How Many Synonyms Should You Have?

Taxonomy Boot Camp
Washington, DC, November 14, 2016

Heather Hedden, Sr. Vocabulary Editor, heather.hedden@cengage.com
John Magee, Director, john.magee@cengage.com
Indexing and Vocabulary Services
Gale | Cengage Learning

Outline

Introduction to synonyms

- Definition, purpose, examples, designations and models, and implementation

Whether to create more or fewer synonyms, based on:

- Sources and methods of creation
- Types of synonyms
- Implementation of search
- Implementation of automated indexing
- Implementation of the user interface
- Display of synonyms to end-users

Introduction to Synonyms

Synonyms (Alternative Labels, Non-preferred Terms, etc.)

- 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
- Enabling consistent indexing/tagging

Introduction to Synonyms

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 synonyms]

Introduction to Synonyms

Designations and Models

Synonym

- Informal designation in taxonomies
- Not entirely accurate, because most are *not* synonyms (not exact equivalents, not single words).
- Simple, non-expert, widely understood.
- Associated with a *Term.*

Non-preferred Term (NPT)

- Formal designation in thesauri, in accordance with ANSI/NISO Z.39-19 and ISO 25964 thesaurus standards.
- Not intuitively understood by non-experts.
- Associated with a *Preferred term.*

Alternative Label (altLabel)

- Formal designation for SKOS (Simple Knowledge Organization System) (W3C) vocabularies.
- Intuitively understood by non-experts and varied stakeholders.
- Associated with a *Preferred label.*

Introduction to Synonyms

Designations and Models

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 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.”

Introduction to Synonyms

When to implement synonyms

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.

Even if it's called a "taxonomy" and not a "thesaurus," that does not matter.

Introduction to Synonyms

Guidelines for implementing synonyms

- A concept may have any number of (multiple) synonyms, or it may have no synonyms.
- A synonym points to only a single preferred term/label. (Thesaurus standards permit using a “multiple-use” reference, but for simplification, most commercial taxonomy management software does not permit it.)
- Synonyms may or may not be displayed to the end-user.
- Synonyms may point (re-direct) to the preferred term/label, or they can point directly to the content.

Creating More or Fewer Synonyms

More or fewer synonyms, based on:

- Sources for and methods of synonym creation
- Types of synonyms
- Implementation of search
- Implementation of automated indexing
- Implementation of the user interface
- Display of synonyms to the end-users

Synonym creation based on sources: Many

Create numerous synonyms based on numerous sources and methods:

- 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:
  - Synonyms for each word and different combinations
  - Flipping adjective-noun phrases and prepositional phrases

Synonym creation based on sources: Fewer

Create synonyms based on **warrant**

- Verify the candidate variant has significant usage/occurrence in the content repository
- Don’t use every possible variant.
- Don’t pull synonyms out of Roget’s thesaurus.

Synonym creation based on types: Many

Create numerous synonyms based on numerous types:

- synonyms: Cars / Automobiles
- near-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: Unions, labor USE Labor unions
- narrower terms that are not preferred terms:
  Genetic engineering USE Biotechnology

Synonym creation based on types: Fewer

Creating numerous synonyms based on types:

- near-synonyms: Politics / Government
- lexical variants: Selling / Sales

- Possibly OK when tangential to the scope of the taxonomy, otherwise nuanced different meanings are lost.
- By creating synonyms out of lesser-used different terms, they become unavailable for a keyword search, for the user who really wants to retrieve anything on the specific concept which does not have a preferred taxonomy term.

.

Synonym creation based on types: Fewer

Other types:

- variant spellings: Taoism / Daosim; Email / E-mail
- foreign language names: Ivory Coast / Côte d'Ivoire
- antonyms (for characteristics): Flexibility / Rigidity

➢ Situations for these are not common.

- phrase variations: Unions, labor USE Labor unions

➢ Inversions are only used in printed thesauri, as simple find/search on page will get the user to the term.

