Controlled vocabularies: an introduction

Article *in* The Indexer The International Journal of Indexing · September 2008

DOI: 10.3828/indexer.2008.37

1 author:

Fred Leise
ContextualAnalysis
4 PUBLICATIONS **13** CITATIONS

SEE PROFILE

Controlled vocabularies: an introduction

Fred Leise

This article, based on a presentation to the SI Annual Conference 2008, introduces basic concepts and terminology associated with the field of controlled vocabularies (CVs). General topics discussed are indexes versus CVs, an introduction to what CVs are and how they are used, the use of facets with controlled vocabularies, and CV governance and maintenance issues.

Introduction – a terminological problem

As Bella Haas Weinberg is fond of saying, the field of controlled vocabularies has a singular lack of vocabulary control. Although there are a number of glossaries of terms from the field,[1] there is often little consensus about terminology.

One vexing problem is the word 'taxonomy.' Originally describing a type of controlled vocabulary that included broader and narrower terms, the word taxonomy is now used in the corporate world to denote any type of controlled vocabulary. In the interest of precision, this article uses 'taxonomy' in its more limited sense.

Indexes versus controlled vocabularies

Table 1 illustrates some of the principal differences between indexes and CVs. While both indexes and CVs are produced using the basic activities of concept identification and term selection, they differ significantly in their methods of construction and in their uses.

Table 1 Indexes and controlled vocabularies compared

	Indexes	**CVs**
End product	index	term list
Use	content locator	content tagging website navigation search enhancement
Project time	weeks	months
Methodology	reading	research

Yet indexes and CVs are intimately connected. For example, it is often the case that for large or long-term indexing projects, such as journal or newspaper indexes, a CV is created first to ensure consistency in indexing by multiple indexers over long time spans. It is also the case that an index is a CV, although not a type that is usually considered when discussing CVs in general.

About controlled vocabularies

What are controlled vocabularies?

According to Hans Wellisch (1991: 246), a controlled vocabulary is a 'list of terms that may be used for indexing, produced by the operation of vocabulary control.' So you do vocabulary control and get a controlled vocabulary? *Pace* Mr Wellisch, I don't find that definition helpful.

Heather Hedden, in her article in the March 2008 issue of *The Indexer* (Hedden 2008a), notes that:

> At a minimum, a controlled vocabulary is a restricted list of words or terms used for indexing or categorizing. It is controlled because only terms from the list may be used for the subject area covered by the controlled vocabulary. It is also controlled because, if it is used by more than one person, there is control over who adds terms or how terms can be added to the list.

I define a CV as:

> A list of terms and term relationships designed to: (1) collect similar information, (2) assist content authors in consistently tagging content, and (3) enable users to find the information they need by translating their language into the language of the information store.

Any time we choose to use one word over another, we are creating a CV. If a website lists 'men's clothes,' then someone has chosen that term rather than 'men's clothing' or 'clothing for men.' When a series of those choices is made and the resulting vocabulary displays certain characteristics, then a CV has been created.

Most important of the CV characteristics is that of relationship. That is, the terms in the CV are related in certain ways, as will be discussed in the next section.

Term relationships

Synonymy

The most basic term relationship is synonymy: that is, having the same (or nearly the same) meaning. It is important to note that context is important in determining synonymy. In one context, a cookie may be the same as a biscuit (if the cookie is American and the biscuit is English). In another context, a cookie may not be the same as a biscuit (if both terms reflect American usage, for instance). The following are examples of synonyms.

> Country = Nation
> Chief of Government = Prime Minister
> Brunei = Sultanate of Brunei = Negara Brunei Darussalam = سلطنة بروناي = برني دارالسلام

Hierarchical relationship

Terms display a hierarchical relationship when one term is broader in meaning than its child term (which has a narrower meaning). The hierarchical relationship can be of several types.

Whole–part relationships obtain when the broader terms represents the whole of something and the narrower terms represent parts of the whole. For example, the following terms display a whole-part relationship.

