Mapping, Merging, and Multilingual Taxonomies

Heather Hedden

Taxonomy Consultant
Hedden Information Management

SLA 2012 Conference Presentation

Heather Hedden

- Taxonomy consultant, Hedden Information Management
- Continuing education instructor with Simmons College Graduate School of Library and Information Science
- Author of *The Accidental Taxonomist* (Information Today, 2010)

Previously worked as:

- Controlled vocabulary editor, IAC/Gale/Cengage Learning
- Internal taxonomy manager for an energy company
- Taxonomy consultant with consulting firms
- Taxonomist in product development at a search software vendor

Agenda

- Background
- Mapping Taxonomies
- Merging Taxonomies
- Multilingual Taxonomies

Agenda

- Background
- Mapping Taxonomies
- Merging Taxonomies
- Multilingual Taxonomies

Background: Taxonomies

Controlled Vocabulary/Taxonomy/Thesaurus

- An authoritative, restricted list of terms (words or phrases)
- Each term for a single unambiguous concept (synonyms/nonpreferred terms, as cross-references, may be included)
- Policies (control) for who, when, and how new terms can be added
- Typically has structured relationships between terms
- To support indexing/tagging/metadata management of content to facilitate content management and retrieval

Hierarchical taxonomy

- Agriculture
- Applied technologies
- Business
- Communications
  - Intercultural communication
  - Journalism
    - Broadcast journalism
    - Electronic journalism
    - Photojournalism
    - Print journalism
  - Mass communication
  - Mass media
  - Nonverbal communication
  - Oral communication
  - Propaganda
  - Public relations
  - Social commentary
  - Social communication
  - Subliminal communication
  - Telecommunication
  - Visual communication
- Computer and information science
- Economics
- Education
- Family and consumer sciences
- Geography
- Health and wellness
- History
- Language arts
- Languages
- Literature and drama

Thesaurus

patients
RT human beings
human pathology
therapy

Patriot missile
DEF Surface to air, antiaircraft missile.
GS missiles
. surface to air missiles
. . **Patriot missile**
RT missile configurations
∞ rockets
weapons

patrols
RT reconnaissance

pattern distribution
USE **distribution (property)**

pattern method (forecasting)
GS management methods
. **pattern method (forecasting)**
predictions
. forecasting
. . technological forecasting
. . . **pattern method (forecasting)**
RT Delphi method (forecasting)
estimating
∞ methodology
operations research
planning
probe method (forecasting)
technology assessment

pattern recognition
DEF The identification of shapes, forms and configurations by automatic means.
UF *automatic pattern recognition*
feature extraction
GS recognition
. **pattern recognition**
. . character recognition
. . graphology
RT change detection
clumps
cluster analysis
computer vision

Faceted Taxonomy

Narrow by

Category

Select category(s) Clear

- Banquet Tables (4)
- Bistro Table (2)
- Bistro Tables (5)
- Counter-Height Table (1)
- Counter-Height Tables (6)
- Dining Table (10)
- Dining Tables (52)
- Folding Table (8)
- Folding Tables (12)
- Kitchen Table (1)
- Kitchen Tables (1)
- Nook Table (1)
- Nook Tables (1)
- Pub Table (7)
- Pub Tables (29)

Material

Select material(s) Clear

- Hardwood (29)
- MDF Composite (1)
- Metal (28)
- Plastic (1)
- Wood (48)
- Wood Composite (35)

Finish

Select finish(s) Clear

- Cherry (4)
- Dark Cherry (1)
- Ebony (1)
- Espresso (14)
- Mahogany (5)
- Natural (7)
- Oak (7)
- Painted (8)
- Unfinished (1)
- Walnut (8)

Color

Background:
Mapping, Merging, & Multilingual Taxonomies

Taxonomies/Controlled Vocabularies (CVs) are:

1. Designed
2. Built
3. Maintained/Managed

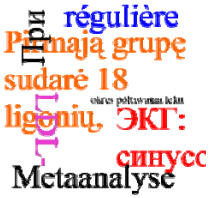
But in time, a taxonomy may gain additional uses, and may need to be:

- Mapped or merged with another taxonomy
- Translated into another language or localized

Background:
Mapping, Merging, & Multilingual Taxonomies

Mapping, Merging, and Multilingual Taxonomies:

- Methods of combining taxonomies
- Different methods > Different purposes

  - Mapping 
  - Merging 
  - Multilingual 

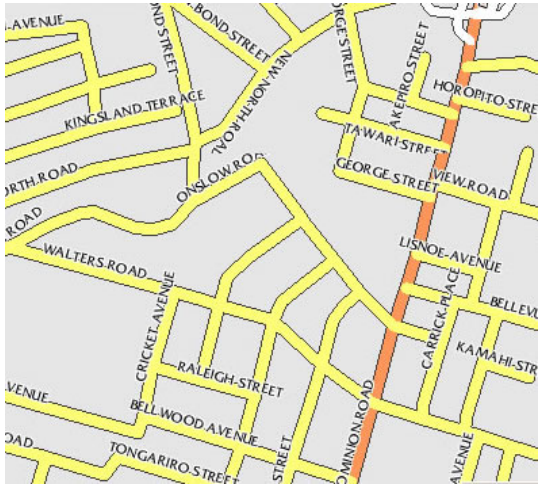
Agenda

- Background
- Mapping Taxonomies
- Merging Taxonomies
- Multilingual Taxonomies

Mapping Taxonomies

Mapping:
Enabling one controlled vocabulary (CV) to be used for another in the 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 CV.
- Mapping tables also called “crosswalks”



Something representing something else

Mapping Taxonomies

Situations:

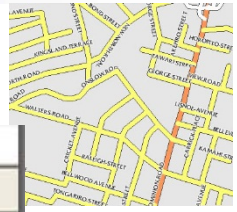
- Selected content with an enterprise taxonomy is made available on a public web site with a different public-facing taxonomy
- A content provider with a CV partners with a third-party information vendor with its own CV
- A provider of scientific/technical/medical content with a technical CV creates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

Mapping Taxonomies

- *From* a CV indexed to content *to* a retrieval/user-interface CV
- Use a software tool or scripts to compare vocabularies, to obtain matches in succeeding passes.
- Human review confirms and approves automatically proposed matching terms.
- Unmatched terms cannot be utilized.
- Narrower-to-broader matches are fine.
- Set automatic matches to also include matches of words/phrases of the retrieval taxonomy *within* a term from the indexing CV.

Indexing taxonomy	**Retrieval/UI taxonomy**
HDTV Television sets	Television sets

Mapping Taxonomies



F29

	A		C
1	Programmable logic controller		mable 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" means second term is broader so also ok, and "n" is narrower or otherwise not acceptable.)

