Controlled vocabularies, taxonomies, and thesauruses (and ontologies)

Last updated: October 30, 2013

When people maintain a vocabulary of terms—and sometimes, metadata about these terms—they often use different words to refer to this vocabulary. One man's taxonomy may be another woman's thesaurus, and what is an ontology, anyway? The following definitions of each are not official, standard definitions, but have worked well for us in practical situations and will help you to understand their relationships.

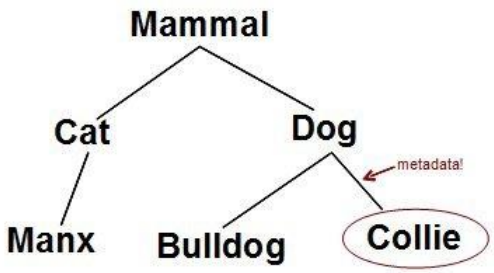
Controlled vocabulary

A controlled vocabulary is a set of terms that you have to pick from. "Yes, No" is one controlled vocabulary; so is "Mr., Ms., Miss, Mrs., Dr."; another could be "Cat, Poodle, Mammal, Collie, Dog, Manx, Bulldog." To be a truly *controlled* vocabulary, there should be some sort of governance, or agreed-upon procedures to follow if the list needs to be changed—for example, if we want to add "Not sure" as a new member of the "Yes, No" vocabulary. You could consider the "folksonomies" used by websites such as flickr and Amazon.com, in which anyone can make up and add any metadata tag to any page they like, to be uncontrolled vocabularies.

Controlled vocabularies and folksonomies are lists of words, and nothing more. Their definition does not include any specific order, although a web form might display long vocabularies alphabetically or short ones in order of popularity to make it easier for people to find the terms that they need. If there was a specific ordering to these lists, that would constitute metadata about their relationships, so we'd be moving away from controlled vocabulary territory toward a taxonomy. As we'll see, the amount and nature of metadata we store about a vocabulary's terms will be a theme as we learn about taxonomies, thesauruses, and ontologies.

Taxonomy

Taxonomies organize controlled vocabulary terms into a hierarchy. For example, if we took the example animal controlled vocabulary mentioned above and said that Cat is a broader term for Manx, that Dog is a broader term for Collie and Bulldog, and that Mammal is a broader term for Dog and Cat, we'd have a simple taxonomy. The "broader" relationships of a taxonomy are often represented visually as a tree:



While this is certainly not enough information for a computer to understand what a collie is, a system can use this little bit of semantics about collies to add value to a data collection so that you can get more out of it. For example, let's say Tempo Publications employee Jane stores a picture of Lassie in Tempo's Digital Asset Management system and tags it as "Collie." Several months later her co-worker Jim needs a picture of a dog for an article about bringing pets to hotels. He searches the DAM for "Dog," and although the picture of Lassie is not tagged with this term, a search engine that's aware of the taxonomy metadata knows that, as a collie, the picture of Lassie is also a picture of a dog, and returns that picture to Jim. The metadata helped Jim to more quickly get value out of one of their information assets.

A large ecommerce website's menu system is often a taxonomy of their products. If this taxonomy is designed well, customers can find what they need easily; if not, a customer may give up and go to a competitor's website, so a well-designed taxonomy can have a direct effect on a company's revenue.