Automobile
- Air bags
- Engine
- Seats
- Steering wheels

The *instance relationship* reflects narrower terms that are examples of the broader term, for example:

Buildings
- Great Pyramid of Giza
- Madison Square Garden
- Petronas Towers
- Sears Tower
- Taipei 101

Associative relationship

The third, and most complex, relationship that terms may exhibit is the *associative relationship*. The associative relationship, while easy for us to grasp, is often difficult to define with precision. There are many ways in which terms may be related. For example, a vase and a drinking glass may be related because both are made out of the same material, glass. The vase may also be related to a planter, since both hold plants. Or the vase may be related to a violin because they both happen to be owned by the same individual.

The following examples of types of associative relationship are taken from Jean Aitchison's *Thesaurus construction and use* (2000: 63–6):

Associative type	**Example**
operation/agent	turning : lathes
occupation/person	social work : social worker
causal dependence	friction : wear
agent/counteragent	pests : pesticides
concept/opposite	tolerance : prejudice
concept/origin	water : water wells

Controlled vocabulary types

The three term relationships discussed above are the defining features of the three main types of CV.

Synonym ring

A synonym ring consists entirely of terms that are synonymous:

chips = French fries

Here again, it is easy to see that context is important, since the above synonymy holds in the United Kingdom, but not in the United States.

Authority file

If one of the terms in a synonym ring is designated as a 'preferred term,' then the CV is called an *authority file*. In this case, the preferred term may be used for tagging content, so as to improve findability. The non-preferred terms are also called 'entry' terms or 'variant' terms.

An index that contains see references can be considered a variety of authority file. The locators are given at the preferred term and non-preferred terms are linked to the preferred term by 'see' cross-references.

Authority files can be presented in two different ways. In the first, all of the terms are presented in alphabetical order. In the second, the terms are presented in a spreadsheet, with the preferred terms in one column and the entry terms in a second column.

In the following alphabetical display of an authority file, 'USE' indicates the preferred term used instead of an entry term, and 'UF' indicates the entry terms that a preferred term is used for. Think of the USE entries as *see* references: x *see* y. Boldface indicates preferred terms.

community USE neighborhood
health and safety UF safety
levy USE tax
neighborhood UF community
parks UF recreation
rebate USE refund
recreation USE parks
refund UF rebate
safety USE health and safety
tax UF levy

Table 2 is the same authority file displayed in tabular form.

Table 2 An authority file in tabular form

Preferred term	**Variant terms**
health and safety	safety
neighborhood	community
parks	recreation
refund	rebate
tax	levy

Although it is easier to see the relationships in the spreadsheet presentation, the alphabetical format makes it easier to find any specific term (either preferred or entry) in the CV.

Taxonomy

Take all of the features of an authority file and add the hierarchical relationship of broader terms (BTs) and narrower terms (NTs) discussed above. The resulting structure is a taxonomy, also called a *hierarchy*.

While an index contains certain features of a taxonomy, namely that each heading/subheading pair exhibits the broader/narrower relationship, the index as a whole does not necessarily contain a single hierarchical structure. Usually there are many top-level terms, while in a strict taxonomy, there is a single top-level term under which the entire remainder of the taxonomy lives. So-called *orphan* terms not part of the overall hierarchy are not allowed.

As with authority files, taxonomies may be displayed in two main ways, as an alphabetical listing of terms with term relationships or as a hierarchy. The hierarchy may be presented as an indented list of terms that indicates broader/narrower relationships by degree of indentation, or a diagram (often referred to as a tree diagram).

Figures 1, 2 and 3 show these three kinds of taxonomy displays. (The dots in the indented display help indicate the number of levels of indentation.) Here again, the alphabetical display makes is easier to find individual terms, while the indented and tree diagrams more clearly indicate the hierarchical relationships of the terms.

In terms of implementing a taxonomy, many IT departments favor the indented display prepared in Excel, as it is easy for them to convert it to an appropriate database file for use in a content management system or other software.

