Synonyms, Alternative Labels, and Non-preferred Terms

Taxonomy Boot Camp London
18 October 2016

Heather Hedden, Senior Vocabulary Editor Indexing and Vocabulary Services Gale | Cengage Learning



About Heather Hedden

- Controlled vocabulary editor at a library database vendor, Gale/Cengage Learning, 1996 – 2004, 2014 – present
- Previous taxonomy consultant
- Continuing education instructor
 Simmons College School of Library and Information Science
- Author of The Accidental Taxonomist (Information Today, Inc.)
- Freelance back-of-the-book indexer
- SLA Taxonomy Division board member
- American Society for Indexing board member



Outline

- Introduction: Definition, Examples, Usage
- Different Designations and Models
- 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



Introduction

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.



Synonym

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



Alternative Label

- Formal designation for SKOS (Simple Knowledge Organisation 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.



Non-preferred Term

- Formal designation in thesauri, in accordance with ISO 25964 and ANSI/NISO Z.39-19 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.



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



Thesaurus non-preferred/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).
- Both preferred terms and non-preferred terms are "terms."

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 and hidden labels are part of a concept's attributes, not equivalent terms and not connected by "relationships."



Thesaurus standards/guidelines

- ISO or BS ISO 25964-1 Information and documentation -Thesauri and interoperability with other vocabularies
 Part 1: Thesauri for information retrieval [2011]
- ANSI/NISO Z39.19-2005 (R2010) Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies

http://www.niso.org/apps/group_public/download.php/12591/z3 9-19-2005r2010.pdf

SKOS model recommendation

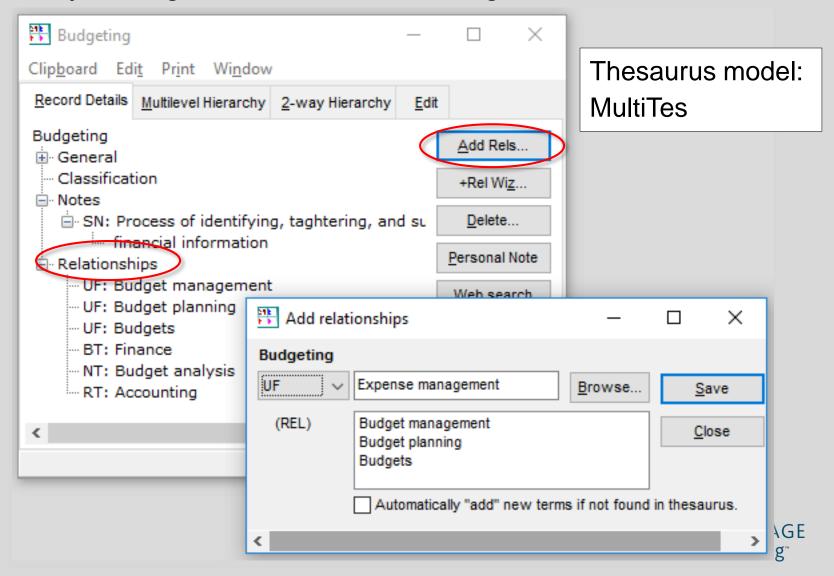
A World Wide Web (W3C) recommendation

"A common data model for sharing and linking knowledge organization systems via the Web"

