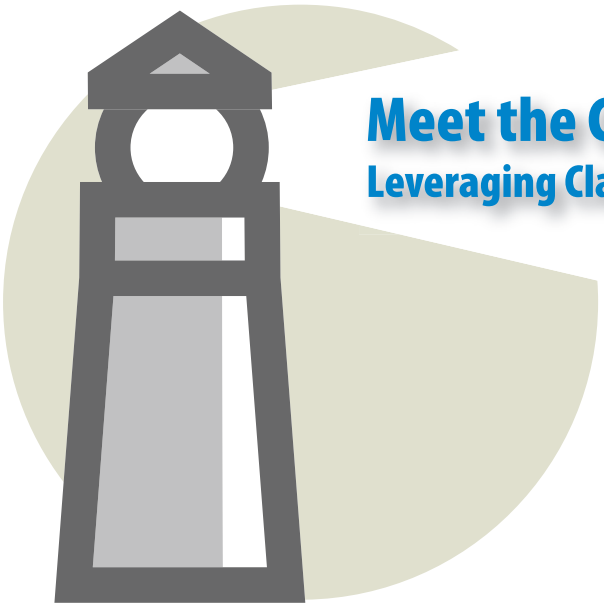
White Paper



Meet the Content Tsunami Head On

Leveraging Classification for Compliant Information Management

Prepared by
Cohasset Associates, Inc.

June 2009

Abstract

The world is awash in information. Exponential growth in the creation and storage of e-mail and other unstructured electronic business content has engulfed workers and swamped information systems, resulting in data and productivity losses, security breaches, and bloated infrastructures. Organizations are struggling to manage their business records amid increasingly complex legal and regulatory compliance pressures.

This white paper is an urgent call for the promulgation of enterprise standards, strategically developed information architectures, and widespread adoption of advanced classification best practices and technologies that can scale and adapt to meet the information governance challenges at hand.

Table of Contents

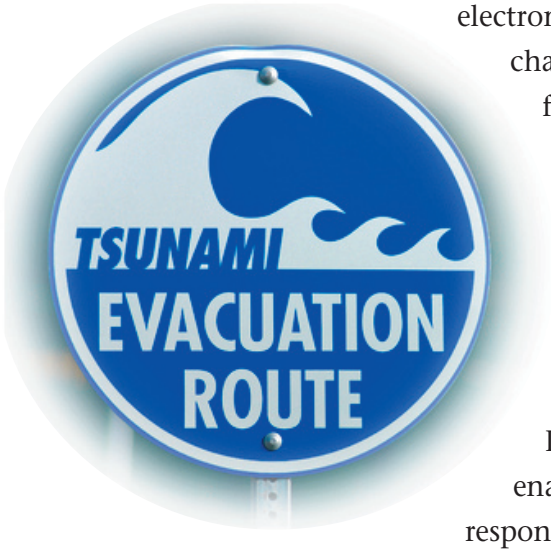
Prepared by:

Cohasset Associates, Inc.
3806 Lake Point Tower
505 North Lake Shore Drive
Chicago, IL 60644 USA
www.cohasset.com

312-527-1550

1.0

1. The Perfect Storm



The first decade of the 21st century has been characterized by a tsunami of electronically produced and stored information ("ESI") and sweeping changes across the legal and regulatory landscape. High profile corporate failures, an alphabet soup of new regulations, and stunning court decisions[1] have converged in heightened demands for increased control and accountability from government regulators, industry and standards groups, as well as the general public. Information privacy, security, and records compliance have become critical, high risk business issues for executive management and government leaders alike.

In the United States, revisions to the Federal Rules of Civil Procedure enacted in December 2006, raised the threshold and shortened the response time for organizations to proactively identify the whereabouts of relevant ESI regardless of format or storage media. More than 300 court decisions involving ESI have been rendered in the last two years, bearing witness to serious inadequacies in many current information management and search practices. And while awareness of the importance of the processes by which electronic records are managed to future litigation is high,[2] performance still falls short of acceptable levels.