Mapping Taxonomies

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_Terr	CV_Terms_ Y
Makes	GVX	y	
Type of Vehicles	4 Wheel Drive	y	y
Type of Vehicles	Four Wheel Drive	y	y
Type of Vehicles	4x4	y	
Type of Vehicles	4 X 4	y	
Type of Vehicles	4x4s	y	
Type of Vehicles	4WD	y	
Type of Vehicles	All Wheel Drive	y	y
Type of Vehicles	AWD	y	
Type of Vehicles	Classic	y	
Type of Vehicles	Vintage	y	
Type of Vehicles	Antique	y	
Type of Vehicles	Commercial Vehicles	y	y
Type of Vehicles	Commercial Trucks	y	y
Type of Vehicles	Commercial Vans	y	y
Type of Vehicles	Fleets	y	
Type of Vehicles	Convertibles	y	y
Type of Vehicles	Coupes	y	y
Type of Vehicles	Diesel	y	
Type of Vehicles	Domestic	y	

Mapping Taxonomies

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

Agenda

- Background
- Mapping Taxonomies
- Merging Taxonomies
- Multilingual Taxonomies

Merging Taxonomies

- 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

Merging Taxonomies

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 purchased/licensed CV

Merging Taxonomies

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 Taxonomies

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

Merging CV *(will go away)*	Primary CV *(Keep and grows)*	Taxonomist Reviews
Exact matches of:		
Preferred term: Cars	*Preferred term: Cars*	no need
Preferred term: Automobiles	*Nonpreferred term: Automobiles* *USE Cars*	no need
Nonpreferred term: Cars *USE Automobiles*	*Preferred term: Cars*	yes
Nonpreferred term: Cars *USE Automobiles*	*Nonpreferred term: Cars* *USE Autos*	yes
Inexact matches of:		
Preferred term: Automobile	*Preferred term: Automobiles*	yes

Merging Taxonomies

Can create rules for automatic inexact or "fuzzy" matches, then subject to human review:

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

Merging Taxonomies

Tools for merging

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

Mapping and Merging Summary

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

- Merging
  - Overlapping Controlled Vocabularies combined permanently, removing duplicates

Mapping and Merging Summary

- Compare two closely redundant vocabularies side-by-side, term-by-term
- First pass is automatic, followed by taxonomist review of matches
- Taxonomy software may have the feature, 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.

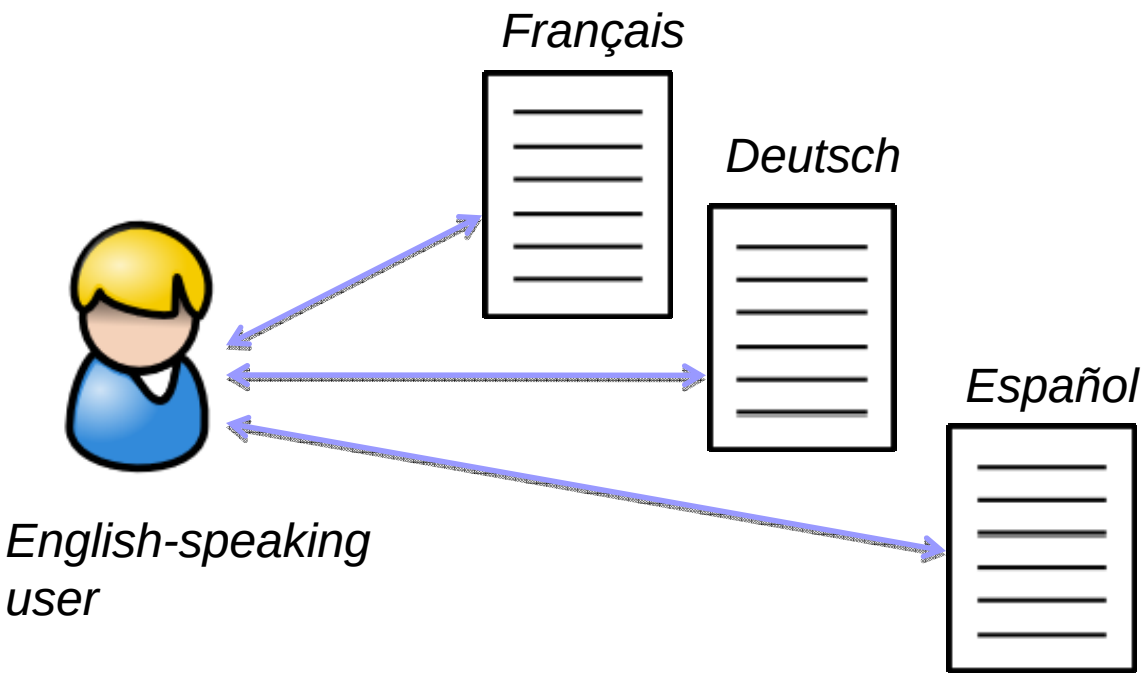
Agenda

- Background
- Mapping Taxonomies
- Merging Taxonomies
- Multilingual Taxonomies
  1. Multilingual Taxonomy Goals
  2. Multilingual Taxonomy Design
  3. Taxonomy Translation Management

Multilingual Taxonomy Goals

Bilingual/Multilingual Taxonomies can enable:

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



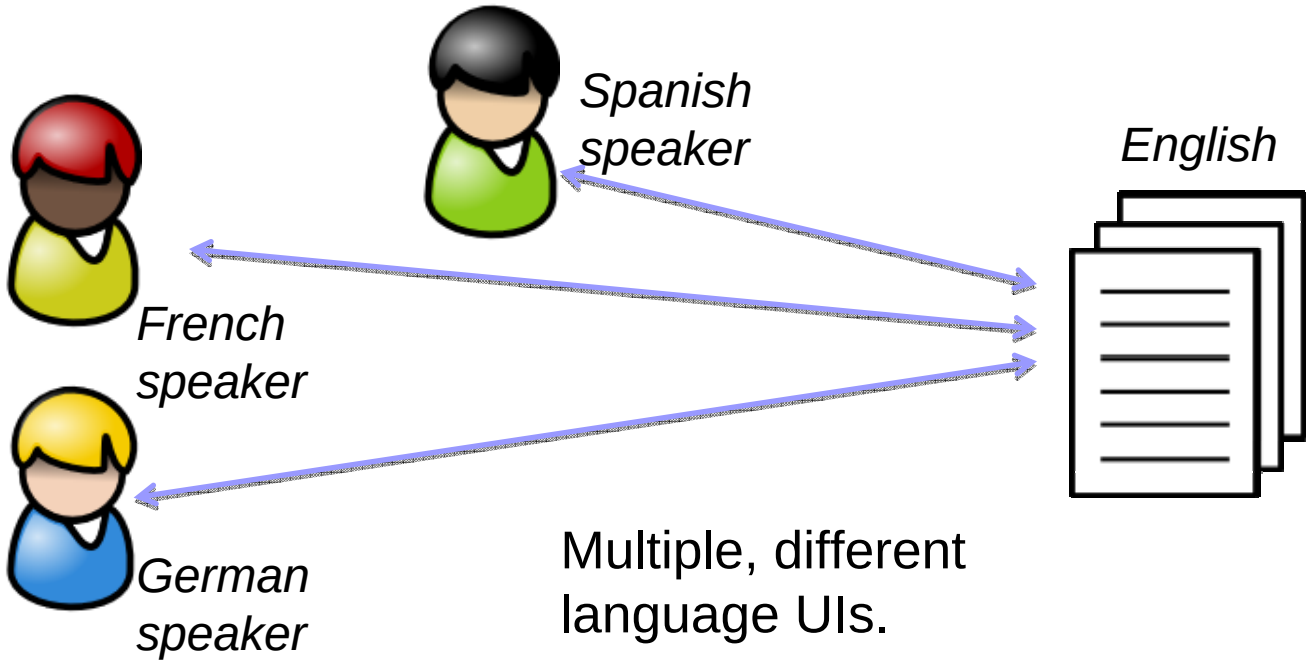
Taxonomy: Single-language user interface (UI).

