Taxonomies for Human vs. Auto-Indexing

Heather Hedden

Hedden Information Management

heather@hedden.net

Background

Heather Hedden's taxonomy-creation experience

- For human indexing
  - Developed controlled vocabularies for periodical article index databases (Gale)
- For auto-indexing
  - Developed taxonomies for integration within an enterprise search software product for corporate content and web page searching (Viziant)
  - Matched controlled vocabulary to keywords for consumer online products/services directories (various "yellow pages" clients)
- For either
  - Created enterprise taxonomies for corporate web sites and intranets for site navigation (Earley & Associations)

Outline

- Taxonomies & Indexing Background
- Choosing Human vs. Auto-indexing
- Taxonomies & Human Indexing
- Taxonomies & Auto-indexing
- Taxonomy Creation Comparison
  - Differences in taxonomy terms
  - Differences in term relationships
  - Differences in definitions & notes
  - Differences in synonyms/variants
- Additional Work for the Taxonomist
- Resources

Taxonomies & Indexing

Types of Taxonomies

1. Organization, classification, navigation support
   – more emphasis on hierarchies
2. Search and retrieval support
   – more emphasis on synonyms

- For indexing, use #2 above:
  Search & retrieval support taxonomies

Taxonomies & Indexing

Search & retrieval taxonomies:

- Connect users to desired content by means of a common nomenclature/terminology/vocabulary
- Matching between:
  1. the vocabulary of the users
  2. the vocabulary of the content
- Taxonomies interface with
  1. the users
  2. the content

- Indexing/tagging/categorization deals with #2 in each case above: the connection of taxonomy to content

Taxonomies & Indexing

Indexing/tagging/categorization:

- Indexing
  – done by (trained) indexers
  – creating a (browsable) index
- Tagging
  – done by any person
  – applying labels, metatag descriptors to documents to be picked up by database or search software
  – may not require a taxonomy/controlled vocabulary
- Categorization
  – done more systematically/automatically
  – putting documents into (pre-defined) categories
  – often within facets

Choosing Human vs. Auto-indexing: The Content

Human indexing

- Manageable number of documents
- Includes non-text files
- Varied and undifferentiated document types/formats
- Varied subject areas

Auto-indexing

- Very large number of documents
- Text files only
- Common document types/formats (or pre-tagged types)
- Focused subject areas (legal, medical, etc.)

Choosing Human vs. Auto-indexing: The Culture

Human indexing

- Higher accuracy in indexing
- Invest in people
- Low-tech: can build your own indexing UI or buy
- Internal control, or outsourcing vendor relationship

Auto-indexing

- Greater volume indexed
- Greater speed in indexing
- Invest in technology
- High-tech: must purchase auto-indexing software
- Software vendor relationship

Taxonomies & Human Indexing

Who are indexers?

- Specialists or not
  - the taxonomist and/or other information specialists, librarians
  - dedicated hired indexers (with or without prior indexing) experience
  - supplemental work for other staff (editors, writers, administrators)
- One person or multiple people
- Usually in-house but could be contracted out

Taxonomies & Human Indexing

Indexing software/module

- Indexing user interface optimized for ease, speed, and accuracy in indexing
- Method for indexers to nominate new taxonomy terms

Training & documentation for indexers

- Indexing policy guidelines
- Method to communicate new and changed taxonomy terms to indexers
- Method for checking and quality control

D - Subject Descriptors

D Censorship

all xRef from xRef

Military censorship

Censorship
- NARROWER TERM: Book burning
- NARROWER TERM: Military censorship
- SEE ALSO: Artistic freedom
- SEE ALSO: Banned art
- SEE ALSO: Banned books
- SEE ALSO: Decency standards
- SEE ALSO: Freedom of expression
- SEE ALSO: Freedom of information
- SEE ALSO: Freedom of speech
- SEE ALSO: Freedom of the press
- SEE ALSO: Information ethics
- SEE ALSO: Internet filtering software
- SEE ALSO: Literature and morals
- SEE ALSO: Obscenity
- SEE ALSO: Political culture
- SEE ALSO: Pornography
- SEE ALSO: Prior restraint
- SEEN FROM: Self censorship

Censure
- BROADER TERM: Punishment
- SEE ALSO: Misconduct in office
- SEE ALSO: No confidence motions

Census
- SEE: Censuses

Census districts
- BROADER TERM: Special districts (Local government)

caching periodical vocab only

Search Form

either enter a search phrase

censorship

Smart ◉ Begins ○ Contains ○ Exact ○

or select an alphabetical range

Range ○ a thru. z

or enter a specific item uid

Item UID ○

select search criteria

Obj Subjects

Cat All Categories

Act Active

display batching

1000 Items per batch

Start Search

Advanced Search Form

- Censorship (Subjects)
- Military censorship (Subjects)
- Self censorship (Subjects) (NPT)

