Synonyms, Alternative Labels, and Nonpreferred Terms

SLA Taxonomy Division Webinar
February 7, 2017

Heather Hedden
Senior Vocabulary Editor
Metadata Standards and Services
Gale | Cengage Learning
About Heather Hedden

- Previously, taxonomy consultant
- Author of *The Accidental Taxonomist* (Information Today, Inc.)
- Instructor of online taxonomy workshops (Hedden Information Management, American Society for Indexing, Simmons, College)
- SLA Taxonomy Division former chair of Mentoring Committee and Membership Committee. American Society for Indexing board member
About Gale, a Cengage Learning Company

- Subscription databases to libraries: GVRL ebooks, In Context, Academic OneFile, Business Collection, Literature Resource Center, etc.

- Web products to the public: Questia, Books & Authors, HighBeam Research, Encyclopedia.com

- Gale Research reference books, directories, and other book imprints (Greenhaven, Thorndike, St. James Press, etc.)

- Primary Source Media digital archives (Artemis)

Legacy library database vendor companies: Information Access Company, Predicasts
Outline

- Introduction: Definition, Examples, Usage
- Different Designations and Models
- Different Models in Taxonomy Management Software
- Creation and Implementation
- Different Types
- How Many to Create
- User Interface and Search
- Variations and Customizations
Introduction

Synonyms, Alternative Labels, Non-preferred Terms

- **Defined:** Approximately synonymous words or phrases to refer to an equivalent concept, **for the context** of the taxonomy and the set of content.

- **Purpose:** To capture different wordings of how different people might describe or look up the same concept or idea.
  - Differences between that of the author and the user/reader
  - Differences between that of the indexers and the end-users
  - Differences among different users/readers
- Serving as “multiple entry points” to look up and retrieve the desired content.
- Enabling consistent indexing/tagging
Examples (from Gale Subject Thesaurus)

Conflict management
  Conflict resolution
  Managing conflict

Wills
  Codicils
  Last will and testament
  Testaments (Wills)

Influenza
  Flu
  Grippe

Movies
  Cinema
  Films (Movies)
  Motion pictures
  Movie genres

Telecommunications industry
  Communications industry
  Digital transmission industry
  Interexchange carriers
  Telecommunications services industry
  Telephone holding companies
  Telephone industry
  Telephone services industry

Environmental management
  Adaptive management (Environmental management)
  Environmental control
  Environmental stewardship
  Natural resource management
  Stewardship (Environmental management)

Piano music [no variants]
Introduction

When to Use

Not needed:
• A very small, browsable taxonomy, where all can be seen or easily scrolled to (such as in facets) and tagging is manual

Needed:
• If taxonomy is too large to be all seen in one view with minimal scrolling.
• If taxonomy will be searched upon and not just browsed.
• If automated indexing/auto-classification/auto-categorization is implemented.

Whether it’s called a taxonomy or thesaurus does not matter.
## Introduction

### Controlled Vocabularies - Complexity

<table>
<thead>
<tr>
<th>Pick List</th>
<th>Synonym Ring</th>
<th>Authority File</th>
<th>Taxonomy</th>
<th>Thesaurus</th>
<th>Ontology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ambiguity control</td>
<td><strong>Synonym control</strong></td>
<td>Ambiguity control</td>
<td>Ambiguity control (Synonym control)</td>
<td>Ambiguity control</td>
<td>Ambiguity control (Synonym control)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Hierarchical relationships</td>
<td>Hierarchical relationship</td>
<td>Semantic relationships</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Associative relationships</td>
<td>Classes</td>
</tr>
</tbody>
</table>
Different Designations and Models

Synonym

- Simple, non-expert, widely understood.
- Associated with a *Term*.
- May use this designation with varied stakeholders.
- Not entirely accurate, because most are not synonyms (not exact equivalents, not single words).
- If used, better to use *in combination* with a more accurate term, such as alternative label or non-preferred term.
Different Designations and Models

Non-preferred Term

- Formal designation in thesauri, in accordance with ANSI/NISO Z.39-19 and ISO 25964 thesaurus standards.
- Shortened as *NPT*.
- Associated with a *Preferred term*.
- Not intuitively understood by non-experts.
- Understood and preferred by taxonomists trained on the thesaurus model.
Different Designations and Models

