About the Speaker

Heather Hedden

Data and Knowledge Engineer

Semantic Web Company

Over 25 years of experience in developing and managing taxonomies, metadata, and other knowledge organization systems for various organizations and applications.

Instructor of self-paced online taxonomy courses.

Prior taxonomy consultant and staff taxonomist.

Author of *The Accidental Taxonomist*.

Semantic Web Company (SWC) and PoolParty

SWC is developer / vendor of **PoolParty Semantic Suite**

Most complete and secure **Semantic Middleware / Semantic AI platform** on the Global Market

W3C standards compliant

ISO 27001:2013 certified

First release in 2009

Current version **8.1**

On-premises or **cloud-based**

Over **200** installations world-wide

Semantic AI: Fusion of Graphs, NLP, and Machine Learning

Named as Visionary in **Gartner's Magic Quadrant** for Metadata Management Systems 2019, 2020

KMWorld listed PoolParty as one of the **Trend-Setting Products** 2015 - 2020 and listed SWC in the **AI 50** list of companies in 2020

Outline

- Introduction to taxonomies
- How to build taxonomies
- Licensing or acquiring taxonomies

Introduction to Taxonomies

Introduction to Taxonomies

Why taxonomies?

- Concepts/terms are used to tag/index/categorize content to make it easier to be found and retrieved
  - ▷ supporting better findability than search alone
- The taxonomy is an intermediary that links the user to the desired content.

Content   Taxonomy   Users
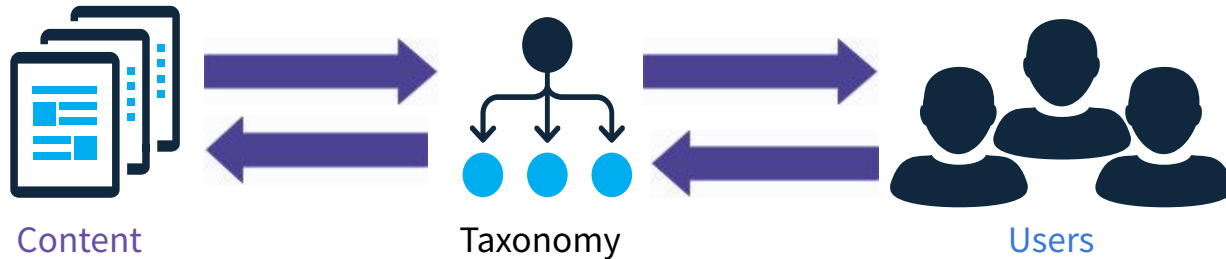
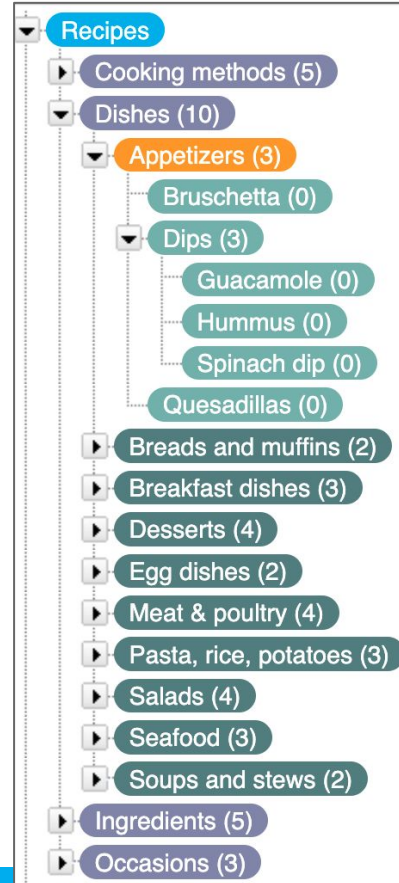- Taxonomies usually link to content; ontologies can also link to data.