Multiple language translations, not displayed.

Multilingual Taxonomy Goals

Bilingual/Multilingual Taxonomies can enable:

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

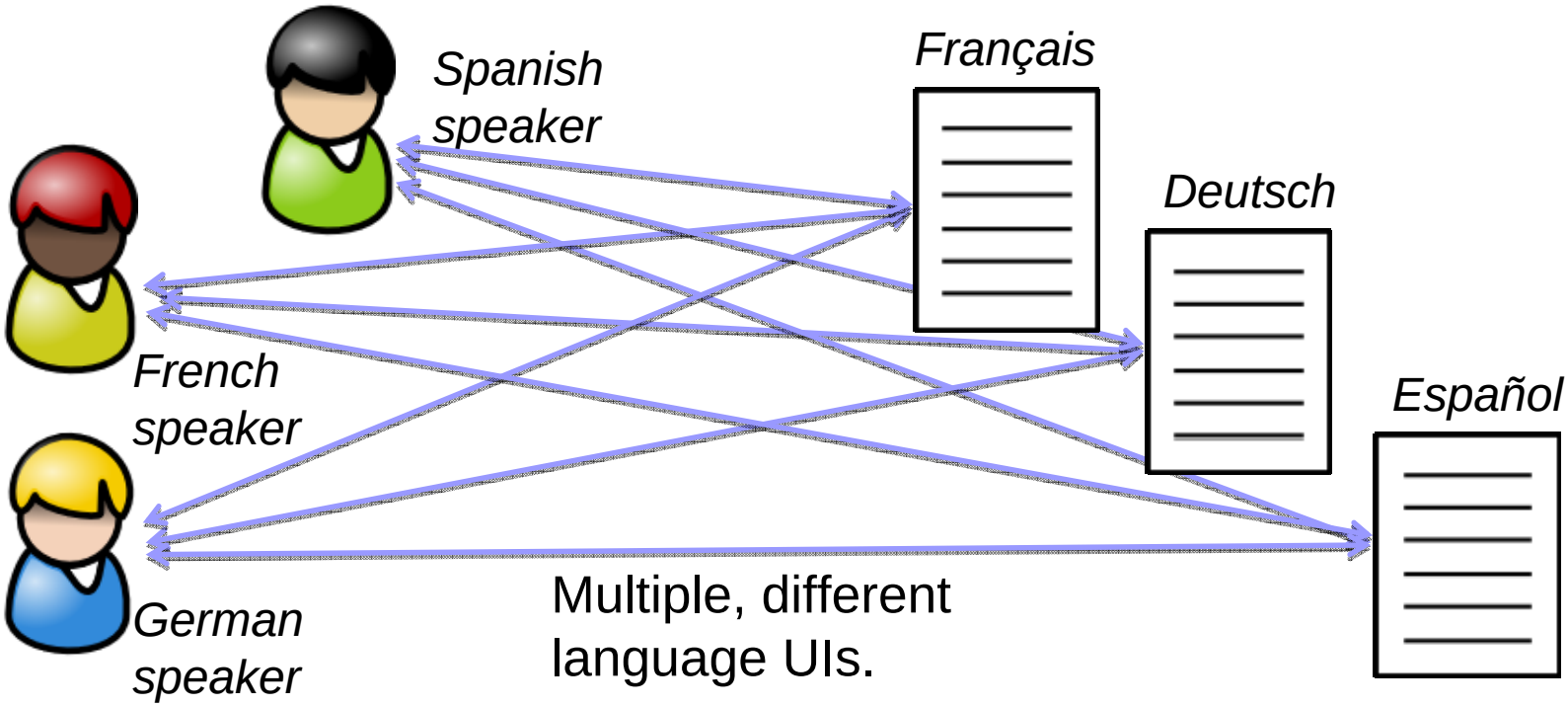


Multiple, different language UIs.

Multilingual Taxonomy Goals

Bilingual/Multilingual Taxonomies can enable:

3. Different users who speak different languages to search the same body of content that is in multiple languages.



Multilingual Taxonomy Goals

Goals #1 or #2: *Users of one language can access content in a different language.*

