

Getting Started with Controlled Vocabularies, Taxonomies and Thesauri

BY RON ROSKIEWICZ

Implementing a metadata taxonomy or controlled vocabulary in a business can be confusing at best for all except those engaged in day-to-day implementations.

Introducing a metadata schema to your company doesn't have to be overly complicated. Using the basic functionality included in most desktop applications, companies can execute simple implementations involving a bit of digital rights or file characterization information. But implementing a robust, scalable controlled vocabulary in a company where search and retrieval is a value-added service requires an exponential increase in subject matter expertise and tools to support the installation.

Four basic components contribute to such an effort: a core controlled vocabulary, a customization strategy, an application to manage and integrate a controlled vocabulary into a asset management workflow, and a certain amount of professional guidance to tie it all together.

Controlled Vocabularies. Most companies are already using a controlled vocabulary without even knowing it. The terms they use to define products and services, topics and events all form a core set of vocabulary terms.

If the controlled vocabulary is merely used to support a predictable, closed set of files, the requirement might simply be to organize the existing terms and use them as the basis for the installation. This approach is pretty typical in the structured world of database forms, where the record is the asset. In this world, the metadata becomes part of the application and is defined during its development.

In the unstructured world of multimedia files and creative content developed on desktop systems, metadata definition and attribution is chaotic at best. This is where the identification and rationalization of legacy metadata that exists on the current Web site or in an analog process is more difficult.

For many schemes with a lot of unstructured data, the best approach is often to combine a core set of controlled vocabulary with extensions based on existing terminology and custom additions. One good source for taxonomies, controlled vocabularies and thesauri is the **Taxonomy Warehouse** (www.taxonomywarehouse.com). Listed on its pages are a broad offering of core sets from government, educational and commercial sources. They are available for licensing, often with an option to subscribe and receive

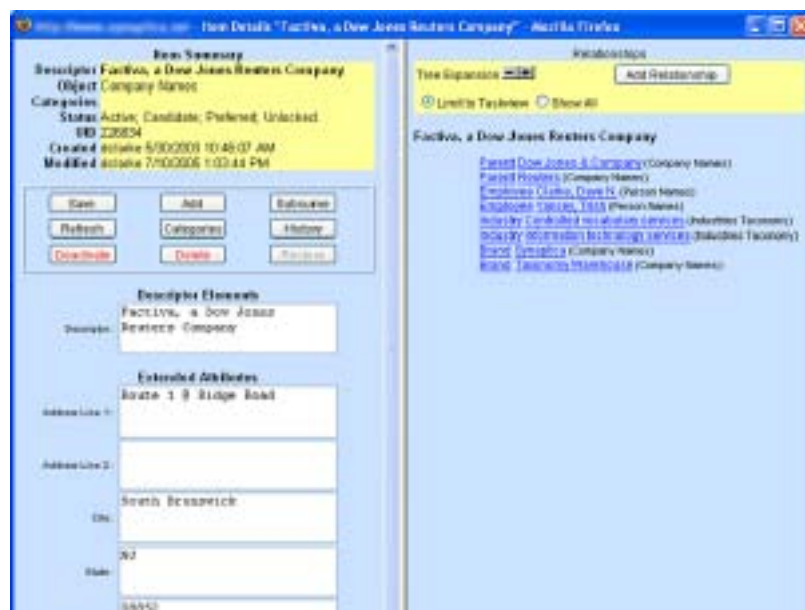
updates when they become available. The license fees for these sets range from a few hundred dollars to \$100,000.

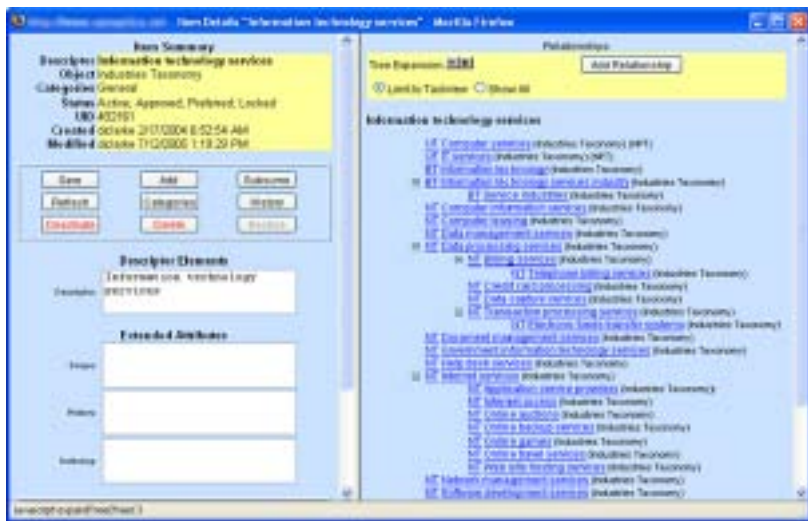
From Public Domain to Commercial. In every case, the buyer must be aware of what he is purchasing and whether it is deep and broad enough to satisfy the company's needs. When in doubt, consult a specialist for advice.