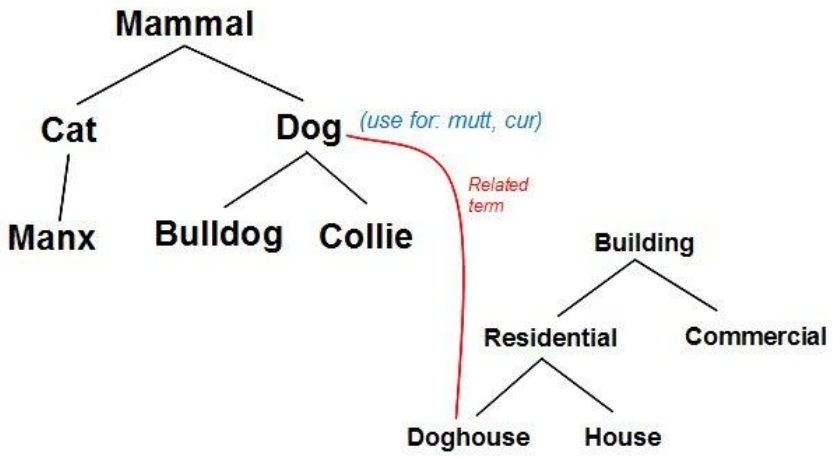
The "broader" relationship in a taxonomy may represent a part-of relationship—for example, to indicate that a steering wheel is part of a car. It may represent an instance-of relationship—for example, to indicate that the car with VIN number 50392345 is a Honda Accord. The most common use of the broader relationship, though, is the "broader generic" relationship, showing that one set of things is a subset of another—for example, that all Accords are Hondas, or that all collies are dogs.

A taxonomy used for serious business purposes often stores more than just broader-than relationships. These can include alternative terms to assist search (for example, "auto" as an alternative to "car"), translations of the term to foreign languages, metadata about who last edited the term and when, and notes about what exactly the term applies to if there is potential confusion—what taxonomists call "scope notes."

Thesaurus

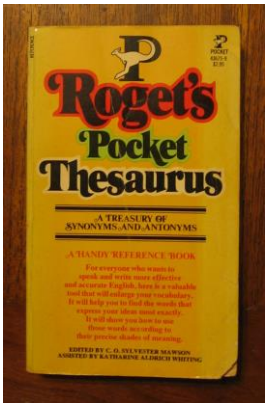
A thesaurus stores even more metadata than a taxonomy. It might store relationship information about opposite terms—for example, that the opposite of Yes is No. It might store what taxonomists call a "Use For" relationship, so that users searching for a particular term that isn't considered to be the best one can

be redirected to the preferred term. For example, if someone is trying to tag an article about summer footwear with the term "Flip Flops," a system that takes advantage of "Use For" metadata can direct them to use the term Sandals instead.



"Broader" relationships are one kind of relationship metadata, and a thesaurus often stores other kinds of relationship metadata as well. These can even connect a term to another term in a different vocabulary. For example, a thesaurus might store metadata indicating that the term Dog in an animal taxonomy is Related To the term Doghouse in a taxonomy of shelter types, or to a particular veterinary product in a pharmaceutical company's taxonomy of products.

If a taxonomy can store more than just broader relationships, and a thesaurus is usually arranged as a taxonomy with additional metadata, then how do you know when a hierarchical controlled vocabulary with metadata has gone beyond being a taxonomy to become a thesaurus? You don't. The terms are sometimes used interchangeably, and specialists in the area, who often have degrees from library and information science schools, call themselves "taxonomists" even when they manage thesauruses. No one calls themselves a thesaurist; perhaps they want to avoid confusing people who think of thesauruses like Roget's Thesaurus, the book of synonyms that their grade school writing teacher made them consult if they used the same word too often in an essay.



The metadata properties associated with a taxonomy's terms fall into two categories, which we can call relationship properties and attribute properties. Relationship properties indicate a term's relationship with another term, such as that Dog is a broader term than Collie or that Yes is the antonym of No. Attribute properties are typically pieces of text entered as metadata about a term, such as the Spanish word for that term or the name of the staff member who last edited it.

Ontologies

Even if we ignore the philosophers and stick to the world of information processing, we can still find a confusing variety of definitions for the word "ontology." When it comes to managing information, you can think of it this way: when you develop a thesaurus, there are various standard, generally applicable

relationship and attribute properties that you can use to store more information about the terms in that thesaurus. When you develop an ontology, you can define your own relationships and attributes, as well as classes of things that are characterized by these relationships and attributes. Where thesauruses and taxonomies use generic relationships such as broader, related and “use for” that can be applied to any term, ontologies define relationship and attributes that are specific to a particular business area. For example, a project management ontology might include a relationship to show that one event is a precondition of another event, and a medical ontology might have a relationship to show that one symptom contraindicates a particular treatment.