Synonym creation based on types: Fewer

Other types:

- acronyms/spelled out: GDP / Gross domestic product

  - Acronyms alone can be ambiguous.

    It's better to include both acronym and spelled out together within the same term.

    GDP (Gross domestic product)

    OR

    Gross domestic product (GDP)

    Depending on preferred style.

Synonym creation based on types: Fewer

Other types:

- narrower terms that are not preferred terms - Examples:
  Genetic engineering USE Biotechnology
  Laptops USE Computers

➢ Correct, because the preferred term is used for the narrower concept, which it fully encompasses.

➢ Can be problematic if:

1) the redirecting non-preferred/preferred term relationship is not displayed to the end-users, *and*
2) there are multiple narrower concepts as synonyms, e.g.:

Computers
- *Laptops*
- *Desktops*
- *Servers*
- *Supercomputers*

Synonym creation based on types: Fewer

Narrower terms as synonyms (continued)

Problematic scenario:

1. Indexer indexes document on **Supercomputers** with **Computers**.
2. End-user looks up term **Laptops**, and is taken directly to result set of documents indexed with **Computers**.
3. Result set includes documents on supercomputers and other computers that are not laptops, in addition to documents on laptops.
4. End-user thinks the indexing is wrong by retrieving documents on other kinds of computers besides the selected laptops.

Implementation of search: Many

If users may input text in search box,

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

  Ethnic groups

  UF **Ethnic communities**

Search boxes are almost universal, so more synonyms are needed.

Implementation of search: Fewer

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

- Do *not* include simple inversion or words within phrases.

Debt financing

~~UF **Financing debt**~~

Health care products industry

~~UF **Health products industry**~~

Tax credits

~~UF **Tax credit**~~

Implementation of search: Fewer

If system supports “smart” search with additional grammatical stemming,

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

Epidermal Cyst

~~UF **Epidermal Cysts**~~

Gatehouses

~~UF **Gate houses**~~

Agricultural facilities

~~UF **Agriculture facilities**~~

Implementation of automated indexing: Many

With automated indexing / auto-categorization

More synonyms are needed than for manual indexing.

- Human indexers will hunt and try different synonyms.
- Machines need exact matches (if not stemming rules).
- Both statistical and rules-based auto-categorization make use of synonyms.
- Synonyms should anticipate possible text strings in 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

