Tools for Taxonomies

Heather Hedden
Taxonomy Consultant
Hedden Information Management

© 2010  Heather Hedden
About Heather Hedden

- Taxonomy consultant, Earley & Associates
- Indexer, Hedden Information Management
- Instructor, Continuing Education, Simmons College Graduate School of Library & Information Science
- Formerly taxonomist at Viziant Corporation and Thomson Learning (Gale)
What are Taxonomy Tools?

- No authoritative industry list of taxonomy software
- “Taxonomy software” can mean different things
  - Auto-categorization vs. taxonomy management
- Existing Web lists are miscellaneous taxonomy-related tools or out-of-date
  - http://taxocop.wikispaces.com/TaxoTools
  - www.taxotips.com/resources/tools
  - www.searchtools.com/info/classifiers-tools.html
  - www.willpowerinfo.co.uk/thessoft.htm
Taxonomy Tool Types

1. Thesaurus/ontology management software
2. Other software with thesaurus/taxonomy modules
3. Auto-categorization/text mining software
4. Other software supporting creating taxonomies
   - mindmapping or concept modeling tools
   - Cardsorting tools
   - Web analytics
Taxonomy Types & Software Types

1. Controlled vocabularies, authority files, or synonym rings
2. Hierarchical or faceted taxonomies
3. Thesauri
4. Ontologies
1. Controlled vocabularies, authority files, or synonym rings
   - Support correspondence between term and synonyms (equivalencies)
   - No other relationships between terms
   - Optional additional data for terms

   - Spreadsheet software suffices
2. Hierarchical or faceted taxonomies
   - Broader/narrower (parent/child) relationships between terms
   - Optional synonyms (equivalencies) or additional data for terms

   - If very small and simple, spreadsheet software
   - If including synonyms and other data, then use thesaurus management software
3. Thesauri

- Broader/narrower (parent/child) relationships between terms
- Associative (symmetrical reciprocal related/“See also”) relationships between terms
- Synonyms (equivalencies) for terms
- Other data for terms

- *Thesaurus management software*
4. Ontologies
   - Classes for terms
   - Customizable semantic relationships between terms (hierarchical and associative), dependent on class
   - Optional synonyms (equivalencies) for terms
   - WC3 or RDF standard outputs

   - *High-end thesaurus management software (with ontology support features)*
   - *Ontology software*
### Tools Used by “Taxonomists”

<table>
<thead>
<tr>
<th>Tool Description</th>
<th>Percentage</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Commercial dedicated thesaurus/taxonomy management software</td>
<td>38.1%</td>
<td>24</td>
</tr>
<tr>
<td>An internally developed thesaurus/taxonomy management system</td>
<td>23.8%</td>
<td>15</td>
</tr>
<tr>
<td>Commercial software, of which taxonomy management is a feature, module or component</td>
<td>15.9%</td>
<td>10</td>
</tr>
<tr>
<td>Other commercial software that is not intended for taxonomy creation</td>
<td>19.0%</td>
<td>12</td>
</tr>
<tr>
<td>Open-source ontology/taxonomy management software</td>
<td>3.2%</td>
<td>2</td>
</tr>
</tbody>
</table>
Thesaurus Management Software Basics

- Maintains terms and their relationships (equivalencies, hierarchical, and associative)
  - As reciprocals
  - When renaming, merging, subsuming, or deleting terms
- Disallows invalid relationships (according to standards)
- Supports term notes and other attributes for terms
- Supports candidate/approved terms; includes term creation and update dates
- Generates reports in various thesaurus display formats (hierarchical, alphabetical)
- Exports data in interoperable formats for importing into a content management, indexing, search, retrieval system
- Supports thesaurus standards: ANSI/NISO Z39.19 or ISO 2788
Thesaurus Management Software Feature 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

© 2010 Heather Hedden
Thesaurus Management Software

Lower-end, single-user desktop software (may also have multi-user versions):

- MultiTes Pro (Windows)
- Cognatrix (Mac)
- One-2-One (Windows)
- TheW32 (Multi-platform)

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

MultiTes Pro
Multisystems (Miami, FL)
www.multites.com

- $295 single user; $1295 for 5 users
  $2495 for 10 users; $3950 enterprise deployment
- Add-on products: web development kit, enterprise development kit
- Imports text; exports text, HTML (as a web page), XML, CSV
- Free limited-time downloadable trial and online tutorial
- Online discussion group for tips
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
SC
SN
UF
USE
BT
NT
RT