This article will not explore the details of creating taxonomies. Suffice it to say that a large, enterprise-wide taxonomy that contains thousands or tens of thousands of terms will take months to create and must be appropriately maintained to ensure continued usefulness.

401(k) plan BT benefits
base salary BT compensation UF salary
benefits BT total compensation NT 401(k) plan; health benefits
compensation BT total compensation NT base salary; deferred payments; variable pay
deferred compensation USE deferred payments
deferred payments BT compensation UF deferred compensation
dental plan BT health benefits
disability insurance BT health benefits
health benefits BT benefits NT dental plan; disability insurance
salary USE base salary
total compensation NT benefits; compensation
variable pay BT compensation

Figure 1 Alphabetical taxonomy display

total compensation
- compensation
  - base salary [salary]
    - deferred payments [deferred compensation]
    - variable pay
  - benefits
    - 401(k) plan
    - health benefits
      - dental plan
      - disability insurance

Figure 2 Indented taxonomy display

Thesaurus

A thesaurus contains all of the features of a hierarchy, plus the associative relationship of RTs, previously discussed. Because of the multiple term relationships they include, thesauri are the most complex CVs to create and maintain.

Just as with the previous types of CV, thesauri may be displayed alphabetically or in indented/tree format. In the latter two displays, some method must be found to show the related terms. In an indented format, this can be done in an additional column: see Figure 4. A tree display might use special typography for indicating related terms.

Obviously, displaying the entire structure of a large taxonomy or thesaurus in a single tree or indented structure is almost impossible. It is therefore customary to show only part of the hierarchy, with links to broader and narrower parts of the taxonomy. The Getty Art and Architecture Thesaurus,[2] which comprises over 131,000 terms, also includes a search function to enable users to locate specific terms. (The largest

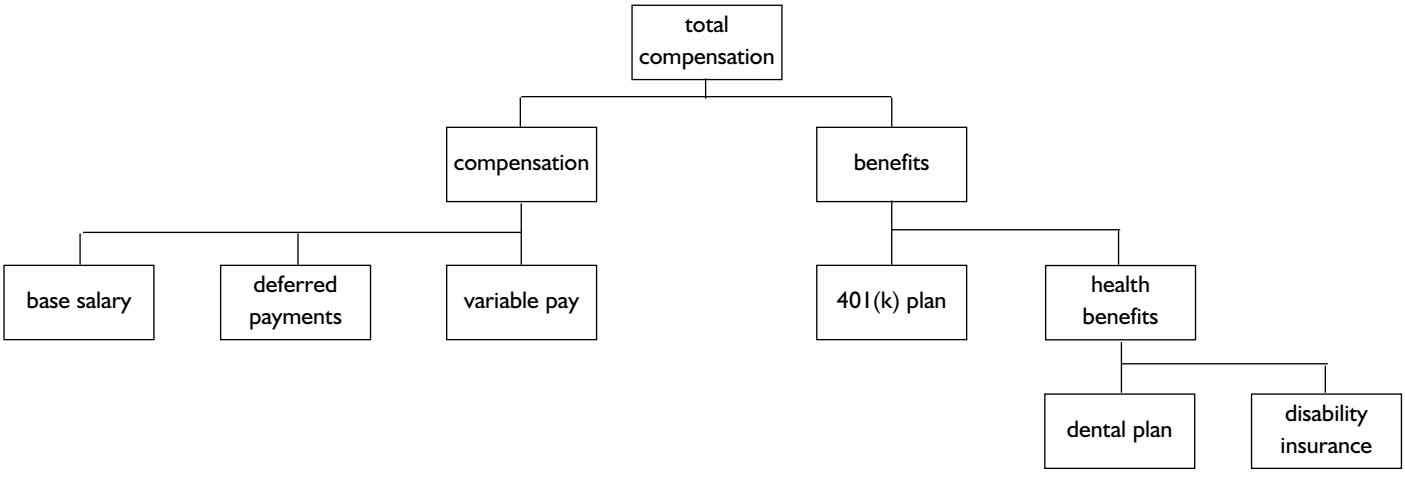

Figure 3 Taxonomy tree display

Vocabulary terms	Related terms
Licenses, Permits & Taxes	
• Fees	Taxes
• Licenses	Permits
• • Business Licenses	
• Permits	Licenses
• • Building Permits	
• • Operating Permits	
• Taxes	Fees
• • Business Taxes	

Figure 4 An indented display for a thesaurus

Getty vocabulary is the Getty Thesaurus of Geographic Names, which includes over 1.1 million entries.)