Regulatory agencies and courts now have tougher enforcement options at their disposal, especially with regard to data security and privacy concerns. Identifying and integrating compliance-based requirements into the enterprise infrastructure has become a driving force for many Information Technology (IT) operations.

Organizations of all sizes need a robust information governance capability for organizing structured and unstructured information assets into meaningful and sustainable categories without a significant reliance on users – especially in an environment of reorganization and downsizing. Taxonomies and classification schemes must reflect the working worlds of the organizations for which they are developed and implemented. Working worlds (which combine records and information management requirements, stakeholder interests, internal codes of conduct, etc.) continue to change and evolve and so must our strategies and tool sets.

2. Tipping Point

Rather than stoop to the lowest convenience level of information governance and hope for the best, every organization should reach towards high level semantic-based means of categorizing, searching and leveraging their business content.

Our reliance on large complex supporting infrastructures and electronic business processes to transact commerce and deliver government services has had widespread and unintended consequences. This includes debilitating challenges to identifying, tracking and managing records which serve as evidence of business activity. It is this evidence of decisions made, actions taken, and goods bought and sold, that serve as the "institutional memory" that informs operations and decision-making and ensures accountability to current and future stakeholders. Awash with information, our large scale organizational memory is at risk. Government and business leaders must determine how to harness the power of computers to help fix the serious problems that it has enabled us to create.

Rather than stoop to the lowest convenience level of information governance and hope for the best, every organization should reach towards high level semantic-based means of categorizing, searching and leveraging their business content. To manage information in a sustainable manner that doesn't impede those who produce business records while assisting those who need to access the information over time, an enterprise approach to organizing information is imperative. The heavy lifting of categorizing and classifying information must become integrated into the working digital infrastructures upon which modern business and government depend.

Businesses across the globe have reached the tipping point where legacy practices from the paper world simply do not scale to required levels of control over ESI due to cost and consistency. Tentative forays into electronic records management (ERM) reveal some hopeful trends but still fall far short of the mark[3] of demonstrating good faith efforts to establish and execute defensible policies for the lifecycle management of records and information. Saving everything is expensive, inefficient and extremely risky.

It's time to renounce user-based classification and strive towards a new world order where the role of the end user is minimized for records management, classification decisions and information archiving. This means that organizations must embrace a new set of best practices, tools and technologies that can deliver cost-effective, consistent and comprehensive compliance enforcement against the background of dynamic requirements defined by the courts, federal and state legislatures, regulatory agencies and other key stakeholders.

3. The Role of Classification in Compliant Information Management

The current regulatory and economic environment brings added pressure to bear on our understanding of how classification structures fit within metadata models as government jurisdictions around the world have begun to impose mandatory metadata requirements.

Requirements for records to be created and managed vary according to the type of organization and the legal and social context in which it operates. Strategies for how to categorize and manage information assets should be based on a risk and value-based assessment of the regulatory environment as well as operational and accountability requirements associated with routine business activities. In addition to mandatory requirements, best practice and other voluntary protocols may apply. Lifecycle requirements will pertain to records, archives, security/privacy, electronic commerce, rights of access as well as data protection.

One key component of an information architecture which helps to provide stability and consistency to the job of organizing information is **taxonomy.** Taxonomy is a hierarchical or polyhierarchical listing of topics or subject categories that provides the means to:[4]

1. Know where to file information correctly,
2. Retrieve information easily when needed, and
3. Meet legislative, compliance and business objectives.

Taxonomy offers a structured path to navigate through a content collection within a collaborative environment such as shared directories, providing an intuitive way for users to locate and access information. Rather than having to design a query and then review the results, users drill down through the categories and subcategories of the taxonomy until they find the relevant concept or document. The taxonomy can be used to limit access to specific categories and sub-categories depending on security rights and authorities. Within a single organization, multiple taxonomies are likely to be used.

The systematic identification and arrangement of information into categories (taxonomies) is referred to as **classification.**[5] Classification relies on logically structured conventions, methods, and procedural criteria. Common criteria used in record and information classification schemes include business activity, subject matter, time, document types and case-based terms such as investigations, matters, claims, or accidents. Classification criteria or facets are a subset of recordkeeping metadata – that is structured data that describes the context, content and structure of information and allows its use, location, management, control and preservation over time.

