# Revisiting, Reviewing, and Revising Taxonomies

IA Conference
April 22, 2022

**Heather Hedden**
Data & Knowledge Engineer
Semantic Web Company

# 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*.

# Outline

- Introduction to reviewing taxonomies
- Taxonomy quality review
- Concept and label issues
- Hierarchy issues
- Taxonomy integration, linking, and mapping

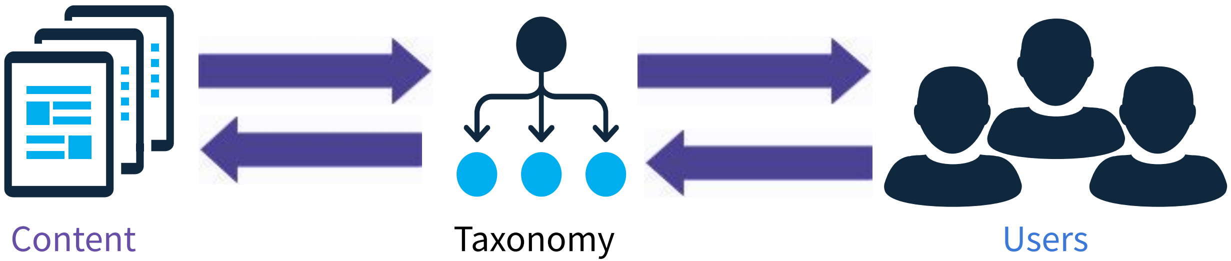
poolparty®

# Introduction to Reviewing 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.



- Taxonomies are a kind of controlled vocabulary or knowledge organization system

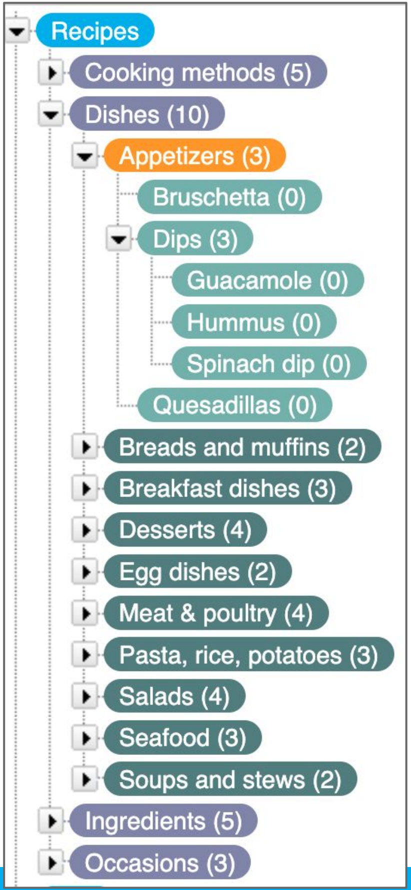
# Introduction to Taxonomies

## What is a taxonomy?

*Controlled* and *organized*

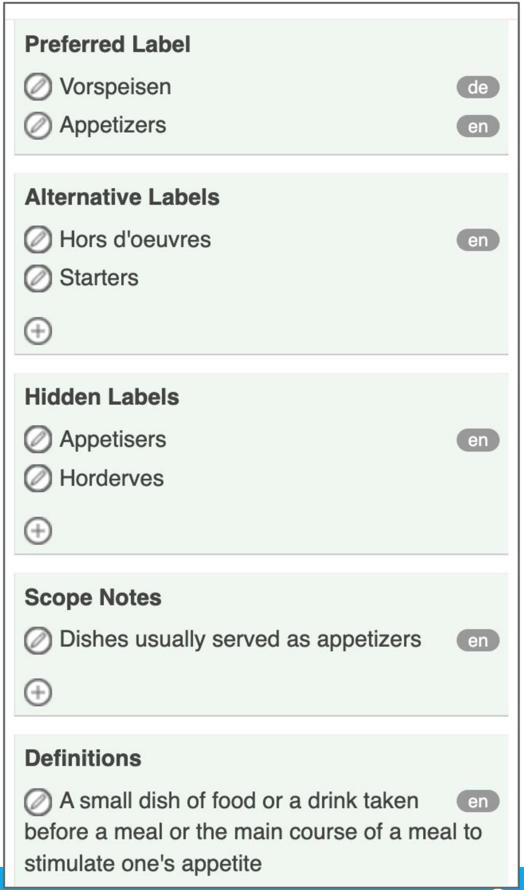
1. A kind of controlled vocabulary or knowledge organization system, based on unambiguous concepts, not just words:
   **things, not strings**
2. Concepts are arranged in a structure of hierarchies, categories, or facets to organize them.

