Quality Metrics for Taxonomies

A CONVERA Technical White Paper

PREPARED BY

CLAUDE VOGEL
Chief Scientist - CONVERA

BASED ON HIS ORIGINAL BOOK *Cognitive Engineering*

CONVERA™

TABLE OF CONTENTS

Introduction

With ever-growing amounts of information available in an enterprise today, simple keyword search is no longer sufficient to provide guidance to the information people need. As a result, many intranet and extranet portals as well as web sites provide online directories or classification of information that users can browse.

There are several approaches to building these taxonomies, ranging from manual, labor-intensive methods to topic search strategies and neural or conceptual networks. Each method takes a very different approach to the problem.

The quality of the resulting taxonomy is the central issue, particularly when comparing automated methods for creating taxonomies. Several factors work against quality results – these include frequent updates to content sources, multiple user needs, and the continually growing mass of information that must be considered as source materials.

This paper does not propose a comprehensive quality assurance plan, but addresses instead the characteristics most unique to taxonomies. It proposes a set of attributes and characteristics contributing to taxonomy quality, specifically along the dimensions of nomenclature and terminology.

Definitions

Although directory, ontology, thesaurus and taxonomy are terms that are used fairly interchangeably, both by vendors of solutions and by users themselves, they are in fact slightly different things, and the terms deserve further definition.

Taxonomy A taxonomy is a hierarchical system describing genera and species. Species derive from a common genus and are hierarchically represented according to their essential characteristics and differences. For example, animals are categorized with the "Taxonomy of Life" which separates mammals from birds and spiders from insects, based on proper features and relative differences. This genus to species nomenclature is highlighted by terminology which moves from generic terms to binomial terms through lexical derivation and compounding.

Thesaurus

A thesaurus is a list of related word groups organized by a combination of taxonomic, ontological and dictionary attributes. A thesaurus represents taxonomic, ontological, and other kinds of relationships. Categories can come from either a taxonomy or an ontology. [2]

Directory

A directory lists associations between pieces of information; the purpose is to let the user access one piece of information using the other. Directories are fairly flat classifications that allow a user to look at lists of related objects. A directory might be an alphabetical list of names and addresses, or a list of products sold. These classifications can have taxonomy-like or ontology-like relationships, but are neither as deep nor as consistent as those classification systems.

Ontology

An ontology is a hierarchical system of classification representing a view of the world. An ontology reflects the commonly used and trusted breakdown of elements of classification. For example, the breakdown of news items into categories of ‘World’, ‘Sports’, ‘Politics’, etc. is ontological.

THE SOFTWARE QUALITY MODEL

In assessing the quality of a taxonomy, we need to have a model to which to refer. In this case, the software process management model works well. [3, 5, 6, 7] By applying the process and practices of software quality assurance to taxonomies and taxonomy creation, we can follow a well-defined path and create a specific quality assurance plan for taxonomies and their ongoing updates and maintenance.

In software process improvement, the first step is to understand the characteristics of the process and the factors that affect quality. The same is true for taxonomies – we need a way to understand and measure the performance of a taxonomy before we can undertake to assure or improve that quality.

This paper proposes a number of measurable attributes, or metrics, specific to taxonomies. In selecting these attributes, we have chosen attributes that can be identified and measured easily, that vary between different taxonomies, and that provide a significant diagnostic value.

The process improvement process itself is to first analyze these metrics, to diagnose based on these findings, and to take action based on the diagnosis. This paper provides a number of quality indicators or metrics; it explains what metrics are useful and what patterns (or indicators) you may find in your taxonomy. It also defines what these metrics mean (diagnostic value), and what actions or steps you might take to improve quality based on these patterns.

To formalize and continue this process across all attributes is to create a comprehensive quality assurance plan for the taxonomy.

QUALITY INDICATORS

There are several dimensions along which to measure the quality of a taxonomy. These include the following:

- Corpus – the source materials for the taxonomy
- Coverage – how well the taxonomy covers the source materials
- Nomenclature - the arrangement of the taxonomy's classes
- Terminology – the terms used to name the various classes of the taxonomy
- Dependency – cross references within the taxonomy
- Maturity – the stability of the taxonomy over time
- Performance – performance compared with other taxonomy solutions

Some of these dimensions, such as performance, can be measured with traditional software quality assurance tools. Discussing each of these sets of indicators in detail is beyond the scope of this paper. Instead, we will focus in more depth on indicators that are truly unique to taxonomies: nomenclature and terminology.