Alternative Label

- Formal designation for SKOS (Simple Knowledge Organization System) vocabularies.
- Shortened as *altLabel*.
- Associated with a *Preferred label*.
- Intuitively understood by non-experts and varied stakeholders.
- May be used in non-SKOS vocabularies, but could confuse information experts who associate it with SKOS.
Different Designations and Models

Even more designations:

- Aliases
- Alternate labels
- Alternate terms
- Alternative terms
- Cross-references
- Entry terms
- Equivalent terms
- Non-descriptors
- Non-postable terms
- See references
- Use for terms
- Use references
- Used for terms

Variants

Found mostly in:

- Taxonomies
- SKOS vocabularies
- SKOS vocabularies
- SKOS vocabularies
- Indexes in print
- Thesauri
- Thesauri
- Thesauri
- Thesauri
- Indexes in print
- Thesauri
- Thesauri
- Thesauri
- Taxonomies
Different Designations and Models

Thesaurus standards/guidelines


- ISO ISO 25964-1 Information and documentation - Thesauri and interoperability with other vocabularies
  Part 1: Thesauri for information retrieval [2011]

SKOS model recommendation

A World Wide Web (W3C) recommendation
“A common data model for sharing and linking knowledge organization systems via the Web”
https://www.w3.org/TR/skos-reference/
Different Designations and Models

Thesaurus non-preferred term/preferred term model

- Considered a kind of “relationship” of the Equivalency type.

- Reciprocity of relationship, pointing in both directions: **USE** and **UF** (use and used for/use for/used from).
  
  Non-preferred term **USE** Preferred term
  
  Preferred term **Used for** Non-preferred term

- Both **Preferred Terms** and **Non-preferred Terms** are “terms.”
Different Designations and Models

SKOS vocabulary model

- Instead of terms, there are Concepts.
- Concepts have multiple labels.
- Concepts have a Preferred Label (for each language).
- Concepts have any number of Alternative Labels and Hidden Labels (for each language).
- Alternative Label and Hidden Labels are part of a concept’s attributes, not equivalent terms and not connected by “relationships.”
Different Models in Taxonomy Management Software

Thesaurus Model: MultiTes
Different Models in Taxonomy Management Software

Thesaurus Model: Synaptica

- **Descriptor**: Budgeting
- **Object**: Hedden_Thes
- **Categories**
- **Status**: Active; Candidate; Preferred; Unlocked
- **UID**: 18665
- **Created**: hhedden 10/7/2016 7:57:34 PM
- **Modified**: hhedden 10/7/2016 8:01:10 PM

**Budgeting**

- **Top Level Parents**
  - TT **Finance** (Hedden_Thes)
- **Parents**
  - **BT Finance** (Hedden_Thes)
- **Children**
  - **NT Budget analysis** (Hedden_Thes)
- **Associations**
  - **RT Accounting** (Hedden_Thes)
- **Variants**
  - **UF Budget management** (Hedden_Thes) (NPT)
  - **UF Budget planning** (Hedden_Thes) (NPT)
  - **UF Budgets** (Hedden_Thes) (NPT)
Different Models in Taxonomy Management Software

Thesaurus Model: Synaptica
Different Models in Taxonomy Management Software

SKOS model: PoolParty
Different Models in Taxonomy Management Software

SKOS model:
- Smartlogic
- Semaphore
- Ontology Editor
SKOS model: Alternative labels and other languages
Guidelines for implementing variants

- A concept may have any number of (multiple) variants, or it may have only a single preferred name (no variants).

- A variant points to only a single preferred term/concept. (Thesaurus standards permit using a “multiple-use” reference, but for simplification, most software does not permit it.)

- Variants may be displayed to the end-user or they may not be.

- If displayed to the end-user, variants may point (re-direct) to the preferred term, or they can point directly to the content.
Creation and Implementation

Sources for variants

- Same sources as for concepts and preferred terms
  - Survey/audit of the content and terms used
  - Search query logs and other internal usage data
  - External sources: websites, Wikipedia, other taxonomies and controlled vocabularies, book tables of contents, etc.

