



WHITE PAPER
October 2005

Best practices in data classification for information lifecycle management

ABSTRACT

This paper describes best practices in classifying data from the perspective of information lifecycle management (ILM). It is intended as a high-level introduction to issues and considerations for IT organizations that are considering implementation of a more formal approach to data classification.

1 Executive summary	2
2 Introduction	3
3 Why classify data?	3
4 Data classification — an information lifecycle management perspective.....	3
5 Data classification criteria	4
6 Does data classification change with data use?	5
7 Classifying your data — the process	6
8 Data classification benefits	8

1 Executive summary

Information lifecycle management (ILM) is a sustainable storage strategy that balances the cost of storing and managing information with its business value. A well-executed ILM strategy will result in a more agile organization, reduce business risk and drive down both storage unit and storage management costs.

Data classification is a key underlying activity that enables an economically feasible implementation of ILM. It is a process that defines, at a minimum, the performance, recovery and retrieval characteristics of an enterprise's different sets of data, grouping them into logical categories to allow storage management activities to be mass-customized. Without well designed and carefully managed data taxonomy, the cost reductions and service improvements promised by an information lifecycle management strategy will not be broadly achievable.

The best practices data classification approach outlined in this paper will provide other benefits. The approach recommended will clarify the linkage of business processes to data, helping align IT with business priorities, and will help to delineate the worth of information by providing clear categories of information value. In the long run, a well thought out data classification initiative will begin to move the overall business toward the capability to use knowledge as a competitive advantage.

While we have identified several best practices in this paper, a few key behaviors appear most important in providing lasting business impact:

Aligning the data classification process with business processes

An effective data classification initiative forces the alignment of storage management with business processes. If tight alignment already exists, it facilitates the data classification initiative. If not, the data classification project forces the alignment. In either case, the alignment is enhanced with additional information linking business priorities to the management of information assets.

Recognizing the value of data classification within, and beyond, information lifecycle management

Data classification has intrinsic value, enables ILM and can provide a basis for an organization's maturity in information asset management. An organization obtains the maximum business impact of a data classification initiative through the careful balancing of these perspectives.

Understanding the dynamic nature of data classification

Data classification is dynamic at several levels. Classifications change as data objects move from one class to another. The taxonomy will likely change as business strategies, business structures and external forces change. It is important for the team engaged in developing the initial architecture to have a clear understanding of the potential for change and to imbed the appropriate processes to manage it.

Executing a well thought out initial project, coupled with effective change management

In some ways, data classification is like any other IT/business project. It requires careful planning, methodical implementation and ongoing process management. Commitment to the project, assignment of appropriate resources, executive oversight and clarity in expectations are critical to the success of the initiative.

Data classification is becoming a necessity in controlling the cost of storage and storage management. The initiative can reap significant business rewards and can enable a business to be more agile and competitive. Doing it right counts.

2 Introduction

It is somewhat obvious that all of your data does not have the same importance to your business. Some of it is indeed mission critical; some of it is of limited or temporary value. One thing that is certainly true about data is that it is growing at very high, often exponential, rates. Your IT organization must have a process to align the value of data with the cost of storing and managing it. This process starts with a clear understanding of the business uses of data and provides a mechanism for storing it according to requirements established by business priorities. Optimizing the relationship between the cost of storing and managing data and the delivery of service levels for data access, recovery and discovery¹ is the objective of implementing an information lifecycle management (ILM) strategy. The foundation of an ILM implementation is the taxonomy established to classify data.

- **Information lifecycle management (ILM)** is a sustainable storage strategy that balances the cost of storing and managing information with its changing business value. ILM provides a practical methodology for aligning storage costs with business priorities.
- **Data classification** is a process that defines the access, recovery and discovery characteristics of an enterprise's different sets of data, grouping them into logical categories to facilitate business objectives.

3 Why classify data?

A data taxonomy serves multiple purposes. Effectively implemented, the taxonomy provides the cornerstone of an ILM strategy, supports linkage to other IT infrastructure issues (such as security) and can impart broader enterprise value as a basis of information asset management.