Introduction to Taxonomies
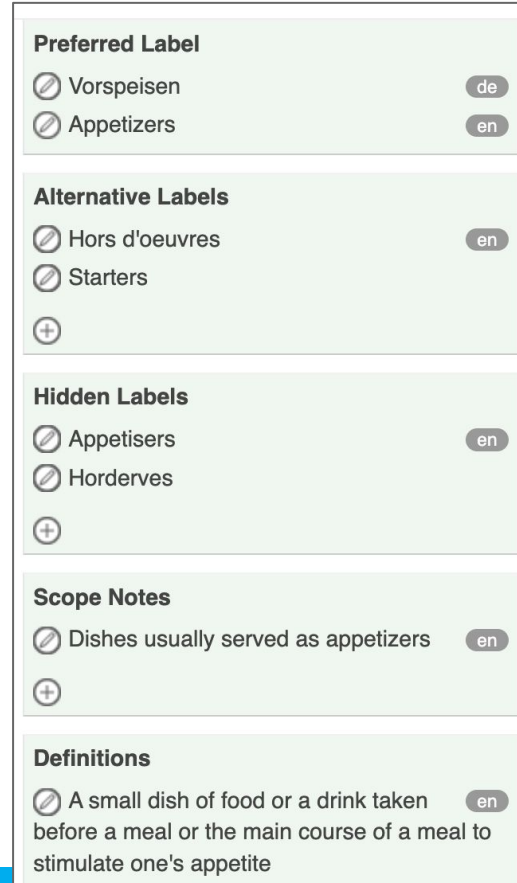
What is a taxonomy?

Controlled and *organized*

1. A kind of controlled vocabulary or knowledge organization system, based on unambiguous concepts, not just words:
things, not strings

2. Concepts are arranged in a structure of hierarchies, categories, or facets to organize them.

organized

- Recipes
  - ▶ Cooking methods (5)
  - ▼ Dishes (10)
    - ▼ Appetizers (3)
      - Bruschetta (0)
      - ▼ Dips (3)
        - Guacamole (0)
        - Hummus (0)
        - Spinach dip (0)
      - Quesadillas (0)
    - ▶ Breads and muffins (2)
    - ▶ Breakfast dishes (3)
    - ▶ Desserts (4)
    - ▶ Egg dishes (2)
    - ▶ Meat & poultry (4)
    - ▶ Pasta, rice, potatoes (3)
    - ▶ Salads (4)
    - ▶ Seafood (3)
    - ▶ Soups and stews (2)
  - ▶ Ingredients (5)
  - ▶ Occasions (3)

controlled

Preferred Label
- ✎ Vorspeisen — de
- ✎ Appetizers — en

Alternative Labels
- ✎ Hors d'oeuvres — en
- ✎ Starters
- ⊕

Hidden Labels
- ✎ Appetisers — en
- ✎ Horderves
- ⊕

Scope Notes
- ✎ Dishes usually served as appetizers — en
- ⊕

Definitions
- ✎ A small dish of food or a drink taken before a meal or the main course of a meal to stimulate one's appetite — en

Introduction to Taxonomies

Benefits of taxonomies and other controlled vocabularies

1. ### Controlled vocabulary
   Brings together different wordings (synonyms) for the same concept
   - ▷ Helps people search for information by different names

2. ### Classification and structure
   Organizes information into a logical structure
   - ▷ Helps people browse or navigate for information
   - ▷ Provides context and meaning for concepts for indexing and retrieval

Scope issues: "Taxonomy" sometimes refers to any kind of controlled vocabulary

Building Taxonomies

How to Build a Taxonomy

1. Identify use cases and implementations
2. Define the taxonomy scope
3. Determine taxonomy type(s)
4. Design the structure, top categories (facets, concept schemes) and top terms
5. Gather terms/concepts
6. Edit concept labels and their relationships and attributes
7. Review, evaluate, test or validate, and revise
8. Document and plan for governance

