The more I talk with other indexers, the more I hear of the frustration that they face with how to create a well-shaped taxonomy. One of the most prevalent causes of this frustration has many names—associated terms, use for, reference terms, facets, variants—whatever name you choose, the different facets of taxonomy terms are often difficult to build. Using linguistic mechanisms, I will outline how to identify facets for creating taxonomies linguistically and engineer a linguistic analysis tool that you can use to gauge whether you have fully harnessed your term facets.

Facet and Concept Creation

Lexicology of Terms

Linguistics is defined as the study of language and language structure through many applications, such as morphology, syntax, phonology, semantics, etymology, and lexicology.1 All of these sub-disciplines of linguistics can be used to analyze and provide improvement to your taxonomy. I will be focusing on the morphological aspects of lexicology—the study of the forms, meaning, and use of words. There are differences in the analysis approach; morphology is based on terms and their arrangement while lexicology is based on terms and their

Facet Example 1: Facets within a taxonomy hierarchy

<table>
<thead>
<tr>
<th>Term: Engineering</th>
</tr>
</thead>
<tbody>
<tr>
<td>Material (facet of engineering)</td>
</tr>
<tr>
<td>Research (facet of engineering)</td>
</tr>
<tr>
<td>Engineering Devices (facet of research)</td>
</tr>
<tr>
<td>Engineering Industry (facet of research)</td>
</tr>
<tr>
<td>Automotive (facet of Eng. Industry)</td>
</tr>
<tr>
<td>Aerospace (facet of Eng. Industry)</td>
</tr>
<tr>
<td>Civilian (facet of Aerospace)</td>
</tr>
<tr>
<td>Military (facet of Aerospace)</td>
</tr>
<tr>
<td>Commercial (facet of Aerospace)</td>
</tr>
</tbody>
</table>

Facet and subfacets of Engineering.

Facet Example 2: Facets within a scope note

<table>
<thead>
<tr>
<th>Term: Liability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scope Note: The state of being responsible for something especially by law</td>
</tr>
<tr>
<td>Facets: Use for - product liability; legal liability; financial responsibility; liability; liable; liableeness; responsabilidad (Spanish); Verpflichtung (German)</td>
</tr>
<tr>
<td>Misspellings: liable; liable; lieable; lyable</td>
</tr>
</tbody>
</table>

Concepts: run the chance; expose oneself; on the cards; stand a chance

Many consider hierarchical and faceted taxonomies separate entities; however, if combined they become a very powerful indexing tool. By creating a hierarchical taxonomy and using frequent facets as narrow terms, or building complex facets into scope notes, taxonomies will not only create a coherent organization of parent terms but will apply a multifaceted support structure on which parent terms may rely. Because a taxonomy is by the subject matter it incorporates, deciding on the scope and breadth of your taxonomy and what parent terms to include cannot be quantified here. However, once you decide these key components, linguistic analysis may be introduced.

Facet and Concept Creation

Lexicology of Terms

Linguistics is defined as the study of language and language structure through many applications, such as morphology, syntax, phonology, semantics, etymology, and lexicology.1 All of these sub-disciplines of linguistics can be used to analyze and provide improvement to your taxonomy. I will be focusing on the morphological aspects of lexicology—the study of the forms, meaning, and use of words.

There are differences in the analysis approach; morphology is based on terms and their arrangement while lexicology is based on terms and their

processes. The foremost linguistic analysis method that can be utilized for facet selection is a cross between the two methodologies. First, to understand the creation of facets you must analyze the morphemes, or the subset of lexemes, meaning the single word analyzed in parts. The second methodology is to create linguistic concepts.

Most often when linguistics is mentioned in conjunction with taxonomy, it is with creating a multilingual vocabulary in mind. While creating a multilingual taxonomy is admirable, and can be achieved with these same linguistic principles, the base language of your taxonomy must be dealt with first. Analysis can be conducted on existing taxonomies or used in the creation of taxonomies. Either scenario will result in empirical evidence for demonstrating the potential of your taxonomy. This results in introducing depth to your taxonomy and expanding users’ understanding of applied indexing terms.

Common Lexeme Analysis of Terms

To understand the term facets on a part basis, there are three common lexemes to identify: roots, meaning the main parent term; derivational, meaning the formation of new words from the root word; and desinence of a term, meaning to add a suffix. This is a form of conjugation and can also be used when developing multilingual taxonomies.