The objective in all cases is control, consistency and customizability. The control comes from the organization that the digital asset management system provides. Consistency is the result of everyone relying on the same dictionary for the values they use in the defining the assets. Customization occurs when the core controlled vocabulary has to be extended to satisfy the unique conditions of the user.

Management Software. Regardless of which controlled vocabulary is licensed, an application will be required to manage and integrate the controlled vocabulary or taxonomy into the search and repository workflow. A capable application provides the functionality to organize and edit metadata properties and values, can be accessed by external systems, and has a range of fea-

Employing a single easy-to-use graphic interface, Synaptica users can create and edit term relationships, as well as term attributes and other details.





Users have complete control over metadata elements and semantic relationships.

Users have complete control over metadata elements and semantic relationships. Each taxonomy specialist will have a favorite application in addition to favorite controlled vocabulary sets for a given industry.

Most graphic arts practitioners work in an unstructured environment and without a controlled vocabulary. Creating a solution is beyond their realm of understanding, so their intellectual property remains in a state of semi-contained chaos.

To have a viable solution to a searchable database, the following components must be in place: some form of a metadata strategy, a controlled vocabulary and a controlled vocabulary management application.

One such application is the Factiva Synaptica Vocabulary and Metadata Management System (<http://factiva.com/taxonomy>). This application allows calls in COM or native database languages to be made to rout the controlled vocabulary values into fields for selection or embedding. The list of functions and features for this system provides useful insight into what a highly evolved metadata management application

Synaptica offers robust search, filtering and contextual and hierarchical display capabilities.



Metadata Terms

One of the difficult first steps in putting together a metadata system is figuring out the bewildering terminology that metadata practitioners use. Here are some of the key terms.

- Controlled vocabularies (CVs) are a collection of preferred terms used to categorize content and create database schema.
- A taxonomy is a type of controlled vocabulary that is organized hierarchically to bring structure to navigation and search systems.
- A thesaurus is a complex controlled vocabulary that takes into account alternative spellings and related terms. The objective of any thesaurus is to support navigation and search system by reflecting the usage of terminology in a company.
- Ontologies are multifaceted taxonomies, meaning that there are relationships between terms in a taxonomy, as well as rules governing the specification of terms and the relationships between the terms. In this sense, ontologies are coded to make connections in ways similar to how we transform data into information and knowledge.

TSR

needs to be. For example, while the controlled vocabulary might be the starting point, it will also be necessary to manage associated glossaries, dictionaries, lexicons and thesauri, and name authority files.

Synaptica supports metadata and taxonomy standards such as ANSI/NISO Z39.19; ISO 2788 and 5964. It also supports customizable semantic relationship types that include vocabulary clusters and cross-walk mapping. There is contextual metadata value support, such as approval and candidate term statuses, and many more that involve context, behavior and rules-based logic.

Another popular application is Protégé-2000, an open source application available from Stanford University (<http://protege.stanford.edu/>). This application is free and is designed as an ontology editor and knowledge-base framework. The application, written in Java, is extensible and supports Frames, XML Schema, RDF and OWL. Owl is a plug-in for Protégé that supports the Web Ontology Language (OWL). The OWL plug-in allows for OWL and Resource Description Framework (RDF) ontologies to be loaded and saved in Protégé, and many other approaches to customizing access to the ontology via an open-source Java api.

Implementing a controlled vocabulary. Finding a core-controlled vocabulary by searching the resources at taxonomywarehouse.com is a good first step. Its aggregation of CV resources is one of the best, and through it you can link to the sources of the CV for licensing and terms. Your first trip to the site might be overwhelming, though, since moving from not knowing

Controlled Vocabulary Software and Consulting

The following companies develop applications for managing taxonomies and/or provide consulting services to support the integration of controlled vocabularies in your workflow:

Advanced Document Sciences www.adocs.com

Association of Independent Information

Professionals www.aiip.org-index.html

Autonomy www.autonomy.com

Controlled Vocabulary

www.controlledvocabulary.com

Data Harmony www.dataharmony.com

Electronic Scriptorium

www.electroniccriptorium.com

Factiva www.factiva.com

Fast www.fastsearch.com

iQuest Analytics www.iquestanalytics.com

Molecular www.molecular.com

SchemaLogic www.schemalogic.com

Software and Information Industry Association

www.sii.net

Teragram www.teragram.com

Verity, Inc. www.verity.com

Vivismo www.vivismo.com

Wordmap www.wordmap.com

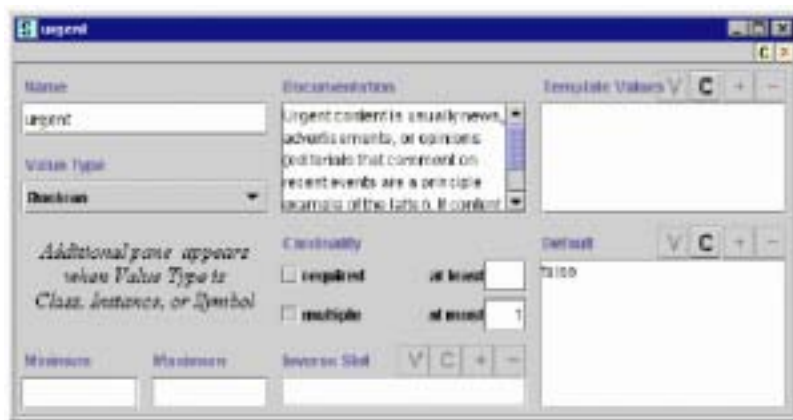
TSR

about any CVs available to being exposed to a great number of industry and domain-specific sets can be a bit much. The good news is that many of them are from authoritative sources. Navigating to the right one might still require you to get some help and guidance from someone with experience implementing a similar system with similar requirements. You can do this by engaging companies such as Factiva (<http://factiva.com/taxonomy>) or Electronic Scriptorium (www.electroniccriptorium.com), which have the experience and tools to construct a CV set suited to your business. (More resources are listed above).

Few controlled vocabularies exist independently, outside of a workflow and in a vacuum. Often the set is dropped in the middle of an existing workflow that already has a rudimentary set of keywords or an industry standard schema filled with unique properties and values.

In some cases, the asset repository that the controlled vocabulary is serving must be reconciled with an existing database management system that is being used to control values related to rights or project data. In such cases, a host of additional, related issues will arise that might require support from the controlled vocabulary application or through professional services from a taxonomy specialist. If you are interested in the specialized terminology used by ontology experts and the process they must go through to create or edit a controlled vocabulary, download the Protégé User's Guide (<http://protege.stanford.edu/publications/UserGuide.pdf>).

The illustration above is an excellent example of



the process used by professionals to identify the needs of users, build and adapt a taxonomy to meet their requirements, and implement it as part of an information storage and search infrastructure.

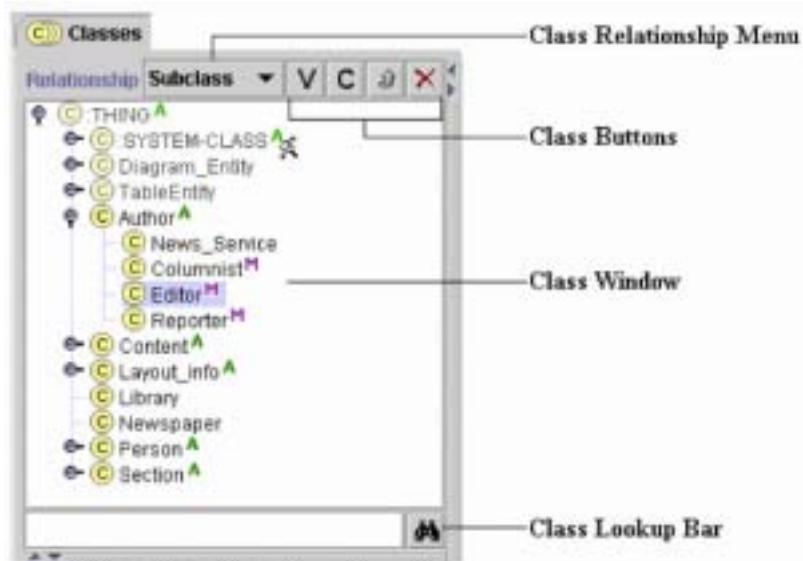
Whatever application you or your consultant chooses to use might have to support one or more of the following, depending on the nature of the legacy data being imported or the type and location of the system being installed.

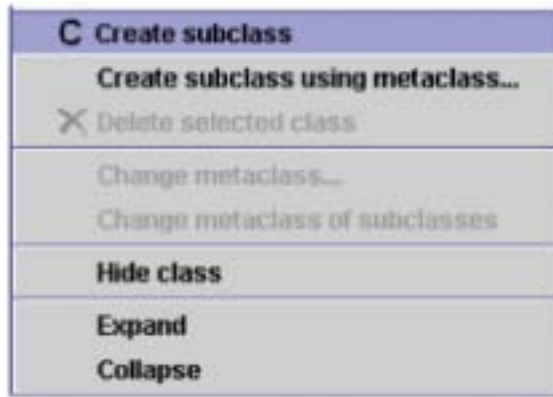
Unicode UTF-8 Support. More and more applications are being rewritten to conform to extended, 2-byte character sets of Japanese, Chinese and Korean languages. An application and the metadata management system that supports it must be unicode-compliant for these languages to be recognized.

