Archiving & Retention Risk Management

Michael McCreary
Chief Operating Officer
Rational Retention, LLC
mmccreary@rationalretention.com

January 4, 2009

Introduction

The vast majority of information in today's corporation is generated and stored electronically; a significant and rapidly growing percentage is unstructured. To date, companies have had little ability to control how unstructured information is created, saved, copied and distributed inside or outside the firm. The lack of visibility and control represents an expanding risk to the organization. In addition, the December 2006 amendments to the Federal Rules of Civil Procedure have broadened what information may be discoverable to all Electronically Stored Information (ESI), regardless of location or form. This event was the tipping point that finally captured the attention of the GC's office. The corporate liability for improper information management has moved well beyond simple business disruption to potentially hampering a company's ability to defend itself. Missteps in document preservation have cost organizations tens of millions of dollars in sanctions, fines, reputational capital and have turned the tide of litigation. It's not surprising that beleaguered General Counsels and Chief Information Officers have turned to retention policies ranging from "archive everything," which is rapidly becoming cost prohibitive, to "delete as much as possible as quickly as possible". Such policies do not effectively manage risk nor do they ensure information that is important to the corporation is retained long enough to deliver its value. Fueled by technology vendors selling both storage hardware and archiving software, the former "archive everything" approach has been gaining favor beyond just the most risk adverse and litigation weary firms. Until recently, archiving everything appeared to be the only method that offered a safe haven from sanctions. However, recent advances in search, categorization and document management technology have made it possible for companies to move beyond the archiving everything mindset to a model that is far more targeted and ensures that the right documents are kept for the right amount of time, and that documents which have reached their end-of-life are fully and defensibly deleted. Doing so enables companies to defend themselves effectively and efficiently without increasing costs and risk.

Policy Drivers

Information risk management policy must be grounded in the law and informed by business reality. For many firms, information risk management has been fragmented with limited central governance, disparate policies and inconsistent procedures. Departmentally driven, with IT focused on costs and security and records managers focused on paper and retention schedules, scant attention has been paid to the broader electronic information lifecycle. However

since 2006 and the subsequent focus on ESI, such a disjointed approach is simply no longer supportable. Information risk management policies and procedures must now go beyond just security and basic retention schedules – they must now include defensible processes for holding, collecting and producing documents for investigation or litigation. Organizations wanting to attack the root causes of information risk must now take a holistic, long term view of retention governance, processes and tools. Strategy and policy clearly must address the challenges of today; however, in doing so, they must not inadvertently create untenable long-term costs and risks. Keeping every document and every email with no real strategy for eventual destruction is a perfect example of short- term, reactive, policies that may well prove very costly in the years to come.

Legal Requirements

For many types of corporate documents the laws and regulations are silent with respect to retention and the value of keeping those documents beyond their immediate use is questionable. For those documents that firms are required by statute or regulation to keep the retention periods are generally well defined and only a small number are required to be kept in perpetuity. In the seminal case Zubulake v. UBS Warburg, LLC, 220 F.R.D. 212, 217 (S.D.N.Y. 2003), it was held that "A corporation need not preserve every shred of paper, every e-mail or electronic document, and every backup tape." While in Wiginton v. Ellis, No. 02 C 6832, 2003 WL 22439856, at *4 (N.D. Ill. October 24, 2003), the court stated that "To hold that a corporation is under a duty to preserve all e-mail potentially relevant to any future litigation would be tantamount to holding that the corporation must preserve all e-mail... Such a proposition is not justified."

Beyond specific case law, the Sedona Conference (www.thesedonaconference.org), arguably the most widely respected legal organization focusing on issues of the law and management of Electronically Stored Information (ESI), has issued several guidelines for organizations facing the increasingly complex challenges of records and information management compliance and litigation readiness. The November 2007 "Sedona Guidelines for Managing Information & Records in the Electronic Age" specifically recommended that:

An organization need not retain all electronic information ever generated or received.

a. Destruction is an acceptable stage in the information life cycle; an organization may destroy or delete electronic information when there is no continuing value or need to retain it.
b. Systematic deletion of electronic information is not synonymous with evidence spoliation.
c. Absent a legal requirement to the contrary, organizations may adopt programs that routinely delete certain recorded communications, such as electronic mail, instant messaging, text messaging and voice-mail.
d. Absent a legal requirement to the contrary, organizations may recycle or destroy hardware or media that contain data retained for business continuation or disaster recovery purposes.
e. Absent a legal requirement to the contrary, organizations may systematically delete or destroy residual, shadowed or deleted data.
f. Absent a legal requirement to the contrary, organizations are not required to preserve metadata; but may find it useful to do so in some instances.