Stay on top

Browse...
Save
Close
Thesaurus Management Software

Cognatrix
LGOSystems Pty. Ltd. (Australia)
www.cognatrix.com

- For Mac OS X 10.4.5 and later
- US $499, or $199 for an “Education” version limited to 500 terms.
- Imports from plain text with tab separations.
  With CognatrixImporter add-on, imports from various XML schemas: Cognatrix, MultiTes, Term Tree, and Zthes.
- Exports to XML and HTML.
- Free limited-time downloadable trial and manual
Thesaurus Management Software

One-2-One
This to That Pty. Ltd. t/a A.C.S.
Active Classification Solutions (Australia)
www.acs121.com

- Price: $800 Australian (approximately $700-750)
- For thesauri and classification systems, but also has features and connectors for records management
- Replaces a previous thesaurus-only product, Term Tree
- Drag-and-drop hierarchy feature
- Free limited-use and limited-size trial

© 2010 Heather Hedden
TheW32

Tim Craven Freeware (Ontario, Canada)

http://publish.uwo.ca/~craven/freeware.htm

- Started as thesaurus feature in his TEXNET auto-abstracting tool in early 1990s.
- Free
- Also provides web site indexing software XRefHT and machine-aided indexing and abstracting software
Thesaurus Management Software

TheW32 Interface
Thesaurus Management Software

High-end, multi-user client-server, large scale systems ($3000/single user - $75,000+; or annual hosted options)

- Data Harmony
- Synaptica
- Smartlogic
- Wordmap
- SchemaLogic
- STAR/Thesaurus
- SoutronTHESAURUS
- Mondeca ITM T3
- PoolParty
- TheMa
- a.k.a.
Thesaurus Management 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.
- All standard thesaurus displays types as view options
- User defined associative and equivalence, but no user-defined hierarchical relationships
- Sold separately or combined with M.A.I. (Machine Aided Indexer) as MAIstro. Other software extensions available.
Term: Health care facilities

Broader Term: + - V

Health sciences

Narrower Term: + - V

Hospices
Hospitals
Mental health facilities
Nursing and retirement homes
Rehabilitation centers

Status: Candidate [ ] Accepted [ ]

Related Term: + - V

Non-Preferred Term: + - V

Scope Note

Editorial Note

Facet+

History+
Synaptica
Synaptica Software LLC (Franktown, CO)
http://synapticasoftware.com
- Web browser-based, priced per user, per year, per vocabulary
- 12 graduations of permission levels
- Can assign relationship weights
- Global term and relationships editor, creating a list of terms to edit
- Side-by-side editor with drag-and-drop
- Imports: CSV, text, MS Excel, XML (including schemas of ZThes, RDF, SKOS, and OWL)
- Exports: CSV, HTML, MS Word, MS Excel, XML (including schemas of ZThes, RDF, SKOS, and OWL)
There are also browse options including a Tree Browse and Alpha-numeric browse option to review terms.

© 2010 Heather Hedden
Thesaurus Management Software

Semaphore Ontology Manager
Smartlogic Semaphore Ltd. (London, UK) + US office
www.smartlogic.com

- Supports creating thesauri according to ISO 2788 standard
- Supports creating ontologies, through customizable relationships and user-created classes
- User-defined term attributes and metadata
- Multiple user access/privilege levels
- Imports/export in CSV, XML, Zthes, SQL databases, and MultiTes files
- Related products:
  - Classification Server for automated classification
  - Ontology Service for a navigation system

© 2010 Heather Hedden
Thesaurus Management Software

Wordmap Designer
Wordmap Inc. (Concord, MA)
www.wordmap.com

- Multi-platform java-based
- One of a suite of products: Wordmap Navigator, Wordmap Content Author (for manual indexing), Wordmap Intelligent Text Classifier, Wordmap Taxonomy + Search.
- User-defined relationships; can also turn on or off relationship name display.
- Can display two taxonomies side by side and drag and drop.
- User access/privileges can be set at the individual node level.
- Imports: text, CSV, XML; Exports: CSV, XML
### Members

<table>
<thead>
<tr>
<th>No</th>
<th>Text</th>
<th>Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Consumer affairs</td>
<td>en</td>
</tr>
<tr>
<td>2</td>
<td>Consumer law</td>
<td>en</td>
</tr>
<tr>
<td>3</td>
<td>Shopping</td>
<td>en</td>
</tr>
</tbody>
</table>