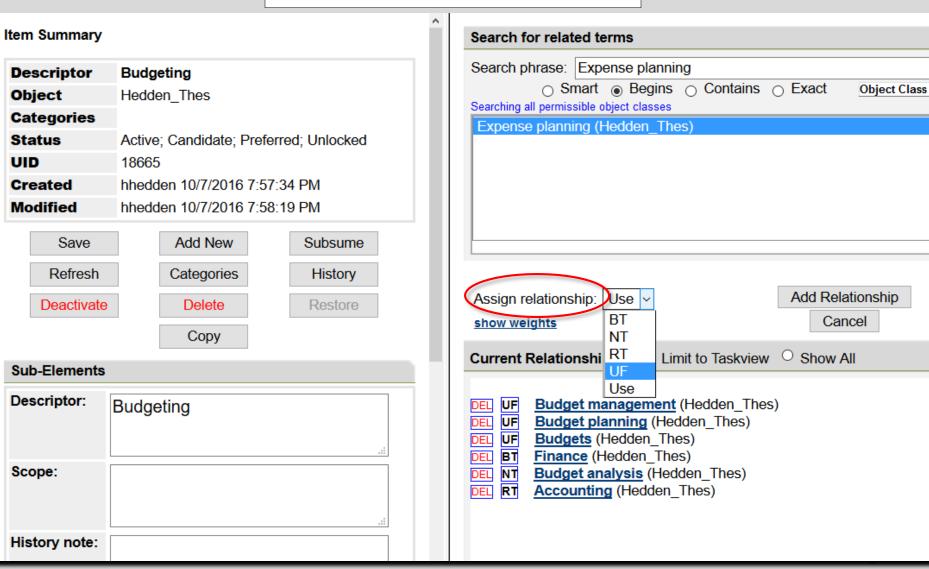
https://www.w3.org/TR/skos-reference/



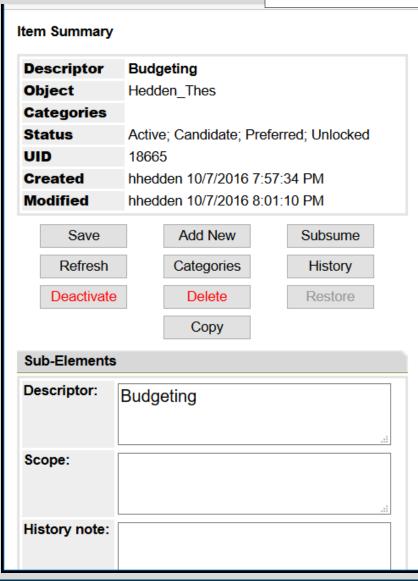
Taxonomy management software's management of variants