*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

1. As a controlled vocabulary

   Brings together different wordings (synonyms) for the same concept

   ▹ Helps people search for information by different names

2. Having classification and structure

   Organizes information into a logical structure

   ▹ Helps people browse or navigate for information

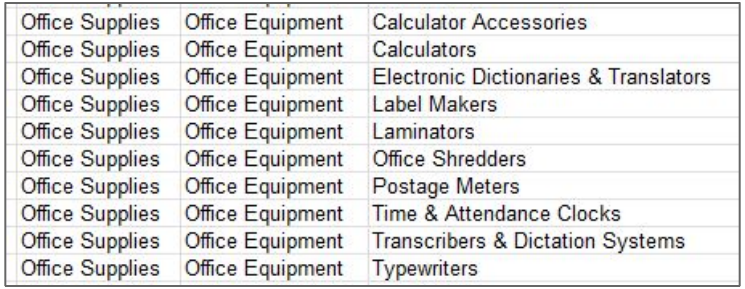
   ▹ Provides context and meaning for concepts for indexing and retrieval

# Reviewing Taxonomies

## Reviewing  and Updating Taxonomies Currently in Use

Changes impacting taxonomies:

- *Certain types of concepts*
  - Terminology changes (e.g. reducing bias)
  - New concepts (e.g related to technology)
  - User feedback suggesting improvements



| | | |
|---|---|---|
| Office Supplies | Office Equipment | Calculator Accessories |
| Office Supplies | Office Equipment | Calculators |
| Office Supplies | Office Equipment | Electronic Dictionaries & Translators |
| Office Supplies | Office Equipment | Label Makers |
| Office Supplies | Office Equipment | Laminators |
| Office Supplies | Office Equipment | Office Shredders |
| Office Supplies | Office Equipment | Postage Meters |
| Office Supplies | Office Equipment | Time & Attendance Clocks |
| Office Supplies | Office Equipment | Transcribers & Dictation Systems |
| Office Supplies | Office Equipment | Typewriters |

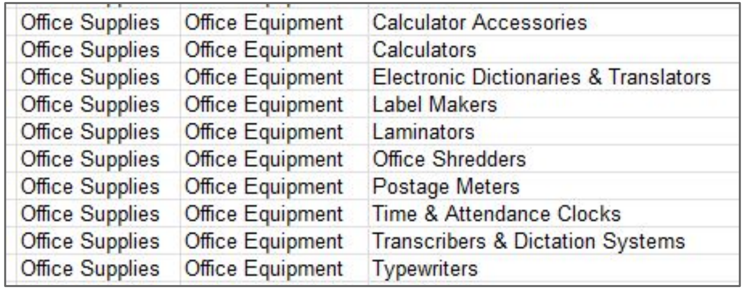
- *Broader, taxonomy-level changes*
  - New content, bringing up new concepts
  - Content  sets that get dropped
  - New requirements, users, needs, trends, markets, etc.
  - Examples:
    - Adding related-concept relationships, definitions, scope notes, etc.
    - Integrating multiple taxonomies, or with a terminology or glossary
    - Implementing a new taxonomy management system
    - Extending the taxonomy to a new application or website

# Reviewing Taxonomies

## Reviewing  and Updating Taxonomies Currently in Use

Changes impacting taxonomies:

- *Certain types of concepts*  - usual governance policies
  - Terminology changes (e.g. reducing bias)
  - New concepts (e.g related to technology)
  - User feedback suggesting improvements



| | | |
|---|---|---|
| Office Supplies | Office Equipment | Calculator Accessories |
| Office Supplies | Office Equipment | Calculators |
| Office Supplies | Office Equipment | Electronic Dictionaries & Translators |
| Office Supplies | Office Equipment | Label Makers |
| Office Supplies | Office Equipment | Laminators |
| Office Supplies | Office Equipment | Office Shredders |
| Office Supplies | Office Equipment | Postage Meters |
| Office Supplies | Office Equipment | Time & Attendance Clocks |
| Office Supplies | Office Equipment | Transcribers & Dictation Systems |
| Office Supplies | Office Equipment | Typewriters |

- *Broader, taxonomy-level changes*  - special taxonomy review projects
  - New content, bringing up new concepts
  - Content  sets that get dropped
  - New requirements, users, needs, trends, markets, etc.
  - Examples:
    - Adding related-concept relationships, definitions, scope notes, etc.
    - Integrating multiple taxonomies, or with a terminology or glossary
    - Implementing a new taxonomy management system
    - Extending the taxonomy to a new application or website

# Reviewing Taxonomies

## Evaluating and Revising Other Taxonomies for Re-Use

- Taxonomies licensed, acquired, adopted on the subject area
    - From published sources or as linked open data
- Taxonomies from different internal systems
- Legacy controlled vocabularies from former projects or purposes
- Taxonomies from acquired/merged companies, businesses

