



Combining Controlled Vocabularies

Heather Hedden

Hedden Information Management

www.hedden-information.com



Heather Hedden

- Taxonomy Consultant, Earley & Associates
- Continuing Education Workshop Instructor, Simmons College Graduate School of Library and Information Science
- Freelance Book Indexer, Hedden Information Management
- Author, *The Accidental Taxonomist*, Information Today Inc., forthcoming, May 2010
- Previously
 - Information Taxonomist, Viziant Corporation
 - Senior Vocabulary Editor, Thomson Gale (Cengage Learning)

Combing Controlled Vocabularies

- Increase in the adoption and number of controlled vocabularies/taxonomies
- More is not always better

- Combine
- Reduce
- Re-use
- Simplify

The collage illustrates the process of combining controlled vocabularies. It features three overlapping Microsoft Excel spreadsheets, each representing a different level of classification. The top spreadsheet shows a hierarchical structure with categories like 'Business & Trade', 'Computers', and 'Entertainment'. The middle spreadsheet shows a more granular view of 'News & Current Events' with sub-categories like 'Top Stories', 'World', and 'Nation'. The bottom spreadsheet shows an even more detailed view of 'News & Current Events' with sub-categories like 'Politics', 'Metro, State, Region', and 'Business'. A green sticky note is placed over the bottom spreadsheet, listing various leisure and culture terms such as 'Arts and entertainment venues', 'Museums and galleries', 'Children's activities', 'Culture and creativity', 'Architecture', 'Crafts', 'Heritage', 'Literature', 'Music', 'Performing arts', 'Visual arts', 'Entertainment and events', 'Gambling and lotteries', 'Hobbies and interests', 'Parks and gardens', 'Sports and recreation', 'Team sports', 'Cricket', 'Football', 'Rugby', 'Water sports', 'Winter sports', 'Sports and recreation facilities', 'Tourism', 'Passports and visas', and 'Young people's activities'.

Integrating, Merging or Mapping

- **Integrating:**
Combining separate controlled vocabularies (CVs) into a single, larger master controlled CV for combined use



Integrating, Merging or Mapping

■ Merging:

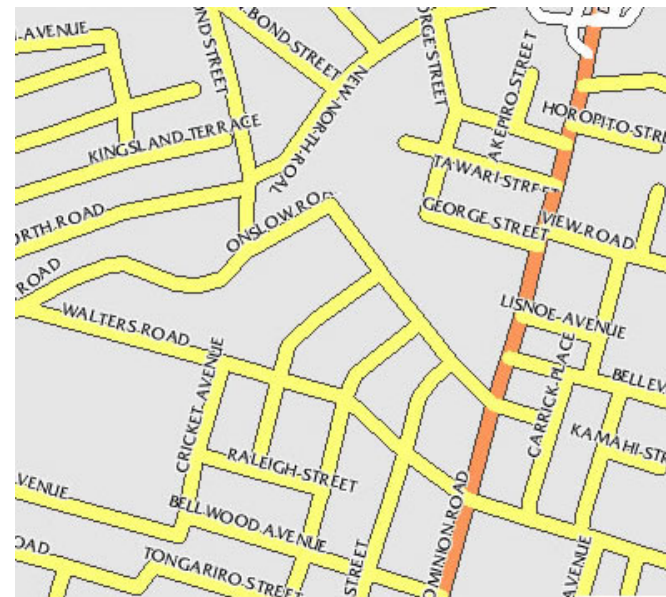
Combining two or more redundant vocabularies in same subject area into one

- Without any longer retaining them as distinct
- Legacy content is retrieved through added equivalence relationships



Integrating, Merging or Mapping

- Mapping:
 - Enabling one CV to be used for another in same subject area
 - Retain them both as continued distinct vocabularies.
 - A CV continues to be used to retrieve its content as before plus additional content associated with the other.



Something representing something else

Integrating Vocabularies



- Combines related, but not redundant taxonomies
- Taxonomies supplement each other in the same area
- Increases (multiplies) the number of preferred terms
- Involves importing an additional taxonomy/hierarchy/thesaurus into a taxonomy management system to be used with other taxonomies
- Concerned with issues of interoperability
- Taxonomist deals with integrating structures more than integrating individual terms.

Integrating – Situations



- An enterprise taxonomy is built via combining existing departmental taxonomies.
- An additional facet is added to an existing faceted taxonomy.
- A new product line is added, requiring a new product/topic hierarchy within a product facet.
- An organization acquires/merges with another organization, and their CVs in different specialty areas are integrated.
- An internally created CV is supplemented by a purchased/licensed CV in a complementary subject area to expand its scope.

