

Taxonomy Updating, Combining, and Translating

Heather Hedden
Senior Taxonomy Analyst

Taxonomy Tuesday, Library of Congress, Washington, DC, April 19, 2011



About Heather Hedden

- Taxonomy consultant full time with Project Performance Corporation
- Continuing education instructor with Simmons College Graduate School of Library and Information Science
- Author of *The Accidental Taxonomist* (Information Today, Inc., 2010)
- Background in indexing

Broad experience creating taxonomies for:

- corporate web sites and intranets (Project Performance Corporation)
- document management in SharePoint (First Wind)
- integration within an enterprise search software product (Viziant)
- periodical index databases (Gale)
- consumer web site taxonomies (Demand Media)



Taxonomies are:

1. Designed
2. Built
3. Maintained/Managed

But they also evolve, change, adapt, move on...



Introduction

The collage features several key elements:

- Excel Spreadsheet:** A Microsoft Excel window titled "Microsoft Excel - Level 1&2" showing a list of categories in column A and corresponding news items in column B. The categories include Business & Trade, Computers, Entertainment, Education, Government, Health & Medicine, News & Current Events, Products & Services, Recreation & Sports, Reference, Regional, Science, Social Sciences & Humanities, Society, Culture & Family, and Technology.
- Hierarchical Tree:** A tree diagram showing a hierarchy of terms under "Information and communication technology". The tree includes categories like Information and communication, Science, technology and innovation, Communication channels, Computer applications, e-Learning, Electronic media, Hardware, ICT services, Networking, Radio, Smartcards, Software, Systems development, Telecommunications, and Television.
- News Table:** A table titled "News-Nation" with columns for "Event", "Nation", and "Category". The categories listed include Accidents, Crime, Demographics, Disasters, Energy, Environment, Human interest, Law, Law enforcement, Military, Natural resources, Religion, Security, Social issues and, and Transportation.
- Term List:** A list of terms with their types, categorized into Hierarchical, Associative, and Equivalence.

Type	Term
BT	Information and communication
BT	Science, technology and innovation
NT	Communication channels
NT	Computer applications
NT	e-Learning
NT	Electronic media
NT	Hardware
NT	ICT services
NT	Networking
NT	Radio
NT	Smartcards
NT	Software
NT	Systems development
NT	Telecommunications
NT	Television

Type	Term
RT	Communications industries
RT	Data security
RT	e-Commerce
RT	Information management

Type	Term
UF	ICT (information and communication technology)
UF	Information Technology
UF	IT (information technology)



In evolving, taxonomies may need to be:

- Adapted to new indexing or search implementations
- Updated and revised
- Integrated, merged or mapped with another taxonomy
- Translated into another language or localized



- Taxonomy updating
- Taxonomy combining: integrating, merging, mapping
- Taxonomy translating



Maintaining

- Ongoing, routine
- Part of the governance process
- More reactive
- Involves changes to individual terms only
- A regular responsibility: an employee's job (or part of a job)

Updating

- One-time or periodic overhaul
- Possibly beyond the scope of a maintenance plan
- More pro-active
- Involves changes to terms and often also structure
- Requires temporary additional resources: possibly a contractor or consultant



Maintaining

Responds to:

- User feedback suggesting improvements
- User search logs and query path reviews
- Indexing quality control
- Minor changes in content
- New trends, buzzwords, and terminology arise in existing content

Updating

Responds to:

- New audiences, users, markets
- New strategies, purpose, vision
- Added implementations
- Major additions of new content
- Changes in indexing methods (i.e. manual to automated)



Maintaining or Updating involves:

- Creating new terms
- Deleting terms
- Splitting existing terms (creating two new terms and removing an old one)
- Merging terms (including deleting a term and making it nonpreferred)
- Changing the wording of a term (perhaps keeping an old name as nonpreferred)
- Adding relationships (hierarchical or associative) between existing terms
- Deleting relationships between existing terms
- Changing a relationships between a pair of existing terms

and may also involve:

- Adding nonpreferred terms to existing terms
- Adding scope notes to existing terms
- Adding other new information, attributes, categories, etc. to existing terms
- Moving a branch/sub-hierarchy within a hierarchical taxonomy to a new location