Before reaching this discussion, we will touch briefly on the first two issues of corpus quality and coverage.

Corpus Quality

The first and perhaps most direct factor contributing to the quality of the taxonomy is the quality of the corpus of source material. If you feed a taxonomy system a large number of irrelevant email messages, the results you get may be fun but will certainly not be useful.

Aside from the obvious factor of the quality of the content itself, several factors can affect the quality of the resulting taxonomy, including:

The multiplicity of sources	How many different sources are involved? Multiplicity of sources can lead to better (broader) results, depending on the other factors.
Size of sources	Are the sources all of similar sizes, or are large documents weighted equally with very small fragments or documents? In considering size, consistency is often important –source documents of similar size will deliver more consistent results than a few large documents mixed with thousands of small sources.

Homogeneity of sources A homogeneous source collection should yield high quality results within the area of coverage, with deep and consistent results. Heterogeneous sources deliver better results over a broader concept base.

Different genres of source information deliver different characteristics, in terms of the resulting taxonomy. The following table lists some of these characteristics, by genre.

Features	Internet News E-Mail	Reports Patents	E-Trade Logs
Informative content	-	+	+
Number of topics covered	+	+	-
Structured information	-	+	+
Size of records	-	+	-
Number of records	+	-	+

Figure 1: Characteristics of source genres

Coverage Quality

Coverage refers to how well the taxonomy covers or includes the concepts that it should contain. For taxonomies created through automated methodologies, coverage is a significant issue – if users expect to find something within a category but do not, the entire taxonomy is of suspect usefulness.

Two different factors contribute to the concept of coverage

How well the source documents cover the topic area.

How well the taxonomy includes the key concepts in the source documents.

To determine the document coverage, you will need to create a list of the key concepts from the source documents. (This requires either reading and extracting the key concepts or using a lexical analysis tool that can do so.) You then need to compare this concept list to a thesaurus developed for the field of study. The closer the matches, the better the document coverage.

To determine the taxonomy coverage, compare a list of taxonomy entries to the "key concepts" list generated above.

The rest of this paper focuses on the quality attributes for Nomenclature and Terminology.

NOMENCLATURE

A taxonomy is defined by its nomenclature and its terminology. [4] The taxonomy's nomenclature is the arrangement of its classes. Typically, the various levels of a taxonomy are referred to using the following terms:

unique beginner
life forms or supra-generic classes
generic classes
specific classes
varietal classes

The classes within each level of the taxonomy may be distinguished on the basis of one or more properties. In the lower levels of the nomenclature, very few properties differentiate between the series. In contrast, at the upper level of the taxonomy, classes are identified by considering the identity of the class before its relationships with other classes of the same level.

For example, the unique beginner specifies the highest topics of the taxonomy, such as plants and animals. Moving lower, the first level is life forms, which might include mammals, birds, trees, etc.

"Generic" classes follow – such as maple and oak under trees. Specific and varietal classes (sugar maple, Norway maple) complete the taxonomy, and always descend from a more generic class. In some taxonomies (with fewer levels), the generic classes may be at the same level as the life forms.