- Creative changes of terms (after verification of variant term usage in search)

- Not to be used as a source:
  Dictionary-type thesaurus, such as *Roget's Thesaurus* or thesaurus-dictionary websites
Creation and Implementation

Synonym Rings

- No preferred term/preferred label; only an associated set of labels/variants for each concept.
- An option only if terms are never displayed to end-users.
- Used to support search, where there is no browsing the taxonomy.
- Sometimes called “search thesaurus.”
Different Types

Types include

- synonyms
- quasi-synonyms
- variant spellings
- lexical variants
- foreign language names
- acronyms/spelled out
- scientific/popular names
- antonyms (for characteristics)
- older/current names
- phrase variations (in print)
- narrower terms that are not preferred terms
Different Types

Types include

- **synonyms**: Cars / Automobiles
- **quasi-synonyms**: Politics / Government
- **variant spellings**: Taoism / Daosim; Email / E-mail
- **lexical variants**: Selling / Sales; Hair loss / Baldness
- **foreign language names**: Ivory Coast / Côte d'Ivoire
- **acronyms/spelled out**: GDP / Gross domestic product
- **scientific/popular names**: Neoplasms / Cancer
- **antonyms (for characteristics)**: Flexibility / Rigidity
- **older/current names**: Near East USE Middle East
- **phrase variations (in print)**: Unions, labor USE Labor unions
- **narrower terms that are not preferred terms**: Genetic engineering USE Biotechnology
Different Types

Narrower terms as variants

Examples:  Genetic engineering USE Biotechnology

Hand gestures USE Body language

Laptops USE Computers

Correct, because the preferred term is used for the narrower concept and fully encompasses the narrower variant term.

Can be problematic if:

1. the non-preferred/preferred term relationship is not displayed to the end-users, and

2. there are multiple narrower concepts as variants

Example: Computers
- Laptops
- Desktops
- Servers
- Supercomputers
Different Types

Acronyms as variants

- Acronyms alone can be ambiguous.
- In large, multi-subject taxonomies/thesauri, it’s better to include both acronym and spelled out together.

Example:

**DRM (Digital rights management)**

USE Digital rights management

Or

**DRM (Digital rights management)**

USE Digital rights management (DRM)

Or

No variant and just **Digital rights management (DRM)**

Depends on search functionality and preferred style.
How Many to Create

How many variants to create depends on various factors. Especially, how the taxonomy is searched or browsed.

If users may input text in search box,

- *Do* include variants that are alphabetically close (unlike in browsable A-Z index).

**Ethnic groups**
**UF** Ethnic communities
How Many to Create

If system supports “smart” search on words within terms,

- Do not include simple inversions or words within phrases.

Debt financing
- Financing debt

Health care products industry
- Health products industry

Tax credits
- Tax credit
How Many to Create

If system supports “smart” search with grammatical stemming,

- Do *not* include simple plurals and lexical variants.

Epidermal Cyst (MeSH)

- UF Epidermal Cysts

Gatehouses (LC Thesaurus for Graphic Materials)

- UF Gate houses

Agricultural facilities

- UF Agriculture facilities
How Many to Create

With automated indexing / auto-categorization

- More variants are needed than for manual indexing.
- Human indexers will hunt and try different variants.
- Machines need exact matches (if not stemming rules).
- Both statistical and rules-based auto-categorization make use of variants.
- Variants should anticipate possible text strings in the content.

Example for the preferred term **Presidential candidates**:

- Presidential candidacy
- Candidate for president
- Candidacy for president
- Presidential hopeful
- Running for president
- Campaigning for president
- Presidential nominee
How Many to Create

Number of variants to create

On average 1.5 variants to each preferred term/concept.
Many have none; many have multiple variants.

Factors for creating more variants:
- Variations in various sources of content to be tagged
- Varied user types (experts/students, internal/external, etc.)
- End-user use of a search box (taxonomy not displayed by default)
- Implementation of automated indexing/auto-classification
Considerations for *limiting* the need for more variants:

- Variants should be created based on usage warrant, not creative possibilities (phrase inversions, permutations of synonyms of words in multi-word phrases).
- Variants should not be created for low-use, especially narrower concepts, lest they be available for keyword searching.
- “Smart” search or stemming will pick up the concepts without variants with minor differences.
- If the variants are displayed to the end-user, then fewer is better so as not to clutter the display.
User Interface and Search

Need to know how the user interface will display variants

- Are there search options to choose from? 
  *Exact, Begins with, Words within the term, Fuzzy/Smart search*

- Are the search options different for indexers vs. end-users?

- Are the search *capabilities* different for indexers vs. end-users?

- Is there stemming on words? If so, to what extent?

- Is there a type-ahead/auto-suggest display of preferred terms?