Item Summary

Descriptor	**Censorship**
Object	Subjects
Categories	Humanities, Law, Literature, OVRC Issues, Social Sciences, Tuned
Status	Active; Approved; Preferred; Locked
UID	18878
Created	mobrien 04/10/2000 03:15:00 PM
Modified	WDIRN ancillary 12/12/2007 01:40:23 AM
Sort Key	censorship

Save | Delete | Combine

Refresh | Deactivate | Make NPT

Descriptor Elements

Descriptor: Censorship

Extended Attributes

Scope:

[Add Relationship]

Relationships

UF Self censorship (Subjects)
NT Book burning (Subjects)
NT Military censorship (Subjects)
RT Artistic freedom (Subjects)
RT Banned art (Subjects)
RT Banned books (Subjects)
RT Decency standards (Subjects)
RT Freedom of expression (Subjects)
RT Freedom of information (Subjects)
RT Freedom of speech (Subjects)
RT Freedom of the press (Subjects)
RT Information ethics (Subjects)
RT Internet filtering software (Subjects)
RT Literature and morals (Subjects)
RT Obscenity (Subjects)
RT Political culture (Subjects)
RT Pornography (Subjects)
RT Prior restraint (Subjects)

Taxonomies & Auto-Indexing

Technologies

- Entity extraction
- Text mining and text analytics
- Auto-categorization or auto-classification utilizing taxonomies:
  1. Machine-learning and training documents
  2. Rules-based categorization

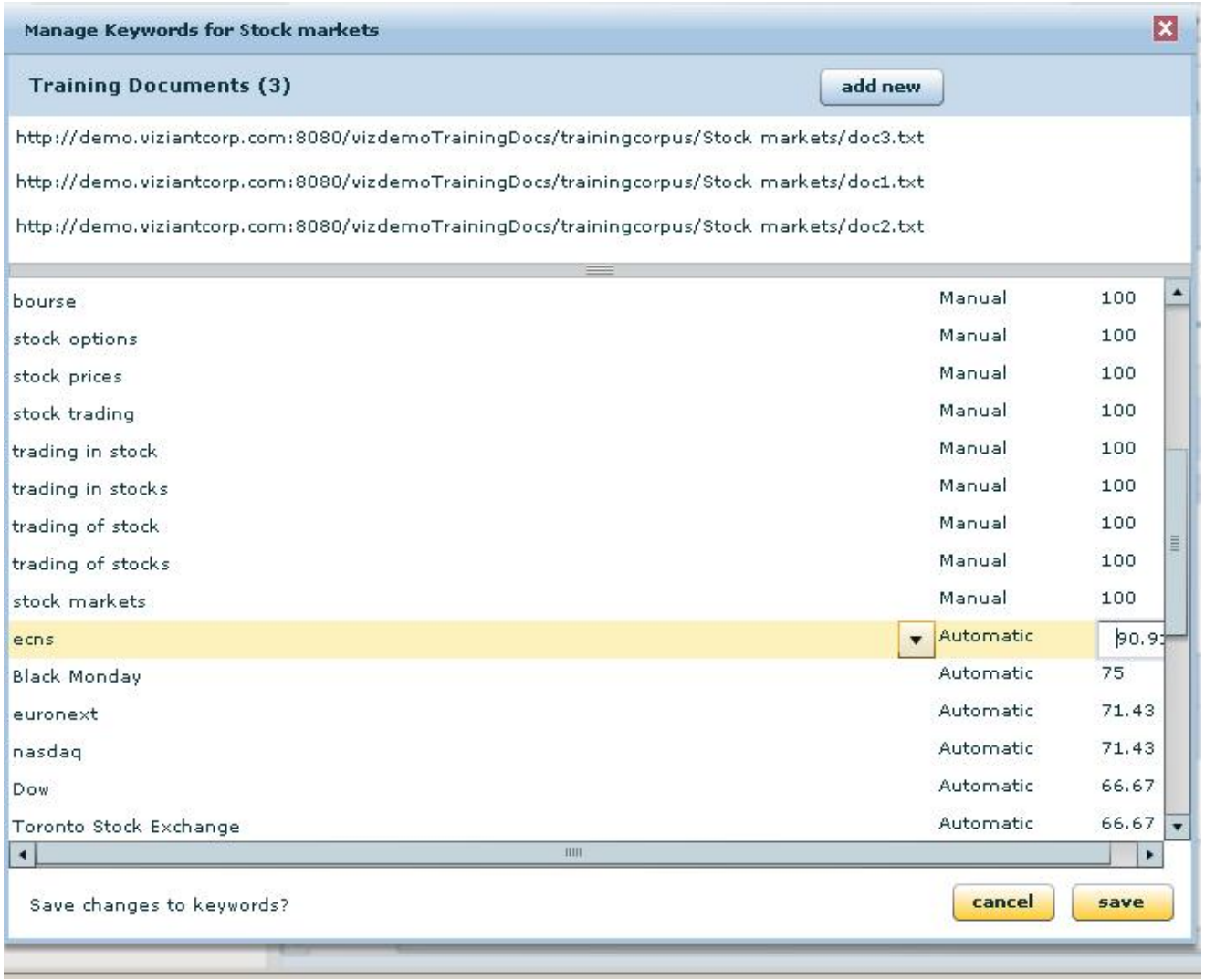
Taxonomies & Auto-Indexing

Machine-learning auto-categorization:

- Complex mathematical algorithms are created
- Taxonomist must then provide several (at least 5-10) representative sample documents for each taxonomy term to “train” the automated indexing system.
- If only using only 5-10 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.