### Features

<table>
<thead>
<tr>
<th>Feature</th>
<th>Value</th>
<th>Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alias</td>
<td>Z</td>
<td></td>
</tr>
<tr>
<td>Asked in version</td>
<td>1.10</td>
<td></td>
</tr>
<tr>
<td>Category</td>
<td>true</td>
<td></td>
</tr>
<tr>
<td>Concept Id</td>
<td>850</td>
<td></td>
</tr>
<tr>
<td>Id</td>
<td>850</td>
<td></td>
</tr>
<tr>
<td>Last updated in version</td>
<td>1.10</td>
<td></td>
</tr>
<tr>
<td>Obsolete</td>
<td>false</td>
<td></td>
</tr>
<tr>
<td>Preferred</td>
<td>true</td>
<td></td>
</tr>
<tr>
<td>Scope nodes</td>
<td></td>
<td>LOCL4045</td>
</tr>
<tr>
<td>URL</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Children

- [a] Energy conservation
- [a] Energy efficiency
- [a] Energy and fuel
- [a] Environmental processes licences
- [a] Combustion processes licences
- [a] Electricity supply
- [a] Electrolyte supply
- [a] Gas supply
- [a] Gas supply
- [a] Nuclear energy
- [a] Renewable energy

### Other Locations

IPSV > Business and industry > Energy and fuel
IPSV > Environment > Energy and fuel
Parent Child Relationship

Define custom associative relationships through simple drag & drop
Synonyms

Translation
Thesaurus Management Software

SchemaLogic Enterprise Suite
SchemaLogic Inc. (Kirkland, WA)
www.schemalogic.com

- Provides thesaurus management according to ANSI/NISO standards, plus broader structural metadata support
- Can create customizable relationships
- Can assign various permission levels to vocabularies or terms
- Classification module supports 3rd party auto-indexing
- Connectors to SharePoint, EMC Documentum, and FAST ESP
- Can import CSV or XML files
Thesaurus Management Software

STAR/Thesaurus
Cuadra Associates, Inc. (Los Angeles)
www.cuadra.com

- Stand-alone or integrates with STAR family of products for records mngmt, collections mngmt, archives mngmt, DAM
- Supports standard thesaurus relationship but not customizable relationships
- Supports unlimited user-defined notes and categories
- Various output report display formats
- Import/export ASCII text and CSV, but not XML
Thesaurus Management Software

Soutron THESAURUS
Soutron Ltd. (United Kingdom)
www.soutron.com

- Markets in the U.S. through partnership with InMagic
- Stand-alone or integrates with SoutronGLOBAL or SoutronSOLO library management systems, or with InMagic Presto social knowledge management software
- Supports standard thesaurus and user-defined relationships
- Supports term merging
- Imports from XML; exports to XML or CSV
Thesaurus Management Software

Mondeca ITM T3 (Intelligent Topic Manager: Thesaurus, Taxonomies, Terminologies)
Mondeca S.A. (Paris, France)
www.mondeca.com

- Since 2008, addition to Intelligent Topic Manager set of products for knowledge management, semantic portals, and e-catalogs
- web-based collaborative application
- conforms to both SKOS vocabularies and OWL-standard ontologies
- connectors to text mining, classification, and search tools
- Imports/exports XML, RDF, and SKOS
PoolParty
punkt. netServices GmbH (Vienna, Austria)
http://poolparty.punkt.at

- A new 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
- Wordpress plugin to build glossaries for blogs
- Imports ZThes XML, CSV
- Integrated text extraction and semi-automatic tagging to enable semantic search
Ontology Software

Tools for ontologies, not thesauri

- TopBraid Composer, www.topquadrant.com
- Protégé, http://protege.stanford.edu
- SWOOP, http://code.google.com/p/swoop
- CMAP Tools Ontology Editor, http://cmap.ihmc.us
Other Software with Taxonomy/thesaurus
Creation & Editing Components

- Metadata or cataloging software, especially for archives and libraries

- Content management and document management systems
  - Open Text Collections Server Webtop Thesaurus Manager, www.opentext.com

- Records management systems

- Auto-categorization and enterprise search software
Auto-categorization and Text Mining

Auto-categorization

- Algorithms, statistics, and training documents – utilize a large set of a sample documents per taxonomy term to “train” the system to learn to index
- Rules base – generate and edit or write “rules” for each term based on co-existing words, proximity, Boolean logic, etc.