Ontologies can be used by software systems to infer new information, such as class membership. For example, if Jack has a playsInstrument property value of "guitar" and the ontology says that anyone with a playsInstrument value is a musician, we can infer that Jack is a musician even if there is no explicit data saying that he is a member of that class. Inference in more sophisticated ontologies can do things like enable medical researchers to infer that a particular protein falls into a class of possible treatments for a given disease.

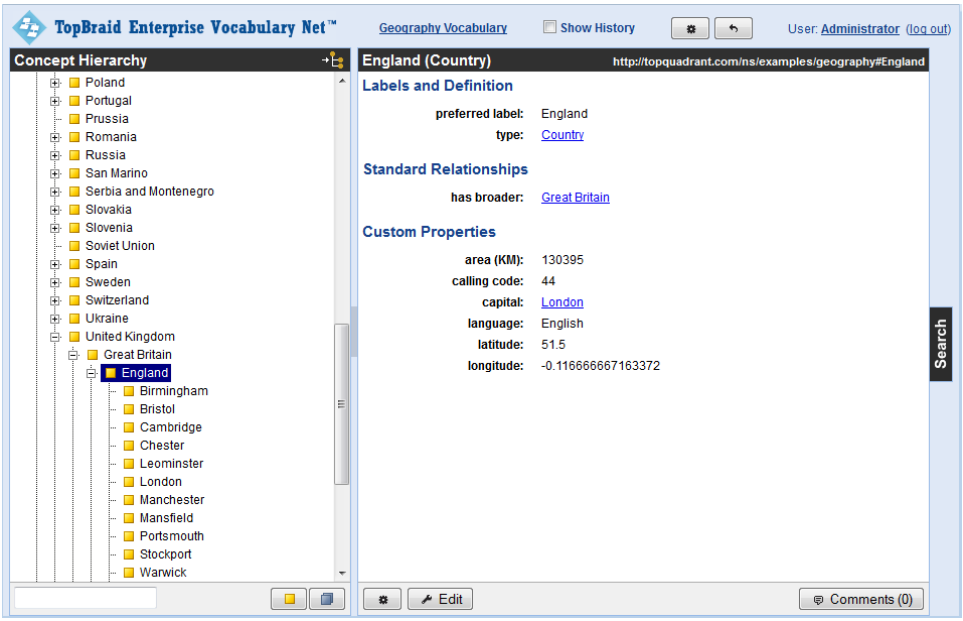
In many systems, ontologies and taxonomies work together. For example, if all the potential values of playsInstrument must come from a specific list whose membership and metadata are controlled by a governance process, that may be stored in a taxonomy.

The W3C standard for defining ontologies is OWL, a key component of semantic web technologies. When using this standard, you can develop, use, and share ontologies without being tied to a particular vendor's file format or syntax. The leading tool in the industry for working with OWL ontologies is TopQuadrant's TopBraid Composer, which makes it easy develop to new ontologies or combine and customize existing ones by pointing, clicking, dragging, and filling out dialog boxes with an intuitive graphical user interface. More powerful editions of TopBraid Composer let you connect to a broader array of data sources, build and deploy applications, and take advantage of more powerful RDFS and OWL features such as inferencing and the use of restriction classes.

SKOS and TopBraid Enterprise Vocabulary Net

Another W3C standard is SKOS, the Simple Knowledge Organizing System. SKOS uses the OWL standard to define the relationship and attribute properties that are most commonly used for vocabulary management, such as broader, related, alternative label, and scope note.

Being an OWL ontology, SKOS can easily be edited or extended using TopQuadrant's TopBraid Composer. TopQuadrant also offers a networked, web-based product with a simpler interface to let teams of people work together in the development of controlled vocabularies, taxonomies, thesauruses, and ontologies: Enterprise Vocabulary Net, or EVN.



