The advantages of using a controlled vocabulary (such as a taxonomy or thesaurus) in a database or website project can seem mysterious. What does it do? How does it work? And why should I use one?

Let’s take a look behind the scenes to find out why utilizing controlled vocabularies is so valuable. Figure 1 shows basic software components of a controlled vocabulary.

First of all, the taxonomy or thesaurus must be in a digital format. It can be kept either as a separate document file (a spreadsheet, for example) or as it exists in a specialized software application (such as a taxonomy tool).
management tool). The screenshot in Figure 2 is from the editorial user interface of a thesaurus software application.

The left panel shows the taxonomy in the hierarchical view. This hierarchy organizes terms and concepts in branches. The broadest subjects are located at the top of this hierarchy. Depending on the size of the thesaurus, each of these broad subjects often contains thousands of terms.

Each term has its own intricate set of relationships, which are found in the term record. The right panel shows a term record containing the broader term, narrower terms, status, related terms and other fields such as synonyms, history, scope notes and so forth.

This amounts to quite a bit of information stored as an object. In this example, the taxonomy term-object is Heating, cooling, and ventilation. Treating terms as objects is a useful and easy-to-use way to access your taxonomy, as the object is the term along with all its pertinent and related data.

The thesaurus terms are all organized into a hierarchy or other preferred format. But how are these thesaurus terms connected with, for example, a website? Most often, it will be through controlling the terms used as metadata for the objects and pages on the site.

As shown in Figures 3 and 4, metadata can be found by going to any website, choosing View and selecting the Source view or Page Source View (depending on which Internet browser is used).
When working with a relational database management system (RDBMS), the taxonomy terms are placed in a table somewhere (Figure 5). This table of terms is then related to the primary key or main records; this table will subsequently be linked to the records directly.

Whether using an object system or relational database management system it is vital to have a place to put those terms. Whoever is building or maintaining the database must find a place for them. In object-oriented code, a very similar kind of model applies (Figure 6). Again, it is extremely important that the data transfers over from the thesaurus to the primary records.

The terms and their connections must be defined in the relational database. In the various relational database models, there are a lot of options for how to carry this out. See, for example, Figure 7.

In the case of an XML-based database system, new text can be input and the system will have a way to suggest the terms automatically and add them to the system (Figure 8).
When looking at the Mediasleuth site, below, for example, the hierarchical list shown on the website is directly connected to the site from the hierarchical list of the original taxonomy (Figure 9).

Oftentimes the narrower terms in a term record become the narrower terms in the search interface (Figure 10); the related terms from the term record may also be posted in the search interface. All of this integration illustrates that there is a fairly direct connection between the original taxonomy and the website, the user interface and the search experience.

Integrating the taxonomy with the content and user interface enhances the findability of the terms on the website or database (Figure 11).

The terms are used as labels in search as well as for tagging the records behind the scenes. Rather than merely having simple terms connected to a webpage, all of the intertwining relationships that define each concept are linked directly to the search.

When the taxonomy terms are attached to the record and loaded into the search system, while using a variation of that same taxonomy on top of the search system, the taxonomy is being used at the same time to search and to tag. Then when the search is being used, the results are vastly improved.

It doesn’t matter whether a relational database management system, MySQL, Lucene, Autonomy or Google is being used as the search software if the taxonomy term is attached to the term record, and the taxonomy terms are placed in the inverted file for search. When choosing a taxonomy term on the user interface, it will go to that inverted index and pull back the appropriate records regardless of the search software.

The Figure 12 workflow diagram shows that it may be necessary to have a lot of raw data placed into a data repository. The taxonomy terms will be added to the records in that repository. That repository could then be stored as an SQL file for e-commerce, in an XIS repository or in a search system. The system may or may not use a presentation layer for performing search. So, from the original repository where the terms have been added to the records, they can also be spun out to all of these different places for storing the records. Use of this feature is not required, but is certainly available, and often times valuable.

Finally, the same set of taxonomy terms and relationships can be inserted in many places on a website. Taxonomies are easily accessible, easily edited, easily stored and easily utilized.