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Root</th>
<th>Derivation affix</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adjective/verb</td>
<td>Concept</td>
<td>-ize</td>
<td>Conceptualize</td>
</tr>
<tr>
<td>Adjective/adjective</td>
<td>Acquisitive</td>
<td>-ness</td>
<td>Acquisitiveness</td>
</tr>
<tr>
<td>Adjective/adverb</td>
<td>Absent</td>
<td>-ly</td>
<td>Absently</td>
</tr>
<tr>
<td>Adjective/noun</td>
<td>Happy</td>
<td>-ness</td>
<td>Happiness</td>
</tr>
<tr>
<td>Noun/adjective</td>
<td>Danger</td>
<td>-ous</td>
<td>Dangerous</td>
</tr>
<tr>
<td>Noun/verb</td>
<td>Electric</td>
<td>-ify</td>
<td>Electrify</td>
</tr>
<tr>
<td>Verb/adjective</td>
<td>Sing</td>
<td>-able</td>
<td>Singable</td>
</tr>
<tr>
<td>Verb/noun</td>
<td>Run</td>
<td>-er</td>
<td>Runner</td>
</tr>
</tbody>
</table>

*The derivation affixes represented here are not a complete list and taxonomy creators are encouraged to create their own lists of suggested derivation affixes.

Figure 3. Example of Derivation Devices and Lexicological Breakdown of Terms

When creating a mapping of derivation devices and lexical breakdown of terms there are also combinations that can apply affixes to other affixes, such as the example act-ive-ate-ion. Common derivation affixes must follow grammatical rules and thereby are limited in how many facets they can produce. This type of general faceting may also be used to support the progression from a taxonomical-controlled vocabulary to a thesaurus-controlled vocabulary. By identifying these lexemes you will be able to categorize the most common facets of the terms you are analyzing. However, there are more compound applications that can be utilized for facet analysis.

Complex Lexemes Analysis of Terms

Language is made of subtle social differential disparities that create different, more complicated, lexeme facets. Again, for individual terms, facet lexicalization devices have been shaped over time by both demographics and societal axioms. Taking English as my example, the language is spoken officially in roughly 75 countries, all of which have their own vocabulary to the language even though it is the same root language.1 These facets of the language can be considered fractures to the root trunk of the language. For instance, in the United States version of English the front section of your car is called the hood while in French it means entrance and, therefore, would be more appropriate to use for an appetizer. This is called “borrowing” as it means to borrow from another language.

The most irksome example of a complex lexeme in the English language is the word like which, as Alexandra D’Arcy's linguistic work on the word like states, is “one of the most ubiquitous and multifunctional” lexemes in the English language.2 The word like can be considered the personification of a morphology example and can be categorized into many lexemes. The word like also demonstrates how complicated facet creation and analysis can become without linguistic tools.

These examples cover the different forms and definitions that a word can take on, but there are other factors to take into account as well. Some examples would be common misspellings (amateur versus amateur), common alternate spellings (defence or defense), slang (dawg), compounded words (ain’t), abbreviations (LOL), acronyms (NATO), and many others. The root term should always be classified as the preferred term if you are using a thesaurus, and the parent term if using a taxonomy. Facets should be categorized as use for if a thesaurus is being used and a child term if using a taxonomy. It is not suggested that lexemes be used as parent/preferred terms.

Concept Creation

The second application of linguistics to taxonomy creation and validation is through concepts. When analyzing a taxonomy through concepts many avenues for future expansion can be created. While all taxonomies should have a process for adding new terms, it is a very different matter when adding new concepts.

Concepts serve as links between terms, adding both delineation value and customization to a hierarchy. You must decide what concepts your subject matter and scope will allow because certain terms can be considered concepts; for example, electric vehicles can be a term or it can be a concept of all the parts and engineering that amalgamate to form an electric vehicle. When looking at a hierarchical taxonomy, you can move through the terms in a strict linear fashion while, if concepts are created, you can move through a taxonomy like a ripple where terms pick up similar terms to form connections between subject headings and the hierarchy contained in each. By understanding the links that can be created among terms you will be integrating expansion points for the evolution of your taxonomy. There are six types of controlled vocabularies: flat lists, synonym rings, taxonomy, thesaurus, ontology, and semantically-linked data. The progression of a controlled vocabulary flows through the six different types in the sequence in which they are listed above. Therefore, if linguistics is used to create facets, whatever form of controlled vocabulary you are using now can evolve to the next level of controlled vocabulary. Each type transcended will improve your indexing as well as your controlled vocabulary.

When adding multidimensional facets to a taxonomy, you are facilitating the possible progression from a taxonomy to an ontology, which is defined as “explicit formal specifications of the terms in a domain and the relationships between them.” You are not required to take the leap to the next controlled vocabulary form if you do not wish to, but by taking these linguistic steps in the creation and analysis of your taxonomy you will be building in the materials necessary to make that step if you so choose.

