A provider of knowledge mining and discovery systems for enterprises and government
  – Integrates base taxonomies into its system, which users can enhance and expand
  – Autocategorizes documents to taxonomy terms

• Heather Hedden
  – Viziant’s information taxonomist
  – Continuing Education instructor at Simmons College Graduate School of Library and Information Science
  – Previously: independent taxonomy consultant, senior controlled vocabulary editor at Gale
Overview

• What are controlled vocabularies and taxonomies
• How they aid search
• How they aid discovery
• What is algorithm-based autocategorization
• How algorithm-based autocategorization aids discovery
• How non-taxonomists can work on taxonomies
• Resources on taxonomies
Taxonomies and Controlled Vocabularies

Controlled Vocabulary (CV):

- An authoritative, restricted list of terms (words or phrases) used for indexing/tagging/categorizing content to support retrieval
- “Controlled” in who and when new terms can be added
- Often includes synonyms that point to correct, “preferred” terms
- May or may not have structure/relationships between the terms
- More general and broad; includes taxonomies
Taxonomies and Controlled Vocabularies

Taxonomy

1. A controlled vocabulary with broader term/narrower term (parent/child) relationships that includes all terms to create a hierarchical structure (e.g. Automobiles/Minivans)
   With an emphasis on categories and classification

2. Another word for “controlled vocabulary” in general, especially in corporate or enterprise applications.
Taxonomies and Controlled Vocabularies

Taxonomy broader term/narrower term (parent/child) relationships:
Asymmetrical reciprocal relationships

Some fruits are apples.
All apples are fruits.

Three types:
1. Generic – Specific
2. Common noun – Proper noun (instance)
3. Whole – Part
A controlled vocabulary gathers synonyms, acronyms, variant spellings, etc.

– Documents not missed due to use of different words (e.g. **Automobile**, instead of **Car**)
– User does not need to guess the spelling of unusual or foreign names (e.g. **Qaddafi**)

**Controlled Vocabularies Aid Search**
Controlled Vocabularies Aid Search

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

Text may contain:

Oil and gas industry
Oil companies
Big oil
Oil producers
A controlled vocabulary (if used with sophisticated auto-categorization or with human indexing) indexes concepts not words.

– Documents excluded for mere text-string matches (e.g. Bush as president, not bush as a shrub)
– Human indexers discern the different meanings.
– Autocategorization can be based on rules written for each term.
– Autocategorization can also be based on algorithms and sample “training” documents, which analyze other words in the document texts.
Taxonomies Aid Search

- A hierarchical taxonomy provides guided search.
  - Users can browse and locate narrower (more specific) subjects of interest.
  - Taxonomies reflect natural categories.
Controlled Vocabularies vs. Taxonomies for Search

• Hierarchical aspect of a taxonomy is not necessary for the retrieval benefits of a controlled vocabulary.
• A CV with sufficient and appropriate variants/synonyms/keywords is what brings the most benefits in retrieval.
• Less than perfect taxonomy hierarchical structure is better than no taxonomy at all.

➢ An subject matter expert (not necessarily a taxonomist) can create the needed variant terms, synonyms, keywords, etc.
Controlled Vocabularies/Taxonomies Aid Discovery

Discovery vs. Search

- **Search**: User knows a specific question to ask
  - Search is about *information retrieval*
- **Discovery**: User starts with a general line of inquiry to explore which questions are most pertinent, returning useful information without knowing specifically what to search for
  - Discovery is about *browsing and investigation*
- Discovery might be just as important as search for the user seeking information.
Controlled Vocabularies/Taxonomies Aid Discovery

How discovery works

– A user searches on one term. If the term is in a controlled vocabulary, it can have links/relationships with other CV terms.

– The search result displays suggested related terms *(See also)* for the user to explore.

Links between related taxonomy terms can be:

1. Taxonomist created
2. Automatically created
1. Taxonomist-created links between terms:

A more structured kind of taxonomy, that not only has hierarchical broader/narrower (parent/child) links between terms, but also associated term links across hierarchies.