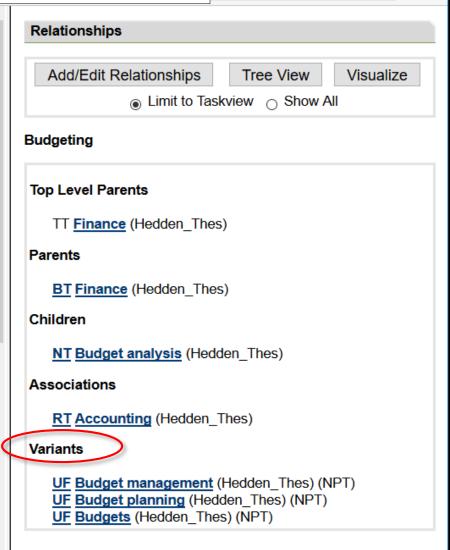


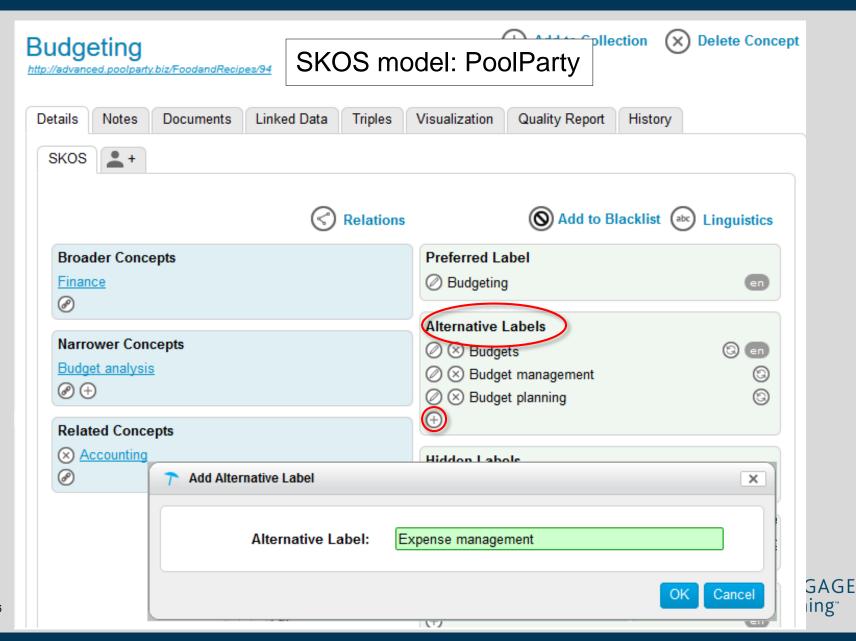
Thesaurus model: Synaptica

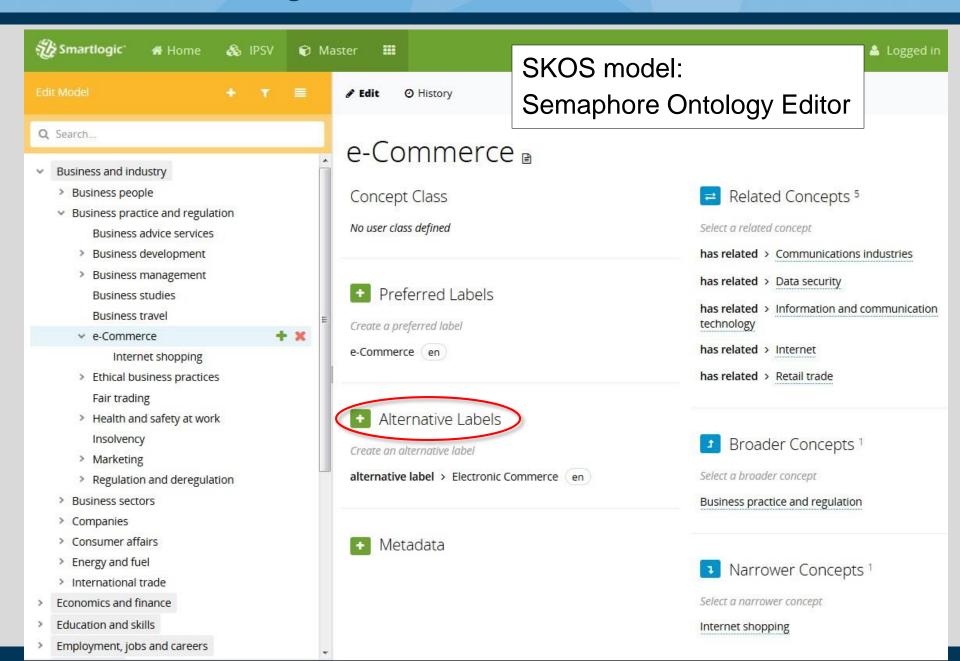


Thesaurus model: Synaptica

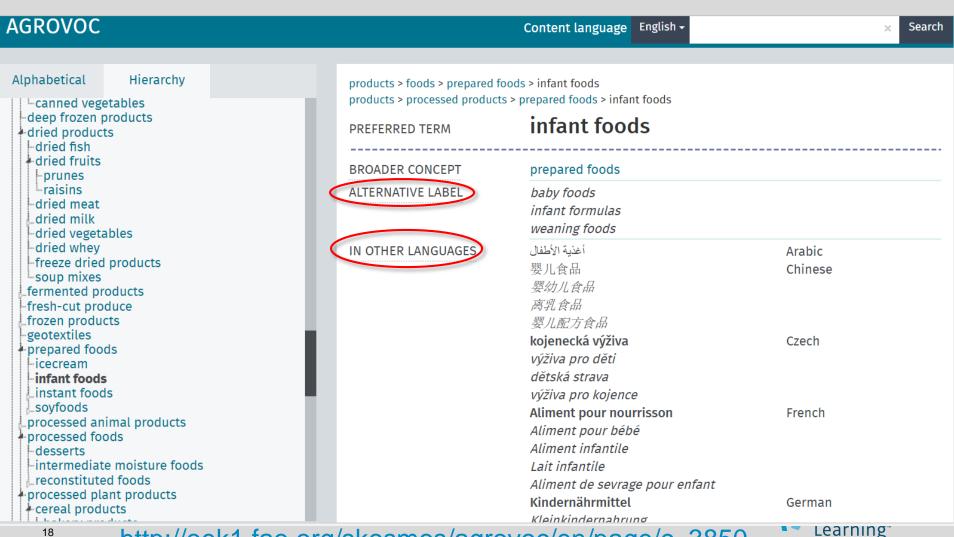








SKOS model: Alternative labels and other languages



Creation and Implementation

Guidelines for using 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.
 (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.
- 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 of English Words and Phrases 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."



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



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, labour USE Labour unions
- narrower terms that are not preferred terms: Genetic engineering USE Biotechnology



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:

- the non-preferred/preferred term relationship is not displayed to the end-users, and
- there are multiple narrower concepts as variants

Computers

- Laptops
- Desktops
- Servers
- Supercomputers



Narrower terms as variants

Problematic scenario:

- Indexer indexes document on Supercomputers with Computers.
- End-user looks up term Laptops, and is taken directly to result set of documents indexed 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 indexing is wrong by retrieving documents on other kinds of computers besides the selected laptops.



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.