## Taxonomy evaluation comprises:

- Ensuring the taxonomy is suitable for **the content**
    - Check the scope and level of detail
    - Perform sample test tagging and test retrieval
- Ensuring the taxonomy is suitable for **the users**
    - Also evaluate the taxonomy quality

# Taxonomy Quality Review

# Taxonomy Quality Checks

## Taxonomy management system quality reporting

- Duplicate labels (if not enforced upon creation)
- Orphan concepts

## Taxonomy management system metrics

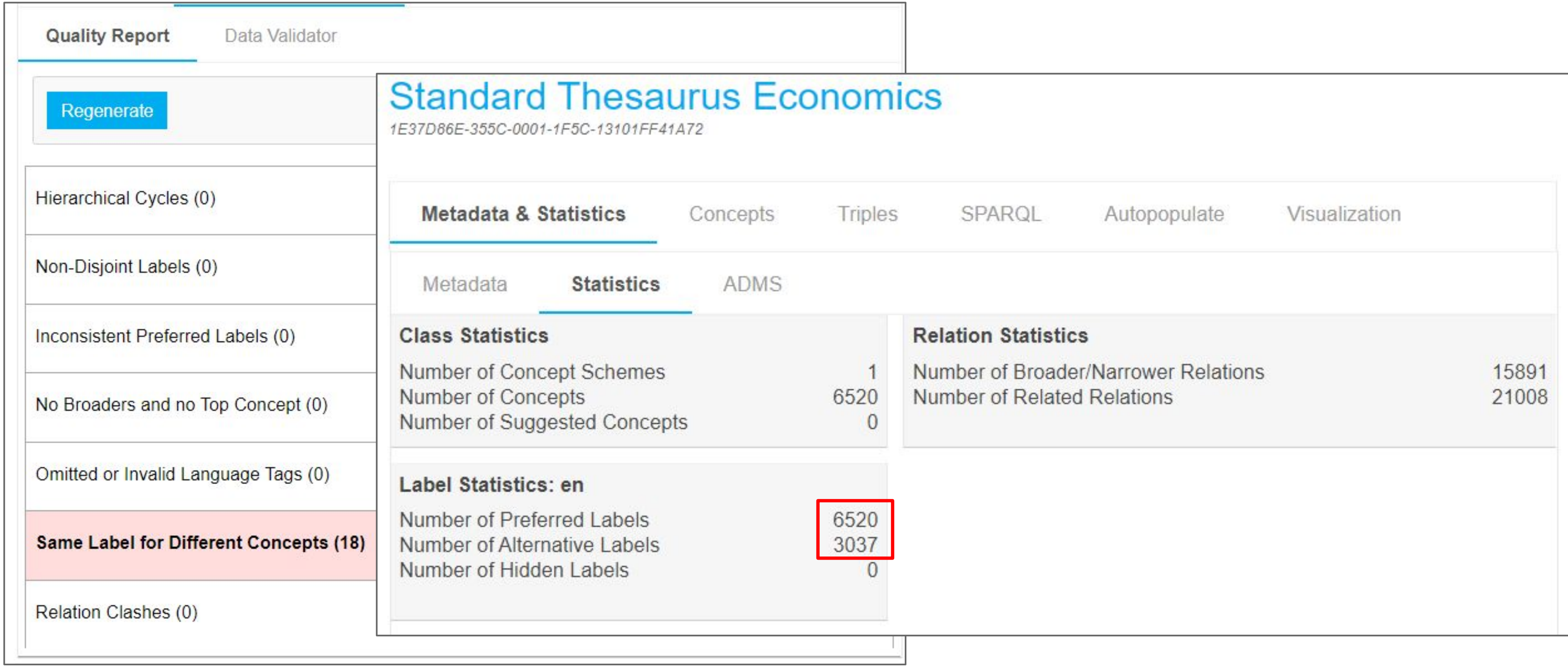
- Numbers of preferred vs. alternative labels
- Numbers of hierarchical and associative relationships
- Number of levels of hierarchy depth

## Taxonomy management system reports or queries

- *Missing* alternative labels, certain relationships, scope notes, etc.,
  - So they can be added
- Only concepts *with* alternative labels, certain relationships, scope notes, etc.
  - So, they can be comprehensively reviewed

## Taxonomy management system spreadsheet exports for review

# Taxonomy Quality Checks



Hierarchical Cycles (0)

Non-Disjoint Labels (0)

Inconsistent Preferred Labels (0)

No Broaders and no Top Concept (0)

Omitted or Invalid Language Tags (0)

**Same Label for Different Concepts (18)**

Relation Clashes (0)

## Standard Thesaurus Economics

*1E37D86E-355C-0001-1F5C-13101FF41A72*

**Class Statistics**

| | |
|---|---|
| Number of Concept Schemes | 1 |
| Number of Concepts | 6520 |
| Number of Suggested Concepts | 0 |