Taxonomies & Auto-Indexing

Machine-learning auto-categorization



Taxonomies & Auto-Indexing

Rules-based auto-categorization:

- Taxonomist must write rules for each taxonomy term
- Similar to 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

Taxonomy Creation Comparison

- Differences in taxonomy terms
- Differences in term relationships
- Differences in term notes, definitions
- Differences in synonyms/variants

Differences in Taxonomy Terms

- For human indexing

  Create terms as specific (granular) as the content will support and users will expect.

- For auto-indexing
  - Cannot have subtle differences between preferred terms:
    International relations; ***Foreign policy***
  - Avoid creating both action and topic terms:
    Investing; ***Investments***

Differences in Term Relationships

- **Hierarchical (broader/narrower) links**
- **Associative (related terms) links**
- For human indexing

  Highly useful to indexer, as is to end-user, in finding the best term. *Consider indexer behavior.*
- For auto-indexing

  Not needed, but could be utilized in search results:
  - Broader terms recursively include narrower term results
  - Related terms display as suggestions
    Consider search results.

Differences in Term Relationships

- Facets

Certain facets may work better with human indexing than with auto-indexing. Automated indexing may not distinguish between different facet meanings of a term.

Examples:

Mergers - Action/Event or Business Topic?

Churches – Place or Organization type?

Differences in Term Notes

Concise explanatory notes (not a dictionary definition) on some terms, as needed:

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

For the end-user, optional aid

For indexing:

- often needed for some terms for human indexing
- never needed for auto-indexing

May have notes for indexers that are not for end-users.

Differences in Term Notes

Scope Notes examples

ProQuest Controlled Vocabulary:

Occupational health

SN: Employer activities designed to protect and promote the health and safety of employees on the job

Inequality

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

Medical Subject Headings (MeSH):

Nonverbal Communication

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

Differences in Synonyms/Variants

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
- antonyms: Misbehavior USE Behavior
- narrower terms and instances that are not preferred terms: Power hand drills USE Power hand tools

- Each preferred term may have multiple non-preferred terms.

Differences in Synonyms/Variants

For human indexing

- "Shortcuts"- unique abbreviations within each facet (2-3 letters) for commonly entered terms
  - For countries, states; industry codes
  - For within a facet of limited size – memorizable

  Examples:
  mna – ***Mergers & acquisitions***
  bnk – ***Banking***
  fr – ***France***

- Phrase inversions for alphabetical browsing
  Example: ***Photography, digital***

Differences in Synonyms/Variants

For Auto-indexing

If machine-learning auto-categorization:

- Need greater number of non-preferred terms
- Can include non-noun phrases

For human-indexing

Presidential candidates

Candidates, presidential

For auto-indexing

Presidential candidate

Presidential candidacy

Candidate for president

Candidacy for president

Presidential hopeful

Running for president

Campaigning for president

Presidential nominee

Taxonomy Creation Summary

Human indexing

- Rich relationships between terms
- Term notes for clarification
- Common-use shortcuts
- Phrase inversions as term variants

Also:

- Browsable (A-Z) display
- Multiple ways to search (beginning of term, word within term, etc.)

Auto-Indexing

- Cannot have subtle differences between terms
- Avoid creating action-type terms
- Be careful with facets
- Need more, varied non-preferred terms, including non-noun phrases

Additional Work for the Taxonomist

Human Indexing

- Inform indexers of newly added terms
- Adjustments based on review of indexers' work:
  - If terms are overlooked (not used):

    - Create more non-preferred terms

    - Create more related-term links
  - If terms are misused:

    - Re-word terms

    - Add scope notes

Auto-Indexing

- Continual update work, for each new term:
  - Add new training documents, or
  - Write new rules
- Adjustments based on inappropriate results:
  - Add, delete, edit training documents
  - Tweak existing rules

Resources

- American Society for Indexing
  www.asindexing.org
- Taxonomies & Controlled Vocabularies SIG of the American Society for Indexing
  www.taxonomies-sig.org
- "Taxonomies and Controlled Vocabularies"
  Simmons College Graduate School of Library and Information Science Continuing Education Program
  - onsite workshop (October 25, 2008, Boston)
  - online workshop (February 2009)

  www.simmons.edu/gslis/continuinged/workshops

Contact

Heather Hedden
Hedden Information Management
98 East Riding Dr.
Carlisle, MA 01741

978-371-0822
978-467-5195 (mobile)
Heather@hedden.net
www.hedden-information.com