Examples:

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)

Depending on search functionality and preferred style.



How many variants to create?

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



If system supports "smart" search on words within terms,

> Do *not* include simple inversions or words within phrases.

Debt financing
UF Financing debt

Health care products industry

UF Health products industry

Tax credits

UF Tax credit



If system supports "smart" search with grammatical stemming,

Do not include simple plurals and lexical variants.

Epidermal Cyst

UF Epidermal Cysts

(MeSH)

Gatehouses

UF Gate houses

(LC Thesaurus for

Graphic Materials)

Agricultural facilities

UF Agriculture facilities



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



Need to know how the user interface will display variants

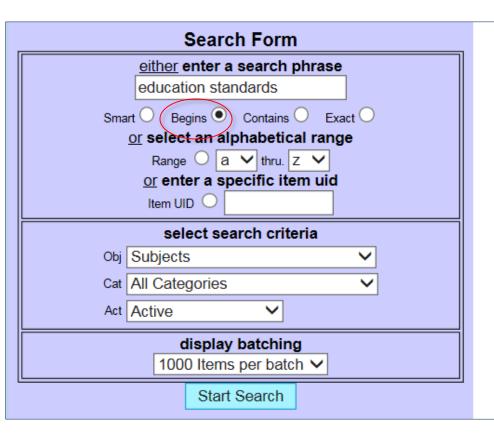
- Are there search options to choose from?
 Exact, Begins, with, Words within the term, Fuzzy/Smart search
- Are the search options different for indexers vs. end-users?
- Are the search capabilities different for indexers vs. end-users?
- Is there stemming on words? If so, to what extent?
- Is there 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 of the taxonomy editor: Begins search



Search Results

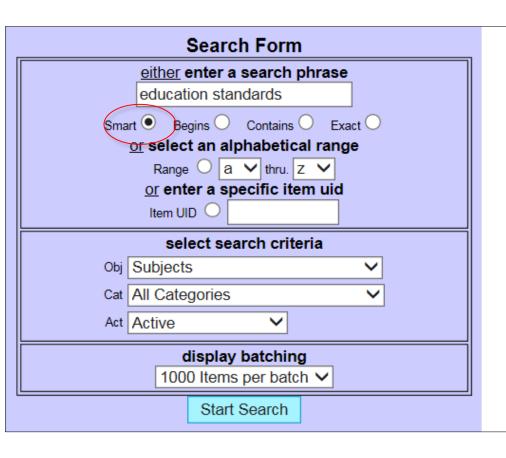
Elapsed Time for Query: 0.019 seconds

1 Items Found

· Education standards (Subjects) (NPT)