**Relation Statistics**

| | |
|---|---|
| Number of Broader/Narrower Relations | 15891 |
| Number of Related Relations | 21008 |

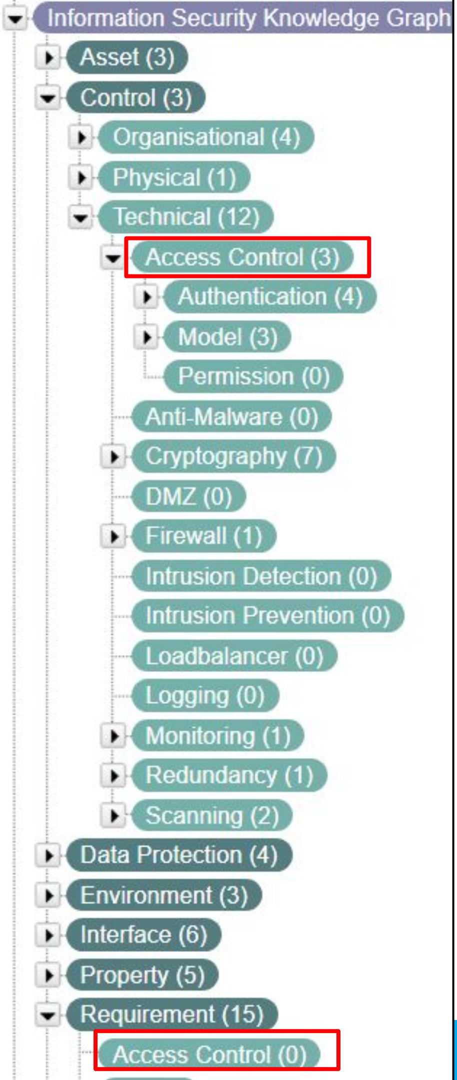
**Label Statistics: en**

| | |
|---|---|
| Number of Preferred Labels | 6520 |
| Number of Alternative Labels | 3037 |
| Number of Hidden Labels | 0 |

# Quality Checks

Duplicated preferred label.

Can also be identified as different concepts, due to presence of narrower concepts.



- Information Security Knowledge Graph
  - Asset (3)
  - Control (3)
    - Organisational (4)
    - Physical (1)
    - Technical (12)
      - Access Control (3)
        - Authentication (4)
        - Model (3)
        - Permission (0)
      - Anti-Malware (0)
      - Cryptography (7)
      - DMZ (0)
      - Firewall (1)
      - Intrusion Detection (0)
      - Intrusion Prevention (0)
      - Loadbalancer (0)
      - Logging (0)
      - Monitoring (1)
      - Redundancy (1)
      - Scanning (2)
  - Data Protection (4)
  - Environment (3)
  - Interface (6)
  - Property (5)
  - Requirement (15)
    - Access Control (0)

# Concept and Label Issues

# Taxonomy Concepts and Labels

## Concept label format and style best practices

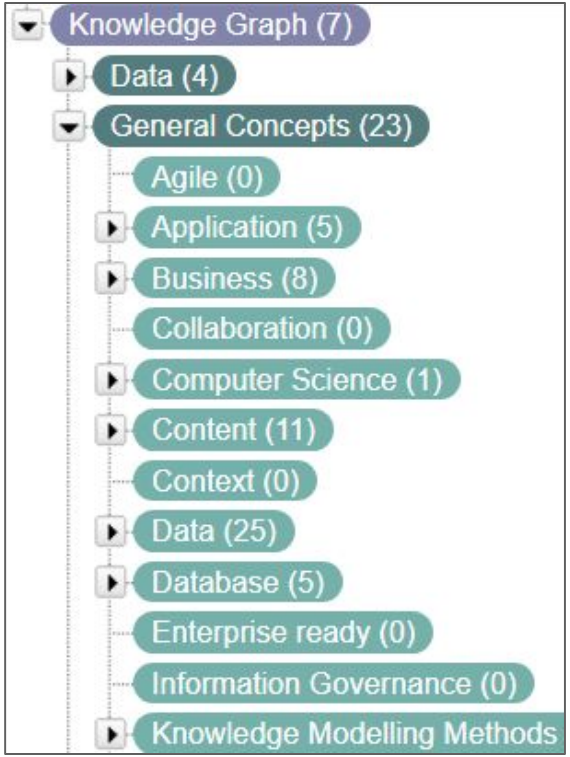
- Unambiguous; understood even out of context of the hierarchy.
  *Example:* **Nursing Certification**, rather than **Certification** as narrower to Nurses
- Consistent capitalization: initial capitalization is recommended.
  *Example:* **Corporate finance**, rather than corporate finance, or Corporate Finance
- Countable nouns are usually plural
  *Example:* **Occupational accidents** (countable), **Occupational health** (not countable)
- Adjectives alone may exist within term lists of characteristics/properties (metadata or attributes), but not within hierarchical taxonomies or thesauri. For example, colors, sizes, status.
- Parenthetical qualifiers may be used for disambiguation, not modification.
  *Example:* **Walnut (wood)**
- Avoid term inversions (e.g. noun, adjective) because labels are searchable
  *Example:* **Racial discrimination**, not Discrimination, racial

# Concept and Label Issues

“Concepts” that are single, vague words:

*Application, Business, Content, Context, Data*

Concepts need to be unambiguous,
and specific for tagging a specific set of content.

