



## Information Classification 101 – Part One: a Primer on Metadata

*As the rate of information growth accelerates, information classification continues to increase in importance as a tool to manage that growth.*

**By, Bob Rogers Chair, SNIA DMF Information Lifecycle Management Initiative.**

Paradoxically, a large percentage of IT people see information classification as difficult, labor intensive, and error prone. This series of articles is intended to clarify some of the issues with the process, identify several of the techniques, and show how each technique applies to the overall objective of classifying or categorizing information objects within the data center. Perhaps one of the more difficult facts to accept is that there is no simple “universal” method of classifying information. The needs of information stakeholders form the basis for the methodology used to get the job done. Although the Information classification process can take many different forms, the one attribute common to every information classification methodology is the necessity for metadata. What is metadata?

The term is used frequently and often without much consideration for its implications, as in “data about data.” The National Information Standards Organization (NISO) defines metadata as “structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource.” The NISO definition introduces the important concept that metadata has a purpose<sup>1</sup>.

There are at least two types of metadata used for information classification; the first type is the criteria or set of rules to apply to the information source; for example, search arguments. The second type of metadata is an index, or an extract, or a summary, of the information source. Both of these types of metadata are “descriptive” because they describe a resource for discovery or identification. A major step forward in business information classification was announced by the Securities and Exchange Commission (SEC) in

December 2008, when it required the adoption of the eXtensible Business Reporting Language (XBRL). XBRL is an example of an international standard for descriptive metadata used to communicate business and financial information. Descriptive metadata is not the only type; NISO has defined other forms including “structural” and “administrative” metadata.

Structural metadata is an interesting concept because it defines the relationships of a compound object. Most databases are a compound object made up of data files, tables, indices, logs, and other entities. (Most storage people figured out the compound nature of a database the hard way, when something was done to one physical component and the whole thing quit working.) Structural metadata may include more than just physical relationships; for example, the relationships between bits of knowledge represented by structural metadata might produce important information without the need to access the data that it describes. Storage professionals are much more familiar with “Administrative metadata.” NISO defines it as “information to help manage a resource such as when and how it was created, file type, and other technical information.”

File system metadata is a good example of administrative metadata. Administrative metadata is not limited to that produced by a system; it could also be the metadata produced by a person such as authorship or other document properties. Yet another example of administrative metadata describes the rights of a user or organization to access information.

Recall that this article began with the assertion that



there is no simple, universal, method of classifying information. Someone, somewhere within your organization, has to define the “rules of identification,” or the descriptive metadata. Equally important, is the metadata for the right to use those rules of identification. (Does anyone remember that passport and customs information of the U.S. presidential candidates were leaked to the press last year? Someone had misused the right to search confidential government data.) Search engines represent the most popular method of classifying information. According to ComScore’s qSearch, 221.2 Europeans conducted 24.6 billion searches during the month of March 2008<sup>2</sup>. However, popular as they are, internet search engines are not the way that corporate information is usually classified.

Few companies knowingly allow public crawlers to index private corporate websites. There are many different brands of search engines designed to traverse corporate file systems; indexing and categorizing everything they find. The descriptive metadata; the rules of interpretation are key to differentiating between valuable information and worthless “noise.”

Content management systems and preservation repositories are usually the choice of enterprises for critical business records. Both create indices of information for search and discovery. However, these are resources that manage only a small fraction of the total of a companies’ electronically stored information<sup>1</sup> or “ESI.” Furthermore, in most cases, an information object is rarely “born” into a content management system or a preservation repository. Today’s enterprise needs to classify an information object from the moment of creation to ensure that it is properly managed throughout its lifecycle. That philosophy requires a well-defined set of descriptive metadata.

No single one of these solutions is likely to address all

the needs of all the information stakeholders. Consider employing multiple methods of classifying information if it seems appropriate.

If information classification is an emerging priority in your organization, the first step should be to obtain “C-Level” sponsorship. As you consider an information classification project, think of who the stakeholders are. You might be surprised by the number people and their scope of responsibility (and make no mistake; the storage team is one of many stakeholders). Can you describe the characteristics of the information to be classified and its management needs? Collaboration is the key to understanding these requirements.

The SNIA DMF Information Lifecycle Management Initiative (ILMI) in conjunction with SNIA’s End User Community has two projects underway to assist in the creation and management of metadata for classification. The ILM Maturity Model is intended to be a guide for end users to interpret and improve IT policies and procedures related to ILM (including the creation and management of metadata). The second project is focused on Storage Service Attributes; its goal is to identify a “standard” set of metrics (i.e., descriptive metadata) for ILM.

The Data Management Forum is also wrapping up work on a White Paper entitled “The Vision of Information-Centric Management” which describes the collaboration process in much greater detail. Visit the DMF ILMI webpage ([www.snia.org/forums/dmf/programs/ilmi](http://www.snia.org/forums/dmf/programs/ilmi)) for more information on our programs and publications. Or join our discussions on these and other topics on the DMF Community site ([community.snia-dmf.org](http://community.snia-dmf.org)). We welcome your experience, input and insight.

Next article: “Building the Collaborative Team”

<sup>1</sup> Understanding Metadata; NISO Press; <http://www.niso.org/publications/press/UnderstandingMetadata.pdf>;2004.

<sup>2</sup> “comScore Releases March 2008 European Search Rankings”; comScore.com; <http://www.comscore.com/press/release.asp?press=2208>