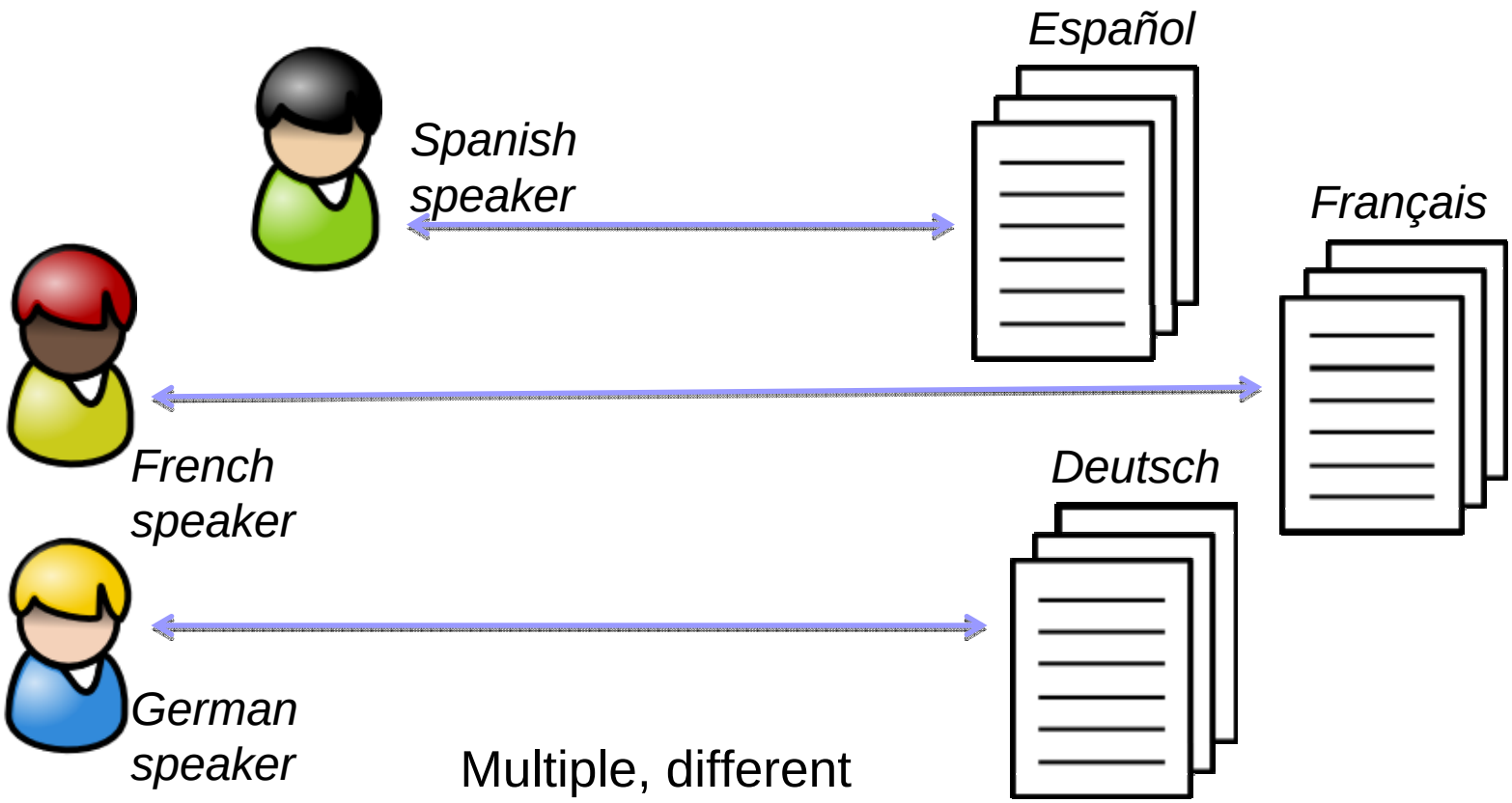
- Taxonomy in one language with equivalent translated terms
- The taxonomy needs to function in only one direction.

Goal #3: *Multilingual users can access multilingual content.*

- *Fully* multilingual taxonomy or distinct taxonomies for each language linked at equivalent-meaning terms
- The taxonomy needs to function in both/all language directions.

Multilingual Taxonomy Goals

Different scenario: Multiple language taxonomies, each connected to its own language content, such as for separate web sites.



Multiple, different language UIs.

Multilingual Taxonomy Design

Design the multilingual taxonomy to meet the taxonomy goals.

- In a one-direction translated taxonomy:
  - The language of the searcher has structure to display.
  - The language of the content may not need structure.
  - Translations may be in one direction (user/display term *may be used for* content/index term, not vice versa).
- For a fully bidirectional multilingual taxonomy:
  - Both language taxonomies need structure.
  - Translations must be exact matches in *both* directions.
- For separate taxonomies in different languages:
  - Taxonomies are not translated but each created and managed separately.

Multilingual Taxonomy Design

Dedicated taxonomy/thesaurus management software tools provide varying multilingual capabilities.

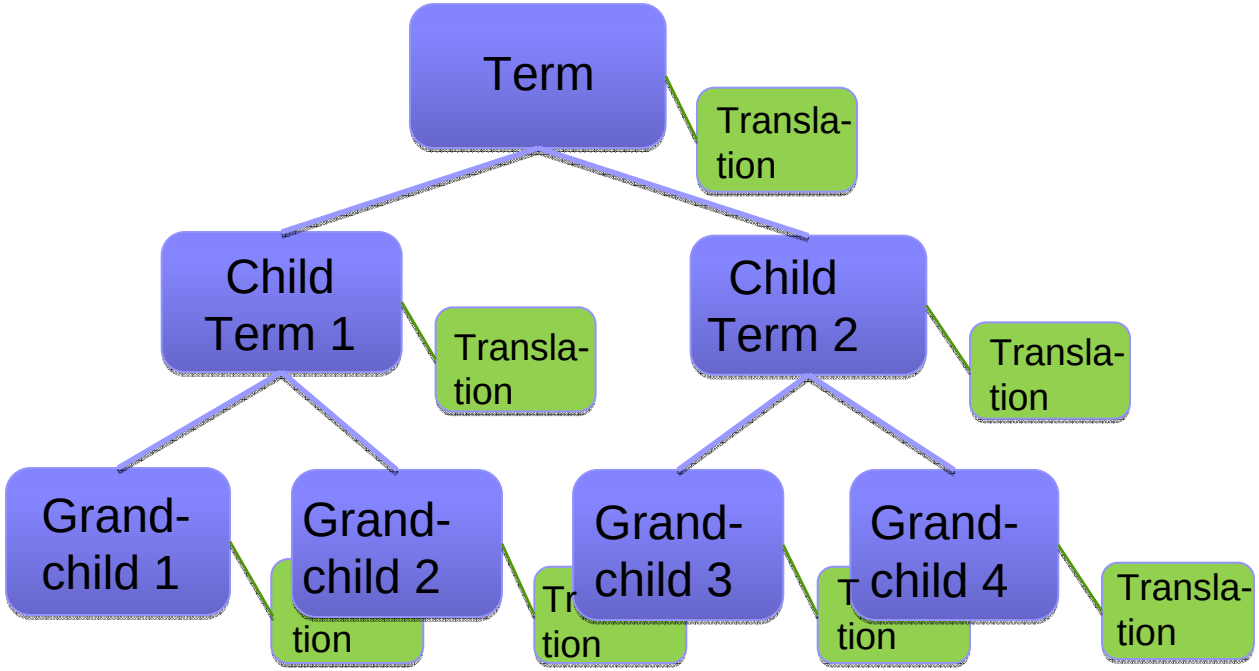
1. Customized text field used for term translations
   - No vocabulary control of second language(s)

2. Second language taxonomy mirroring first, linked at each translated term
   - Vocabulary control of second language(s)
   - Copying taxonomy structure of primary language

3. Multiple taxonomies in different languages linked at equivalent term translations
   - Each language may have its own structure (requires additional work to build)