Updating a taxonomy

may reflect a changes in **editorial policies or style:**

- Permitting polyhierarchies
- Requiring unique specific term labels
- Limiting/expanding the hierarchy depth
- Changing requirements (such as the minimum number of documents) to justify creating new terms
- Changing the general size and granularity of the taxonomy



Updating a taxonomy

may involve **structural changes**:

1. Changing from a simple taxonomy to thesaurus,
or from a thesaurus to an ontology

- Permitting nonpreferred terms
- Permitting associative relations
- Converting simple associative (RT) to customized, semantic relationships
- Incorporating scope notes and other term attributes
- Adding categories or classes



Updating a taxonomy

may involve **structural changes**:

2. Changing from structured indexing, pre-coordinated/post-coordinated indexing, use of facets

- Changing from more precoordinated to more post-coordinated terms
- Creating new smaller taxonomies as facets



Agenda

- Taxonomy updating
- Taxonomy combining: integrating, merging, mapping
- Taxonomy translating



- Increase in the adoption and number of taxonomies/
controlled vocabularies (CVs)
- More is not always better
 - Combine
 - Reduce
 - Re-use
 - Simplify



Different Methods/Different Purposes:

- Integrating
- Merging
- Mapping



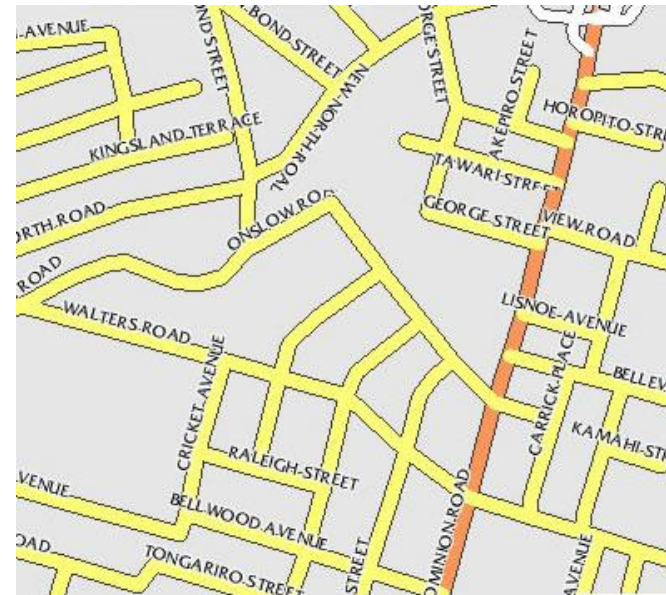
- **Integrating:**
Combining separate controlled vocabularies (CVs) into a single, larger master controlled CV for combined use



- 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



- 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



Taxonomy Combining: Integrating

- 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.



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.



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)



