

# Content Taxonomy— Reduce Your Exposure

By Greg Boyd, President, Results Engineering

Remember when you first started implementing database systems? Before you knew it, one database had grown to five. Data consistency was in question. Years later, you found yourself wondering why you hadn't implemented data modeling and data dictionaries before the problem got out of hand. A Document Taxonomy is the proper architectural first step, and will easily pay for itself, across the organization, through smarter document creation, storage, retrieval, and retention—and decreased audit, compliance, and litigation exposure.

A taxonomy is a classification scheme. Content Taxonomy is a classification of all unstructured content (email, document images, HTML, XML, PC files, computer printouts, audio, and video) into a series of categories. This information is metadata describing an enterprise's unstructured information.

Many organizations build purpose-specific document taxonomies focused on one application or department. Few carry it far enough to fully encompass today's significant knowledge management, workflow, productivity, audit, control, and compliance issues. As a result, organizations are constantly re-examining their own documents and making the same mistakes over and over. Users spend too much time searching for information and not enough time finding.

## Why Taxonomy is Vital

Currently, most public companies are undergoing internal controls, workflow, and document management assessments of financial records and document retention for compliance with the **Sarbanes-Oxley Act of 2002**. If a proper content taxonomy existed, this effort would not have to be done. Also, the information being collected is often so focused on Sarbanes-Oxley that opportunities for process improvement and better knowledge management are missed.

Without an up-to-date Enterprise Content Taxonomy database to understand the interrelationships, controls, and potential liabilities of structured and unstructured content, we suspect regulations such as Sarbanes-

Oxley, and concerns over possible shareholder and other litigation, will result in gridlock over any future document destruction.

Most organizations skip formal taxonomy development and rush into deployment by building their classification system around index keywords for retrieval. This is not a complete taxonomy and will cause many problems later with their document management system, as well as missing many reengineering and process improvement opportunities. Taxonomy incorporates the practice of characterizing the context and relationships among documents and their retrieval aspects.

***“Users spend too much  
time searching for  
information and not enough  
time finding.”***

Without a consistent taxonomy across all areas of the enterprise:

- ◆ Isolated silos of information systems and processes expand, causing significant cost, duplication of effort, and liabilities to the organization. How many times do you have to give your personal and insurance information to different departments of the same hospital?
- ◆ One group calls a form or report one name, another group another name. Both groups file them. Retention rules are applied to the form under one name, but not the other. When litigation occurs and a discovery action results, the information which was properly destroyed in one system is discovered in another and becomes a liability.
- ◆ One department creates a business form with many data fields. Another form already exists with the same fields in a different layout. A third version exists as an electronic web form.

- ◆ Different courts, social services agencies, and prison systems file the same paper documents. Most of this paper is created by other government agencies. When courts or other agencies request the information, it is copied and exchanged on paper. 90% of this paper was originally generated electronically, yet a half-dozen agencies each undertake the labor to scan and index or file this paper in their own systems—and then exchange the data on paper.
- ◆ A bank forecloses on someone's unpaid mortgage, while that same bank issues them a new credit card.

## The Anatomy of a Taxonomy

A complete document taxonomy typically includes: the document source, creator, owner or control point, version number, frequency of update, retention period, effective date of the document for retention calculation, related documents in a process, which version is the official legal copy, indication of content subject to regulatory compliance such as Sarbanes-Oxley, HIPAA, EPA, FDA, personal privacy acts, company confidentiality, and indication of information which potentially could be used in identity theft or corporate fraud.

We do not recommend all of these properties be stored in the document management system for each document, since this additional information overhead could create performance and labor cost issues. Instead, many of these properties are assigned to portions of the classification. For example, most accounting records are subject to a 7-year retention period, based on IRS guidelines, i.e., the retention property is applied to a group of records.

Results Engineering stores these attributes in **Contax**, our web-based document taxonomy system. This system is used to mine relationships in documents, discover duplicate forms, seek process improvement opportunities, assess risks, and avoid compliance penalties. Once the relationships, linkages, and rules are defined, the metadata in Contax is loaded into Records Management, Electronic Document Management, and Workflow systems. It is also used as an on-going reference as active documents evolve, new processes are defined, and new regulatory and compliance requirements arise.

An Enterprise Content Taxonomy facilitates the development and evolution of an organization's Content Management systems, while reducing redundancy, error, labor, and legal exposure. Like doing the occasional inventory, it's a prudent, necessary, and insightful process. ■

Results Engineering is an industry leader in implementing ECM and Workflow Automation solutions. We've been developing tools, techniques, and integrated systems (such as Contax) since the dawn of this industry.