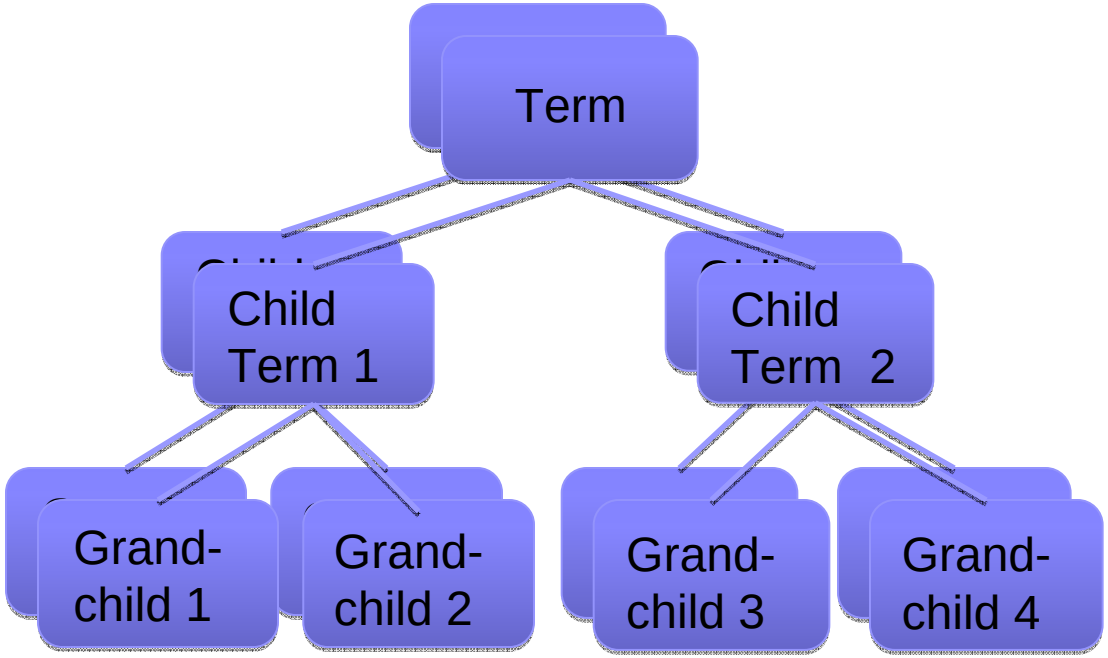
Multilingual Taxonomy Design

1. Customized field used for term translations



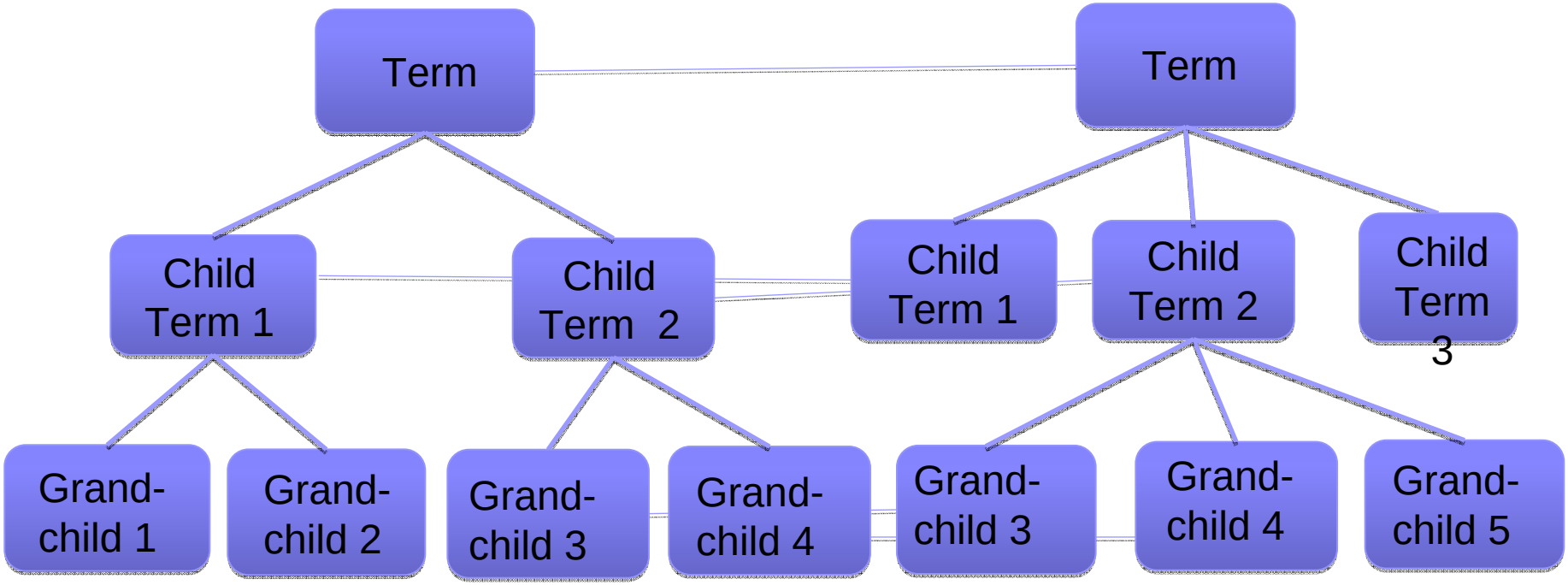
Multilingual Taxonomy Design

2. Second language taxonomy mirroring first, linked at each translated term. Inter-term relationships replicate.



Multilingual Taxonomy Design

3. Multiple taxonomies in different languages linked at equivalent term translations. Inter-term relationships may differ.