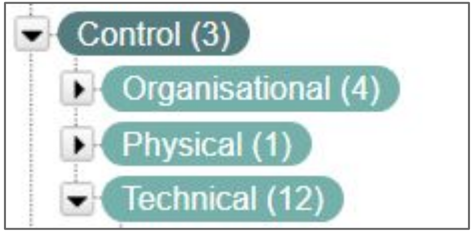


- Knowledge Graph (7)
  - Data (4)
  - General Concepts (23)
    - Agile (0)
    - Application (5)
    - Business (8)
    - Collaboration (0)
    - Computer Science (1)
    - Content (11)
    - Context (0)
    - Data (25)
    - Database (5)
    - Enterprise ready (0)
    - Information Governance (0)
    - Knowledge Modelling Methods

# Concept and Label Issues

Concepts that are adjectives
Adjectives are only suitable as

Attributes, but not in a hierarchical taxonomy



Control (3)
Organisational (4)
Physical (1)
Technical (12)

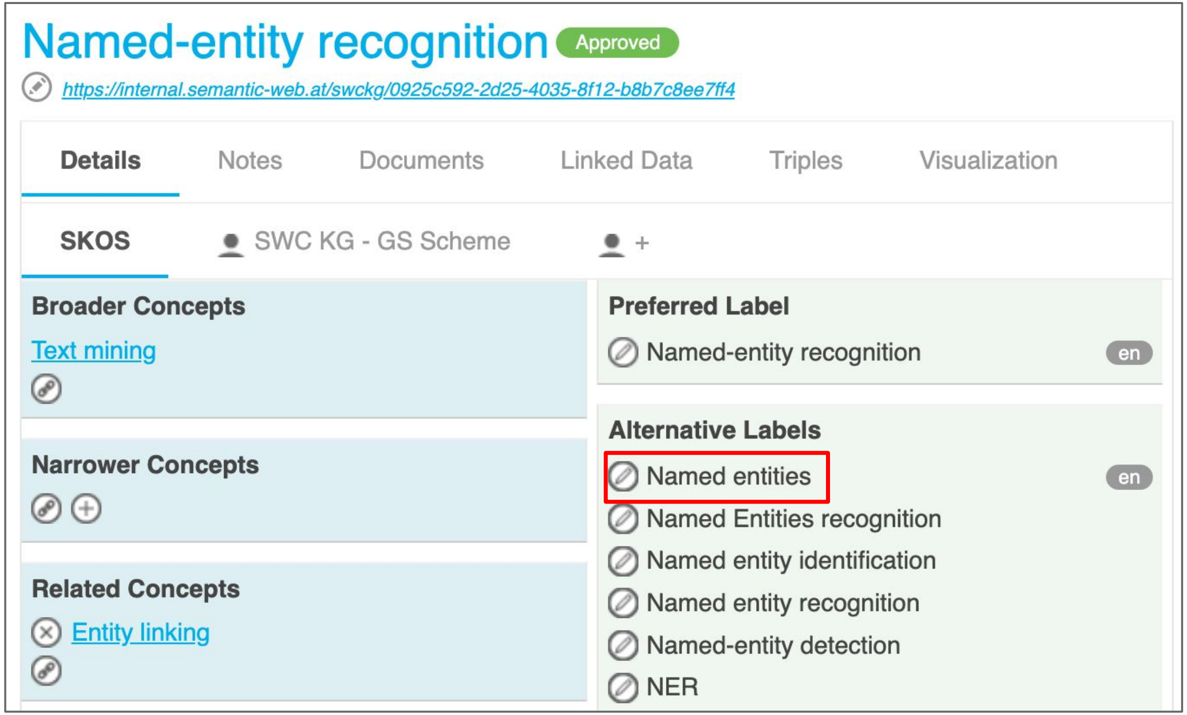
# Concept and Label Issues: Alternative Labels

- Ensure there are *sufficient* alternative labels to match:
  - different possible user search search strings for the same concept
  - different wordings of the concept in the text
- Don’t have *too many* alternative labels that might:
  - match text with different meaning
  - complicate and confuse users when displayed
- Consider converting some to Hidden Labels instead so as not to display.

