
PREPARING A CONTROLLED VOCABULARY FOR CONTENT MANAGEMENT AND ACCESS

By Seth Earley
Earley and Associates, Inc.
www.earley.com

As indexers, you understand the need to have “handles” on information – the key-words that are used to access ideas and concepts in documents. Typically, when creating a back-of-the-book index, content is reviewed for key ideas, subjects and topics, and words representing those topics (along with the synonyms that a reader might think of) are used as the pointers to get to that information. You are generating a bottom-up classification of terms – analyzing content and determining the words that represent key concepts.

(In a purist definition of a taxonomy, terms are arranged in a hierarchical, parent-child relationship. A set of classification terms is not necessarily hierarchical. In many cases, when people refer to taxonomies, they are really talking about classification.)

When deriving the back-of-the-book index, we are relying on the language used by the author to drive our term list. Our job is to give the reader a set of hooks to get to the ideas that the author conceived of in the author’s preferred terms and also in equivalent terms that readers will usually consider.

Indexing is the process of putting more structure around a body of work. Content management systems are a different facet of the same problem. When authors create a document using a content management tool, the content is tagged in smaller chunks (usually at the page or document level) and the tags are preconceived. Rather than looking over a document and saying, “OK, what are the important concepts here?” we are saying, “You wrote something for this web site (or intranet). Here are some categories that we think readers care about. Where do your ideas fit in this scheme?” In the first case, we take information and build a structure around it. In the second, we’ve created a structure and are asking the author (or content manager) to put the information into our preconceived framework.

In effect, we’re asking the content creator to file their information.

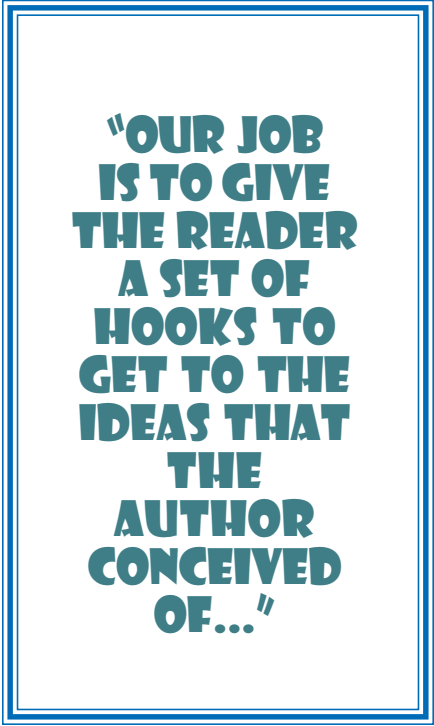
Definitions

In order to discuss some of the challenges and techniques around controlled vocabularies and tagging, let’s define some terms that you already are familiar with.

Keywords – words or phrases that identify a document’s contents. Ideally, users would select keywords from a controlled vocabulary, but in many situations, this is not the case.

Controlled vocabulary – list of subjects from which users pick the subject to describe an item when cataloging. Controlled vocabularies limit choices to an agreed-upon unambiguous set of terms.

Classification – the act of distributing things into classes or categories of the same type. By applying keywords and metadata to documents, we are engaging in the process of classification.



**“OUR JOB
IS TO GIVE
THE READER
A SET OF
HOOKS TO
GET TO THE
IDEAS THAT
THE
AUTHOR
CONCEIVED
OF...”**

Continued on next page

Taxonomy – the science, laws and principles of classification. Most definitions of taxonomy specifically pertain to biology – the classification of organisms. In the strictest sense, taxonomy denotes terms in a hierarchical relationship, either parent child or whole part.

Metadata – the attributes of an information object – document, data set, database, image, artifact, collection, etc. When we apply keywords to a document, we are tagging the document with metadata.

Facets – any of the definable aspects that make up a subject or an object. We can have various types of metadata applied to an object. These form different facets of the object. An article can be classified according to industry and whether it is technical or nontechnical. These are each facets of classification. You can also think of facets as search parameters.

Thesaurus – a list of all the subject headings or descriptors used in a particular database, catalog, or index. A thesaurus helps define preferred terms and relationships between terms.

Preferred term – the term that is the pointer to a document. We would tag a document with the preferred term and have equivalent terms point to the preferred term.

Equivalence – synonyms and quasi-synonyms. Two words can mean the same thing, but two words with opposite meanings can be treated as quasi-synonyms if they can be used to describe the same concept. Terms can also be considered equivalent if they have overlapping meanings.

Related terms – terms may not be hierarchical and they may not be equivalent, but they may be included in a thesaurus because they are conceptually related. (“See also” terms: for example, Vermont – see also *Maple Syrup*)

Deriving and applying a system for classification

Frequently, a client will come to me and say, “We need to create a taxonomy.” What they mean is “We need to solve a problem of navigation and content access.” Providing a taxonomy without the context of user navigation and search scenarios will not solve the problem. Taxonomy is not the same as navigation and the two are frequently confused. There are a number of components to solving problems of search and navigation. Developing metadata standards and applying controlled vocabulary terms to content

combined with complementary navigation structures will help users access the content they need in the context of work tasks. I’ve defined some of the components and now will discuss how to fit the puzzle pieces together.