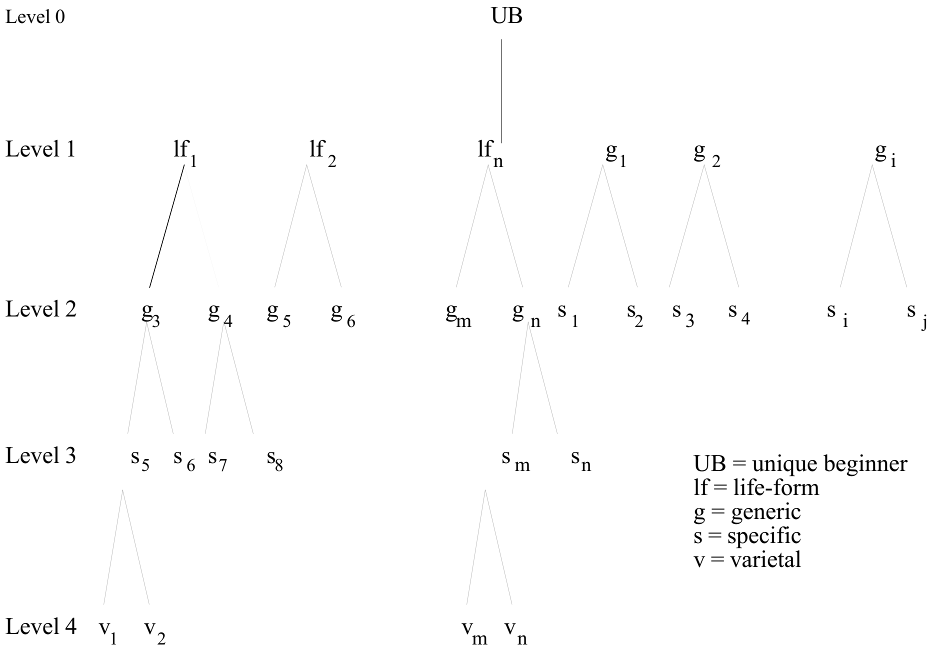


Figure 2: Schematic relationship of the five taxonomic categories [4]

Valid taxonomies have the following attributes:

- There should be at most five levels of nomenclature. A number greater than five may indicate a proliferation of life forms or a binary degeneration of the tree. These cases are described in the terminology section.
- The first level is life forms. If these are named by secondary terms, the different entries do not constitute a series. For example, "nocturnal animals" and "diurnal animals" are in a binary relationship, and as such should not exist at the life form level.
- The second level of the taxonomy is generic classes, corresponding to the physical world.
- The third level is specific classes, which constitute a lexical series. Each secondary level must be homogeneous.
- The last level is of varietal classes.
- Each level is homogeneous.

In addition, there are several metrics for assessing nomenclature; these measure the *depth, width,* and *balance* of the taxonomy. The following sections describe metrics for these indicators. For simplicity, we will sometimes refer to a taxonomy as a tree, and a node and its children as a subtree in the discussions that follow.

Depth

One measure is the depth of the taxonomy. There are several potential metrics for assessing depth. Depth indicators provide a perception of taxonomy depth, either as a single value or as a histogram visualizing the statistical depth distribution.

Depth(y, x) is the *relative depth* of a some node y with respect to a node x. In particular, if y is an descendant of x, then depth(y, x) is the number of links between x and y.

A mean depth indicator is the average depth of a subtree hanging from a given node. The mean depth equals the depth if all branches hanging from the node have exactly the same length. Otherwise, the average depth is less than the greatest depth.

You can also calculate a *histogram of depth* (HDEPTH) in a subtree hanging from node x. This indicator contains a list of depth intervals. For each interval, the HDEPTH indicator displays the number of leaves of node x the depth of which falls into the interval. A histogram is useful in providing a perception of the irregularity in subtree depth.

Width

Another measure is the number of children for each node, which gives a perspective of the tree's width. A numerical measure is the average number of children inside the subtree hanging from a given node (not taking into account the "leaves" or end points, which have no children and would skew the average.)

Balance

Balance (or imbalance) indicators may be applied to branch nodes (not leaves). They indicate to the user whether a given node is balanced (in which case the indicators will be close to 0) or unbalanced (maximum unbalance value normalized to 1).

The definition of balance is quite simple for a node with only two children; you simply compare the weights of both hanging subtrees. At a node with three children and more, the unbalance indicator expresses the deviation from a uniform weight repartition between children.

TERMINOLOGY

The terminology of the taxonomy is defined by the terms used to name the various nodes of the classification.

In discussing terminology, we distinguish between *primary terms* and *secondary terms*. Secondary terms may be broken down (such as *fetal monitor)* while primary terms are irreducible (such as *monitor*). Secondary terms generally designate varietal classes.