Bulk Import and Validation Functions. Moving metadata from legacy systems is best done in bulk, with the resulting data being compliant with industry standards. Validation of data means that imported data is compliant with the format, style, spelling or data type expected in the new system. Data validation is important for imported files and files ingested on a day-to-day basis.

The Slots form in Protégé can be used to edit the attributes of a slot. A slot is an attribute of a class. For example, a physician class might have name, address and phone number as slots.

Class hierarchies as they are represented in Protégé.





The application menu used for creating and editing classes in Protégé.

Metadata Clusters and Crosswalk Mapping. Metadata that is organized in clusters or metamodels will often require some form of crosswalk or bridge to connect them. This crosswalk, in effect, reflects the association of similar metadata existing in different clusters while maintaining the structure of the cluster.

Process model for Developing and Deploying Taxonomies.

Metadata Export. Built-in functionality to export dictionaries, taxonomies, etc. is important. Two com-

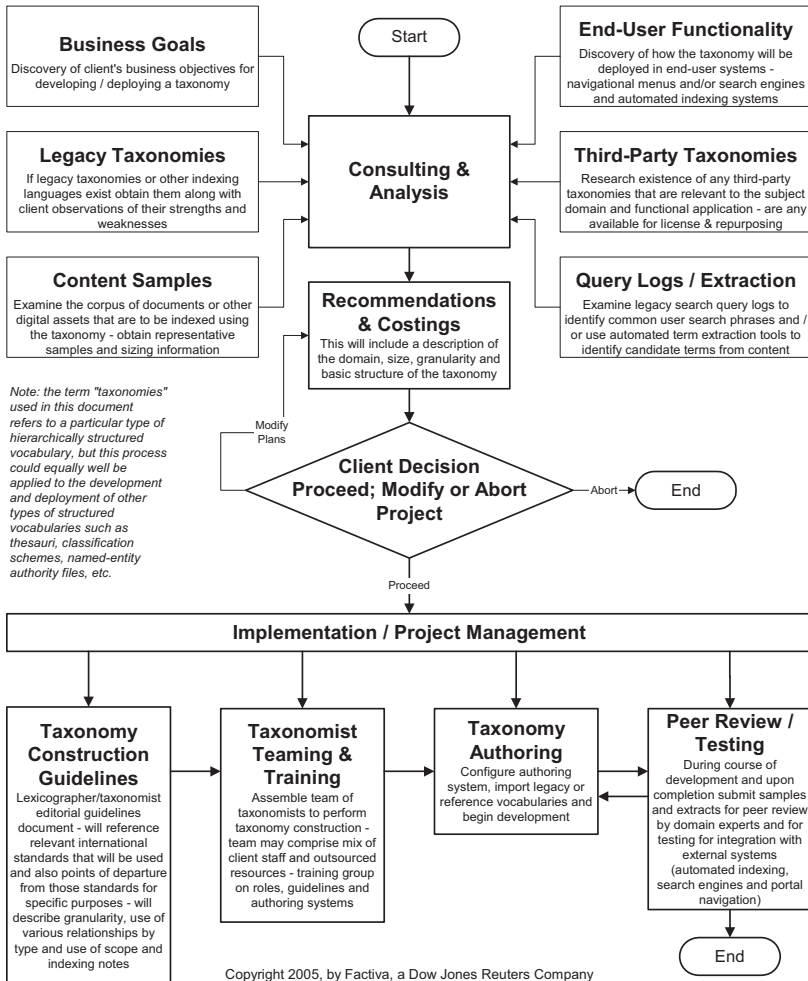
monly used formats are XML and comma separated value (CSV) file. Using XML means that integration into other systems that support XML is straightforward. CSV reflects the common practice of developing and storing controlled vo-cabularies in text editors or Word files.

Security and Permissions. Access to metadata information often depends on a user's permission level. Companies need to be able to change.

Our Take

The tools for developing and maintaining controlled vocabularies and taxonomies and the professional consulting support to implement them are available today. Desktop equivalents of these tools, suited to the rest of us who are not library scientists, are slowly emerging. This emergence is taking place in graphic arts and publishing, as so many other innovations for managing digital assets have in the past. The transition will take more time and will probably result in an acceptance of "just enough metadata," just as we have settled on "good enough color" and "good enough typography" for most of our day-to-day requirements. **TSR**

Process Model for Developing & Deploying Taxonomies



About the Author

Ron Roszkiewicz is a consultant and writer from Leucadia, Calif. He can be reached at roszkiewicz@gmail.com.

Copyright 2005, by Factiva, a Dow Jones Reuters Company

for more information: <http://factiva.com/taxonomy>