The current regulatory and economic environment brings added pressure to bear on our understanding of how classification structures fit within metadata models as government jurisdictions around the world have begun to impose mandatory metadata requirements. Courts are taking an increasing interest in the metadata management for repositories containing preserved content.[6]

Relationships between taxonomy terms and other metadata must be acknowledged for compliant information management to be possible. That means the relationship between classification facets and security/access and disposition rules that apply to business records must be made explicit and enforceable across the enterprise.

Classification systems that organize work processes and their inputs/outputs are powerful technologies vital to the management of commercial and government enterprises. They serve as cornerstones of working infrastructures and become deeply interconnected and embedded into large information systems and global enterprises. These schemas are essential for describing business processes and work practices, quality standards, recordkeeping requirements, ethical behavior protocols as well as the role of people and their value to the organization.

4. Classification as Human Activity

Classification is an inherently human activity. From the informal activity of sorting laundry, creating a grocery list or managing a music or wine collection, all the way up to and including the formalized and highly sophisticated mapping of human genes,[7] people and organizations are driven to categorize and organize information and objects around them.

The key advantage that humans bring to the classification process is their natural ability to discern context – to turn the circumstances related to information or an object into meaning. In the case of records management, this means the "business context" in which the record was created, received, and used, including the business process of which the transaction is part, the date and time of the transaction and the participants in the transaction.[8] Along with content and structure, context is a key element of accountability and imbues authority to the information.

5. Why User-Based Classification and Recordkeeping Is No Longer Good Enough

End users often opt out of the recordkeeping process by not capturing records into ECM/EDRM systems or by not capturing sufficient metadata to describe the information objects for future retrievability and production.

For decades, the classification of records into categories based on organizational units or business functions was the job of the Records Manager. Retention was applied, mostly to hard copy inactive records at the file, folder or box level, upon transfer to Records Management control. Now that nearly all business is transacted electronically, Records Managers no longer see, must less are able to exercise control over, the majority of the enterprise records.

Transferring records management duties to end users, however, has proven unsuccessful. Classifying information is tricky and managing electronically stored information is time consuming and overwhelming for most workers. In the absence of enterprise standards and tools, users have been left largely to their own devices to establish filing plans and naming conventions. Taking into account the exponential growth in electronic communications[9] and other unstructured information, it is not surprising that the result has been clutter and chaos.

End users can also be highly resistant to taking on responsibilities outside their primary business tasks and which don't directly relate to their own business processes. In 2002, the U.S. National Archives and Records Administration (NARA) conducted a six month study of user driven records declaration[10] using technology available at the time. In addition to difficulties that users experienced in determining what constituted a "record" and should be captured, there was a significant participation drop-off rate after the initial training period by users who found the process "burdensome."

This and other more recent studies have shown that end users often opt out of the recordkeeping process by not capturing records into ECM/EDRM systems or

by not capturing sufficient metadata to describe the information objects for future retrievability and production. When users "game" the system, official records are left unprotected outside of authorized repositories and the investment in content management technologies is severely compromised.

Even with the best of intentions, studies have shown that humans lack consistency and accuracy in assigning classification categories except at very specialized knowledge levels or with advanced and repeated training. In a 2005 study conducted by the Department of the Navy, only 12.5% of documents were classified with "Exact" accuracy at least 75% of the time.[11] Taking this result into account against a backdrop of shifting legal and regulatory requirements and business conditions highlights limitations to adapt quickly enough to manage the organization's cost and risk exposure.

This is increasingly clear in the area of electronic discovery where the enormous volumes of digital data to be reviewed for relevance in preparing a legal strategy and defense within the time constraints set by the court is often not humanly possible.[12] In a study comparing automated relevance assessment to relevance assessments made by human reviewers, the software, on average, identified more than 95% of the relevant documents compared to an average of 51.1% for the people.[13]

Some organizations have adopted a "save everything" policy because they don't have value-based systems for assessing information and applying rules and/or because they are under intensive litigation or regulatory pressures. Given the exponential growth in unstructured information, especially with regard to e-mail and attachments, this is a risky and costly gamble. It also places an enormous burden on workers who must sift through growing landfills of irrelevant and outdated information.