Implementation of user interface: Many

“Begins with” or “type-ahead” feature on search box

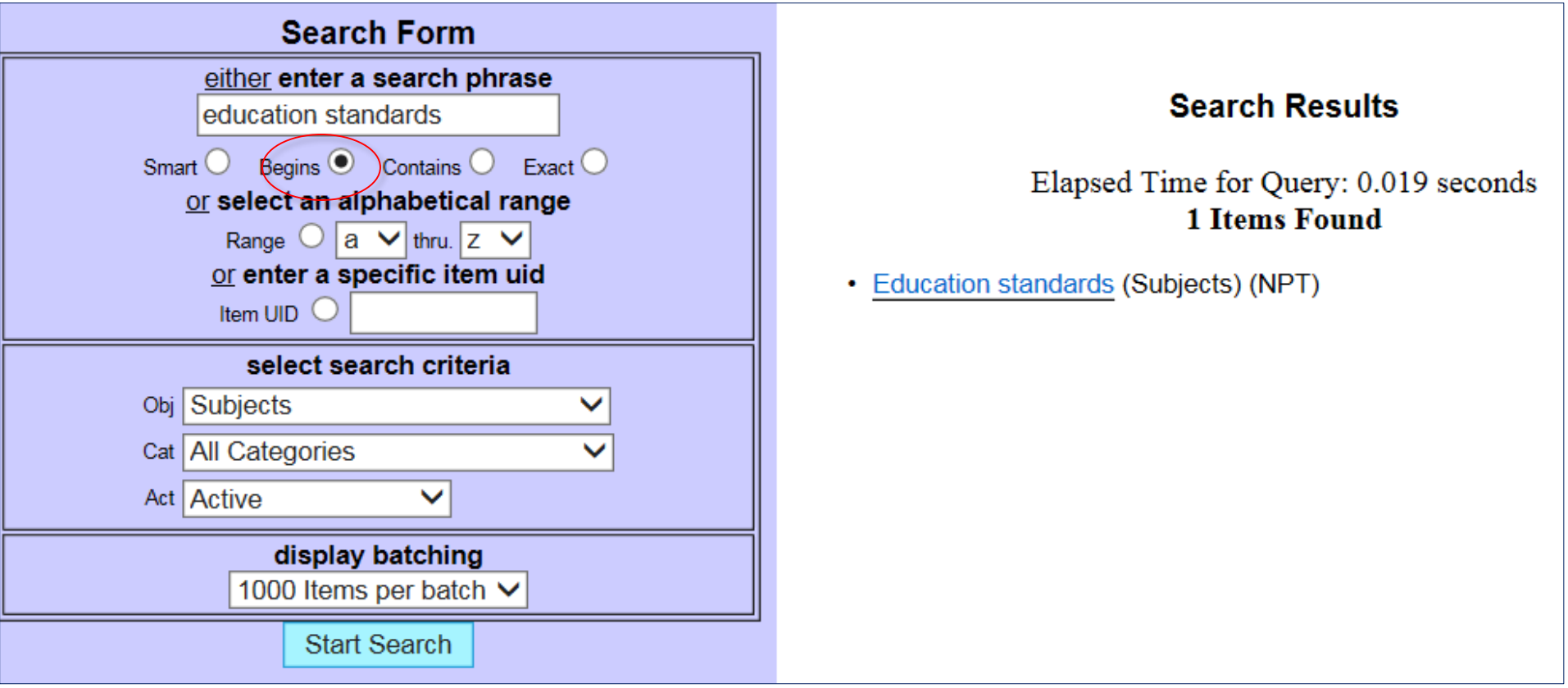
- Only retrieves terms that start with word or phrase.
- More synonyms are needed for different initial words or phrases.

Example on the following screenshot slides:

Education standards USE **Educational standards**

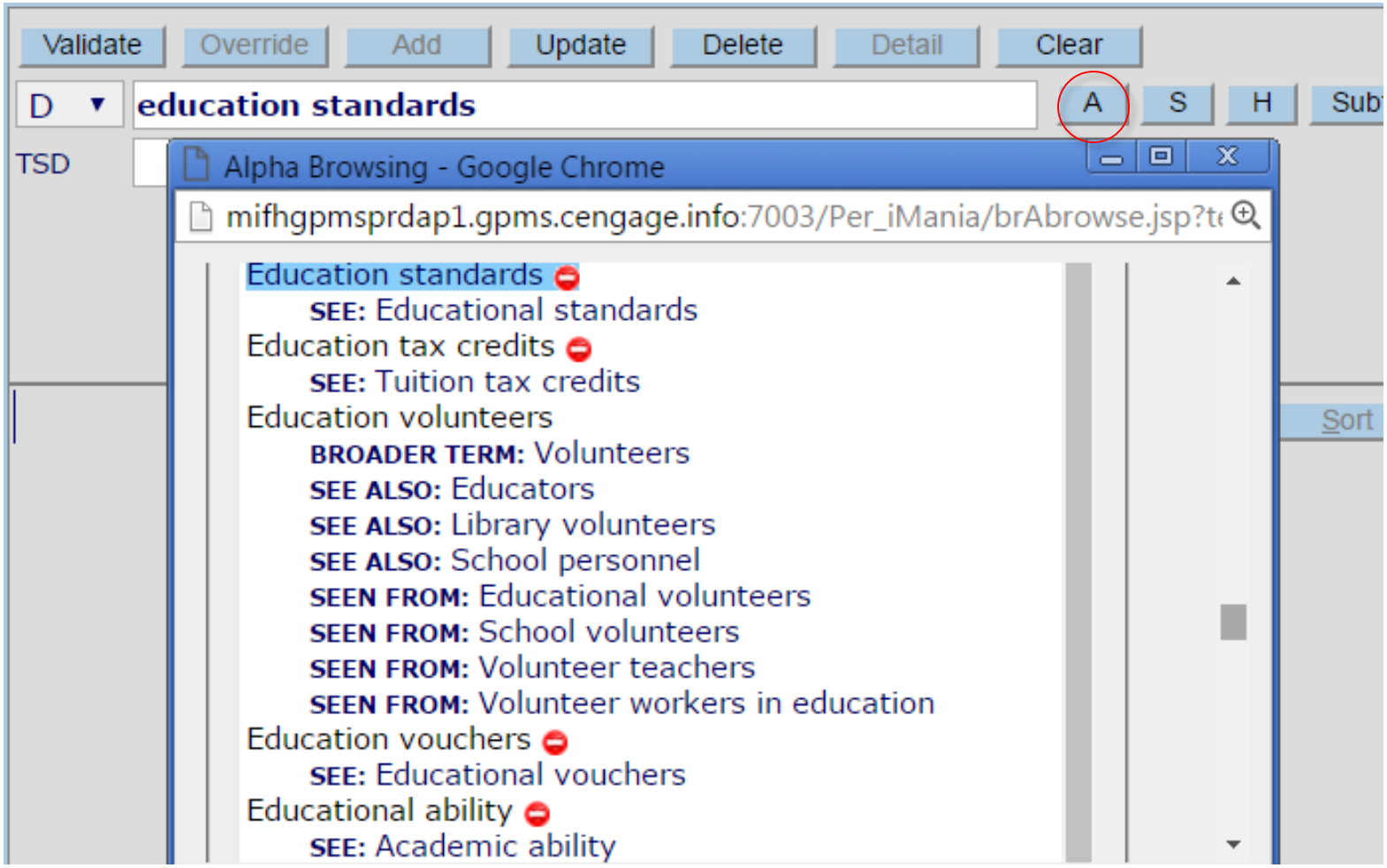
Implementation of user interface: Many

User interface of the taxonomy editor: Begins search



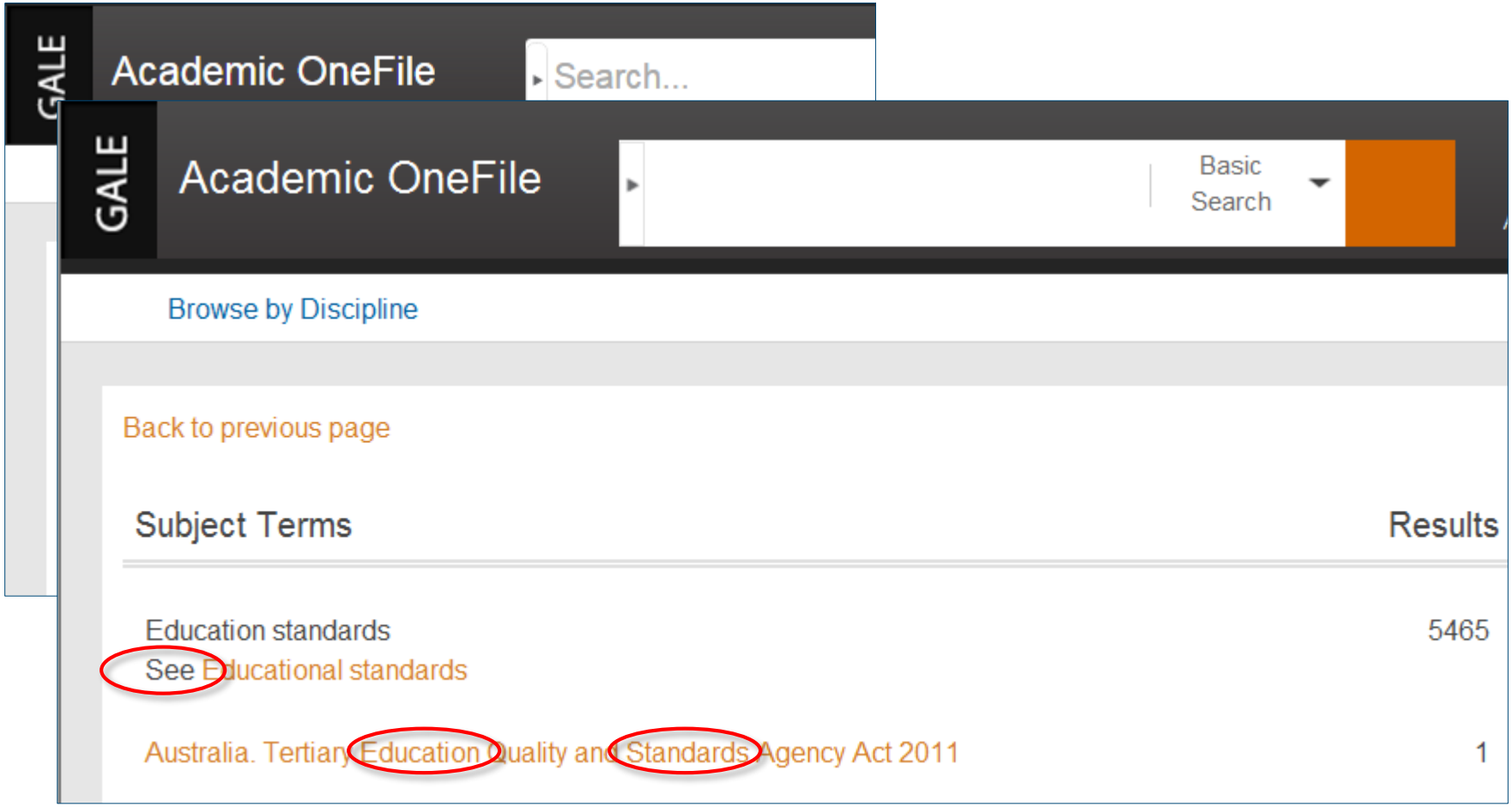
Implementation of user interface: Many

User interface of the indexer: Alphabetical browse



Implementation of user interface: Many

User interface of the end-user: Search on Subjects



Implementation of user interface: Fewer

“Smart” search or “auto-suggest” feature on search box

- Retrieves terms that have the words in them in any order.
- Retrieves terms that have search words within larger words.