Information lifecycle management (ILM) is entirely dependent on effective data taxonomy. The taxonomy defines classes of data, each of which will be managed with a set of rules or policies. Without a solid, business-based taxonomy, the benefits of information lifecycle management cannot be achieved. The reason why data classification is necessary for ILM is relatively simple: the rules that define the information lifecycle management strategy must operate on classes of data to be economically possible. The alternative to classification is having as many rules as there are data objects. In an enterprise, this is a very unattractive proposition. Thus we seek to mass-customize our approach to managing data by grouping it into classes that have similar management requirements.

Organizations creating an initial data taxonomy must carefully balance the number of data classes against the cost of maintaining and managing them. This generally can be handled in a manner similar to any mass-customization trade-off decision.

4 Data classification — an information lifecycle management perspective

In much of the currently available literature the information lifecycle management deployment process is described as having five key steps:

- Classify or categorize data
- Relate business rules to data classes
- Determine service levels
- Establish tiered services
- Choose appropriate products, including management tools and infrastructure

While the steps are fairly straightforward and seem satisfactory, we are very concerned about their order. Our experience indicates that best practice is focused first and foremost on business processes and applications.

¹ Here we borrow a term from the language of compliance. "Discovery" embodies the contextual access to information required to meet a legal request for information about a subject or a business relationship. It is a content-based form of retrieval from an archive.

We believe that the best practice approach involves:

- Providing that service level agreements include data retention/disposal, access, recovery, discovery and cost objectives.
- Developing a horizontal view across all existing service level agreements to establish a “catalog” of standard service offerings. These (for example) might include an offering that provides 100 percent access/availability, sub-second recovery, and seven-year retention with cross-enterprise, one-day discovery on any topic at a fixed service cost.
- Once these “standard service offerings” are established, data classification and rule development is essentially done, and establishing the appropriate infrastructure and management processes to meet service level objectives follows.

Of course, not every enterprise has pervasive service level agreements, and even if they do, some of the criteria recommended above may not be included. So while we believe that the data classification taxonomy (and hence ILM) will benefit from a service-level-agreement-driven approach, we recognize that not every organization can get there quickly. Further, an initial classification based on judgment and experience will bring value to information lifecycle management. As the implementation evolves, we would expect most organizations to evolve to a maturity level where data classification is service-level-agreement-driven.

The current state of data classification is largely a byproduct of hierarchical storage management (HSM) implementations, where the pervasive classification criterion was the age of the data. Today, data is generally classified based on access or availability and recovery requirements, and cost. With the advent of additional compliance requirements, retention, discovery and alterability requirements have changed. In general, these two approaches to classifying data have not been broadly integrated.

The future state of data classification will involve a much broader perspective, motivated by business requirements. The taxonomy will serve several needs (including ILM, enterprise content management, compliance, data mining and decision support, and security) and will emerge as the basis of enterprise information asset management.

5 Data classification criteria

The overriding goal of data classification for ILM is assigning business value to different sets of data. As discussed in the previous section, this is best accomplished in a high-process maturity IT environment, where service level agreements have been mass-customized into standard service offerings. Still, the service level agreements require the benefit of an initial classification model to provide that service level agreements contain the right attributes in the first place.

In regard to data classification and its organizational impact, we propose that business users mostly care about a fairly straightforward issue: “How rapidly and accurately can I get the information I need, and what will it cost?” With the increase in focus on compliance and the related issues of retention, disposal and discovery, this has become only slightly more complex from a user’s perspective. It’s typically up to IT to translate these somewhat simple needs into a set of criteria to manage data, and finally into storage management methodology.

Experience indicates that these considerations should likely include:

Access and initial placement define the performance characteristics experienced by the user of the data in production applications.

- Where is the data placed initially? How rapidly is the information accessible at its point of creation? What are the initial performance requirements?
- What is the trigger event that changes the status of the data and results in a change of requirements or a change of classification (and can result in a change of placement)?

Recovery and protection define what happens when a business data object is damaged or destroyed.

- In the event of a failure of the primary data availability, how quickly can the data be recovered?
- What methods and levels of data protection are required to meet the recovery needs?

Discovery, retention and disposal define the service characteristics of data that has been archived.

- How quickly can archived data be accessed?
- How broadly can contextual search capability be applied?
- For how long and under what conditions (e.g. alterability and migration management considerations) will data be stored?
- What is the trigger event or time frame that initiates disposal? How is disposal accomplished? How is the audit trail managed?