# Concept and Label Issues: Alternative Labels

Alternative label with different meaning:

Named entity recognition vs. Named entities



Named-entity recognition Approved

https://internal.semantic-web.at/swckg/0925c592-2d25-4035-8f12-b8b7c8ee7ff4

**Broader Concepts**

Text mining

**Narrower Concepts**

**Related Concepts**

Entity linking

**Preferred Label**

Named-entity recognition en

**Alternative Labels**

Named entities en

Named Entities recognition

Named entity identification

Named entity recognition

Named-entity detection

NER

# Concept and Label Issues: Alternative Labels

Alternative label with narrower meaning

Problematic scenario:

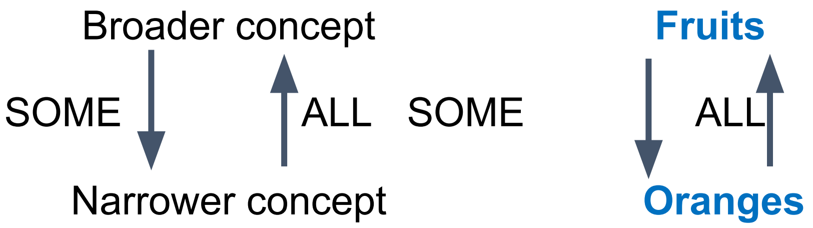
- **Laptops** is an alternative label for **Computers**
- Document on Supercomputers is tagged with **Computers**.
- End-user looks up term “Laptops,” and is taken to result set of all documents tagged with **Computers**.
- Result set includes documents on supercomputers and other computers that are not laptops, in addition to documents on laptops.
- End-user thinks the tagging or taxonomy is wrong by retrieving documents on other kinds of computers besides the selected “laptops.”

# Hierarchy Issues

# Hierarchical Relationship

## Hierarchical relationship from ANSI/NISO guidelines

Reciprocal (bi-directional) relationship, but asymmetrical



**Fruits** (has narrower concept) **Oranges**

**Oranges** (has broader concept) **Fruits**

Three subtypes:

1. Generic – Specific: “is/are a kind of”
2. Generic – Instance: “is an instance of”
3. Whole – Part: “is/are within”

| | | |
|---|---|---|
| **Hospitals** | *Narrower Concept:* | **Children's hospitals** |
| **Hospitals** | *Narrower Concept:* | **Boston Medical Center** |
| **Hospitals** | *Narrower Concept:* | **Emergency rooms** |

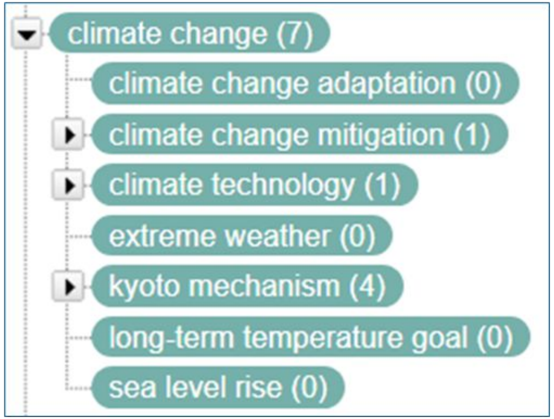
# Hierarchy Issues

Bending the rules for hierarchical relationships is acceptable, if:

1. It is a hierarchical taxonomy, not a thesaurus (esp. without related-concept relationships)

   *and*

2. The incorrect hierarchical relationships are limited to grouping categories, especially at the higher levels of a taxonomy



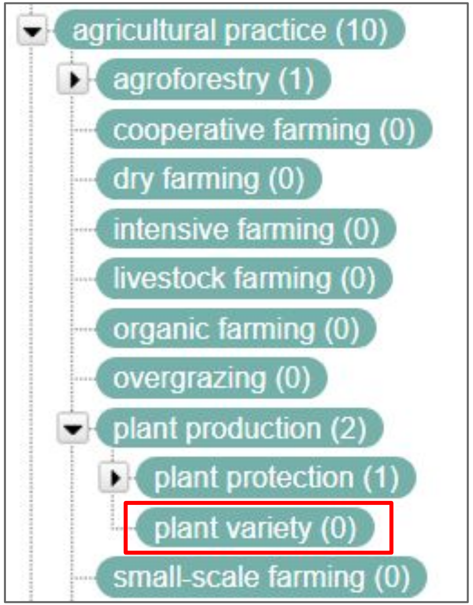
- climate change (7)
  - climate change adaptation (0)
  - climate change mitigation (1)
  - climate technology (1)
  - extreme weather (0)
  - kyoto mechanism (4)
  - long-term temperature goal (0)
  - sea level rise (0)

# Hierarchy Issues

If a concept does not fit the “is a” rule,
Perhaps it’s just a labelling problem,
and the concept is correct.