In the 2008 Workplace Productivity Survey commissioned by LexisNexis, a majority of employees in the legal and professional fields reporting feeling close to a "breaking point" with regard to handling an increase in information flow.[14] In another study, 58% of respondents reported spending more than 25% of their time reviewing irrelevant information as part of their search/locate activities.[15] Users who can not find information in a timely fashion frequently end up re-creating the information they were looking for. This creates an overhead burden, slows down decision-making, and negatively impacts collaboration and communication.

The economic downturn has led to new government regulations, adding pressures to incorporate enforcement protocols into content repositories and workplaces which are suffering from layoffs and low morale. Worker layoffs mean abandoned heaps of unmanaged information and fewer people left to do the remaining work. A 2005 study related to classifying e-mail and other documents left by exited employees revealed a cost of more than $12,000 per employee for manual classification vs. $800-1200 for outsourced auto classification.[16]

Establishing reliable and cost effective means to categorize and ensure the future usability of information assets without a reliance on the time and attention of the average information end user has never been more urgent.

6. The Ascendency of Classification Technologies

As early as 2001 it was reported that the accuracy of automated text classification had reached effectiveness levels comparable to trained professionals and were expected to exceed them.

If we can not rely on the accuracy and consistency of human workers to categorize and organize business records, then we must seek alternative solutions. Significant research and investment in library and information science has been undertaken around the world, and classification technologies have advanced significantly over the past two decades. As early as 2001 it was reported that the accuracy of automated text classification had reached effectiveness levels comparable to trained professionals and were expected to exceed them.[17]

E-mails seized by the Federal Energy Regulatory Commission (FERC) in the Enron investigation provided a rich research data set and an opportunity to compare human (measured by inter-annotator agreement) vs. computer classification. A study undertaken by the University of Sheffield, Department of Computer Science, to sort e-mails into business vs. personal categories and then classify by type produced this finding:

> Given that our inter-annotator agreement statistic tells us that humans only agree on this task 94% of the time, preliminary results with 93% accuracy (the statistic that correlates exactly to agreement) of the automatic methods are encouraging. While more work is necessary to fully evaluate the suitability of this task for application to a machine, the seeds of a fully automated system are sown.[18]

Machine based classification technologies work by systematically analyzing e-mails and documents at the point of creation or receipt, assigning metadata, and filing them into logical and consistent categories. Predefined classes are assigned using taxonomies and defined semantic rules that are based on natural human language. A key advantage of an automated system vs. a manual system is consistency.

Categorizing the same concepts into the same place 24x7 is what computer-based systems do well and which humans do not.

Software solutions that analyze document and message content are able to propose categorizations that can be enhanced through user feedback or through self-training mechanisms. In this way the classification adapts to both the unique nature of each organization's operations and to changing business conditions. Processes to monitor and periodically adjust the categorization parameters should be established at the start so that the automation software can deliver the desired business results and return on investment over time.

Advanced classification solutions provide a unique opportunity for organizations to establish and test enterprise standards for describing their information assets. The data about the information and records created or received and the relationships between records can be routinely analyzed, improved, and leveraged over time to ensure access and proper management.

Enterprise content management (ECM) software providers, such as IBM, have responded to the challenges of information overload and user resistance by integrating varying degrees of computer-based classification and policy enforcement options into their solutions. Options range from fully automated classification to user-based assistance via suggestions and prompts. This enables organizations to select an implementation strategy that meets their risk/cost profile and the maturity of their intellectual architecture (taxonomies, standards, tool sets, retention and disposition rules, etc.).

Electronic recordkeeping applications can also be designed to register or 'declare' records through automatic processes that are transparent to the user of the business system from which it is captured and without the intervention of a records management practitioner. Even where registration is assisted but not totally automated, some of the essential metadata can be automatically derived from the computing and business environment from which the record originates (based on roles, business process, workflow, etc.). The assignment of retention rules based on a predetermined classification or the recognition of specific metadata elements is what makes automated disposition[19] possible.