Text mining

- Extracts relevant terms from texts to generate a candidate taxonomy or supplement an existing taxonomy
Auto-categorization, text mining, and search systems that utilize taxonomies handle these taxonomies in different ways:

- With pre-installed taxonomies that cannot be edited
- With pre-installed taxonomies that the user may edit and extend through the user interface
- Automatically generate a taxonomy that can be edited
- Support the import of taxonomies but do not support the editing of those taxonomies
- Support the import of taxonomies and then the editing of those taxonomies
- Various combinations of above
Auto-categorization/Search Software

Software that can import and use taxonomies but lacking user-interface features to edit those taxonomies includes:
- Microsoft SharePoint
- IBM Classification Module
- Fast
- Endeca
- Temis
- Vivisimo
- Mindbreeze
- Exalead
- PerfectSearch

They collaborate with other vendors that develop taxonomies and/or have taxonomy editing capabilities.
Auto-categorization/Search Software

Examples of tools with *some* taxonomy management capabilities:

- Inxight SmartDiscovery Analysis Server
  www.inxightfedsys.com
- Autonomy Collaborative Classifier
  www.autonomy.com
- Autonomy Interwoven MetaTagger
  www.interwoven.com
- Lexalytics Classifier
  www.lexalytics.com
- Conceptsearching
  www.conceptsearching.com
Auto-categorization/Search Software

Examples of tools with *full* thesaurus management capabilities:

- Data Harmony MAIstro  
  www.dataharmony.com/products/maistro.html
- Smartlogic Semaphore Classification Server  
  www.smartlogic.com
- Wordmap Intelligent Text Classifier  
  www.wordmap.com
- SAS Enterprise Content Categorization  
  (formerly Teragram TK240)  
  www.sas.com/text-analytics
- Nstein Text Mining Engine (part of Open Text)  
  www.nstein.com
Auto-categorization Software combined with Thesaurus Management

Data Harmony MAIstro (combines Data Harmony Thesaurus Master and Machine-Aided Indexer)

- Automatically creates a basic rule for every term and its variants in the Thesaurus Master’s thesaurus
- Rules may be edited and additional rules can be manually written
- Statistics module tracks the editor’s term choices and compares them with M.A.I. term suggestions, sorting them as hits, misses, and noise to guide and prioritize the editor’s fine-tuning of rules
- Can be used for machine-aided indexing or fully automated indexing
- Connectors to Sharepoint and search engines
Auto-categorization Software combined with Thesaurus Management

Smartlogic Semaphore Classification Server
(connects with Semaphore Ontology Manager)

- Automatically creates classification rulebases directly from a taxonomy/thesaurus/ontology, and applies these rules to content as it is received to automatically classify content.
- Rules are based on the term, its equivalencies, and broader/narrower/related terms
- Employs 20 different kinds of rules
- Rules have weights and scores
- Variants based on spelling, plurals, and stemming may also be considered.
- Manual rules can take precedence over generated rules.
Auto-categorization Software combined with Thesaurus Management

Wordmap Intelligent Text Classifier
(connects with Wordmap Designer for thesaurus creation)

- Auto-classification based on statistical method based of Support Vector Machine algorithms and machine learning with training documents
- Pre-packaged with statistical algorithms based on a generic taxonomy, the U.K.’s Integrated Public Services Vocabulary (IPSV), for which each of hundreds of terms have already been “trained” with representative documents
- Wordmap also offers Taxonomy Connectors for taxonomy-driven tagging and search within SharePoint and Endeca
Auto-categorization Software combined with Thesaurus Management

SAS Enterprise Content Categorization (ECC)
(Formerly Teragram TK240 Taxonomy Manager)
Supports taxonomy building or connects with SAS Ontology Manager

- ECC supports equivalent, hierarchical & related relationships
- Ontology Manager supports customized relationships and attributes
- Utilizes both auto-categorization and entity/concept extraction
- Auto-categorization bases on rules
- Rules-writing supported with a graphical tree view of Boolean operators and commands
- User can define weighting of terms
Auto-categorization Software combined with Thesaurus Management

Nstein Text Mining Engine (TME)

- Modules include concept extractor, entity extractor, auto-categorizer, automated abstract creation, sentiment analysis
- Taxonomy Manager module (not sold separately) supports creating & editing hierarchical, associative and equivalence relationships according to ANSI/NISO standard
- Auto-categorization technology based on use of training sets for taxonomy terms, combined with concept extraction technology
- Ships with pre-installed taxonomies already “trained” for auto-categorization
Concluding Remarks

- Some “taxonomy tools” are stronger in taxonomy/thesaurus/ontology management
- Some “taxonomy tools” are stronger in auto-categorization
- A few tools combine both, but vendor partnerships and connectors and also achieve high results
Questions/Contact

Heather Hedden
Carlisle, Massachusetts
978-371-0822
978-467-5195 (mobile)
heather@hedden.net
www.hedden-information.com
www.accidental-taxonomist.com