1. Identify Uses and Implementations

Through stakeholder meetings/interviews

Determine the uses of an implementations for taxonomies. Possibilities are:
- Topic/category browsing
- Search (matching search strings to concepts)
- Consistent tagging/indexing
- Discovery (related concept links, or content sharing the same concepts)
- Filtering results
- Sorting results
- Content management workflow (rights, audience, retention, etc.)
- Consistent metadata for identification, comparison, analysis
- Visualization of topics (importance and/or relations)
- Curated content in feeds or info boxes
- Automatic linking of relevant topics for personalization or recommendation systems
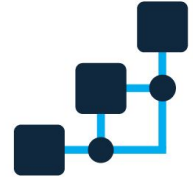- When integrated with ontologies, support for knowledge graphs

2. Determine the Taxonomy Type(s)

As suited for the content, implementation, and use

Possible taxonomy types:

- ▶ Set of term lists / faceted taxonomy
- ▶ Hierarchical taxonomy
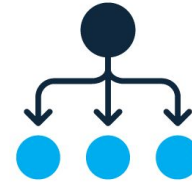- ▶ Faceted + hierarchical combination
- ▶ Thesaurus
- ▶ Taxonomy + ontology

2. Determine the Taxonomy Type(s)

As suited for the content, implementation, and use

Possible taxonomy types:
- Set of term lists / faceted taxonomy (as metadata filters/refinements)
- Hierarchical taxonomy (tree structure of categories and subtypes)
- Faceted + hierarchical combination (hierarchy within facets or concepts with attributes)
- Thesaurus (hierarchical and associative relationships, but less of a tree structure)
- Taxonomy + ontology (with custom relations and attributes)

Multiple taxonomies?
- Different taxonomies for different use cases
- Linked/mapped together, or designated concepts or branches of a single "universal" taxonomy

3. Define the Taxonomy Scope

Considerations:

▸ Taxonomy coverage breadth and depth/level of detail
▸ Types of concepts to include (document types, subjects, places, people, organizations, methods/activities, events, products, etc.)

Based on:

▸ What kinds of content that will be tagged
▸ What is *in* the content that will be tagged
▸ What is the purpose of the taxonomy implementation
▸ Who will be the users

Example for a knowledge base: content management system pages, wiki pages, help documentation, conference and internal presentations, training materials.

4. Design the Structure and Top Categories

Taxonomy should be designed for the users and for the content.

Obtain input and insights to inform the design from both the users and the content.

From users by means of:

- ▷ Stakeholder interviews
- ▷ Brainstorming workshops - such as PoolParty CardSorting

From content by means of:

- ▷ Existing metadata properties
- ▷ Spreadsheet data column headers
- ▷ Manual review and analysis of unstructured content

4. Design the Structure and Top Categories

User sources: Brainstorming workshop option

PoolParty Cardsorting:
Users create "cards" writing down concept ideas, which are grouped, and then candidate concepts can be integrated into the taxonomy.