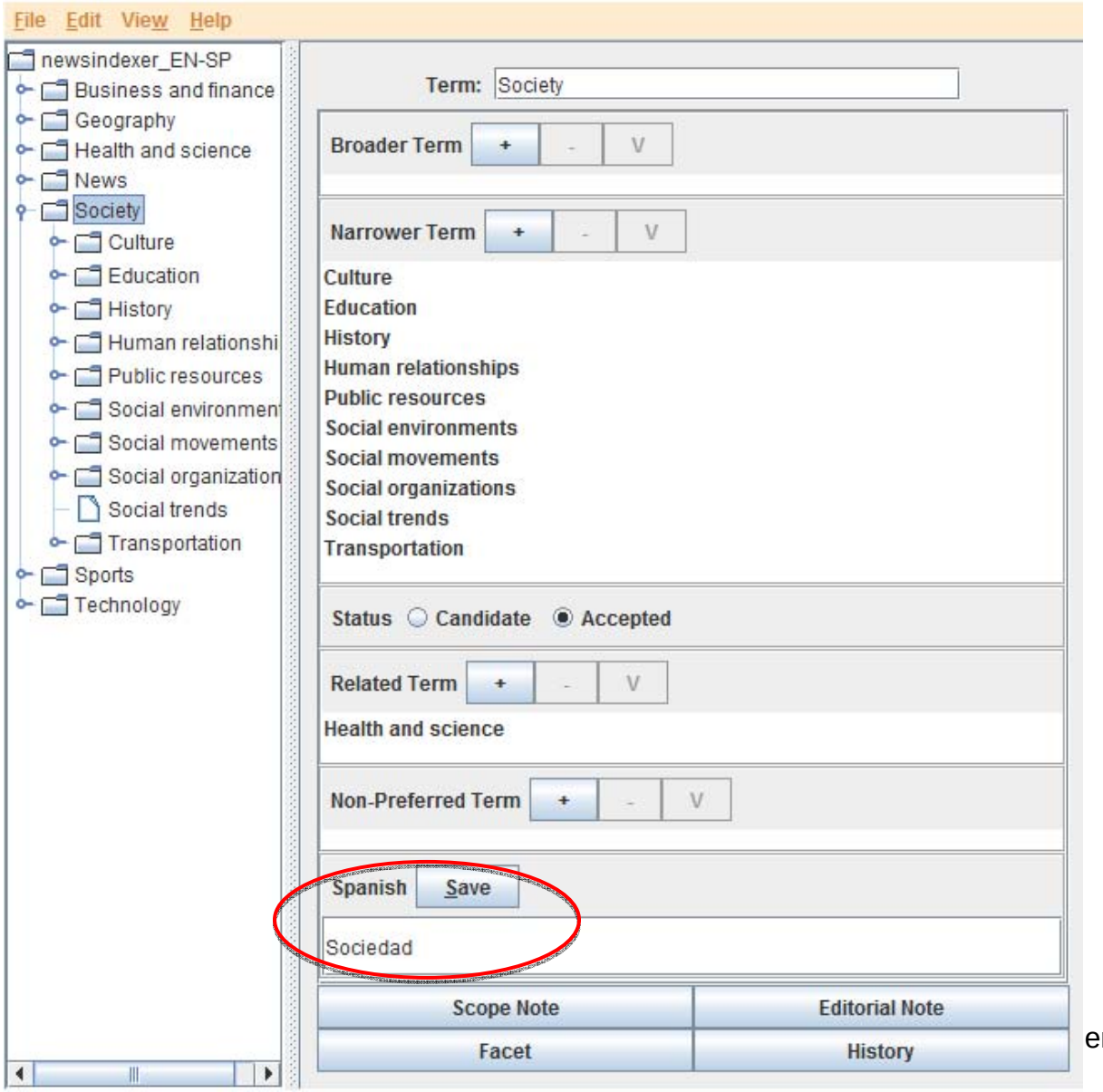
Multilingual Taxonomy Design & Tools

Dedicated taxonomy/thesaurus management software tool screenshot examples from:

- Data Harmony Thesaurus Master (Access Innovations, Inc.)
- Synaptica (Synaptica, LLC)
- MultiTes (Multisystems)
- Semaphore Ontology Manager (Smartlogic)

Additional tools also provide similar capabilities.