Clearly, firms are not required by law or best practice to archive everything; however, many organizations are in fact adopting policies and tools that either explicitly or implicitly support such a position. One example is the rapid growth in email archiving, whereby a company creates a separate archive of every email that comes in or leaves the organization. Absent well defined policies and tools for segmenting the archive by content and assigning a retention period, these companies are, de-facto, moving towards keeping everything forever.

Preservation Requirements

The core driver behind this mindset and approach is the obligation to respond properly to litigation holds. In the event of an investigation or litigation, companies are explicitly mandated to supersede all normal lifecycle related deletions and hold any potentially responsive documents. Archiving is an attempt to comply by proactively creating a copy of all communications or potentially responsive user files. While a blunt instrument, broad, proactive archiving does help ensure near term compliance with hold

obligations, however long term it is likely to create more problems than it solves.

Industry experts W. Lawrence Wescott II, Esq. and Randolph A. Kahn, Esq. address the hold dilemma directly in their article contained in the August 8, 2008, issue of Computer Technology Review, "The Litigation Hold: Why You Don't Have to Hold Everything". The authors state, "A common reaction of those who have not had experience with electronic discovery is to hold everything. This approach is sometimes taken by outside counsel whose concern is making sure that necessary information is preserved; their concern is winning the suit, not with the costs, inconvenience, or business impact that saving everything can impose. In other cases, hold everything is imposed out of fear—companies don't know where all of their information is, how the information is managed, who manages it, and how to preserve it. In still other instances, not knowing what content exists on company computers may also play a major factor in the blanket preservation approach."

Khan and Wescott further state, "There is never a legal requirement for a litigant to save all company information for a lawsuit". Courts have absolutely accepted and understand the need for a company to destroy documents in the regular course of business. The requisite conditions to do so are:

a. No law or regulation requires its retention or its statutory retention period has expired
b. The document is not subject to litigation hold
c. The document no longer has business value to the company

Recent case law has made it clear that even companies in continuous litigation are not required to keep "every shred of paper, every email or electronic and every back up tape". "Such a rule would cripple large corporations", Zubulake v. UBS Warburg, LLC, 217 F.R.D. 212 (S.D.N.Y. 2004). Furthermore, in Andersen v. United States, 544 U.S. 696 (2005), where Arthur Andersen was accused of misconduct for destroying documents relating to Enron, the Court clearly stated; "[i]t is, of course, not wrongful for a manager to instruct his employees to comply with a valid document retention policy under ordinary circumstances."

It is clear that if an organization can reliably separate and hold potentially responsive documents, it is free to delete nonresponsive documents according to normal business and legal retention schedules. Not only can they delete such documents, for many firms doing so will reduce operational and e-discovery costs and reduce the chance that old documents, taken out of context, are used later to build a distorted picture of events.

Risks of Keeping Everything

Archiving all to ensure holding the few creates substantial costs. In short, needless retention is a waste of scarce business resources and assets. It hinders the ability to quickly locate relevant documents to meet the needs of the business and increases the likelihood that an outdated version is used in error. Rework associated with using the wrong version costs professional and manufacturing firms millions each year, while today's knowledge worker increasingly spends time searching for the right information within the enterprise. The challenge has moved from not enough information to too much. By archiving everything, systems are slowed down and organizations incur massive ongoing storage costs. It's true that unit storage costs continue to decline, however those reductions are not keeping pace with the exploding volume of documents retained.