It is difficult to identify classes by descending from the unique beginner to specific classes. Instead, it is simpler to begin by identifying generic classes as follows:

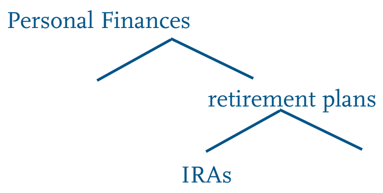
1. Find levels (or classes) that are named by primary terms and followed by secondary terms using the first term as a lexical base. For example:

   insurance ← generic class
       life insurance ← specific class
       health insurance ← specific class

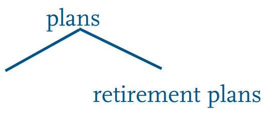
2. Consider the class above the generic class as a potential life form.
3. Consider the class below the specific class as a varietal class.

Life forms are derived from the unique beginner and are generally (but not always) marked by primary terms. (In technical content, a life form may frequently be a secondary term.) Life forms always include generic classes – they are not terminal in the taxonomy.

Life forms are difficult to identify, because there are no formal characteristics for their identification. The life form terms do not have an explicit lexical relationship with the terms used to designate subordinate (generic) classes.



If a life form is a secondary term, we must be careful not to unite it with lexical bases/classes that are no longer useful in the same context:



The life form level of a taxonomy must typically be defined by a human expert.

Varietal classes are derived from specific classes, and are marked by secondary terms whose lexical base is the designated generic class term. Varietal classes are always terminal.

Quality indicators for terminology include *objectivation, genericity, speciation, necessity, and consistency.*

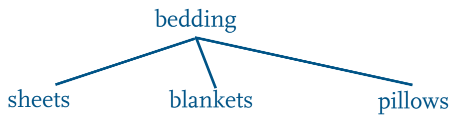
Objectivation

Before examining the characteristics of the properties, we must of course make sure that the object identified is indeed an object. The lexical derivation of the name used to designate the particular object must confirm the physical and static nature of this object. We want to avoid strictly verbal forms in the terminology. In general:

-avoid "tion" derivations: indicating actions in progress
-avoid "ment" derivations: indicating completed actions

Genericity

Other forms of classification are generally collected simultaneously with the taxonomy of the field of knowledge. The taxonomy is constructed on the basis of a relationship of a subordinate species to a superordinate class. Another form of competing relationship is the part to whole relationship. This second form often merges with the first, as follows:

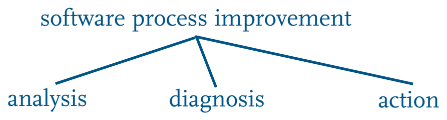


In the example above, sheets and bedding share a *part-to-whole* relationship, not a species-to-class relationship. This compositional relationship may be displayed as a tree of components – it is, however, quite distinct from a taxonomic relationship. (See Rosch’s prototype definition, e.g., [9].)

When considering only a component/compound pair, it can be difficult to identify a compositional relationship. Although this relationship may be represented by a tree, it is very different from a taxonomic relationship.

Although the sheets or blankets could be considered bedding, clearly all three components are necessary to reconstruct all of “bedding.” In a compositional relationship, all descendants maintain a relationship with the class considered to be generic.

Other semantic families complicate taxonomy structure. For example, in a processing relationship, a general action is comprised of a series of actions. This replaces the taxonomy inclusion relationship by changing the basis for classification. For example, consider the software improvement process discussed at the beginning of this paper:



In the example on the previous page, software process improvement involves three successive operations: attribute analysis, diagnosis, and corrective action. To identify compositional relationships:

- Consider the relationship between all of the action's components with the superordinate action. All three steps are required to carry out software process improvement.

- Consider the relationship between the various components of the superordinate action – to carry out software process improvement, you must first analyze, then diagnose, then take action.

Speciation

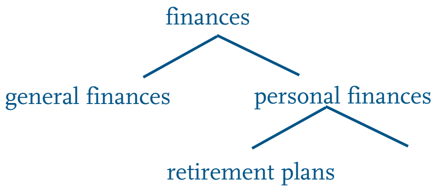
Speciation refers to a proliferation of life forms in the taxonomy. For example, we find the following hierarchy in WordNet:

Person
Unwelcome person
Unpleasant person
Selfish person
Opportunist
Backscratcher

The problem with this is that it is not obvious that an unwelcome person, unpleasant person, and selfish person are very different. There is no progress in the speciation. The speciation should represent a progressive and continuous progress.