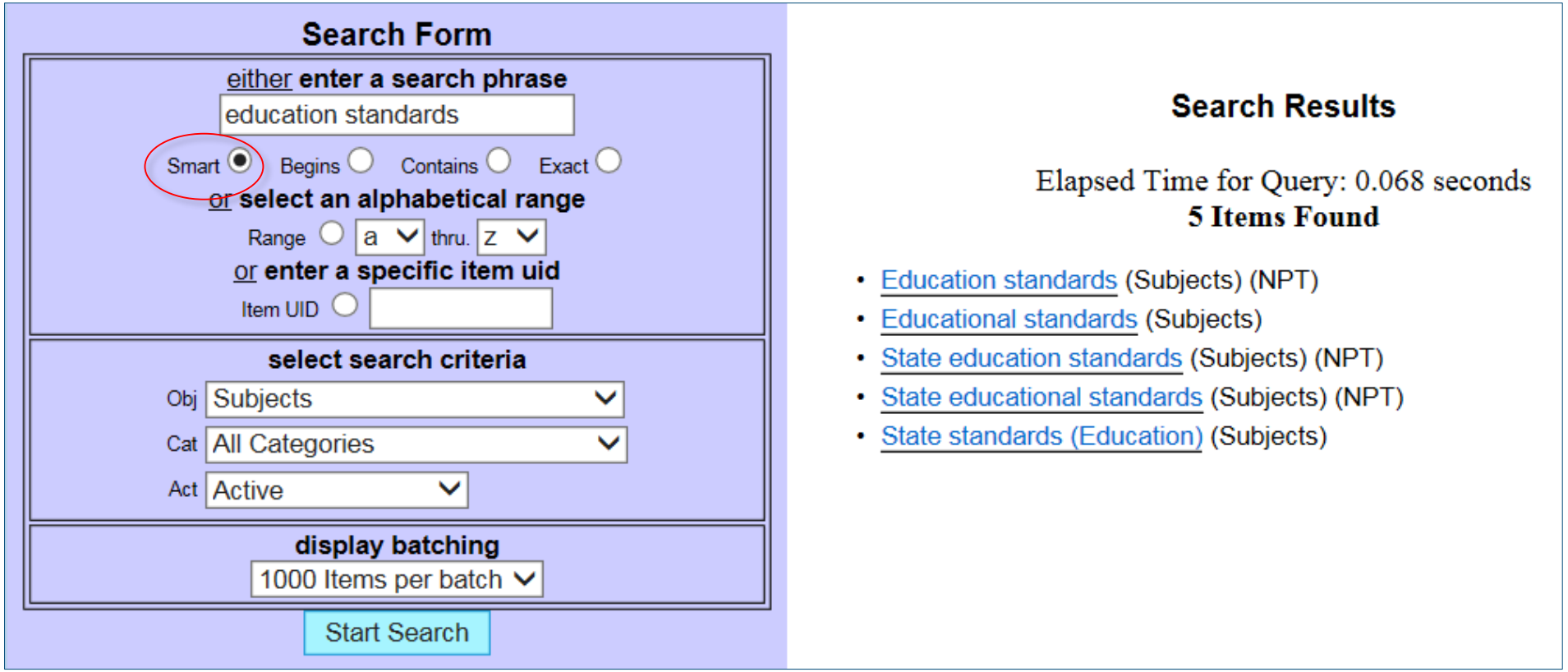
- Synonyms are not needed for simple phrase inversions or shorter words within terms words.
- Too many synonyms can clutter up the list of matching terms.

Example on the following screenshot slides:

Education standards USE **Educational standards**

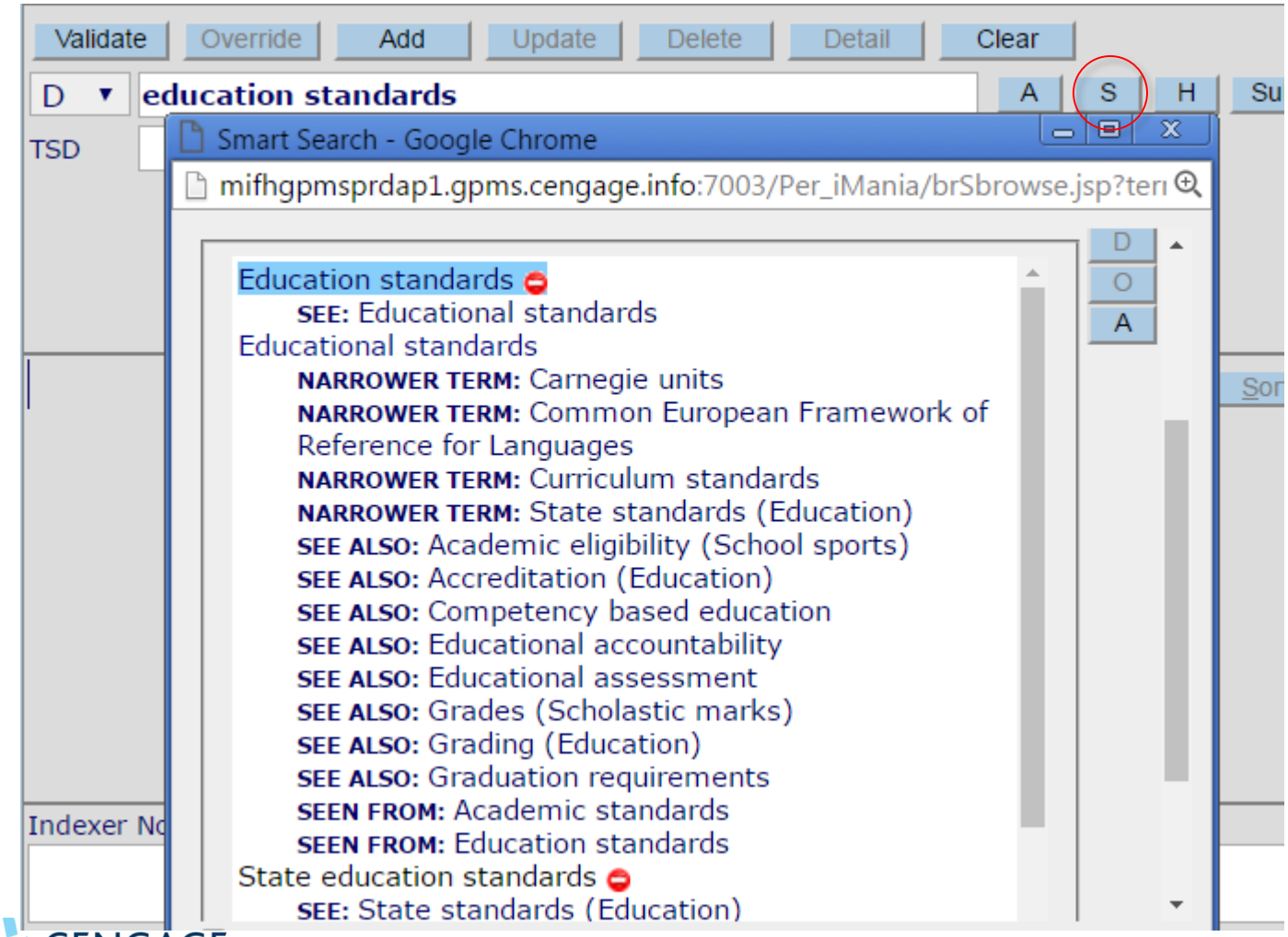
Implementation of user interface: Fewer

User interface of the taxonomy editor: Smart search



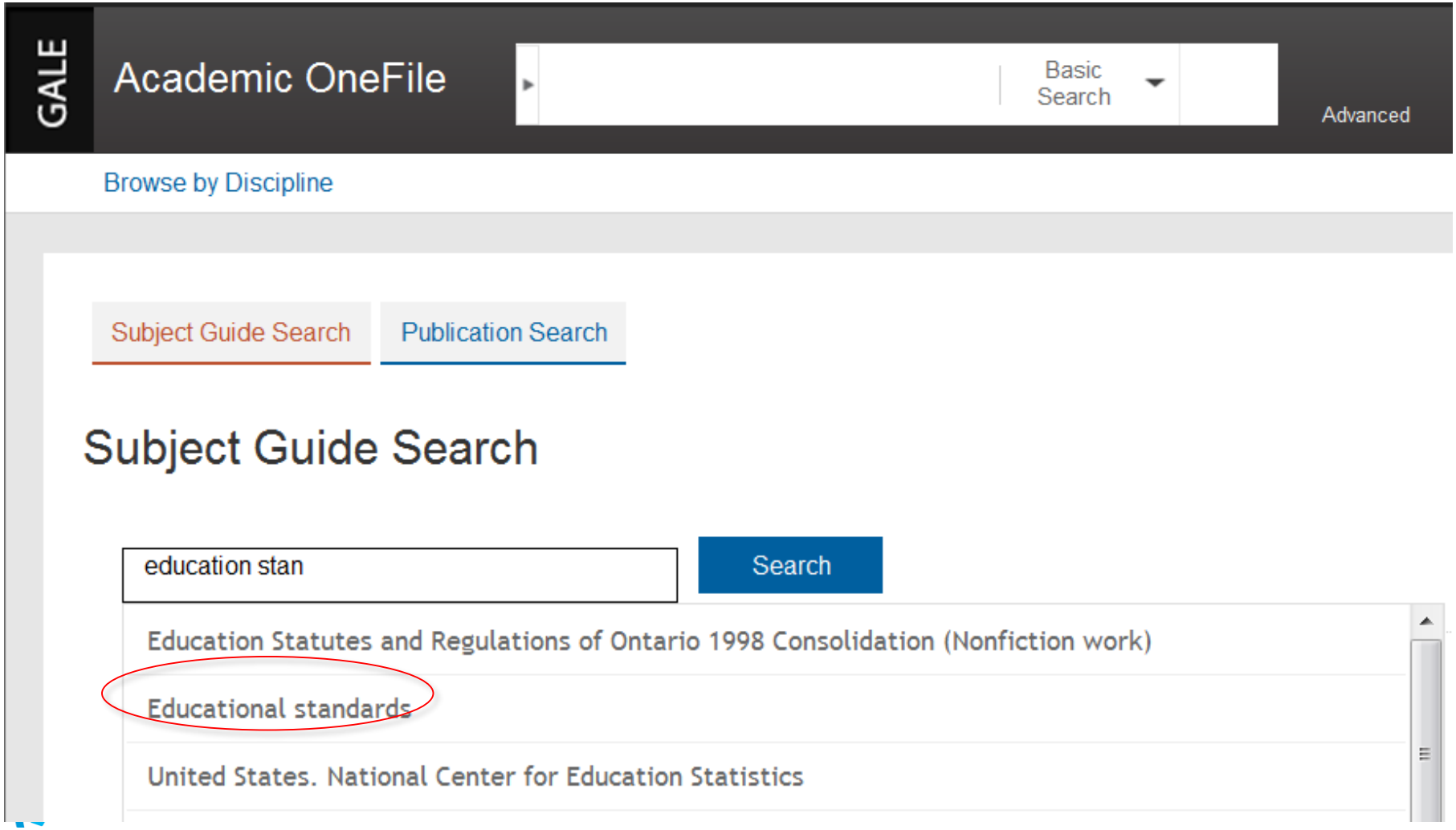
Implementation of user interface: Fewer

User interface of the indexer: Smart search



Implementation of user interface: Fewer

User interface of the end-user: Auto-suggest enabled



Display of synonyms to end-users: Fewer

When synonyms are displayed to the end-users

It's better to have fewer synonyms so as not to clutter the display with every possible variant, especially those not appropriate to display:

- Common misspellings
- Slang, jargon, or potentially controversial/offensive/not politically correct terms
- Deprecated terms
- Commonly entered search strings from search logs that are not good quality terms

Display of synonyms to end-users: Many

When synonyms are displayed to the end-users

Compromise: Designate certain synonyms not to display

- SKOS model also has **Hidden Label** (hiddenLabel) for this.
- 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)

Examples of “internal use” for deprecated arcane terms:

Bars, saloons, etc. IUS Bars (Drinking establishments)

Soap trade IUS Cleaning agents industry

How many synonyms should you have?

Suggested rough guideline ratio of synonyms to preferred terms/concept:

Many: 1.5 : 1

Fewer: 1 : 1 or less

Questions/Contact

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

John Magee
Director, Indexing & Vocabulary Services
Gale | Cengage Learning
27500 Drake Rd.
Farmington Hills, MI 48331
248-699-8091
john.magee@cengage.com

www.cengage.com