There are now a number of good books devoted to taxonomy and thesaurus creation in addition to Jean Aitchison's previously mentioned tome. One recent example is *Organising knowledge: taxonomies, knowledge and organisational effectiveness* (2007) by Patrick Lambe. Another helpful book is Darin Stewart's *Building enterprise taxonomies* (2008), which is unfortunately marred by a particularly bad index. In addition, there exist two international standards[3] on thesaurus development which are available from the International Standards Organization, at a cost of approximately $160 or £80 (as of 15 June 2008).

The BSI Group operates BSI British Standards, which had previously issued copies of the two ISO standards; it currently has available a newer standard,[4] available at a cost of £80 (as of 15 June 2008).

Summary of CV types

Figure 5 provides a quick reference summary of the types of CVs and how they are related.

Polyhierarchy

In addition to the four types of CV described above, there is one more type, namely a polyhierarchy, which can modify the formation of either a hierarchy or a thesaurus. In polyhierarchies, a term may 'live' in multiple categories. That is, it may be part of more than one parent/child relationship.

Synonym ring
+ preferred terms
= **Authority file**
+ broader/narrower terms
= **Taxonomy**
+ related terms
= **Thesaurus**

Figure 5 Summary of CV types

Sultanates	**Countries**
Audhali	Albania
…	…
Brunei	*Brunei*
…	…
Oman	China

When creating a polyhierarchy, it is important to confer with whoever will be implementing the CV, to ensure that whatever tool is used can appropriately handle polyhierarchies.

Using controlled vocabularies

Navigation taxonomy

Controlled vocabularies are often used to organize website or other content. In this case, the CV terms become labels for the browsing categories. Different levels in the browsing hierarchy are mirrored by the levels in the CV hierarchy. For example, on the Fortnum & Mason website (www.fortnumandmason.com), the following navigation hierarchy is used (partial list):

Hampers
Food Hall
    Fortnum's Sale
    Fortnum's Tercentenary
    Fruit & Flowers
    Tea
    Chocolate
        Selection Boxes
        Truffles
        Mints
        Chocolate Bars
    Fresh Food
    Confectionary
    Fortnum's Curiosity
    Pantry
    Preserve, Marmalade & Honey
    Condiments
    Bakery
    Coffee
Wine & Spirits
For the Home
Fashion & Beauty
Tea & Coffee

As is obvious from this example, navigation hierarchies, like CVs in general, must be created to reflect a particular context.

Search enhancement

CVs are often used in conjunction with website or content management system search functionality. Appropriate selections from a hierarchy or thesaurus, for example, may be shown along with search results, allowing users to expand or reduce the scope of the search by choosing broader or narrower terms.

CVs can also be helpfully used to categorize search results. On the 'Marks & Sparks' website (that's Marks & Spencer to you non-Brits: www.marksandspencer.com), a

search for 'trousers' yielded 779 results. However, rather than just showing all 779 results and letting the user plow through them, the website displays the results in a categorized fashion, showing how many results are in each subcategory.

Womenswear (386)
Menswear (146)
Boyswear (28)
Girlswear (27)
…
Gift Shop (1)

Users can click on the appropriate subcategory to narrow their search. (And for those interested, the item in the Gift Shop area is a wooden trouser hanger.)

The inclusion of synonym rings in the search engine is another important way to improve search results. In this case, when one of the words in a ring is searched for, the search engine returns items containing any of the words in the ring. So a search for 'cats' might also return documents that contain the word 'feline.' This helps improve recall, ensuring that more of the relevant results are retrieved.