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>
```



Taxonomy Combining: Integrating

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

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.



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



Taxonomy Combining: Merging

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



Primary CV	Merging CV	Taxonomist Reviews
<div style="display: flex; align-items: center; justify-content: center;"> <div style="border: 1px solid black; padding: 5px; margin-right: 10px;">← ONE WAY</div> </div>		
<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



Taxonomy Combining: Merging

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



Tools for merging



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



Taxonomy Combining: 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

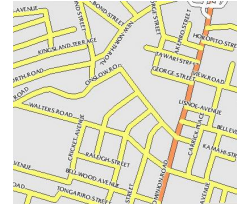


Retrieval taxonomy	Indexing taxonomy
Television sets	HDTV television sets



Situations

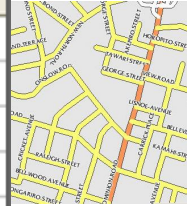
- Selected content with an enterprise CV is made available on a public web site with a different public-facing CV
- A content provider with a CV partners with a third-party information vendor with its own CV
- A provider of scientific/technical/medical content with a technical CV creates a simpler CV 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



Taxonomy Combining: Mapping



	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.)



Taxonomy Combining: Mapping



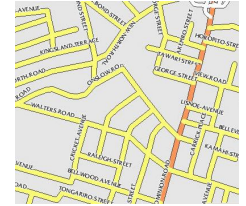
Mapping user-entered search queries (column 2) to terms, in this case the term “Type of Vehicles.”

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_Term	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	



Tools for mapping



- 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)