4. Design the Structure and Top Categories

Content sources: Manual review and analysis of unstructured content

	Article Title	Abstract	Population (Patients)	Medical procedure, intervention, or care	Music interaction or other therapy	Health Aspect or Condition	Music type or genre	Study method	Outcome/ What is measured/ evaluated	Tests	Setting	Other
14	Perioperative music may reduce pain and fatigue in patients undergoing laparoscopic cholecystectomy	BACKGROUND: Acute post-operative pain is a predictor in the development of chronic pain after laparoscopic cholecystectomy. Music has been shown to reduce surgical stress. In a randomized, clinical trial, we wanted to test the hypothesis that perioperative and post-operative soft music reduces pain, nausea, fatigue and surgical stress in patients undergoing laparoscopic cholecystectomy as day surgery. METHOD: The study was performed in otherwise healthy	Surgical patients	Laparoscopic cholecystectomy	Perioperative music listening	surgical pain (post-operative pain), post-operative nausea, post-operative fatigue, surgical	Soft music	Randomiz ed control trial (randomiz ed clinical trial); Longitudin		visual analog score, c-reactive protein test, cortisol test	Hospital, Home	
15	Improved recovery after music and therapeutic suggestions during general anaesthesia: a double-blind	PURPOSE: This study was designed to determine whether music or music in combination with therapeutic suggestions in the intra-operative period under general anaesthesia could improve the recovery of hysterectomy patients. METHODS: In a double-blind randomised clinical investigation, 90 patients who underwent hysterectomy under general anaesthesia were intra-operatively exposed to music, music in combination with therapeutic suggestion or	Surgical patients, Women, Patients under general anesthesia	Hysterectomy	Intra-operative music listening, music in combination with therapeutic suggestion	surgical pain, post-operative nausea, post-operative fatigue, post-operative vomiting (post-		Randomiz ed control trial (double-blind radomized clinical	visual analogue scale, bowel function, duration of hospital stay		Hospital	General anesthesia
16	Preterm infants exhibited less pain during a heel stick when they were played the same music their mothers listened to during	AIM: Playing music during painful procedures has shown inconsistent benefits for preterm infants. This study observed preterm infants during a heel stick procedure to assess whether listening to the music their mothers listened to during pregnancy had any impact on their pain and physiological and behavioural parameters. METHODS: We randomly exposed 42 preterm infants, with a mean gestational age of 31.8 +/- 2.79 weeks, to the music their mothers listened to during	Preterm infants (Premature infants), Pregnant women	Heel stick procedur	Prenatal music exposure, Music listening during heel stick procedure	medical procedure pain	Lullabies, Familiar music		Observed pain	Neonatoal Pain, Agitation and Seation Scale (N-PASS)		
17	Preventive Chair Massage with Algometry to Maintain Psychosomatic Balance in White-Collar Workers	People working at computers often suffer from overload-related muscle pain, and physical and mental discomfort. The aim of the study was to evaluate the effectiveness of chair massage, conducted in the workplace among white-collar workers, in relieving symptoms of musculoskeletal strain related to prolonged sitting posture. The study was conducted in 124 white-collar workers, 55 women and 69 men, aged 33.7 +/- 7.6 years. Subjects were randomly assigned to three	White color workers		Music listening sessions	Musculoskeltal strain (muscle tension, muscle strain)	Relaxing music	Randomiz ed control trial	Algometric neck muscle compression pain, heart rate variability, relaxation		Workplace	Relaxation

5. Gather Terms

Taxonomy should be designed for the users and for the content.

Gather terms, both suggested by users and found in the content.

From users by means of:

- ▷ Suggested term lists, especially from subject matter experts
- ▷ Search query logs

From content by means of:

- ▷ Manual review and analysis of content
- ▷ Automated content analysis and term extraction

Terms from external sources may also be acceptable.

- ▶ For named entities (e.g. names of countries, companies, organizations, etc.)
- ▶ For scientific, medical, and technical concepts (e.g. names of chemicals)

Prepare spreadsheets in format for upload into taxonomy management system.

5. Gather Terms

User sources:
Search log reports

Search log from
help.poolparty.biz

	Search Term	Total Unique Searches	Results Page Views/Se	% Search Exits	% Se
2	import	104	1.25	0.96%	
3	extractor	82	1.43	4.88%	
4	corpus	75	1.16	8.00%	
5	sparql	73	1.48	12.33%	
6	api	65	1.12	6.15%	
7	ontology	64	1.19	6.25%	
8	excel	56	1.21	7.14%	
9	export	43	1.09	11.63%	
10	excel import	40	1.1	2.50%	
11	workflow	39	1.31	2.56%	
12	API	38	1.08	5.26%	
13	graphsearch	38	1.18	13.16%	
14	snapshot	38	1.16	2.63%	
15	language	37	1.54	18.92%	
16	collection	35	1.26	0.00%	
17	blacklist	34	1.85	14.71%	
18	suggest	34	1.41	0.00%	
19	linked data	31	1.23	9.68%	
20	extract	29	1.31	3.45%	
21	extractor api	28	1.18	7.14%	
22	history	28	1.11	3.57%	
23	search	28	1.71	0.00%	
24	custom scheme	27	1.19	0.00%	
25	solr	27	1.7	14.81%	
26	unified views	26	1.15	19.23%	
27	unifiedviews	26	1.23	3.85%	
28	architectur	24	1	54.17%	
29	version	24	1.42	20.83%	
30	mapping	23	1.43	21.74%	
31	deprecated	22	1.41	4.55%	
32	disambiguation	22	1.23	0.00%	

5. Gather Terms

Content sources:

Content term extraction

Term extraction for candidate taxonomy concepts in PoolParty taxonomy/thesaurus management tool.