Example:

Change the label **plant variety**
to **plant diversification**

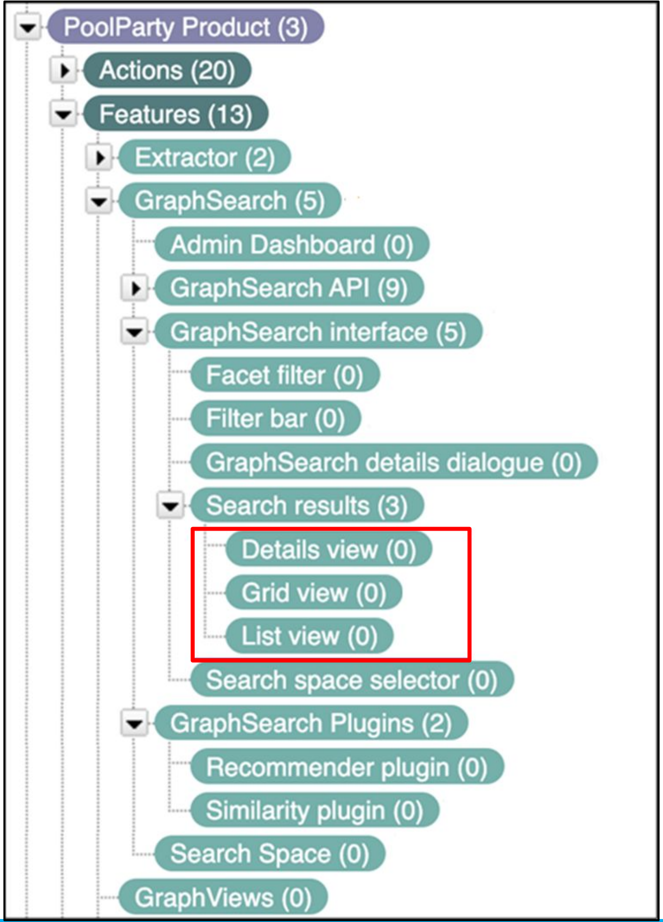


- agricultural practice (10)
  - agroforestry (1)
  - cooperative farming (0)
  - dry farming (0)
  - intensive farming (0)
  - livestock farming (0)
  - organic farming (0)
  - overgrazing (0)
  - plant production (2)
    - plant protection (1)
    - plant variety (0)
  - small-scale farming (0)

# Hierarchy Issues

Error of context-based narrower concepts, which are actually more generic

Example: Details view, Grid view, List view

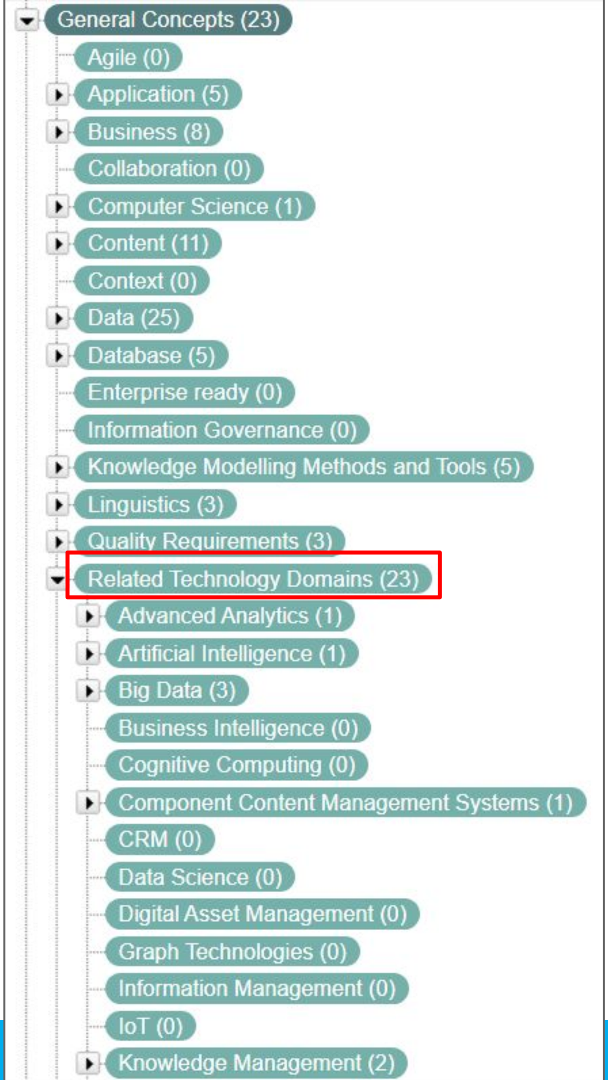


- PoolParty Product (3)
  - Actions (20)
  - Features (13)
    - Extractor (2)
    - GraphSearch (5)
      - Admin Dashboard (0)
      - GraphSearch API (9)
      - GraphSearch interface (5)
        - Facet filter (0)
        - Filter bar (0)
        - GraphSearch details dialogue (0)
        - Search results (3)
          - Details view (0)
          - Grid view (0)
          - List view (0)
        - Search space selector (0)
      - GraphSearch Plugins (2)
        - Recommender plugin (0)
        - Similarity plugin (0)
      - Search Space (0)
    - GraphViews (0)