Ambiguity and Concept Principles

In order to create facets of terms and meaningful concepts, you must test linguistically if you have an ambiguous taxonomy. Ambiguous
terms are major hazards within any taxonomy. Terms that may have different meanings, such as the word nail (anatomy versus hardware), create vagueness that can be distinguished only through context. In a controlled vocabulary there is not the luxury of context unless you are incorporating facets and concepts through scope notes. There are at least 1,521 words that can cross between noun, verb, and adjective, but it has been shown that the frequency of a term is key to understanding how to combat this issue. For instance, if you search in a database for the word bluff, the different definitions of the word will become apparent. The definition that is the most commonly used for this term will occur twice as much as any of the other definitions. Combining this with the scope of your taxonomy will greatly diminish the number of ambiguous terms your taxonomy will encompass.

Another analytic tool that can be used to diminish ambiguity is creating a detailed concept record where terms are pieces of the whole and terms may be combined to form complex concepts. Creating concepts can be complicated, but by following these simple relationship principles you can begin to create and then analyze taxonomy concepts.

Concept Principle | Example
---|---
Application of term | Vehicle propulsion
Effect and cause | Combustion process
Cause and effect | Fuel
Terms strongly associated with term | Motor
Materials to create the product | Aluminum
Complimentary activities | Powertrains
Common opposites | Manual propulsion
Activity and instrument of activity | Automobile
Activity and product of activity | Emissions
Parts of the term* | Cams
*Incorporate term parts with caution because they fluctuate between pieces versus concepts

Figure 4. Concept record

Examples of using principles to create concepts:
Example 1: An engine is used to create vehicle propulsion in automobiles.
Example 2: Diesel fuel engines create fewer emissions.
Example 3: Within the powertrain the cylinders in an aluminum engine rotate to produce combustion.

Facet Application:
Scope: Specificity and Words to Avoid
Using linguistic analysis to identify and validate terms and concepts is the goal, but the same mechanisms can be used to identify terms and concepts that should be avoided. Possible facets to avoid include unpredictable entities such as personal names, individual demographics, dates, company and organization titles, brand names, new slang, unidentified acronyms, and any other facet that has the probability of frequent, and unknown, change. This is not to say that if your scope is focused on demographics, brand names, organizations or any number of other “avoidable” terminologies that these facets should be excluded outright. Your scope should always be the critical factor when deciding what facets should be included. These avoidable facets are a general list of exclusions for taxonomies that do not cover these topics within their scope. As a rule, you should look to your scope to determine which terms would be in an ever-changing state or would drastically change in short timespans.

The specificity of your taxonomy also should depend on your scope. The linguistic devices explained here can lead to an overwhelming number of possible facets for inclusion. The specificity of your facet depth should be decided beforehand. The more specific your facets, the more overlap and ambiguity you may face. Conversely, the less specific your facets are, the more likely that your terms will be so vague as to be ambiguous and, therefore, not meet the indexing needs of your subject matter. A compromise between the two must then be made. Two methods are looking at the content and deriving your specificity boundaries by the depth found within your holdings and, secondly, surveying your indexers and staff to draw on their knowledge of indexing the material and judge specificity in that fashion. Each has their advantages and disadvantages. While having your taxonomy driven by your content may seem like a sound idea, it also is dependent on the language that the content authors use and the modernity present in the contextual vernacular. If outdated or if author-centric language is used, your taxonomy specificity may not be pulling adequate information.

The same can be said of indexers. Indexers are in the trenches of content classification and, therefore, are prime experts in taxonomy and the terms used within the content. However, no matter how pure our actions are, there will always be other influences that have an effect on our understandings and opinions. Either methodology has issues grounded in contemporaneousness, biases, and an individual’s vernacular. I have found that each taxonomy is unique, and therefore the decision to go with either methodology should be considered carefully.

Organizational Placement
The typical organization for a taxonomy is a hierarchy. The hierarchy is set up as parent and child relationships of terms. Specificity also plays a large role in how deep your hierarchy will be. The more depth your taxonomy has, the larger the cascade of terms. When deciding on the structure of the hierarchy, look to your scope for guidance and keep in mind the concepts that you have created or will create. If your concept pulls from terms in the same hierarchical family, or node, then the concept you have created is not sound. It would be better presented as the parent term of the subset you are looking at. In a thesaurus, terms are set into a broad term (BT), narrower term (NT), and associative terms (RT). The visual of the two different schemas are essentially the same.

Facets can be incorporated into these schemas a number of ways. The facets of a term can grow exponentially so it is also important to decide on a set facet depth first, and then move forward with placing the facets. Narrower terms are a form of facets as are associative terms. In Figure 3, chemical, material, and research are facets of engineering. To expand the facet concept in this example, including facets of chemical such as chemically engineered under chemical, components or CO2 under mate-
tials, or researcher under research, you would be creating facets that could be placed as children of these terms. This is just one example of adding lexeme facets.