Multilingual Taxonomy Design & Tools

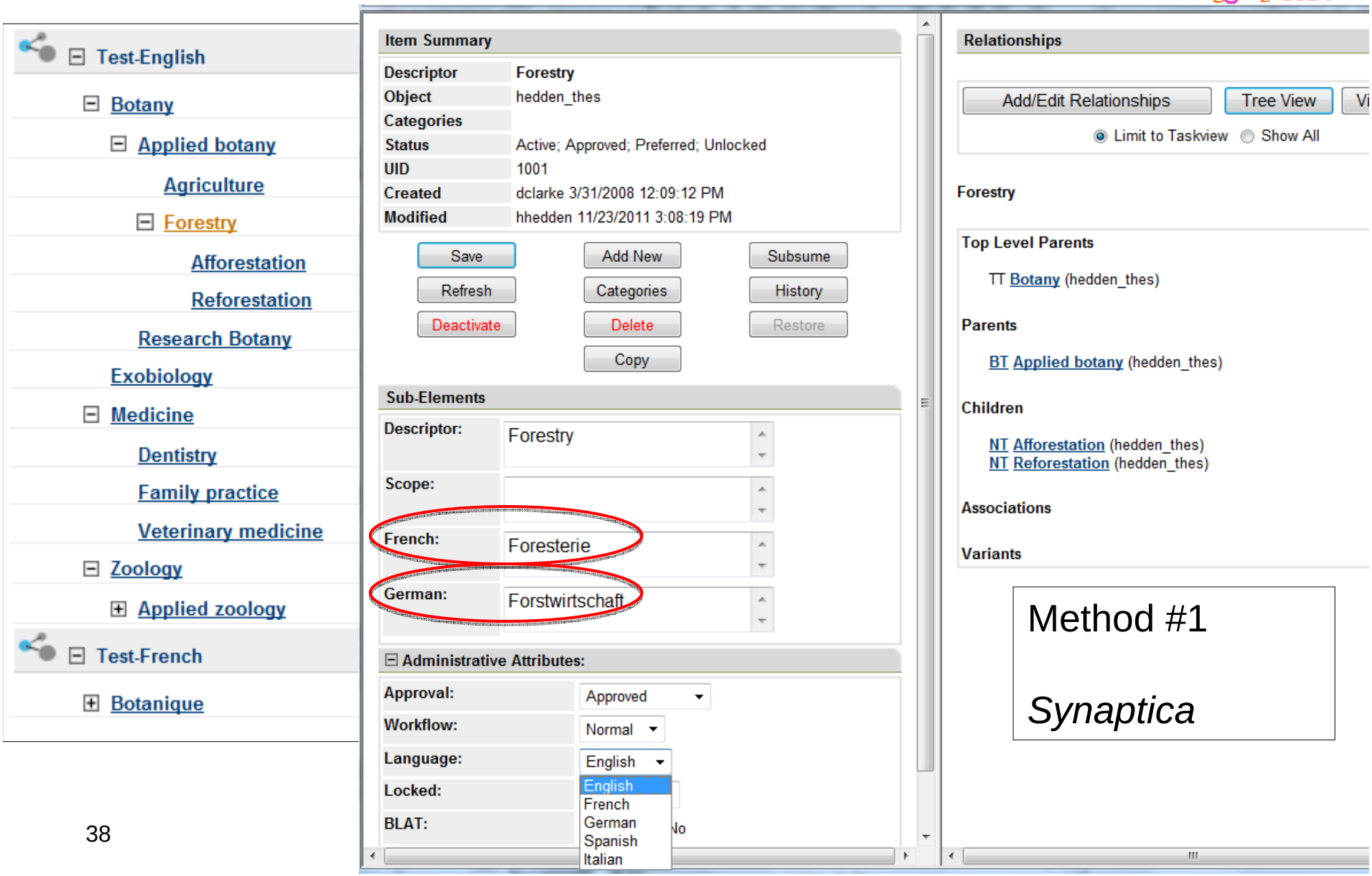


ent

Method #1:
Create user-defined text field and enter translation

Data Harmony Thesaurus Master

Multilingual Taxonomy Design & Tools

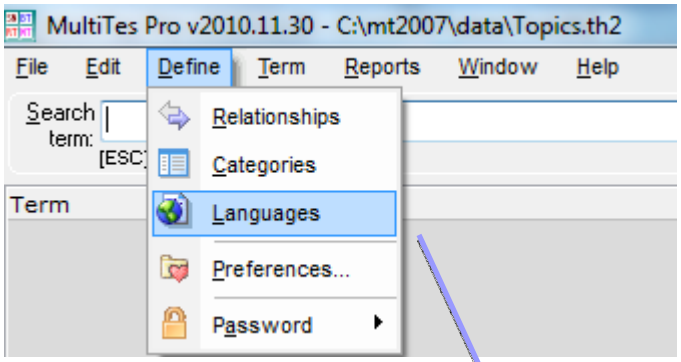


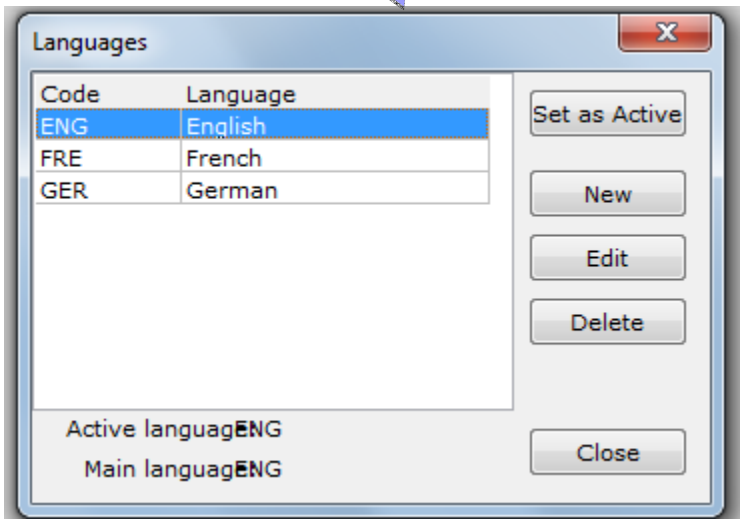
Method #1

Synaptica

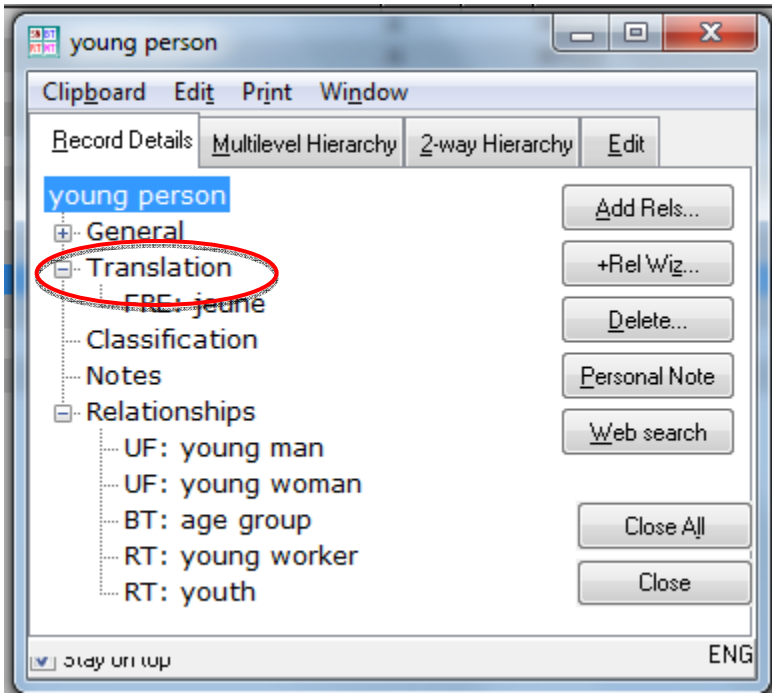
Multilingual Taxonomy Design & Tools

Method #2: Create second language taxonomy mirroring first, linked at each translated term. Inter-term relationships replicate.