# Hierarchy Issues

Problem of non-intuitive hierarchy

Example: Related Technology Domains



- General Concepts (23)
  - Agile (0)
  - Application (5)
  - Business (8)
  - Collaboration (0)
  - Computer Science (1)
  - Content (11)
  - Context (0)
  - Data (25)
  - Database (5)
  - Enterprise ready (0)
  - Information Governance (0)
  - Knowledge Modelling Methods and Tools (5)
  - Linguistics (3)
  - Quality Requirements (3)
  - Related Technology Domains (23)
    - Advanced Analytics (1)
    - Artificial Intelligence (1)
    - Big Data (3)
    - Business Intelligence (0)
    - Cognitive Computing (0)
    - Component Content Management Systems (1)
    - CRM (0)
    - Data Science (0)
    - Digital Asset Management (0)
    - Graph Technologies (0)
    - Information Management (0)
    - IoT (0)
    - Knowledge Management (2)

# Taxonomy Integration, Linking, and Mapping

# Integrating Existing Taxonomies

## Reusing & Extending

- Simplest way to integrate existing taxonomies is reusing them and extending them based on need.

## Linking & Mapping

- Taxonomies are linked at individual concepts, and the taxonomies are retained as distinct, but can be used in combination, extending each other.
  - Mapping is a form of linking for exact or close matches
    - One taxonomy can be *used for* another (not alongside each other)
    - One is the backend, and one is the frontend

## Merging

- Taxonomies are combined permanently, removing duplicates, without any longer retaining them as distinct.
- First step is to link the taxonomies, then incorporate the unlinked concepts.

# Project Linking Relationships

SKOS relationships within a Concept Scheme (a single taxonomy, thesaurus, or controlled vocabulary), also known as “thesaural” relationships:

- Broader concept
- Narrower concept
- Related concept

SKOS relationships across two different taxonomy projects:

- Exact match
- Close match
- Narrow match
- Broad match
- Related match

# Project Linking Use Cases

## Possible reasons to link taxonomy projects:

- Link to a standard, published vocabulary/classification scheme for alignment.
  - Involves **Exact Match** only
- Use one taxonomy in the user interface to retrieve additional content already tagged with a different taxonomy (also called “mapping”).
  - Involves **Exact Match**, possibly **Close Match**, and **Narrow Match** in one direction
- Enrich a taxonomy with concepts from another controlled vocabulary (“mapping”).
  - Involves **Exact Match**, possibly **Close Match**, and **Narrow Match** in one direction
- Combine two or more taxonomies to extend them, but each still remains intact.
  - May involve all match types
- Compare and align taxonomies prior to fully merging them (with one absorbed into the other taxonomy).
  - May involve all match types

# Resources

# Further Information

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

- “Testing Taxonomies” Taxonomy Boot Camp, November 5, 2013
  www.hedden-information.com/wp-content/uploads/2019/07/Testing_Taxonomies.pdf

- “Mapping Taxonomies, Thesauri, and Ontologies” SEMANTiCS conference, September 11, 2019
  www.hedden-information.com/wp-content/uploads/2019/09/Mapping-Taxonomies-Thesauri-Ontologies.pdf

- “How Many Synonyms Should You Have?” Taxonomy Boot Camp, November 14, 2016
  www.hedden-information.com/wp-content/uploads/2019/07/How-Many-Synonyms-Should-You-Have.pdf

- “Evaluating Taxonomies” blog post
  http://accidental-taxonomist.blogspot.com/search/label/Heuristic%20evaluation

# Upcoming Taxonomy Workshops and Tutorials

Heather Hedden will be teaching about taxonomies at:

- 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"

# Q&A / Contact

**Heather Hedden**
Data and Knowledge Engineer

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

[heather.hedden@semantic-web.com](mailto:heather.hedden@semantic-web.com)
[www.linkedin.com/in/hedden](http://www.linkedin.com/in/hedden)
[http://accidental-taxonomist.blogspot.com](http://accidental-taxonomist.blogspot.com)

Semantic Web Company [www.semantic-web.com](http://www.semantic-web.com)

PoolParty software [www.poolparty.biz](http://www.poolparty.biz)