A controlled vocabulary is a set of terms whose meaning and usage we agree upon. Out of all the terms that an author can use to describe information, we’re narrowing things down and saying: “This is it. Here are the terms you can use to describe your ideas.” There are several challenges to this. Meaning is subjective and context-sensitive. It depends on whom we are communicating with and what *they* are trying to accomplish.

Continued on next page

**“FREQUENTLY,
A CLIENT WILL
COME TO ME AND
SAY,
'WE NEED TO
CREATE A
TAXONOMY.'
WHAT THEY MEAN
IS
'WE NEED TO
SOLVE A PROBLEM
OF NAVIGATION
AND CONTENT
ACCESS.'”**

1. Understand user tasks

The first piece to the puzzle is defining what users are looking for and how they think about information. What are the ways in which they label documents? How do they organize information on their hard drives? What search terms do they use when finding information? What are their work tasks and what information do they need at each step of the process?

2. Find representative content

After defining user tasks, get examples of valuable content. What are the types of documents that users store locally? What do they bookmark? How do they organize email? How do they organize their local drive?

3. Review existing folder structures, indexes, web sites

If there are file shares, directories, intranets or industry web sites that people already use, leverage these. Determine what works and what can be repurposed.

4. Build out search scenarios

How do people think about documents and how they can be accessed? If a folder structure is suggested, how many documents will live in that folder? How will they be distinguished? How might they be sorted? How might users want to search against a list of documents? How will they be named? For example, the following structure might be suggested:

Sales Tools

Case Studies

How will case studies be organized? By Product? By Industry? By Solution? By Customer?

What this suggests is that we have metadata for Case Studies that includes facets for industry (pulling from our industry-controlled vocabulary), solution (again, a controlled list of solution keywords), product (perhaps from an accounting or order entry system), and customer (from a sales force automation application).

By tagging case study documents appropriately, we can allow users to search for the types of case studies they need. These user search and navigation scenarios drive development of a taxonomy of document types and derivation of additional metadata.

5. Build a thesaurus

It is difficult to get people to agree on meaning. Terms can have different meanings in different contexts and different words can be used to describe the same ideas.

When deriving taxonomy terms, consider synonyms and policies for applying terms. What constitutes a “customer reference”? How are “benchmarks” defined? Is a “brochure” the same as “collateral”? Define preferred terms and guidelines for usage.

6. Review and maintain

Some aspects of the information architecture will remain more stable: other aspects will evolve. Customer names, products, solutions, markets, etc., will naturally change. It is important to recognize that taxonomies will need to be reviewed and updated and that policies for governance will need to be established.

Continued on next page

The bottom line is that deriving controlled vocabularies, classification systems and navigation are interrelated problems and cannot be addressed in isolation. We need to understand the types of the problems that users are attempting to solve when accessing the web site or intranet and the nature of the information they are seeking. Tagging is the common language used to improve knowledge flow in the organization. Stay focused on user tasks and realize that maintaining controlled vocabularies is a work in progress.

Seth Earley is a renowned expert in assisting businesses derive maximum business value from knowledge management, content management and collaborative systems. He is the author and teacher of courses and workshops on portal development, knowledge management, taxonomy development and content strategy. For more information on his Taxonomy JumpStart class series and on the Taxonomy Community of Practice, visit www.earley.com.

**"TAGGING IS
THE
COMMON LANGUAGE
USED TO IMPROVE
KNOWLEDGE FLOW
IN THE
ORGANIZATION"**

NEW UC BERKELEY INDEXING COURSE OPENS

The new UC Berkeley Extension Online indexing course (Indexing Theory and Application, X477) is taught completely online, and was developed by Sylvia Coates, with many other indexers contributing material. The course covers introductory materials and an overview of the indexing process, the creation of basic indexes, how to select terms, indexing specifications, the business side of indexing, workflow process, and an introduction to embedded indexing and web indexing.

Students can enroll at any time, and have six months to complete the course. The course is self-paced and fits your time frame. Indexing assignments are handed in via email, graded, and returned quickly. The course contains three indexing assignments that gradually get more difficult and also expose the student to a variety of content. Students get reading assignments, research assignments, and exercises. Demo packages of the big three indexing programs (Macrex, Sky, and Cindex) are included, and the students

must use each one to complete assignments.

The course also allows for personal interaction with the instructors and with the other students in the course. The course materials and assignments are posted on the course web site. The course also has a message board, with categories for discussion, and students can ask each other questions and read the instructor's responses to previous issues and thoughts.

Be on the lookout for a new indexing course available from the American Society of Indexers as well, launch date this summer. More information is on their web site at www.asindexing.org.

Course Outline for X477

Indexing: Theory and Application

- Unit 1: Introduction
- Unit 2: Creating a Basic Index
- Unit 3: Term Selection
- Unit 4: Indexing Specifications
- Unit 5: Writing a Full-Length Index
- Unit 6: The Business of Indexing
- Unit 7: Developing Your Own Process and Style
- Unit 8: Putting It All Together
- Unit 9: Embedded Indexing; Indexing a Web Site

The course is currently accepting enrollments in two sections, the first taught by Sylvia Coates, and the second taught by Jan Wright, who developed the web and embedded indexing portions of the course.

For more information, visit the Extension web site at: <http://explore.berkeley.edu/UCExtcourseview.asp?secid=525&value=related&action=Internet>

There is a potential advanced technical indexing course being considered for development. There is a survey posted to collect potential students' opinions and needs. Please go fill one out at <http://www.surveymonkey.com/s.asp?u=28391025766>