Necessity

A content expert, on creating a taxonomy, stresses the properties that contribute most to his overall view of the field of expertise. For example,



The expert's distinctions sometimes do not solve the problem of integrating life forms for the actual taxonomy. In the example above, the objects at the "retirement plans" level are considered life forms, but the "general finances" and "personal finances" terms qualify objects named elsewhere and are not independent of their context.

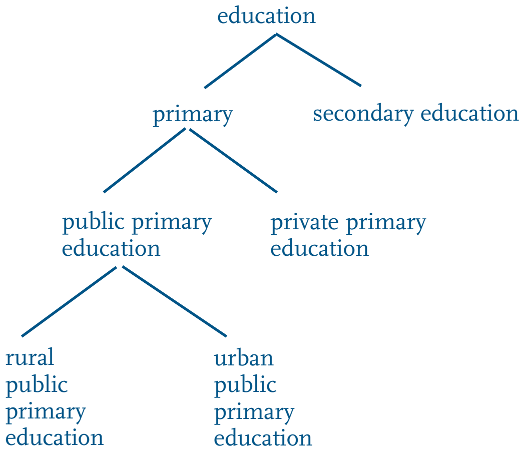
Constructing a classification based on a single discrimination such as this should be avoided:

- This kind of discrimination (frequently binary) is better suited to the lower levels of the taxonomy than the life form level. At the life form level, the potential lack of consensus in the discrimination makes update difficult.

- Discriminating based on a single property often leads to the duplication of classes within the taxonomy; the distinction will not be relevant to some classes.

A critical analysis of the life forms leads to an examination of the form and meaning of the categories between the Unique Beginner and Generic.

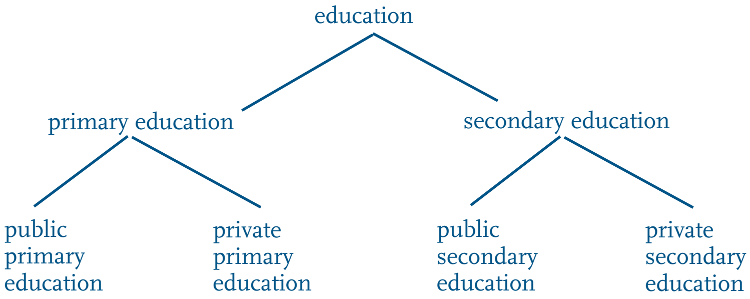
One classic problem in the specific and varietal classes, particularly of taxonomies created by human experts, is the issue of the *binary degeneration* of the tree. The terms used to make distinctions are binary, then tertiary, adding up the dimensions used to classify at each level:


education
primary
secondary education
public primary education
private primary education
rural public primary education
urban public primary education

Other experts may well pose different structures; identifying these binary forms and comparing different approaches to the terminology helps you reduce possible combinations of properties on relevant ascendants.

repetition of the same contrasts in different areas of the taxonomy

Another common problem is the duplication of trees – binary distinctions are repeated in different regions of the classification:


education
primary education
secondary education
public primary education
private primary education
public secondary education
private secondary education

The duplication of the subtrees indicates that this tree as a whole represents a non-hierarchical relationship with logical intersections:

	primary education	**secondary education**
public education	public primary education	public secondary education
private education	private primary education	private secondary education

This is a *paradigm tree* instead of a true taxonomy tree. [8]

Taxonomies built with this kind of paradigm tree lack the quality of *necessity* – you can as easily change the order of node properties, and have the highest level division as public or private education, further subdivided into primary and secondary education. Contrast this with a traditional taxonomy in biology – whale must necessarily be subsidiary to *mammal*.

The problem with taxonomies that lack necessity is that they are inherently unstable. Because it is based on a relatively arbitrary distinction, another expert could rework the taxonomy, and it may be difficult to maintain these distinctions over time and through updates.