An often overlooked dimension of archiving is the cost of e-discovery itself. E-discovery costs are essentially volume driven. The process is a funnel whereby huge numbers of documents are loaded in at the top and responsive items come out the bottom. The more that goes in the top the greater the costs and the more time the process takes. By archiving every e-mail or user created document, an organization can expect to see measurable increases in per custodian based volume. In fact, when DuPont went through an enterprise-wide reorganization of its corporate records, the company discovered that more than 50 % of the documents the company gathered for discovery between 1992 and 1994 should not have been retained. It estimated that it had spent an unnecessary $10-12 million in retention and production costs.

Beyond tangible issues of rising storage and e-discovery costs, there lies the intangible risk of unnecessarily keeping and therefore producing documents that may be used out of context against an organization. Documents are sterile, representing only a snapshot of an event through

a single lens. As years go by and memories fade, people are no longer available; in business often the only thing that truly survives are the documents created. Documents without the context provided by clear recollection may be as misleading as a faulty recollection.

Documents kept in the ordinary course of business are exempt from the hearsay rule of evidence and as such are generally admissible in court. In U.S. v. Freidin, 849 F.2d 716, 719-20 (2d Cir. 1988), it was held that the business-records exception favors admission over exclusion of evidence with "any probative value at all." while in Conoco Inc. v. Dep't of Energy, 99 F.3d 387(Fed. Cir. 1997), the court stated that "Because of the generally trustworthiness of regularly kept records and the need for such evidence in many cases, the business records exception has been construed generously in favor of admissibility."

By retaining documents far beyond the dictates of legal, regulatory or business requirements, companies unwittingly expose themselves to sloppy document management practices of the past. Email is of particular concern. People too often treat emails casually, like a phone call or a water cooler conversation. The tenor of these messages may represent more opinion than fact and when taken out of context can be used to construct a distorted reality of the past, benefiting only an adversary.

Risk Management & Policy

When considering these issues, it is critical that organizations move beyond the narrow confines of traditional records management to a broader perspective of information risk management. The traditional records management view is too often focused more on records retention requirements defined by statute or regulation and on the storage and lifecycle of paper documents than on addressing all ESI wherever it resides and across its entire lifecycle. When seen from a risk management perspective, it is clear that in today's environment the greatest risks lie in the inability to control effectively and fully the rapidly expanding morass of emails and documents on desktops, in file shares or document management systems not in the inability to store paper or retain records in accordance with a retention schedule.

Moreover, as companies craft solutions to address the entirety of ESI, they must be keenly aware of potential impact changes, even minor ones, will have on employee work habits and productivity. Too often we hear that the poor success rate of records management initiatives is due to poor change management. Given to ubiquity, quantity (often in the billions of documents annually for many companies) and speed of information passing through the computers of today's knowledge worker, it is clearly unrealistic to ask them to stop and consider how to appropriately manage every document they come across.

It is, of course, critical that policies and appropriate training are in place to educate all staff on their information management and records keeping responsibilities. However, the reality is that for any program to be comprehensive enough to measurably drive down risk it must look at all ESI not just records and focus on taking users out of the equitation to the extent possible.

Auto Classification

The lynchpin for taking responsibility away from end users is auto-classification. By allowing the system to classify documents you can, for the first time, build tools that fully automate all or part of the document lifecycle. Such systems include auto coding for privilege during review and auto assignment of retention categories. The latter when applied to email is the key enabler for a fundamental shift in email management strategy. Because email volumes are so high solutions requiring end-users to choose how to manage individual messages have met with very limited success. The net of this is the creation and growth of massive archives with no defensible method for deletion of any content contained therein.

By auto-assigning retention to email, we are able to segment by content the messages that need to be kept from those that don't. If we take the next step and enable litigation holds to active messages, we essentially eliminate the need for email archiving for compliance purposes. Within the active email repository, it becomes possible to hold and retrieve messages for litigation, move the few messages that are records to a separate retention management system and destroy the balance as dictated by business needs. In this model, there is no longer the need for a separate compliance archive as the active email system contains only messages that are needed for current business activities or are being held for litigation.

Proven text classifying methods such as Support Vector Machines and Decision Trees are rapidly being adopted for use in records management and litigation review tools. They allow the system to be trained to reliably recognize virtually any type of document. The algorithms are well researched and understood academically and have established track records in fields such as chemistry and bio-informatics. They ensure a consistent, defensible and scale-able approach that is easily auditable and can be improved over time. When compared to human review, the accuracy rates of well trained machine based classification are superior and offer the only feasible approach for addressing the data volumes we face today.