Facets

The idea of facets was first expounded by the librarian S. J. Ranganathan in the early 1930s. He was trying to solve the basic problem of classification, namely, how can one find the *single best category* that will describe an object? This is known as top-down classification. Does a book on the history of French literature belong under France, history, or literature? (Obviously there are cataloging rules to guide that decision, but let's ignore those for the moment.)

Ranganathan attempted to create bottom-up classification, identifying a list of universal basic categories by which anything in the world could be described. He developed the famous list know as PMEST: personality, matter, energy, space, time. These can be described as:

Personality	What is it?
Matter	What is it made of?
Energy	What action is it performing?
Space	Where is it?
Time	When is it?

Not all of these basic facets are necessarily used when describing a particular object.

Ranganathan also developed what became known as the Colon Classification, which used a series of introductory punctuation marks to indicate specific facets:

, = Personality
; = Matter
: = Energy
. = Space
' = Time

The Colon Classification also included an alphanumeric notation scheme for each of the facets. So, for example, an article on 'research in the cure of tuberculosis of lungs by x-ray conducted in India in 1950' would be assigned the Colon Classification number of:

L,45;421:6;253:f.44'N5

which translates to:

Medicine,Lungs;Tuberculosis:Treatment;X-ray:Research.India'1950

or the series

Personality-Personality-Matter-Energy-Matter-Energy-Space-Time

Obviously the use of Ranganathan's facets allowed extremely precise categorization of almost any topic imaginable. At the same time, however, it produced an extremely complex notation. As a result, the Colon Classification itself never became widely used.

To generalize, we can say that facets represent fundamental categories by which an object or concept may be described. The facets used in any particular instance depend on the specific context. A toy ball, for example, might be described using the following facets: size, weight, shape, color, texture, material. But other facets could also be used: owner, manufacturer, location, age. The choice depends on the particular needs of the creators and users of the information.

Although the Colon Classification generally languished, the use of facets emerged early in the 21st century as an important means of describing and providing access to large collections of documents. Facets can be used to provide browsing hierarchies (essentially canned searches) which allow users to follow the path best matching the way they think. There are multiple paths to the same information.

For example, Epicurious (www.epicurious.com) in its recipe browsing section uses the following top-level facets: main ingredient, cuisine, preparation method, season/occasion and meal/course, along with type of dish and dietary consideration. Users can click on any category they wish, depending on what is important to them and can then narrow their search results by using values from the other available facets. A recipe for onion soup might be found, for example, by following any of these paths:

- French cuisine > Onion > Type of Dish > Soup/Stew
- Soup/Stew > Onion > Cuisine > French
- Soup/Stew > Cuisine > French > Onion.

As another example of facet use, in its Laptop Search section, the *PC World* website (www.pcworld.co.uk) allows you to refine your search using any of the following facets: brand, processor, clock speed, RAM memory, hard disk capacity, wireless enabled, minimum value/maximum value. The site also allows you to specify how you would like the results to be sorted, by price–lowest first or price–highest first.

We use facets all the time in daily life, but most likely never think of them as facets. When we decide to see a movie, we might be interested in the location of the theater,

the genre, the director, who the leading actors are, the length of the movie, what the show times are – all facets that describe the movie in particular ways.

Louise Spiteri has developed a model (1998) for facet analysis that can help you create facets for your particular needs.

CV maintenance

Once a controlled vocabulary has been created, it cannot just be shelved and forgotten. It becomes a living thing which must be tended and cared for. Any of the following changes might be required to keep that CV up to date.

Add/delete associative relationships
Revise term
Add/delete term
Reorganize hierarchy (change hierarchical relationships)
Modify facet label (major hierarchy section label)
Add/delete facet or major hierarchy section

The frequency of such changes generally occurs in relationship to the order of the list above. In any vocabulary the terms and term relationship will change much more quickly than will the facets or major sections of the hierarchy.