Consistency

Whether your relationships indicate necessity or not, it is important to maintain consistency of hierarchical relationships throughout the taxonomy.

There may well be situations in which you need to use relationships other than traditional taxonomic relationships in your taxonomy, such as the parts to whole or processing relationships described above. If this is so, then all trees in the taxonomy should be constructed using the same type of relationship. Then at least the user will have a consistent experience of how to work with and find information in the taxonomy. Mixing different kinds of relationships in a taxonomy leads to an inconsistent classification and therefore a less useful, lower quality tool.

CONCLUSION

We find in fact that the software quality assurance model works quite well when assessing overall taxonomy quality. Although this paper has addressed primarily the terminology and nomenclature aspects of taxonomies, which is unique to taxonomical discussions, the other dimensions of taxonomy quality are equally assessable using this model.

At Convera, we have created a complete Taxonomy Development Certification Program that includes taxonomy development and benchmarking applications, training, and support to ensure your success in building and maintaining high-quality taxonomies. To learn more about this program, please visit www.convera.com.

REFERENCES

1 Vogel, Claude. *Cognitive Engineering*, Masson, Paris, France. 1988

2 ANSI/NISO Z39.19-1993, *Guidelines for the Construction, Format, and Management of Monolingual Thesauri,* NISO Press, Bethesda, MD. 1993

3 Baumert, John H. and McWhinney, Mark S. *Software Measures and the Capability Maturity Model,* Software Engineering Institute, Carnegie Mellon University, Pittsburgh, PA. 1992

4 Berlin, Brent. *Ethnobiological Classification,* Princeton University press, Princeton NJ. 1992

5 Burr, Adrian and Owen, Mal. *Statistical Methods for Software Quality,* International Thomson Computer Press, London, 1996

6 Florac, William A. and Carleton, Anita D. *Measuring the Software Process,* Addison-Wesley, Reading MA. 1999

7 Kan, Stephen H. *Metrics and Models in Software Quality Engineering,* Addison Wesley, Reading, MA. 1995

8 Tyler, Stephen A. *Cognitive Anthropology,* Holt Rinhart and Winston, New York NY. 1969

9 Varela, Francisco J, Thomspon, Evan and Rosch, Eleanor. T*he Embodied Mind: Cognitive Science and Human Experience,* The MIT Press, Cambridge, MA. 1999

ABOUT CONVERA

● ● ●

Convera is a leading provider of mission-critical enterprise search, retrieval and categorization solutions. Convera's RetrievalWare solutions maximize return on investment in vast stores of unstructured information by providing highly scalable, fast, accurate and secure search across more than 200 forms of text, video, image and audio information, in more than 45 languages. More than 750 customers in over 29 countries rely on Convera's search solutions to power a broad range of mission critical applications including enterprise portals, knowledge management, intelligence gathering, profiling, corporate policy compliance, regulatory compliance, customer service and more.

CONVERA™
Mission-Critical Enterprise
Search, Retrieval & Categorization Solutions

US	T 800 788 7758	**www.convera.com**	
	T 703 760 4085	info@convera.com	
	F 703 748 1255		
UK	T + 44 1344 781 800	info@convera.co.uk	
	F + 44 1344 781 801		

GLOBAL OFFICES
Carlsbad, CA
Columbia, MD
London, UK
Montreal, QC
Munich, GER
Paris, FRA
San Jose, CA
Vienna, VA

● ● ● VISIT OUR WEBSITE FOR WORLDWIDE OFFICE INFORMATION

Mission-Critical Enterprise
Search, Retrieval & Categorization Solutions

US	T	800 788 7758	**www.convera.com**
	T	703 760 4085	info@convera.com
	F	703 748 1255	
UK	T	+ 44 1344 781 800	info@convera.co.uk
	F	+ 44 1344 781 801	

GLOBAL OFFICES
Carlsbad, CA
Columbia, MD
London, UK
Montreal, QC
Munich, GER
Paris, FRA
San Jose, CA
Vienna, VA

● ● ● VISIT OUR WEBSITE FOR WORLDWIDE OFFICE INFORMATION