Modern statistical algorithms are able to reliability learn core concepts from training examples overcoming the occasional errors introduced by the coders. On the other hand human coders often produce inconsistent and unpredictable results. Different coders categorize the same document into different classes and even the same coder will classify the same document differently at different times. Studies have shown that amount of disagreement between coders can be surprisingly high. Godbole and Roy studied the quality of human classification of natural language texts in the support industry. They found that when different groups of reviewers were asked to review the same set of documents they disagreed on categories for 47% of documents. Furthermore, when same reviewer was given the same document to review on different occasions their labels only agreed in 64% of cases (Shantanu Godbole, Shourya Roy: Text classification, business intelligence, and interactivity: automating C-Sat analysis for services industry. *KDD 2008*: 911-919). In other words, reviewers didn't even agree with themselves in more than one third of cases.

Modern machine learning techniques are significantly more stable and consistent. They are designed to recognize patterns that span over many examples and can therefore overcome occasional human errors in training examples. With a proper design of a learning algorithm and careful selection of example documents it is possible to train a machine algorithm to outperform or work on par with a human classifier. Wai Lam, et.al., observed this when comparing the quality of manual and automatic classification of medical literature with respect to text retrieval (Wai Lam, Miguel Ruiz, Padmini Srinivasan: Automatic Text categorization and its application to text retrieval, *IEEE Transactions on Knowledge and Data Engineering, 1999).* Similar observations were reported in the litigation support industry by Anne Kershaw, a founder of nationally recognized litigation management consulting firm. They compared the results of automatic and manual privilege coding over population of 48,000 documents and found that automatic methods provided much stronger recall minimizing the chance of missing an important privileged document (Anne Kershaw: Automated Document Review Proves Its Reliability. *Digital Discovery and e-Evidence,* Vol. 5, No. 11, November 2005).

It is clear that without leveraging auto-classification firms will find themselves backed into a corner where they are forced to either keep everything forever or place unacceptable burdens on their staff or risk non-compliance with increasingly punitive courts or even end up losing or unnecessarily settling litigation.

Next Generation Solutions

Given that archiving substantially increases risks and costs and that pushing all responsibility for compliance out to the users is entirely unrealistic, it becomes clear that next generation systems must be able to automatically classify and tag documents and be able to utilize that tagging to control the lifecycle of documents in place, i.e., without having to copy them to a separate archive. Moreover, when considering what capabilities are needed for true lifecycle control we must consider both proactive (retention management) and reactive (litigation response) activities. In order to address both side of the coin a system must be able to perform the following functions on all high risk documents in their native locations:

a. Identify and tag
b. Hold
c. Analyze
d. Move and collect
e. Destroy at the end of life

The software industry has been slow to respond to this challenge due the entrenched mindset of enterprise content management vendors. ECM vendors have for years unsuccessfully driven a vision of consolidated enterprise content all under their particular platform versus a more pragmatic

view of heterogonous and distributed data under the control of a homogeneous policy engine.

The good news is that we now are seeing the emergence of vendors that are moving beyond the traditional vision of mass consolation of active data combined with the mass archiving of inactive data. By leveraging auto classification and in place retention management of email and other loose documents these tools obviate the need to push all active documents into a single repository and more importantly they eliminate the need to build expensive, redundant and risky archives.

Conclusion

To truly drive retention risks and costs down firms must reconsider their archive everything forever mindset. Key to success will be leveraging a new breed of technologies that enable companies to defensibly segment and manage their unstructured data along with tools that allow the active lifecycle management of data wherever it resides. Unless firms take a strategic, proactive approach and embrace next generation tools they are in real danger of finding that policies and systems implemented today have inadvertently threatened the company of tomorrow.

Michael McCreary is the Chief Operating Officer of Rational Retention, LLC. Rational Retention is a solution based software company that brings loose enterprise data into compliance with a company's retention policies while providing an immediate ROI by significantly reducing records management and litigation costs. 2 Tower Place, Albany, NY 12208, 518-489-3000