The impact of the changes is in reverse relationship to the order of the list above. Changing a single term or term relationship has a significantly smaller impact on the overall CV than changing or reordering facets or top-level categories.

There are a number of useful software packages for creating and maintaining controlled vocabularies. Heather Hedden has recently written an analysis (2008b) of three such programs.

Change control process

In any corporate or business setting, it will be necessary to have a process in place that governs CV maintenance. For example, who submits proposed changes? Who receives the proposed changes? Are provisional terms allowed as part of the content tagging process? Who decides on changes to be made? Is subject matter expertise required? How often are changes made? Who implements the changes?

CV reviews

In addition to accepting changes from content authors or CV users, the person or persons responsible for the taxonomy should establish a policy on regular CV reviews, including how often reviews occur and who is responsible for carrying them out. There might be particular triggers for a review in addition to regularly scheduled reviews. Such instances might be the introduction of a new product line or a corporate merger, or even a corporate reorganization.

CV ownership

An important part of CV maintenance is establishing who 'owns' the CV. That is, who is responsible for its care and feeding? Is it the IT department? Is it the marketing/ communications department? Who controls the budget for CV maintenance? What personnel are responsible? Answers to those questions will be different depending on the circumstances. But every organization that depends on a controlled vocabulary must have answers, or their CVs will soon become out of date and will impede the efficiency of information finding, costing the organization time and money.

Conclusion

Although there are many new concepts and ideas to learn about in the world of CVs, especially in an enterprise setting, many indexers find the work of creating CVs to be both a natural extension of their indexing skills and an important new income stream.

Notes

1 Examples include: ANSI/NISO, Z39.19-2003: Guidelines for the Construction, Format, and Management of Monolingual Thesauri, pg. 35; CMSWiki, CmsGlossary, available at http://www.cmswiki.com/tiki-index.php?page= CmsGlossary; Leise, Fred, Karl Fast and Mike Steckel, 'Controlled vocabularies: a glosso-thesaurus' available at http://www.boxesandarrows.com/view/controlled_vocabularies_a_glosso_thesaurus; and Taxo-Tips, A Taxonomy Glossary, available at http://www.taxotips.com/resources/glossary/.

2 Available at http://www.getty.edu/research/conducting_research/vocabularies/aat/. The Art & Architecture Thesaurus is an effort of the Getty Research Institute.

3 ISO 2788:1986, *Guidelines for the Establishment and Development of Monolingual Thesauri* and ISO 5964:1985, *Guidelines for the Establishment and Development of Multilingual Thesauri.* These standards are available in both English and French versions from www.iso.org.

4 BS 8723-2:2005, *Structured Vocabularies for Information Retrieval. Guide. Thesauri*, Nov. 2005. Available at: http://www.bsi-global.com/en/Shop/Publication-Detail/?pid=000000000030094114

References

Aitchison, Jean. 2000. *Thesaurus construction and use: a practical manual*, 4th edn. Chicago: Fitzroy-Dearborn.

Hedden, Heather. 2008a. Controlled vocabularies, thesauri, and taxonomies. *Indexer* **26**(1), 33–4.

Hedden, Heather. 2008b. Comparative evaluation of thesaurus creation software. *Indexer* **26**(2), 50–9.

Lambe, Patrick. 2007. *Organising knowledge: taxonomies, knowledge and organisational effectiveness.* Oxford: Chandos.

Spiteri, Louise. 1998. A simplified model for facet analysis. *Canadian Journal of Information and Library Science*, **23**, 1–30 (April-July). Available at: http://iainstitute.org/pg/a_simplified_model_for_facet_analysis.php

Stewart, Darin. 2008. *Building enterprise taxonomies.* n.p. Mokita Press.

Wellisch, Hans. 1991. *Indexing from A to Z*. New York: H. W. Wilson.

Fred Leise, current president of the ASI, has been a freelance indexer since 1995, specializing in scholarly works including East Asian history and civilization, international relations, and politics. He co-authored Indexing for editors and authors: a practical guide to understanding indexes. *Email:* fredleise@contextualanalysis.com