- Is there a type-ahead/auto-suggest display of both preferred and variant terms?

Example on the following screenshot slides:

*Education standards USE Educational standards*
User Interface and Search

User interface of the taxonomy editor: “Begins” search

Search Form

- **either enter a search phrase**
  - education standards

- **or select an alphabetical range**
  - Range: [a, v, thru, z]

- **or enter a specific item uid**
  - Item UID: [input field]

Select search criteria

- **Obj**: Subjects
- **Cat**: All Categories
- **Act**: Active

Display batching: 1000 Items per batch

Start Search

Search Results

Elapsed Time for Query: 0.019 seconds
1 Items Found

- Education standards (Subjects) (NPT)
User Interface and Search

User interface of the taxonomy editor: “Smart” search

**Search Form**
- **Search criteria:**
  - **Obj:** Subjects
  - **Cat:** All Categories
  - **Act:** Active
- **Display batching:** 1000 items per batch

**Smart**
- Begins
- Contains
- Exact
- **Select an alphabetical range**
  - Range: a □ thru. □ Z □
- **Enter a specific item uid**
- **Enter a search phrase**
  - education standards

**Search Results**
- Elapsed Time for Query: 0.068 seconds
- 5 Items Found
  - **Education standards** (Subjects) (NPT)
  - **Educational standards** (Subjects)
  - **State education standards** (Subjects) (NPT)
  - **State educational standards** (Subjects) (NPT)
  - **State standards (Education)** (Subjects)
User Interface and Search

User interface of the indexer: Alphabetical browse
User Interface and Search

User interface of the indexer: Smart search

[Image of a user interface showing a search query for education standards and a list of narrower terms and related standards.]
User Interface and Search

User interface of the end-user: Search on Subjects (“Subject Guide”)
User Interface and Search

User interface of the end-user: “Autosuggest” enabled
User Interface and Search

User interface of the end-user: Default “begins with” type-ahead search (http://vocabulary.worldbank.org/thesaurus.html)
Variations and Customizations

Displayed vs. non-displayed variants

Non-displayed variants are useful:

- For common misspellings, slang, or deprecated, or potentially offensive terms not displayed to users but can match searches
- For auto-categorization support but not intended for manual indexing
- For search support but not intended for type-ahead display

SKOS model also has **Hidden Label** (hiddenLabel) for these uses.

Non-SKOS thesaurus management software allows relationship customization, such as designating a non-displayed USE/UF.

- As a reciprocal relationship, such as IUS/IUF (internal use/internal used for)
Variations and Customizations

Internal Use / Internal Used for (IUS/IUF)

Typically for changed terms to ensure that records indexed with the old term will be retrieved with the new term, but the old term is inappropriate as a variant.

Examples:

Bars, saloons, etc. IUSBars (Drinking establishments)
Mixers (Cookery) IUSMixers (Food preparation)
Pates (Food) IUSPates
Soap trade IUSCleaning agents industry
Spaying IUSSpaying and neutering

Example of two former narrower terms that had been removed:
Proposal writing in public contracting IUSProposal writing
Proposal writing in research IUSProposal writing
Variations and Customizations

More specific customized variations

Most thesaurus software permits full customizing the equivalence relationship for multiple sub-types.

SKOS-based software may also permit customization, but not in accordance with the SKOS model for data exchange.

Examples

- An acronym or abbreviation, corresponding with the spelled out form
- A misspelling or alternate spelling, corresponding with the preferred spelling
- An obsolete/legacy term, corresponding with the current term
Conclusions

- Variants are different wordings that refer to the same general concept, for the context.
- Variants are useful in many taxonomies, not just in thesauri.
- Variants may be of the equivalence model in thesauri or as alternative labels in SKOS vocabularies.
- Variants are of different kinds, not just synonyms.
- Search features and user interface need to be taken into consideration when deciding how many variants to create.
- Consider using the SKOS Hidden Label or customized equivalence relationships in thesauri, if you don’t want all variants to display to all users.
Questions/Contact

Heather Hedden
Senior Vocabulary Editor
Indexing & Vocabulary Services
Metadata Standards and Services
Gale | Cengage Learning
20 Channel Center St., Boston, MA 02210
(o) 617-757-8211 | (m) 978-467-5195
heather.hedden@cengage.com
www.cengage.com

heather@hedden.net
www.accidental-taxonomist.com