- Thesauri - with standard related-term (RT) links
- Ontologies - with custom-specific semantic links

Both suggests related terms (already existing elsewhere within the taxonomy) for the user to explore.
Controlled Vocabularies/Taxonomies Aid Discovery

Taxonomist-created links between terms:

Related-term links (in structured taxonomies/thesauri)

- 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
Controlled Vocabularies/Taxonomies Aid Discovery

Taxonomist-created links between terms: related-term link examples

Between “sibling” terms in the same hierarchy with overlapping meaning:

- Boats – Ships
- Children’s books – Picture books
- Taxonomists – Information architects
- South America – Latin America
- Telecommunications industry – Media industry

Between terms in different hierarchies:

- Process and agent: Programming - Programmers
- Process and instrument: Skiing - Skis
- Process and counter-agent: Infections - Antibiotics
- Action and property: Environmental cleanup - Pollution
- Action and target: Auto repair - Automobiles
- Cause and effect: Hurricanes - Flooding
- Object and property: Plastics - Elasticity
- Raw material and product: Timber - Wood products
- Discipline and practitioner: Physics - Physicists
- Discipline and object: Literature - Books
2. Automatically-created links between terms:
   Any CV with auto-generated “keywords” for each term can provide suggested related taxonomy terms based on shared keywords.

   Auto-generated keywords are both synonyms and related terms.

   The added complexities of thesauri and ontologies are useful but not required for related-term discovery.
Algorithm-based Autocategorization

- Autocategorization can be either rules- or algorithm-based
- Algorithm-based autocategorization makes use of sample training documents to generate additional “keywords” for each taxonomy term. (Rules-based autocategorization does not suggest keywords.)
- Keywords are generated with varying associated relevancies (such as 1-100), based on frequency of occurrence within the training documents.
- The user merely provides training documents, doesn’t need to write rules.

➢ You don’t need a taxonomist to identify and feed training documents.
Algorithm-based Autocategorization for Discovery

• The presence of shared keywords with high relevancies across multiple CV terms leads to the suggestion of closely related terms for the user to discover.

• Generation of keywords can be dynamic, as new training documents are added and ingested.
  – New training documents contain new keywords, leading to new shared-keyword terms to be discovered.
  – New relationships can be discovered.
Algorithm-based Autocategorization for Discovery

- A search on a term can bring up related terms based on shared auto-generated keywords.
- The terms can be grouped by their frequency in retrieved documents.
Example:
Searching on “monetary policy” the user discovers related terms, such as “Stock markets” and “Banks” based on their associated documents.
Algorithm-based Autocategorization for Discovery

Keywords, both auto-generated and manually created, can be viewed and edited.
Building Taxonomies by the non-Taxonomist

- A hierarchical taxonomy is easier to maintain than a CV that is merely a list of terms.
  - Easier to scan the taxonomy to verify appropriateness
  - More obvious where gaps need filling
  - More practical to segment the maintenance work among multiple editor-users.
Building Taxonomies by the non-Taxonomist

• Vendor supplied base taxonomies
  – As hierarchical starting points
    • Create additional specific terms to existing terms
  – As hierarchical examples
    • Design broader/narrower relationships based on existing relationships in parallel hierarchies
  – As examples of types of variants/keywords
    • Create variant/synonym terms for new terms, to the same degree as found for existing terms
Building Taxonomies by the non-Taxonomist

• Vendor supplied documentation and training
  – Need to go beyond how to use the software
  – Provide guidelines in how to create variants/keywords
  – If desired, provide guidelines in how to create correct hierarchical relationships
Building Taxonomies by the non-Taxonomist

• Taxonomy creation software that enforces taxonomy rules
  – Preventing circular references
  – Preventing or alerting upon the creation of keywords that match existing term names
Building Taxonomies by the non-Taxonomist

• Non-taxonomists, but not non-experts
  – Taxonomy is built out by subject matter experts.
  – Taxonomy development work is restricted to certain individuals, not all search users, based on software user access privileges.
• “Knowledge discovery” vs. “Knowledge modeling”
Conclusions

• Hierarchies that are not perfect are OK, because the greatest search & discovery benefits are from the keywords/synonyms.

• A CV with algorithm-based autocategorization can yield shared keywords for automatically supporting discovery.

• Hierarchical taxonomies in one’s field/specialty are not difficult to create, if basic structure is in place as a start.


Resources: Organizations

• American Society for Indexing: Taxonomies & 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
  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, www.earley.com/TaxoCoP.asp

• "Taxonomies and Controlled Vocabularies" workshop
  Simmons College Graduate School of Library and Information Science Continuing Education Program
    – Saturday, October 25, full-day, at Simmons College, Boston, $220
    – Online 5 weeks, next session in November, $250
      www.simmons.edu/gslis/continuinged/workshops

• Taxonomy Boot Camp conference, Information Today Inc.
  www.taxonomybootcamp.com
  September 25-26, 2008, San Jose, CA
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  
Information Taxonomist  
Viziant Corporation  
Boston, MA  
www.viziantcorp.com

Heather.hedden@viziantcorp.com  
978-467-5195 (mobile)