Security defines the overall management of the data from the perspective of protecting it from unauthorized use.

- What access control, physical protection and encryption will be employed?
- How will this (or will this) change as the status of the data changes?

In combination, these “groups” of criteria define the key characteristics of information through its lifecycle, hence through ILM.

6 Does data classification change with data use?

The simple answer is, of course. The objective of the information lifecycle management strategy implies that service levels will change. Dynamic data classification is the term we use to describe the need for adaptive classifications and classification taxonomy. There are at least three levels of change, driven by trigger events, external forces or time, that can make the classification of data somewhat challenging. We categorize these classification changes as follows:

Changes embodied in a classification or a service level

The classifications and service levels themselves will sometimes contain a temporal or event-driven change to the handling of data. (e.g. “Archive the order data objects after the order is fulfilled, but no sooner than 60 days.”) In addition, the business calendar creates management aspects that are temporal, related to monthly processing cycles or year-end processing. These also impact service level requirements, and should be embodied in the classification system.

Changes to the purpose or use of the data

As the usage of data changes, the data may be subject to a different classification. This generally happens when the business process using the data objects changes, and (ideally) is triggered by a different service level agreement (or different requirements defined in agreements). An example would be POS data that is migrated to a data warehouse for business intelligence application; this may be considered a typical change (i.e. active to archive). Perhaps a more interesting change occurs when data is archived under normal operating processes and later becomes more critical. For example, customer transaction data, sent to deep archive after the initial terms of the service agreement, might again become business critical in a new customer self-service application.

Changes to the classification taxonomy

External events, such as changes to regulations, business structure, technology or business strategy may change the basic structure of the data classification system. The change-management system implemented during the development of the classification taxonomy is critical to its long-term viability.

The high potential for change indicates the need for a data classification approach that is adaptive. The process should account for change, and the initial project should focus on uncovering the need for adaptive classification via careful analysis of business workflows. The need for solid change management processes associated with the management of the taxonomy is also indicated. Finally, the IT and user organizations must be sure that they are aware of fundamental shifts in regulations, and business strategies and structure, to provide that the taxonomy is adapted. A formal mechanism and review team should be implemented to provide that the data taxonomy is current with the business environment and external factors.

7 Classifying your data — the process

- 1. Establish a clear goal for the project and ongoing data classification process ownership and stewardship. Create a balance in the goal between driving forward toward information asset management and attempting to do too much.** One of the obvious — but critical — best practices we have found in successful data classification projects is the development of clear goals, expectations and resource assignments. This not only involves the initial responsibilities, but also should include the ongoing responsibilities for data classification process ownership and stewardship. Throughout this paper we have been focused on an information lifecycle management perspective with some extensions, and have advised that a better approach is to look forward to information asset management. However, we must caution against trying to do too much. Best practice is to start with a solid and comprehensive information lifecycle management goal, with a few extensions, and then institute change management processes to allow for adaptation.
- 2. Establish a cross-functional project team involving business process owners and IT. Seek external help, but make sure that process and project ownership is internal.** After establishing the goals, a cross-functional project team is chartered from business and IT with members having an intimate knowledge of business data objects and their linkages to business processes. External resources can be of real benefit in this process. If external resources are involved, it is of extreme importance to provide that their role is clear. Generally, the most successful data classification projects are internal projects, not “consultant” projects. The role of outside resources is best defined as one of assistance. Outside resources can provide guidance, methodology, process methodology, skills transfer and training. It must be clear to all involved that the project is a company project.