User interface of the taxonomy editor: Smart search



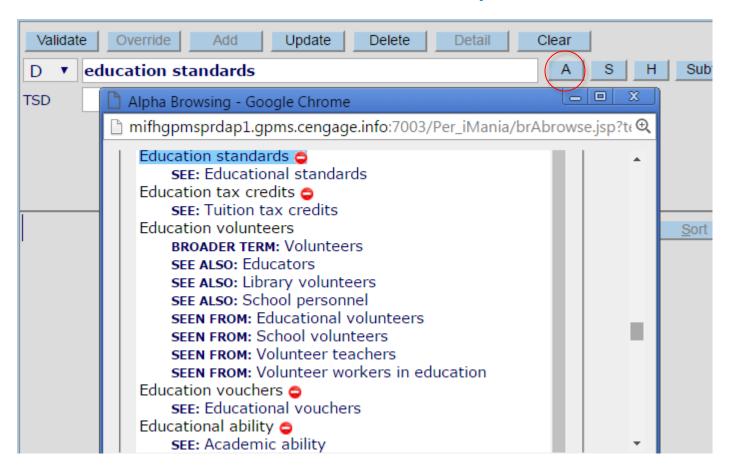
Search Results

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)

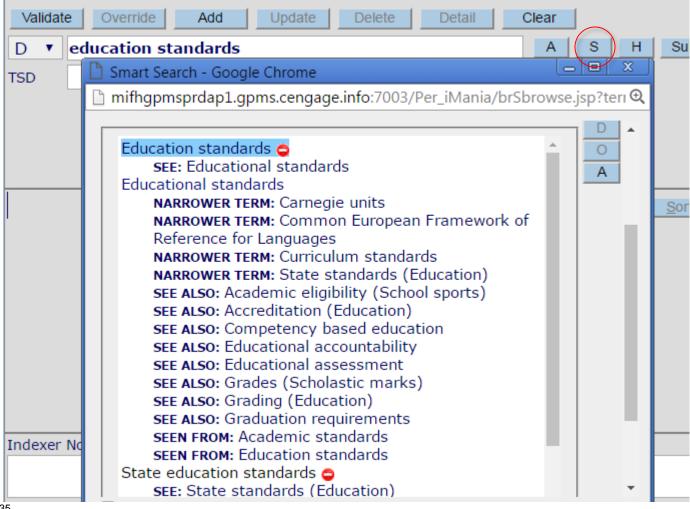


User interface of the indexer: Alphabetical browse



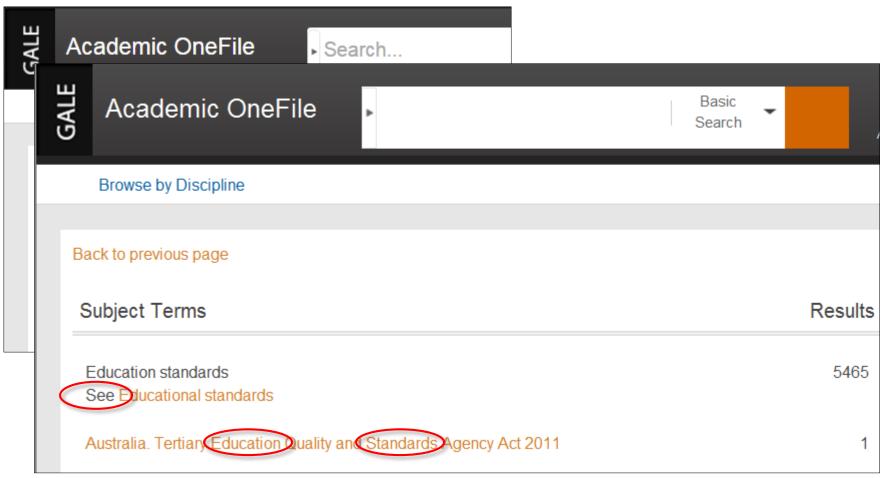


User interface of the indexer: Smart search



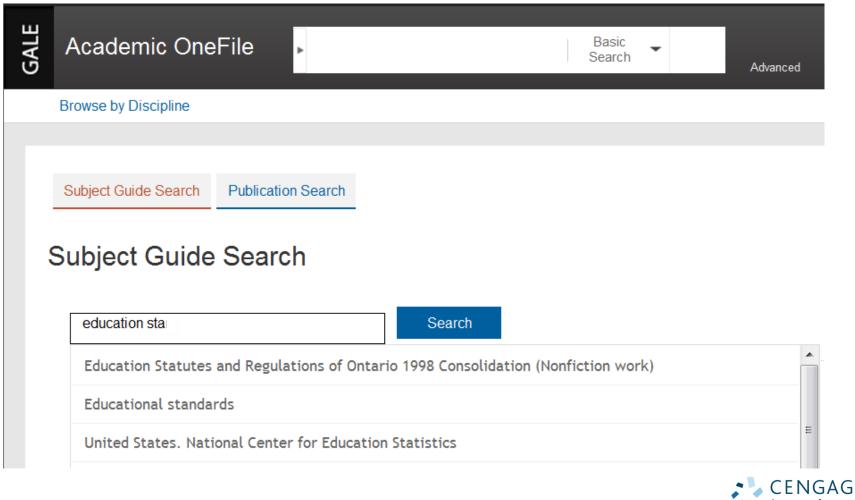


User interface of the end-user: Search on Subjects



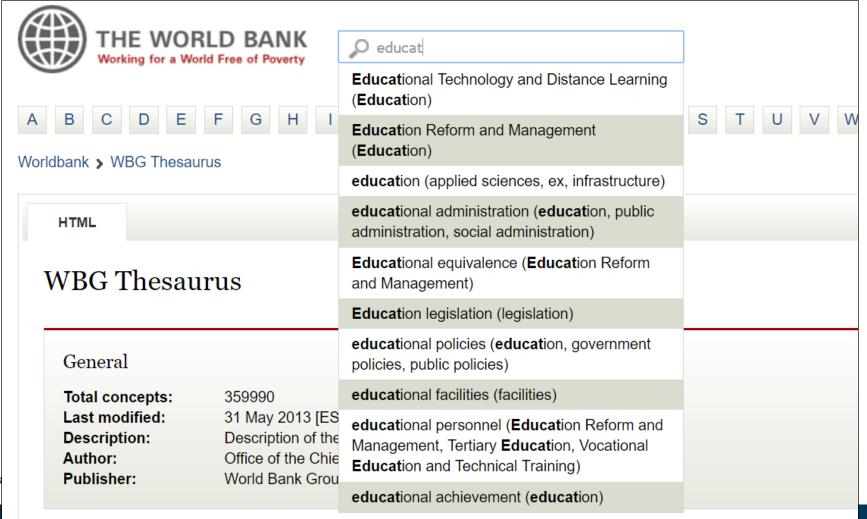


User interface of the end-user: Autosuggest enabled





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. IUS Bars (Drinking establishments)

Mixers (Cookery) IUS Mixers (Food preparation)

Pates (Food) IUS Pates

Soap trade IUS Cleaning agents industry

Spaying IUS Spaying and neutering

Example of two narrower terms that were removed:

Proposal writing in public contracting IUS Proposal writing

Proposal writing in research IUS Proposal 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

Summary

- 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?

Heather Hedden

Senior Vocabulary Editor Cengage Learning 20 Channel Center Street Boston, MA 02210 USA

(o) +1-617-757-8211 | (m) +1-978-467-5195

Heather.Hedden@cengage.com | Heather@Hedden.net

www.cengage.com

www.accidental-taxonomist.com