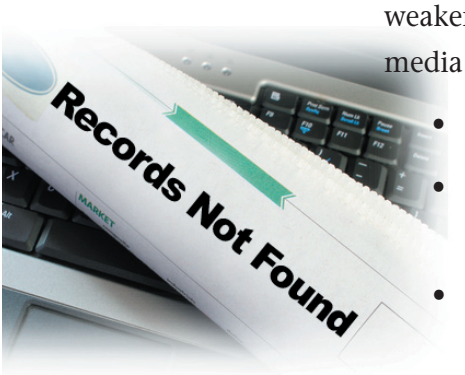
Compliance with all relevant regulations and guidelines is overwhelming. But with the right combination of policies, technologies, and information architecture strategy, organizations can reduce the burden on end users and proactively apply and enforce content-based policies for sustainable records and information management.

Plans, programs and tools must be deployed to speed up the sorting and indexing of content to ensure that compliance requirements are integrated into electronic content producing and recordkeeping systems and the business processes that these systems support.

7. Conclusions

The only good classification is a living classification.[20]

Hardly a week goes by without a records or information management incident or disaster in the news. High profile data losses have shaken consumer confidence, weakened long-standing brands and sullied corporate reputations. Many of these media events have familiar themes:



- The public disclosure of private information,
- The inability to find and produce documents and other information required in litigation or investigations,
- The destruction of records which should have been retained and the retention of records which should have been destroyed.

Trends towards large scale infrastructural systems, distributed work processes and the widespread use of the Internet trends have made compliance harder and increasingly expensive. As new global threats arise and work practices evolve in a volatile economic climate, government and industry will undoubtedly create new regulations and rules to protect business and personal data. Organizations of all types and sizes will continue to operate under immense pressure to demonstrate compliant information practices while minimizing risk and cost for the foreseeable future.

An organization's information architecture should include components which work in combination to provide a systematic and comprehensive capability to organize and access information over time. This capability must accommodate changes in business, legal and regulatory requirements and be able to reach across all domains and organizational units. This demands enterprise-level governance and enforcement.

Line of business managers, systems analysts, as well as records/compliance professionals should continue to educate their executive management about the significant risks and costs associated with continued failures to meet recordkeeping requirements. Whatever the choice of technology or the pace of evolution, every organization must harness the power of computers to appraise, organize, and retrieve content at a pace that can match the rate at which it is received and created.

It's time to meet the content tsunami head on and take immediate action to achieve compliant information and records management. By deploying advanced classification technologies that can apply a structured collection of terms and guidelines based on content, critical business information can be managed in a manner that facilitates its discovery, interpretation and use to the greatest extent possible.

End Notes

1. A representative list includes the American Recovery and Reinvestment Act of 2009, the Fair Credit Reporting Act, the Health Insurance Portability and Accountability Act, the Sarbanes-Oxley Act of 2002, 2006 Revision to the Federal Rules of Civil Procedure (FRCP), and the decisions handed down in Coleman (Parent) Holdings, Inc. v. Morgan Stanley & Co, Inc., and Zubulake vs. UBS Warburg.
2. According to the 2009 Cohasset/ARMA International Electronic Records Management (ERM) Survey, 98% of respondents rate the importance of the process of managing electronic records to future litigation as "Very important" (71%), "Quite important" (19%), or "Important" (8%).
3. A white paper exploring results of the 2009 Cohasset/ARMA International survey on Electronic Records Management (ERM) will be published in July 2009 and available at www.cohasset.com.
4. Robertson, J. (2004). *Rolling out a Records Management System*. Retrieved June 1, 2009, from Step Designs Pty Ltd's Web site: www.steptwo.com.au/papers/kmc_recordsmanagement/index.html
5. The ISO 15489 standard on Records Management defines classification as the systematic identification and arrangement of business activities and/or records into categories according to logically structured conventions, methods, and procedural rules represented in a classification system.
6. Hedges, R. and Gable, J. (2009). *Metadata Management: The New ERM Frontier?* [PowerPoint slides]. Presented at 2009 Managing Electronic Records (MER) Management Conference. For access to streaming video of the presentation go to www.mereducation.com.
7. Learn about the Human Genome Project at www.ornl.gov/sci/techresources/Human_Genome/project/about.shtml
8. See ISO 15489, 7.2.1 b, Characteristics of a record.
9. IDC, Worldwide E-mail Usage 2007 – 2011, (2007). This publication reported worldwide person-to-person business e-mail volumes grew from 2.51 to 6.3 trillion messages annually during the period 2001–2007.
10. *NARA RMA 2000 Data Analysis Report,* December 2001.
11. Winters, J. and DiMartino, A. (2005) Development and Evaluation of a Taxonomy for Human Performance Measures. *Department of the Navy – Human Systems Performance Assessment Capability.*
12. Readers are encouraged to download *The Sedona Conference® Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery, 2007,* from www.thesedonaconference.com.