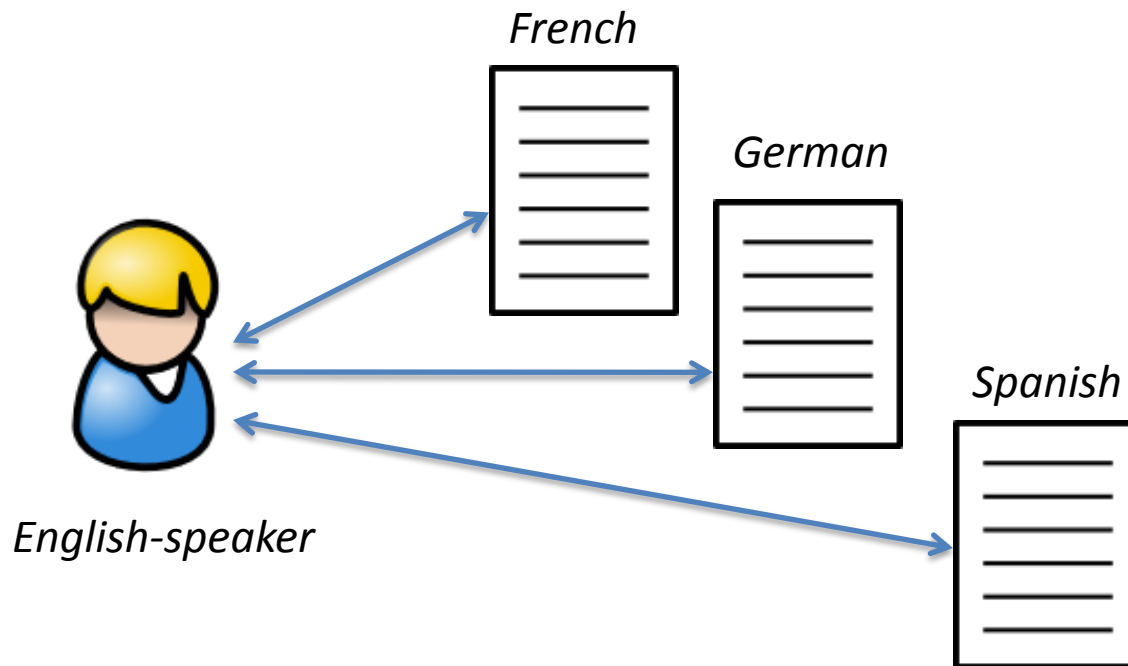


- Taxonomy updating
- Taxonomy combining: integrating, merging, mapping
- Taxonomy translating



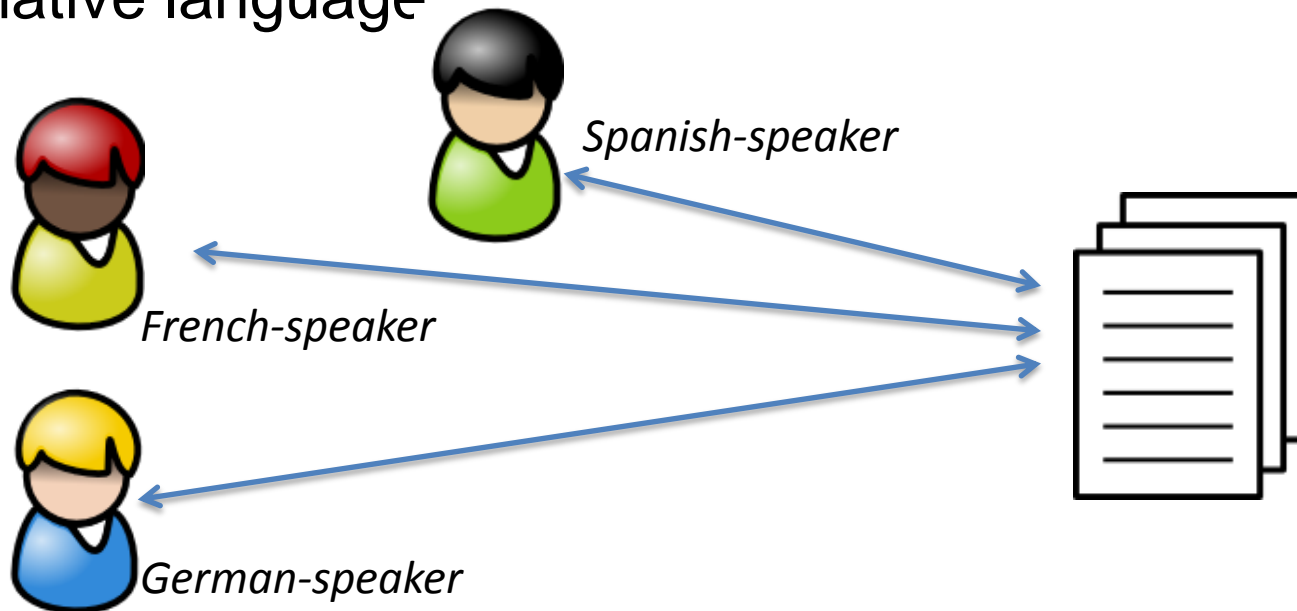
Bilingual/Multilingual Taxonomies enable:

1. A user to search and retrieve content that is in multiple languages through a single taxonomy in their own language



Bilingual/Multilingual Taxonomies enable:

2. Different users who speak different languages to search the same body of content (which may be in one language or more), each using a taxonomy in the user interface in their native language



- User interface taxonomies in one language may be mapped to indexing taxonomies in another language.
 - The retrieval taxonomy is in the language of the searcher.
 - The indexing taxonomy is in the language of the content.
 - The role of the different language taxonomies is typically dynamic
 - depending on the language of the user
 - depending on the language of the content
 - The taxonomy of either language could be the retrieval taxonomy or the indexing taxonomy.
- Mapping has to go in both directions.
- Matches between terms in both languages have to be exact translations.



Taxonomy Translating

- Translations of a term are managed as another kind of relationship.
- Similar to equivalence, but both languages are preferred and none is nonpreferred

young person	jeune
FD: 13. Population	MT: 13. Population
UF: young man young woman	EP: jeune femme jeune homme
BT: age group	TG: groupe d'âge
RT: young worker youth	VA: jeune travailleur jeunesse
FR: jeune	EN: young person

From the bilingual European Training Thesaurus <http://libserver.cedefop.europa.eu/ett>



- Taxonomies translations are typically created from scratch, translating each term.
- It is also possible to map an existing foreign language taxonomy to another, if their coverage is nearly identical.