**Figure 6. Thesaurus Organizational and Taxonomy Hierarchy Schema**

<table>
<thead>
<tr>
<th>Thesaurus Organizational Schema:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Engineering (BT) ←(RT)→ Engineering devices (BT)</td>
</tr>
<tr>
<td>Material (NT) → Engineering Industry (BT)</td>
</tr>
<tr>
<td>Research (NT)</td>
</tr>
<tr>
<td>Automotive (NT)</td>
</tr>
<tr>
<td>Aerospace (NT)</td>
</tr>
<tr>
<td>Civilian (NT)</td>
</tr>
<tr>
<td>Military (NT)</td>
</tr>
<tr>
<td>Commercial (NT)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Taxonomy Hierarchy Schema:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Engineering</td>
</tr>
<tr>
<td>Material</td>
</tr>
<tr>
<td>Research</td>
</tr>
<tr>
<td>Engineering Devices</td>
</tr>
<tr>
<td>Engineering Industry</td>
</tr>
<tr>
<td>Automotive</td>
</tr>
<tr>
<td>Aerospace</td>
</tr>
<tr>
<td>Civilian</td>
</tr>
<tr>
<td>Military</td>
</tr>
<tr>
<td>Commercial</td>
</tr>
</tbody>
</table>

It is important to keep in mind that facets that are used more frequently can be used as narrow terms and associative terms, while the more complex lexemes should be used in a definition or scope note setting.

Concepts should be included in a separate file with the terms associated with the concept indicated and the concept definition. This concept document could be a duplicated listing of your taxonomy and its facets with definitions or it could be housed in a scope note format as well, either through supporting software or manually through the document itself. Concepts are fluid due to their sociological and demographic nature and, as such, need to be reviewed regularly to ensure that they are still in concordance with the taxonomies scope, standards, and linguistic integrity.

When organizing your taxonomy with scope notes, attempt to linguistically analyze the relationships of parent-child terms. Utilize the mechanisms that we have explored in this article for analyzing terms and apply the same devices to analyze taxonomy term and facet placement. Specific devices to use are: to ensure that all relationships have continuity; there is not overlap in facets or concepts; and above all, make efforts to review whenever possible. Reviewing your relationships and facets linguistically on a regular basis will ensure continuity remains, that the linguistics are current, and that if any new term or facet is added that it will be in harmony with the taxonomy.

**User Validation**

It is imperative that after selecting and validating facets through linguistic analysis that you test the selections and the organizational schema on the users of the taxonomy. First, ask the indexers using the taxonomy to review and make any recommendations that they feel are necessary for the utilization of the taxonomy. Keeping your scope and audience in mind, proceed to then query the external users about the taxonomy. The external users will offer the most weighted recommendations that you can receive. Taking the earlier example of demographic and societal morphological terms, the query base for the indexers will have more denominators in common because they share more of the same demographic and societal commonalities than a random pool of external users. External users are more likely to have more factors that influence their interpretations and opinions than internal users and will, consequently, have a higher number of diverse recommendations.

Validation from internal and external users will never be 100% and anything greater than 75% is encouraging. Once more, a taxonomy is a living document and will never be able to achieve full accuracy because terms, concepts, and user input will always be changing. Additional avenues for facet analysis are using the simplistic five class methodology introduced by S. J. Ranganathan, the expanded thirteen class Classification Research Group methodology, ISO 2788:1986, BS 8723:2005, or any combination of these options. Each new addition to a taxonomy can be analyzed and created through linguistic analysis. Facets will therefore be more likely to be an efficacious integration rather than a disruptive one if validation is successful.

Languages go through morphological changes, creating pockets of unique yet similar vernaculars within one root language. Taking the fractures of a language and applying it to the terms that are in a taxonomy enables the development of term facets. Understanding the variants of a term in all of its irregularities, social impacts, and definitions will insure that not only are you selecting terms that are correct in the sense of your indexing standards, but that you also ensure that the greatest quantity of users will be able to understand the taxonomy vocabulary.

Taking all of the linguistic mechanisms outlined for identifying and analyzing facets in this article, you may apply this knowledge to examine your existing taxonomy terms in order to gauge whether you have fully harnessed your taxonomy term facets and validate your term’s functionality. If you are creating a taxonomy, you can employ these linguistic techniques to help you in your pursuit of linguistically sound facet creation. Either way, facets enable the indexer as well as the user, to find what they are really searching for. Through validation, organization, and review, you are also ensuring continuity and strength of your taxonomy facets. Most important of all, by having a diverse and complex facet set, you will enable better indexing and better content search.

The New England Chapter has produced some great bumper stickers and is selling them for $5 each. They may be ordered from Wendy Catalano, wencat5@yahoo.com or 781-438-0609.
LINGUISTIC ANALYSIS OF TAXONOMY FACET CREATION AND VALIDATION

References