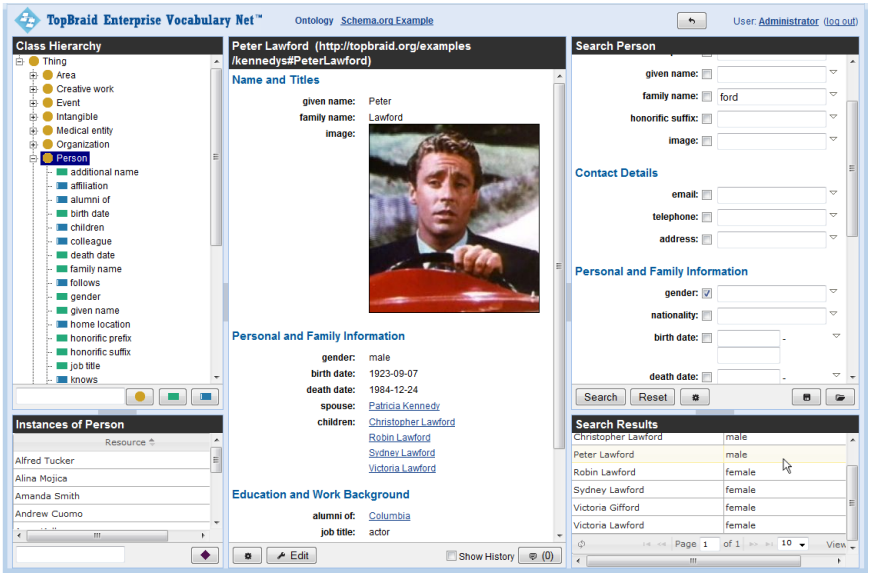
TopBraid EVN and controlled vocabularies

EVN lets people with no knowledge of semantic web standards edit controlled vocabularies and associated metadata using a simple, web-based interface. Relationships between terms can often be defined by clicking and dragging nodes on the hierarchical tree of terms. Data is stored using the SKOS standard, making it interoperable with a growing number of tools and freely available datasets such as those from the New York Times and the U.S. Library of Congress. Custom classes and properties can be created using both the EVN web-based interface and TopBraid Composer, a copy of which is included with the purchase of EVN.

Many find that EVN's basis in SKOS and related semantic web standards, in addition to letting them define relationships between their own vocabulary and publicly available ones, lets them define relationships between vocabularies from different divisions in their enterprise. The "web" part of semantic web gives you the "Net" part of Enterprise Vocabulary Net, by using granular, globally unique identifiers to connect terms in different vocabularies, taxonomies, and thesauruses in your enterprise as soon as they are created or long after they've been in production. Building on related W3C standards such as RDF, EVN features such as change history reports, multi-user support, and vocabulary working copies are all implemented with data stored using standards-based models that can interoperate with other tools or be queried with the W3C standard SPARQL query language.

TopBraid EVN and controlled vocabularies

In addition to letting you manage controlled vocabularies, taxonomies, and thesauruses, TopBraid EVN now lets a workgroup do collaborative editing of any RDFS schema or OWL ontology that you like, as well as instances of that schema or ontology. With a simpler, more streamlined interface than TopBraid Composer, it still has all the workgroup features that the vocabulary editor has such as change history reports, support for constraint rules, and working copies. Just as you can use TopBraid EVN to customize and combine external and internal vocabularies, you can now do this with external and internal ontologies as well.



Your vocabularies and ontologies

If you have a vocabulary that you want to organize and maintain to add value to your business processes or data collections, or if you want to turn an uncontrolled folksonomy into a taxonomy or other richer set of data and metadata, or if you want to do collaborative editing of controlled vocabularies or ontologies with change tracking in a networked environment, TopBraid EVN and the standards that it builds on combine ease of use with interoperability and scalability to help you get the most from your vocabulary and ontology assets. If you currently maintain vocabularies and ontologies using spreadsheets or another simple tool that doesn't offer standards support, extensibility, and support for multi-user workflows in a networked environment, EVN will take your management of these semantic assets to a more professional level. See http://www.topquadrant.com/evn or contact info@topquadrant.com to find out more.

Lassie picture from the Florida Memory Project hosted at the State Archive of Florida
http://commons.wikimedia.org/wiki/File:Lassie.jpg