при регулярной
Первая группа
сударè 18
лiгенiя, ЭКГ:
Metaanalyse
синусс



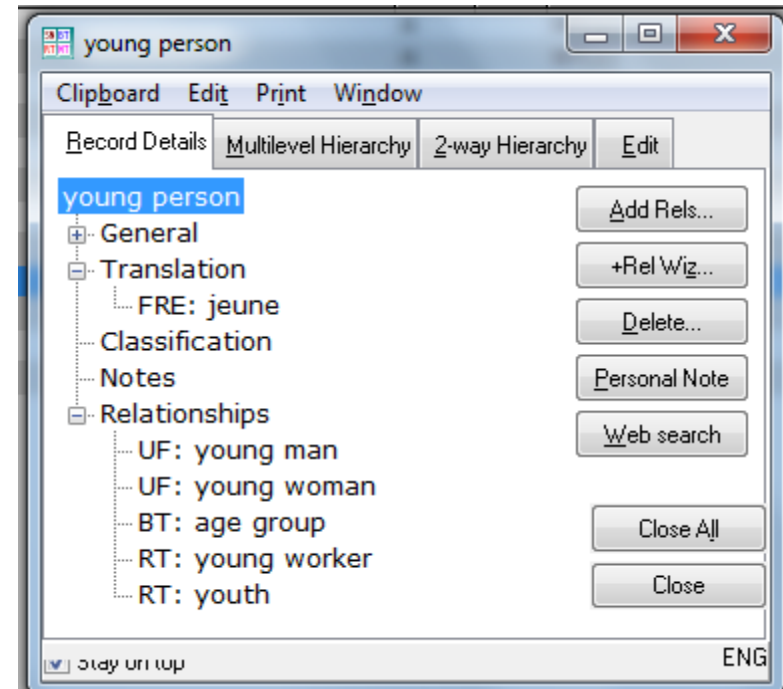
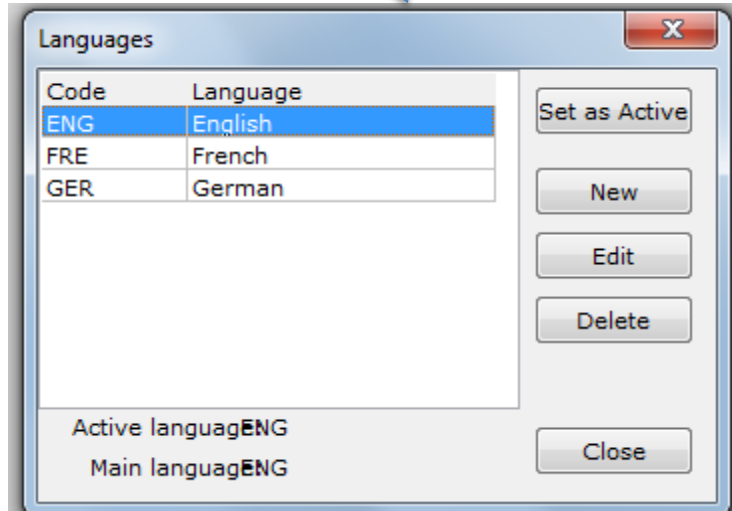
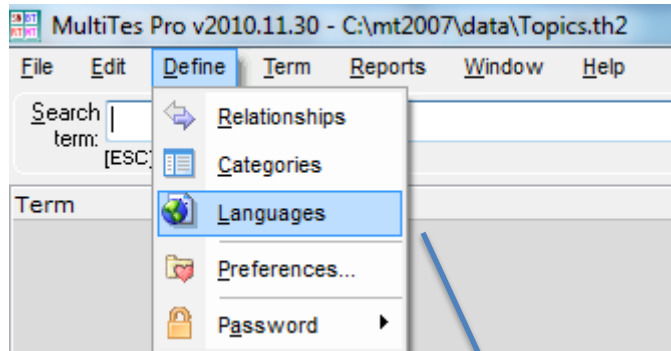
- Matches are for concepts, not terms.
 - Translations are for the concept and not necessarily for the preferred term.
- All hierarchical and associative relationships ideally should match in both languages.
- Nonpreferred terms may vary.
 - They don't all have to be translated, and probably shouldn't be.



Taxonomy Translating

Thesaurus Management Software may manage multilingual taxonomies

Screenshots from MultiTes



Translating taxonomies/thesauri is different from translating documents

- Pay by hour/project, not by word
- Translators should have experience with translating in both directions
- Translators should be familiar with using taxonomies
- Have taxonomist/information specialist native speaker of target languages review the translated taxonomy



Taxonomy Translation Issues

- Support for characters in the writing systems of other languages
- Translations of the user interface
 - menus, instructions, pop-ups, help files, etc.
- Other language issues
 - use of plural
 - Use of capitalization
 - Alphabetizing rules
 - “logical” sort orders of terms.
- Cultural issues
 - Could impact choice of facets



- Taxonomy updating
- Taxonomy combining: integrating, merging, mapping
- Taxonomy translating

In all cases:

- Need to be pro-active and anticipate and plan for the future
- Need to bring in additional experts: subject matter experts, technology experts, translators



Questions

Heather Hedden
Senior Taxonomy Analyst
Project Performance Corporation
Heather.Hedden@ppc.com
703-462-3746 (cell)
978-371-0822 (home office)
Corporate office: McLean, VA