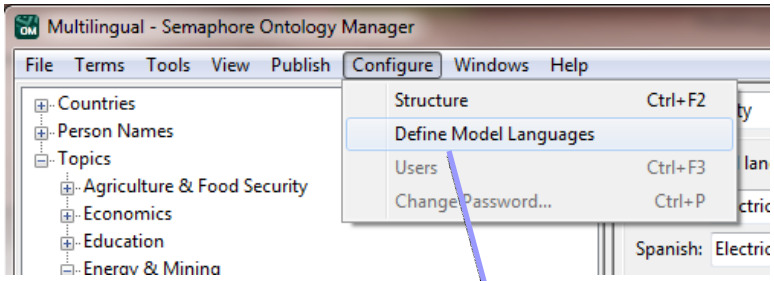




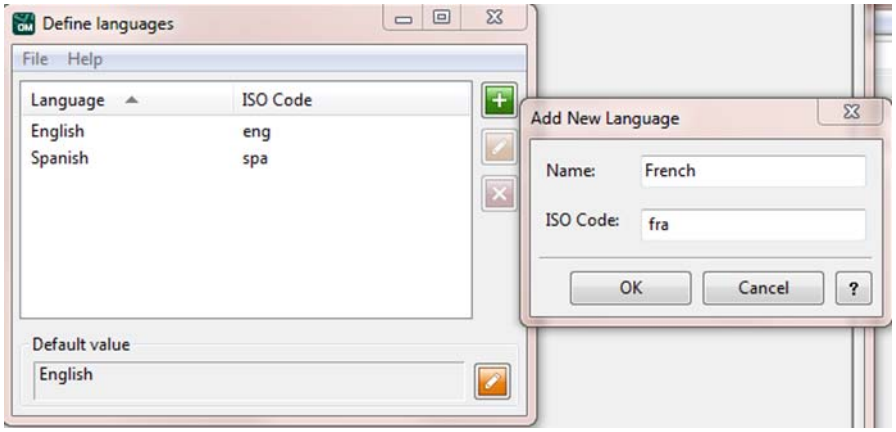
MultiTes

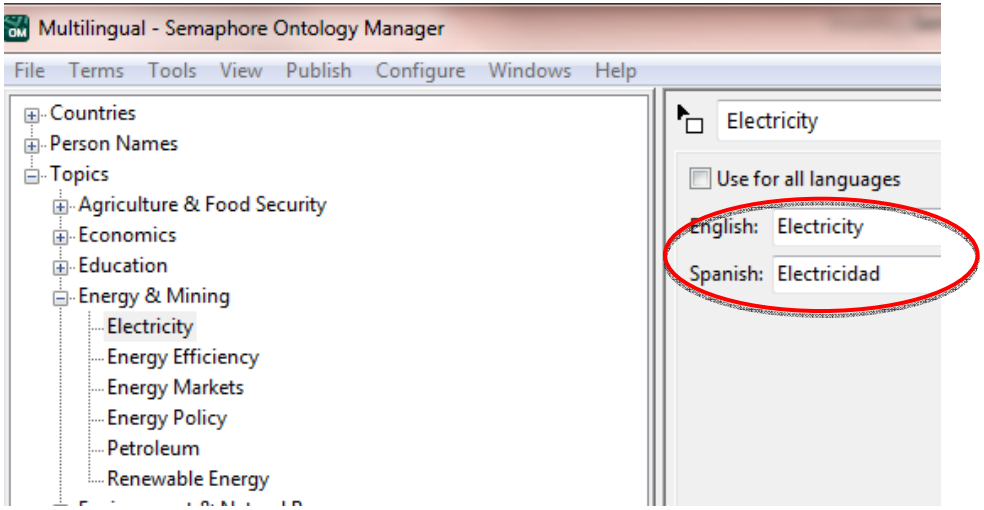


Multilingual Taxonomy Design & Tools

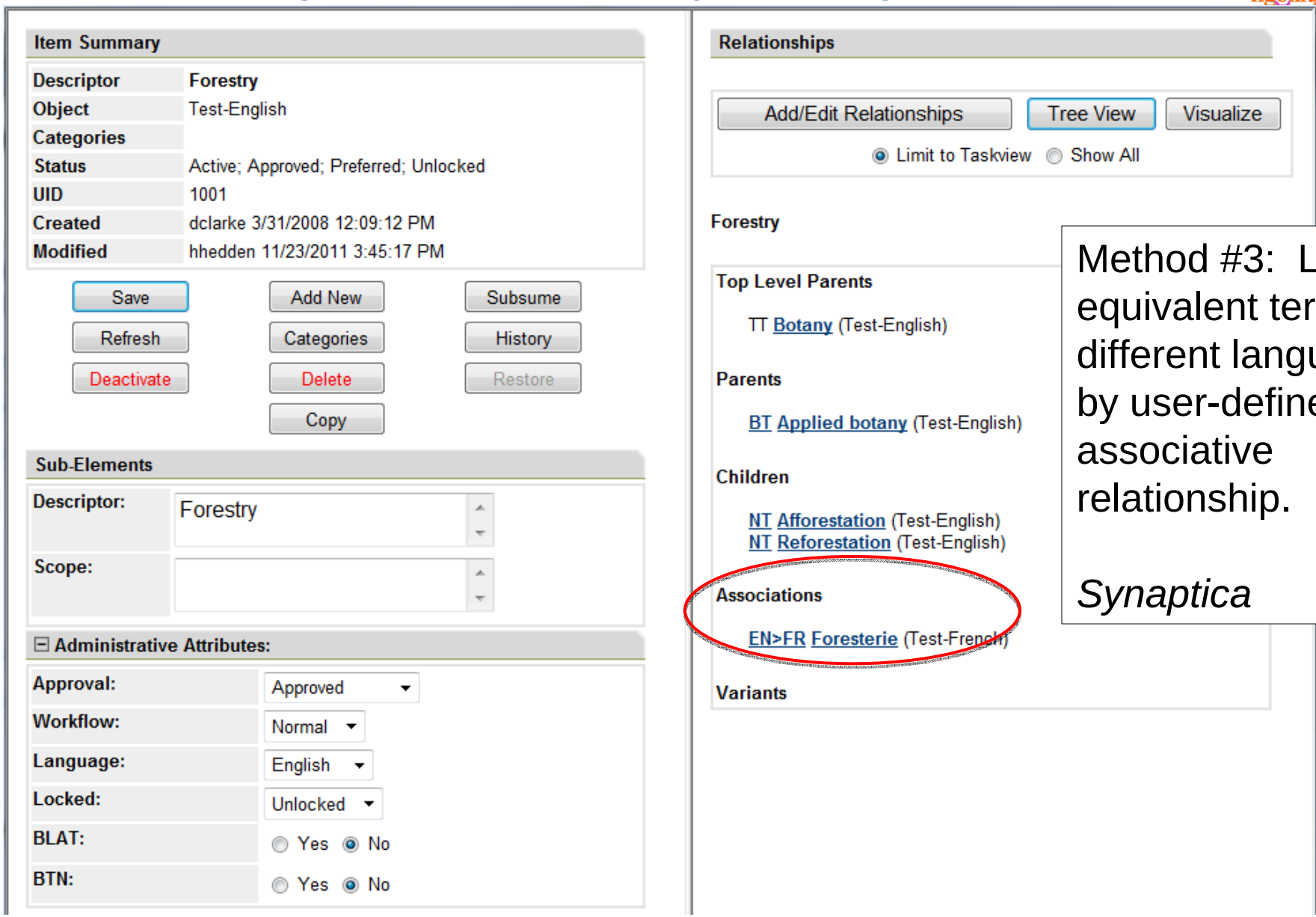


Method #2:
Smartlogic Semaphore Ontology Manager





Multilingual Taxonomy Design & Tools



Method #3: Link equivalent terms in different language by user-defined associative relationship.

Synaptica

Multilingual Taxonomy Design

- Translations of a term may display 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

Taxonomy Translation Management

- Taxonomy translations are typically created from scratch, translating each term.
- It is also possible to map and existing/separately created foreign language taxonomies to another, if their coverage is nearly identical.

- For Goals #1 or #2 (*Users of one language accessing content in a different language)* translations may suffice
- For Goal #3 *(Multilingual users accessing multilingual content)* mapping separately created taxonomies in each language is better.

Taxonomy Translation Management

- 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 Translation Management

- Matches are for concepts, not terms.
  - Translations are for the concept and not necessarily for the preferred term.

- Nonpreferred (variant/synonym) terms may vary.
  - Some can be translated
  - Some cannot be translated
  - Additional nonpreferred terms may be created in the second language(s)

Taxonomy Translation Management

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, if not also taxonomists.
- If not using a translator who is also a taxonomist, have a taxonomist/information-specialist native speaker of target languages review the translated taxonomy.

Taxonomy Translation Management

Taxonomy Translation Issues

- Lack of an equivalent translation
- A term in one language having two meanings with two terms in another language
  (e.g. seguridad = safety or security)
- Term length
- Use of definite articles
- Use of abbreviations
- Use of plural
- Use of capitalization
- Alphabetizing sorting rules

Taxonomy Translation Management

Translation projects end, but taxonomy management does not.

Taxonomy management issues:

- Taxonomy growth
- Taxonomy change
- Taxonomy management/ownership responsibility
- Merging or combining additional taxonomies

➢ Translations/additional language versions will need frequent reviewing and updating.

Conclusions

- Mapping Taxonomies
- Merging Taxonomies
- Making Multilingual Taxonomies

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

Additional Taxonomy Resources/Training

Book: *The Accidental Taxonomist*
2010, Information Today, Inc.
www.accidental-taxonomist.com

Taxonomies & Controlled Vocabularies 5-week online workshop
Simmons College Graduate School of Library & Information Science
Starting November, 2012, and January, 2013
http://alanis.simmons.edu/ceweb

SLA Taxonomy Division
http://taxonomy.sla.org

Contact

Heather Hedden
Hedden Information Management
Carlisle, MA
heather@hedden.net
www.hedden-information.com
accidental-taxonomist.blogspot.com
Twitter: @hhedden
978-467-5195