- 3. Complete a broad audit of current data classification practices both from a logical (relating business data objects to business processes) and physical (current policy, placement and technology) perspective.** Engage the team in a broad effort to document the current state from two perspectives: a physical perspective classifying business data objects based on their current technology and policy, and a logical perspective relating business data objects to business processes. Depending on your maturity in the business process/infrastructure management continuum, this effort should provide (or you should already have) a model tying business processes to business data objects. This process may involve interviews, group discussions or the use of modeling tools to provide a complete and comprehensive audit is accomplished. Visualization tools are frequently used to aid in understanding complex data and process relationships. Several organizations have service offerings that can provide assistance for this effort.
- 4. Learn from external sources — associations and industry analysts have done a good job of addressing the process of data classification.** Learn from those who have gone before. There are many sources of experience and information on the subject of data classification and the more general academic science of taxonomy. These include associations (like SNIA), analysts and vendors that offer general classification models. While all of these are valuable, it is important that you tailor your model to your specific business needs. While 80 percent of the work of defining the taxonomy may be generic, the tailored 20 percent holds much of the leverage for your organization.
- 5. Create a proof of concept classification taxonomy and optimize the number of classes in the taxonomy.** Create an initial (proof of concept) model to provide linkage to service level agreements and business process alignment. During this phase of deployment, the optimum number of data classes is initially determined. We have found organizations that function quite well with five to 10 data classes; most seem to operate in the 10 to 15 range and a few have 30 or more. The important thing is not the number but the mass-customization trade-off between having a more granular taxonomy and the cost of maintaining it. Add a data class when the cost of storing and managing the relevant business data objects in an existing class which exceeds service requirements, is greater than the cost of implementing and managing a new class (and a new set of policies and procedures to support the new class). As a rule, data taxonomies should be as simple as possible, but no more.
- 6. Link the initial (proof-of-concept) model to the rules and infrastructure plans to provide that the classification system is actionable.** Engage in a careful test of the proof-of-concept classification model and rules to provide that the service levels are actionable. Link the initial model to the rules and infrastructure plans. It is costly and wasteful to have categories you will never serve or to have multiple classes that will be treated the same forever.
- 7. Classify data objects, building a repository of the classification information and linkage to business processes.** Classify data objects and build a repository relating the data objects to the classes, and the classes to policies/service objectives. Implement the data classification in a stepwise fashion to provide that the process meets expectations and is performing to specifications. Monitor the performance of the taxonomy in production, carefully assessing its implementation, the application of the appropriate rules and the business impact of the classification.

8. **Establish an ongoing review process and team to provide that the taxonomy remains relevant.** Implement a change-management process for the taxonomy and the data classification repository. Establish maintenance and governance processes, automating change management as possible. Keep a history of changes to the taxonomy, and build a tracking process for the data class assignments themselves. Be aware that change management is often the most expensive part of an ongoing classification initiative. As mentioned, a review mechanism and team should be implemented to provide that the data taxonomy is current with the business environment and external factors.
9. **After the data classification project is completed, address the next steps of ILM.** Specifically, these include designing and implementing a rule-based storage management process that enables serving the data categories, and the implementation of information lifecycle management tools and appropriate infrastructure.

8 Data classification benefits

In a broad sense, data classification enables information lifecycle management and information asset management. One could easily argue that neither is economically possible without the solid underpinning of a high-quality data classification taxonomy, coupled with solid data classification project implementation and ongoing change management. Data classification is the enabler of a mass-customized approach to implementing information lifecycle management or storage optimization rules. As such, data classification drives organizational agility by facilitating a set of processes that are regularly reviewed and refined. It enables a categorized approach to managing to the requirements of disparate business processes, and business processes that share business data objects. Further, from an information lifecycle management perspective, data classification can reduce business risks by providing that the appropriate data is managed with the appropriate standards for compliance, retention, protection and security.

Key benefits

- Mass-customization of rules — acting on a limited number of data categories
- Critical to implementing an information lifecycle management (or storage optimization) strategy — resulting in a more agile organization, reducing business risks and driving down storage management costs
- Providing linkage of business processes to business data objects
- Delineating the worth of information by providing clear categories of value
- Moving the organization toward the capability to use knowledge as a competitive advantage

In a narrower sense, data classification and the execution of a data classification project have several intrinsic benefits. These include:

- Providing a clear visualization of data and information categories and providing that the linkage between business processes and business data objects is understood
- Providing a clear view of the value of the data and information your organization stores and manages, enabling the IT organization and the business to begin to understand the value of information in a quantitative sense
- Potentially lowering the cost of storage and storage management, while providing service levels that meet or exceed business process requirements by proactively identifying the actual management needs of all data

Finally, the act of data classification itself is a step toward organizational understanding of the value and management of knowledge — one of the most important, yet largely unrecognized, assets in any company's drive for competitive advantage.