Integrating – Interoperability Issues



- The 2 vocabularies may have different:
 - Editorial style (caps, plural, abbreviation use)
 - Relationships (associative, semantic, polyhierarchies)
 - Additional term attributes/details (notes,
 - Use of unique numerical IDs
 - Interoperability format (XML, ZThes, RDF, OWL, SKOS)

Integrating – Interoperability Issues



XML for a term record as exported from MultiTes

```
<CONCEPT>
  <DESCRIPTOR>Firewalls</DESCRIPTOR>
  <BT>Intrusion prevention systems</BT>
  <NT>Application firewalls</NT>
  <NT>Network firewalls</NT>
  <N-TYPE>Subject Subject</N-TYPE>
  <TAXONOMY> Risk Management</TAXONOMY>
  <UF>Firewall</UF>
  <UF>packet filtering</UF>
  <UF>packet filters</UF>
  <UF>packet inspection</UF>
  <SN>A device or software configured to permit, deny,
    encrypt, or proxy all computer traffic between different
    security domains based upon a set of rules and other
    criteria</SN>
</CONCEPT>
```

Integrating – Interoperability Issues



XML for a term record exported from Smartlogic Semaphore

```
<term name="Child protection" status="Approved" id="57" type="preferred">
<relationships>
  <relationship type="hierarchical" name="Broader Term" termId="163">Care</relationship>
  <relationship type="hierarchical" name="Narrower Term" termId="1554">Children at risk</relationship>
  <relationship type="hierarchical" name="Narrower Term" termId="1555">Children in need</relationship>
  <relationship type="equivalence" name="Use For" termId="5534">protecting children</relationship>
  <relationship type="associative" name="Related To" termId="650">Child abuse</relationship>
  <relationship type="associative" name="Related To" termId="2805">Child safety</relationship>
  <relationship type="associative" name="Related To" termId="382">Domestic violence</relationship>
  <relationship type="associative" name="Related To" termId="2478">Sales to children</relationship>
  <relationship type="associative" name="Related To" termId="387">Sex offences</relationship>
  <relationship type="associative" name="Related To" termId="51">Child care</relationship>
</relationships>
<notes>
  <note name="Scope Note">Safeguarding children from neglect or physical, emotional or sexual abuse</note>
  <note name="Added In Version">1.00</note>
  <note name="Last Updated In Version">2.00</note>
</notes>
<attributes>
  <attribute name="Use for classifying content" />
  <attribute name="Use for concept mapping" />
  <attribute name="A-Z Entry" />
</attributes>
</term>
```

Merging and Mapping Vocabularies



- Compares two closely redundant vocabularies side-by-side, term-by-term
- First pass(es) automatic followed by taxonomist review of matches
- Taxonomy software may have the feature (Synaptica, Wordmap), or do your own scripting
- Taxonomist reviews, discerns distinction between equivalent, broader/narrower, related terms to approve matches
- Taxonomist deals with terms more than structure.

Merging – Situations



- An enterprise taxonomy replaces multiple CVs of separate administrative departments
- An organization acquires or merges with another organization, and their redundant vocabularies are merged
- A folksonomy is incorporated into a CV
- An internally created CV is combined with a purchase/licensed CV

Merging – Which Direction?



Designate a dominant/primary CV into which to merge the other:

- If an organization acquires another, then the acquirer's CV is dominant.

Or choose:

- The larger CV
- The CV with greater breadth
- The CV with greater depth
- The more structured CV
- The “better” CV



Merging – Automated matching



Use a software tool to compare vocabularies, to obtain matches in succeeding passes:

Primary CV	Merging CV	Taxonomist Reviews
Exact matches of:		
<i>Preferred term: Cars</i>	<i>Preferred term: Cars</i>	no need
<i>Nonpreferred term: Automobiles USE Cars</i>	<i>Preferred term: Automobiles</i>	no need
<i>Preferred term: Cars</i>	<i>Nonpreferred term: Cars USE Automobiles</i>	yes
<i>Nonpreferred term: Cars USE Autos</i>	<i>Nonpreferred term: Cars USE Automobiles</i>	yes
Inexact matches of:		
<i>Preferred term: Automobile</i>	<i>Preferred term: Automobiles</i>	yes

Merging – Automated matching



Inexact, “fuzzy” matches to automatically match and then human review:

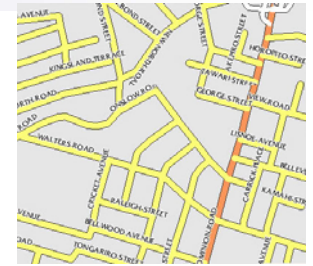
Match Type:	Examples:	
<i>hyphens, parentheses, punctuation, and spaces</i>	Healthcare	Health care
<i>plural/singular</i>	Teaching method	Teaching methods
<i>common abbreviations and acronyms</i>	and Dept.	& Department
<i>Word order</i>	Photography, digital	Digital photography
<i>Addition of specified words (industry, services, etc.)</i>	Healthcare industry	Healthcare services
<i>Grammatical endings</i>	Production	Producing

Merging – Tools



- Commercial thesaurus/taxonomy software with merge vocabularies feature
 - Synaptica
 - Wordmap
- Custom scripting (Perl, etc.) to compare vocabularies

