Integrating, Merging and Mapping Controlled Vocabularies

Heather Hedden
Hedden Information Management
www.hedden-information.com
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.

© 2009 Hedden Information Management
Integrating Vocabularies

- Combines related, but not redundant taxonomies
- 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.
- An organization acquires/merges with another organization, and their CVs in different areas are integrated.
- An internally created CV is supplemented by a purchased/licensed CV in a complementary subject area to expand its scope.
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 – 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

© 2009 Hedden Information Management
Merging – Automated matching

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

<table>
<thead>
<tr>
<th>Primary CV</th>
<th>Merging CV</th>
<th>Taxonomist Reviews</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exact matches of:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Preferred term</td>
<td>Preferred term</td>
<td></td>
</tr>
<tr>
<td>Nonpreferred term</td>
<td>Preferred term</td>
<td></td>
</tr>
<tr>
<td>Nonpreferred term</td>
<td>Preferred term</td>
<td>yes</td>
</tr>
<tr>
<td>Nonpreferred term</td>
<td>Nonpreferred term</td>
<td>yes</td>
</tr>
<tr>
<td>Inexact matches of:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Preferred term</td>
<td>Preferred term</td>
<td>yes</td>
</tr>
</tbody>
</table>

© 2009 Hedden Information Management
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

© 2009 Hedden Information Management
Mapping

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

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

© 2009 Hedden Information Management
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.
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
Questions/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

More Info:

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

© 2009 Hedden Information Management