5. Gather Terms

External sources:
Linked open data

Linked data harvesting feature in PoolParty taxonomy/thesaurus management tool with the function to generate a seed thesaurus.

Generate Seed Thesaurus

Search Term: countries

LOD Source: EnDBPediaCategories

Additional Information

Depth: 2

Add Alternative Labels: ☑

Add Definitions: ☑

Add Relations: ☐

Available Categories

Category
Countries ↗
Countries articles needing attention ↗
Countries articles needing infoboxes ↗
Countries in Melanesia ↗
Countries in Micronesia ↗
Countries in Polynesia ↗
Arabic-speaking countries ↗
Balkan countries ↗
Baltic countries ↗

Selected Categories

Countries ⊗

Generate   Cancel

5. Gather Terms

Import spreadsheets

Format for importing a hierarchical taxonomy or subtree of a taxonomy into Poolparty software.

	A	B	C	D	E	F	G	H
1	concept	concept	concept	concept	concept	altLabel@en	definition@en	
2	Technology products					Software products		
3						Software tools		
4		Digital asset management systems				DAM		
5			Bynder					
6			Brandfolder					
7			Widen					
8		Customer relationship management systems				CRM		
9			Hubspot					
10			Sugar CRM			Sugar		
11						SugarCRM		
12		Content management systems				CMS	A content management system (CMS)	
13						content management software		
14						ECMS		
15						WCMS		
			Alfresco					
			CCMS			Component content management systems		
						DITA-based structured content management		
				RWS Tridion		Tridion		
					Tridion Sites			
					Tridion Docs			
22			Drupal			Acquia Drupal	Drupal is a free and open-source conte	
23						Civic Space		
24			Contentful					
25			AEM			Adobe AEM		
26						Adobe Experience Manager		
27			Confluence			Confluence administration		
28			WordPress					
29			SharePoint			Microsoft Office	SharePoint is a web based application	

6. Edit Concepts

Use taxonomy management tool to:

▸ Manually create concept schemes and some top concepts to reflect the determined high-level structure.

▸ Import spreadsheets of gathered terms.

▸ Check quality, remove duplicates and near duplicates, etc.

▸ Add more alternative labels, scope notes and definitions, as desired

▸ Adjust the hierarchy (move concepts and branches), as needed.

PoolParty Thesaurus Management

Benefits & Challenges of Building a Taxonomy

Benefits of building your own taxonomy

- Best suited for your content and your users
  - ▷ Achieving better results and high user satisfaction
  - ▷ Especially when content and users are internal
- Engages stakeholders for long-term buy-in and commitment

Challenges of building your own taxonomy

- Lack of resources, expertise, and time
- Risk of wasted effort and resources if not done right

Licensing or Acquiring Taxonomies

Licensed Taxonomy Uses

A challenge to build taxonomies?

▸ A taxonomy does not always need to be built from scratch.

▸ Parts of a taxonomy (or set of taxonomies) may be better customized

▸ Parts of a taxonomy can be off-the-shelf/licensed

Use cases for a licensed taxonomy

▸ As a starting point, to be developed and customized further

▸ For tagging external content in certain subjects

▸ A term list for a single metadata field or facet (filter) in a faceted taxonomy

Licensed Taxonomy Uses

As a starting point

When you need a large (perhaps 500+ terms), detailed taxonomy/thesaurus quickly

Examples: news topics, health topics, information technology topics

Conditions

- ▶ A good controlled vocabulary exists
- ▶ The license allows unrestricted modification/enhancement
- ▶ You have software that manages the controlled vocabulary
- ▶ You have skills to update end edit the controlled vocabulary

Licensing is to save time, not substitute for expertise.