Mapping

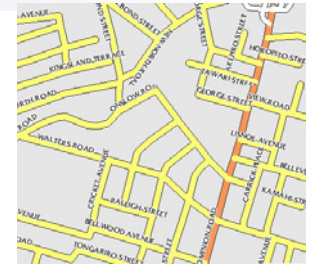


- Between a retrieval/user-interface CV and a CV indexed to content
- Unmatched terms cannot be utilized.
- Narrower-to-broader matches are fine.
- Same kinds of matches as in CV merging plus: matches of words/phrases of the retrieval taxonomy *within* a term from the indexing CV

Retrieval taxonomy	Indexing taxonomy
Television sets	HDTV television sets



Mapping - Situations



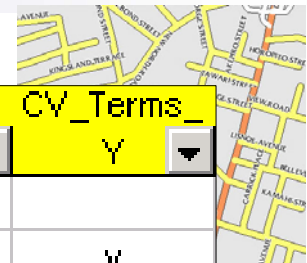
- Selected content with an enterprise CV is made available on a public web site with a different public-facing CV
- A content provider with a CV partners with a third-party information vendor with its own CV
- A provider of scientific/technical/medical content with a technical CV creates a simpler CV aimed at laypeople
- Search log query terms need to be integrated into the CV as additional nonpreferred terms.
- To support “federated search” that involves taxonomies



	A	B	C
1	Programmable logic controllers	ok	Programmable controllers
2	Programmable logic devices	ok	PLDs (Programmable logic devices)
3	Programming (Computers)	ok	Computer programming
4	Progressivism (United States politics)	b	Progressive movement
5	Prohibited books	ok	Banned books
6	Project method in teaching	ok	Project method (Education)
7	Projectile points	ok	Projectile points (Archaeology)
8	Projection	n	Projection (Drawing)
9	Projection televisions	ok	Projection television sets
10	Prolactin	n	Prolactin test
11	Proletariat	ok	Working class
12	Prolog (Computer program language)	ok	Prolog (Programming language)
13	Promethazine hydrochloride	b	Promethazine
14	Promoters (Entertainment)	b	Promoters
15	Promotion (School)	ok	Student promotion
16	Pronghorn antelope	ok	Pronghorns
17	Propaganda, American	ok	American propaganda

Indexing CV in column A, retrieval CV in column C,
taxonomist notes in column B

ok is equivalent, b is broader so also ok for upward posting, and n is not acceptable.

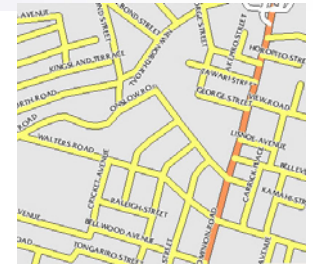


Mapping user-entered search queries (column 2) to terms, in this case the term “Type of Vehicles.”

If terms could be (narrower) examples of automobiles, put a “y” in the CV_Terms_Y column. Some terms are too broad and vague.

		Candidate_ CV_Terms_ CV_Terms_Y	CV_Terms_Y
<i>Makes</i>	GVX	y	
<i>Type of Vehicles</i>	4 Wheel Drive	y	y
<i>Type of Vehicles</i>	Four Wheel Drive	y	y
<i>Type of Vehicles</i>	4x4	y	
<i>Type of Vehicles</i>	4 X 4	y	
<i>Type of Vehicles</i>	4x4s	y	
<i>Type of Vehicles</i>	4WD	y	
<i>Type of Vehicles</i>	All Wheel Drive	y	y
<i>Type of Vehicles</i>	AWD	y	
<i>Type of Vehicles</i>	Classic	y	
<i>Type of Vehicles</i>	Vintage	y	
<i>Type of Vehicles</i>	Antique	y	
<i>Type of Vehicles</i>	Commercial Vehicles	y	y
<i>Type of Vehicles</i>	Commercial Trucks	y	y
<i>Type of Vehicles</i>	Commercial Vans	y	y
<i>Type of Vehicles</i>	Fleets	y	
<i>Type of Vehicles</i>	Convertibles	y	y
<i>Type of Vehicles</i>	Coupes	y	y
<i>Type of Vehicles</i>	Diesel	y	
<i>Type of Vehicles</i>	Domestic	y	

Mapping – Tools



- In commercial thesaurus/taxonomy software, designate a custom equivalence relationship:
 - Example: USE-Map / UF-Map (in place of USE/UF)
- Import CSV mapping tables, such as created in Excel

Summary



■ Integrating

- Non-overlapping CVs combine & supplement each other, to create a larger CV



■ Merging

- Overlapping CVs combine, remove duplicates, but increase non-preferred terms



■ Mapping

- Overlapping CVs remain distinct, one used for the other in a specific application (indexing vs. retrieval CVs)

Questions/Contact



Heather Hedden

Carlisle, MA

978-371-0822

978-467-5195 (mobile)

www.hedden-information.com

More Info in:

Heather Hedden, *The Accidental Taxonomist*
forthcoming book from Information Today Inc., May 2010