13. Kershaw. A. (2005) Automated Document Review Provides Its Reliability. *Digital Discovery and e-Evidence,* Vol .5, No. 11.

14. Lexis Nexis (2008). *Lexis Nexis Workplace Productivity Survey.* Retrieved from www.lexisnexis.com/literature/pdf/Workplace_Productivity_Survey_Results.

15. AIIM International (2008). *AIIM Findability Study,* from Intelligence Quarterly, Q2 2008.

16. Session Number T027 (2005). *Classifying e-Records of Exited Employees: Case Study Using an Auto-Classification Tool.* 2005 ARMA International Conference.

17. Sebastiani, F. (2001). *Machine Learning in Automated Text Categorization.* Consiglio Nazionale delle Ricerche, Italy. Retrieved from arxiv.org/abs/cs.IR/0110053.

18. Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, Sydney, July 2006.

19. The ISO 15489 standard on Records Management defines disposition as the range of processes associated with implementing records retention, destruction or transfer decisions which are documented in disposition authorities or other instruments.

20. Bowker, Geoffrey C. and Susan Leigh Star. (2000). Sorting Things Out: Classification and its Consequences. Cambridge, MA: MIT Press. 326.

About Cohasset Associates, Inc.

Cohasset Associates (www.cohasset.com) is recognized as one of the foremost consulting firms specializing in document-based information management. Now in its fourth decade of serving clients throughout the United States, Cohasset Associates provides award-winning professional services in three areas: management consulting, education and publishing.

Consulting: The focus of Cohasset Associates' consulting practice is improving the programs, processes, and systems that manage document-based content. This ranges from establishing effective corporate records management programs to planning state-of-the-art electronic content and records management systems. With its unique combination of records management, legal and technical skills, together with its extensive problem-solving experience, Cohasset works to provide clients with cost-effective solutions that will achieve their business objectives and meet their legal/regulatory responsibilities.

Education: Cohasset Associates is renown for excellence in education. Its focus is organizing and presenting the annual national conference on Managing Electronic Records (MER) – providing special emphasis on legal, technical and operational issues (www.merconference.com).

Publishing: Cohasset Associates authors thought-leadership papers (www.cohasset.com/whitepapers.html) and conducts the definitive survey research on electronic records management (www.merresource.com/whitepapers/survey.htm).

Other IBM White Papers Prepared by Cohasset Associates, Inc.

IBM: The Legality of Digital Image Copies of Paper Records
October 2008

This white paper addresses key legal issues and provides guidance to help organizations make informed decisions about converting paper documents to digital image copies.

IBM: No Paper Weight – ROI Assessment
May 2008

This white paper presents the results of a review by Cohasset Associates of a Return on Investment ("ROI") analysis titled, *No Paper Weight*. The analysis was originally researched by IBM Corporation in collaboration with one of its customers, a large regional bank.

Cost Effective Electronic Records Management –
IBM FileNet Records Manager
November 2007

This white paper provides Cohasset Associates' assessment of FileNet Records Manager software and associated products with regard to the capabilities these products provide for improving the effectiveness and efficiency of managing electronic records.

White papers are available for download at www.cohasset.com.