Licensed Taxonomy Uses
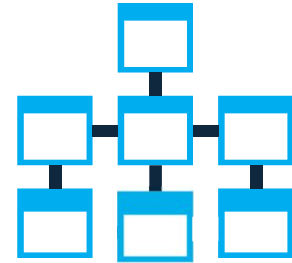
For tagging *external* content

▶ When content comes from multiple sources, there is no need to customize the taxonomy for a particular content set.

▶ Is practical, if there exists a taxonomy or thesaurus for license in the particular subject domain/discipline (e.g. MeSH for medicine, ERIC for education)

▶ Such suitable taxonomies exist usually just for academic disciplines

▶ Suitable for subject-matter-expert users who may already be familiar with the thesaurus.

▶ Taxonomies may be published by international or nongovernmental organizations, trade or professional organizations.

▶ Would not necessarily require any editing/modification.

Licensed Taxonomy Uses

For a single metadata property, concept scheme, or facet

- ▶ When you need a generic controlled vocabulary, part of a larger taxonomy set.

- ▶ But linked open data options are not suitable.

- ▶ Could be for one or more of: **Geographic places**, **Industry types**, **Product categories**, **Genres**, **Chemicals**, etc.

- ▶ Controlled vocabularies of other facets can still be custom created.

- ▶ Modification of licensed vocabulary will still be needed, but usually just of adding or deleting concepts.

- ▶ More often the controlled vocabulary will need editing down.

Licensed Taxonomy Issues

Licensing taxonomy suitability conditions

A taxonomy for license should…

- exist for the specific desired subject domain,
- cover the subject domain sufficiently (not too small),
  but not have too many irrelevant terms (not too large),
- be of the right type (e.g. a hierarchical taxonomy and not a thesaurus),
- permit commercial reuse and modification.
  - Free/open-source nonprofit published vocabularies may prohibit commercial reuse or modification.

Taxonomy Sources

Non-profit and governmental organizations with taxonomies available:

▶ NAL Agricultural Thesaurus agclass.nal.usda.gov/download.shtml (XML, SKOS-XML Word, MARC)

▶ DTIC (Defense Technical Information Center) Thesaurus discover.dtic.mil/thesaurus/ (Excel, HTML, TXT, XML, SKOS TTL)

▶ NASA Thesaurus www.sti.nasa.gov/nasa-thesaurus/ (SKOS, OWL, ZThes, CSV/Text)

▶ USGS Thesaurus www2.usgs.gov/science/about/ (RDF-XML, SQLite)

▶ NLM Medical Subject Headings (MeSH) www.nlm.nih.gov/databases/download/mesh.html (XML, ASCII, MARC 21, RDF)

▶ ERIC Thesaurus (Education Resources Information Center, U.S. Dept. of Education) eric.ed.gov/?download (XML)

Taxonomy Sources

Taxonomy usage rights

"As a work of the U.S. Government, the **NAL Agricultural Thesaurus and Glossary** are in the public domain within the United States. Additionally, the U.S. Government waives copyright and related rights in this work worldwide through the CC0 1.0 Universal Public Domain Dedication."

"The **DTIC Thesaurus** is not copyrighted. No license is needed to use it."

"**NASA** does not grant exclusive use rights with respect to this product or the data contained therein. This product or the data contained therein are intended for the sole use of the customer. The data **may not be installed on any system with public access**. The customer may not reproduce the data for the distribution to any third party."

"**USGS**-authored or produced data and information are considered to be in the U.S. Public Domain. When using information from USGS information products, publications, or websites, we ask that **proper credit be given**"

"Government information at **NLM** Web sites is in the public domain. Public domain information may be freely distributed and copied, but it is requested that in any subsequent use the National Library of Medicine (NLM) **be given appropriate acknowledgement**."

"The **ERIC** thesaurus and ERIC database are available for use by the general public."

Taxonomy Sources

Directories of taxonomies and other controlled vocabularies

General:

- ▶ BARTOC (Basic Register of Thesauri, Ontologies & Classification) bartoc.org
- ▶ Research Vocabularies Australia vocabs.ardc.edu.au
- ▶ Open Metadata Registry metadataregistry.org/vocabulary/list.html

Domain specific examples:

- ▶ NCBO BioPortal – biomedical ontologies bioportal.bioontology.org/ontologies
- ▶ Heritage Data (UK cultural heritage) www.heritagedata.org/blog/vocabularies-provided

Issues:

Controlled vocabularies tend to be academic, scientific, or medical, not for business uses.
Controlled vocabularies may not be in a format easy to import
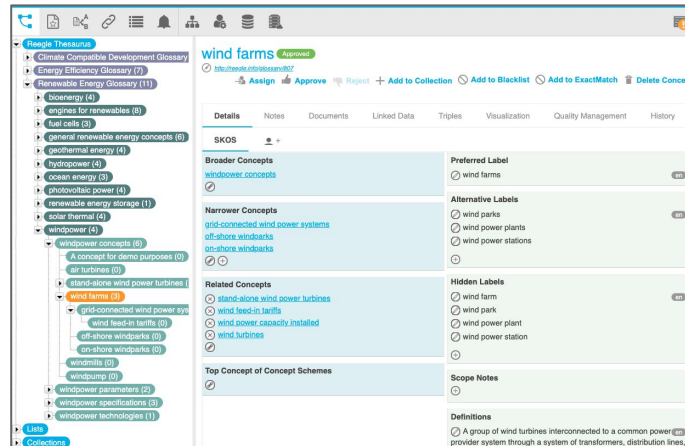
Taxonomy Sources

Commercial taxonomy publisher

▶ WAND

Taxonomies bundled with
taxonomy management
software

▶ PoolParty +
WAND taxonomies

Editing Licensed Taxonomies

- Add more detailed concepts in certain areas, as needed.
- Delete (or merge) unneeded specific concepts.
- In some cases, rename preferred labels for your audience.
- Add/delete alternative labels to reflect your context.

Taxonomy Licensing Conclusions

▸ It's not a simply build vs. buy question.

▸ You may build some of your vocabularies and buy/license others.

▸ Use different sources for different facets, metadata properties.

▸ Consider internal vs. external use of taxonomy.

▸ You may license taxonomies and then modify them.

▸ You may license for a temporary period, while you build your own.

Resources

Upcoming Taxonomy Workshops and Tutorials

Heather Hedden will be teaching about taxonomies at:

▶ Information Architecture Conference (IAC), 18 April, online www.theiaconference.com
   "Revisiting Taxonomies: Topics in Taxonomy Design"

▶ Knowledge Graph Conference, May 3, New York, NY (hybrid) www.knowledgegraph.tech
   "Foundation for a Knowledge Graph: Taxonomy Design Best Practices"

▶ Data Day Texas, June 13, Austin, TX https://datadaytexas.com
   "Introduction to Taxonomies for Data Scientists"

▶ SEMANTiCS conference, September 13-15, 2022, Vienna (hybrid) https://2022-eu.semantics.cc
   Tutorial: "Knowledge Engineering of Taxonomies, Thesauri, and Ontologies"

▶ LavaCon, October, October 23 - 25, 2022, New Orleans, LA https://lavacon.org
   "Using Taxonomies and Tagging to Connect Content Across the Enterprise"

Further Information

▶ Other PoolParty webinar recordings and white papers
https://www.poolparty.biz/resource-library

▶ ANSI/NISO Z39.19-2005 (2010) *Guidelines for Construction, Format, and Management of Monolingual Controlled Vocabularies*.
ww.niso.org/publications/ansiniso-z3919-2005-r2010

▶ Taxonomy Boot Camp conference, November 7-8, 2022, Washington, DC
www.taxonomybootcamp.com

▶ The Accidental Taxonomist Blog
http://accidental-taxonomist.blogspot.com

Q&A / Contact

Heather Hedden
Data and Knowledge Engineer

Semantic Web Company Inc.
One Boston Place, Suite 2600
Boston, MA 02108 USA
+1 857-400-0183

heather.hedden@semantic-web.com
www.linkedin.com/in/hedden

Semantic Web Company www.semantic-web.com

PoolParty software www.poolparty.biz

Semantic Web Company GmbH
Neubaugasse 1/8
1070 Vienna, Austria
Tel.: +43-14021235