WHITE PAPER

Is It Time to Consider Automated Classification?

Part One: A Better Approach is Needed

Abstract
This white paper is part one of a two-part series. Part one outlines the factors that are compelling enterprises to consider automated records management tools, including solutions that offer automated classification for effective retention of needed records and defensible disposition of expired records and non-record materials. Part two provides an overview of the key factors that organizations must consider when evaluating potential approaches and strategies for automated classification.

Sponsored by:

opentext™

contoural

READY. COMPLIANT. IN CONTROL.

Introduction

The holy grail of electronic records management is automated classification—taking people out of the records-retention process and having computers automatically decide which electronic documents should be classified as specific record types. With today's emerging cloud and mobile computing capabilities, the need to classify records for privacy, security, information rights management and file sharing, the scope of this opportunity has only grown. Automated classification is becoming more mainstream, and organizations are adopting autoclassification approaches more widely as the available technology tools develop and mature. Meanwhile, the scale of the problem continues to grow, with expanding data volumes and user controlled repositories, making traditional approaches increasingly costly and painful. Going forward, many organizations will find it worthwhile to take a fresh look at autoclassification.

Why Classify Records? Compliance, Costs and Risks

Enterprises retain business records to comply with legal and regulatory requirements for records retention. They also retain documents to support internal processes, customer needs and stakeholder expectations. Effective classification of records—and of non-records—is an essential tool for compliance and cost control. Accurate classification can enable effective retention—and appropriate disposal of records and information that are no longer needed. New privacy regulations, the explosion of data breaches and the importance of protecting proprietary data mean that not just records need protection, but indeed non-records as well.

However, today's increasing volumes of electronic documents threaten to overwhelm traditional approaches to Records and Information Management (RIM) and Information Governance (IG), which have largely relied on human beings to classify and file all documents—paper and electronic—in compliance with the company's records retention schedule as well as applicable privacy and security policies.

The growing volumes of retained documents lead to rapidly increasing costs for data storage and information governance. And the costs are further multiplied when documents and data are retained for too long, or duplicated in many locations.

Legal risks are also increasing—especially in the United States, in the wake of the updated Federal Rules of Civil Procedure (FRCP), and in Europe with the introduction of General Data Protection Regulation (GDPR) requirements. When companies cannot control the growth of unnecessary documents and data, the data storage costs pale in comparison with the potential risk and cost burdens of litigation discovery. Those costs can include legal hold management, extensive search and collection, and expensive legal review. Failure to meet these obligations can prove even more costly. Potential legal consequences include court-ordered sanctions, adverse legal judgments, unfavorable settlements—or simply the inability to pursue needed legal remedies, when faced with the burden of expensive legal discovery processes applied to millions of potentially relevant files, documents and messages. Security and privacy breaches can cause similar sanctions, fines, loss of goodwill and even decreased company valuation or reputation.

The Deluge of Electronic Documents

Industry estimates indicate that the electronic information retained by most organizations is growing by 30% to 60% per year, or even faster in some cases. Employees, along with computer applications and communication systems, create and save electronic records and information in many different formats. In many cases, there is no effective process to classify and manage these files and messages.

For example, email volumes are steadily increasing. Emails have largely replaced hardcopy letters and memos, as the standard method of business communication. Employees send and receive more and more emails per day, and many employees simply keep most or all of their email—going back to the day they joined the company. The result is an ever-growing collection of email messages and attachments.

Similarly, unstructured data repositories—such as file shares or SharePoint sites full of document files, images and voice recording—continue to fill up with additional information, which is rarely reviewed or purged when it expires.

The advent of file sharing utilities controlled by the end user means this data can now be easily moved within and outside of company boundaries, unless it is tagged or marked so that firewalls, data loss prevention (DLP) tools and information rights management tools can prevent it from being shared in accordance with company policies.

It is important to recognize that these repositories do generally contain some percentage of business records that are subject to records retention requirements—as well as working documents that employees rely upon when responding to inquiries or when creating new documents. These business records must be retained in compliance with applicable regulations and company policies. They cannot simply be deleted when available space fills up. They must be classified, and then managed, according to the applicable retention policies and retention schedules. In some cases, information in these repositories may need to be deleted on a specific schedule to meet applicable privacy regulations.

These electronic repositories also contain documents and information that are retained for productivity reasons. Employees need to retain a variety of working documents, including work products and reference materials that they rely upon when responding to inquiries or when creating new documents.

Why Manual Classification of Electronic Documents Is Often Ineffective

One traditional strategy is to have all employees manually classify all electronic records, either in their native application view, or in a separate document management application. Typically this can run into problems as employees do not consistently comply with the required procedures. Deluged with electronic documents, people may feel that it takes too much time to classify their documents. Even well-intentioned employees may put it off until a later day.

In Record Management assessments across a number of companies, we have found that traditional employee-driven retention programs experience fairly low program compliance, especially for electronic documents. This is a big problem, as electronic records typically constitute 90% of an organization's records.

In some situations, manual classification may be the only option—at least in the near term. In such cases, organizations may increase compliance by making the classification process simple, quick and easy to perform.

For example, when classifying email messages, we find that companies achieve good compliance only when employees can reliably classify their messages quickly, in less than five seconds per message. An "email file plan" is essential, as it enables rapid and accurate classification of emails into a few "big bucket" filing categories—e.g., working documents, official records and non-records. Moving beyond email, the problem becomes harder to solve. Employees must deal with a much more complex variety of document types, repositories, and business processes. At the same time, business requirements may mandate more granular filing categories, and the capture of more complete metadata for each document.

Where automated classification tools can make the process easier and faster -- replacing or minimizing human decision-making—such tools may enable companies to achieve more consistent classification of records, combined with a substantial reduction in cost.

The challenge is to achieve autoclassification results that are as accurate as human- driven classification, and to establish adequate controls and transparency to ensure that the approach remains accurate and defensible over time. In some cases, this may be achievable with automated classification alone. In other cases, the best solution will involve some combination of automated classification assistance, with the results reviewed and confirmed by appropriate human behavior. This semi-assisted classification is now available in multiple platforms and represents a good compromise between automated and manual approaches. Manual, Automated, or hybrid approaches will require a well-orchestrated strategy that combines people, process and technology.

Monolithic Retention Policies May Invite Risk

To avoid the difficulties encountered with manual classification, some companies attempt to apply a monolithic retention strategy: one size fits all. However, a long retention period creates problems when applied to all documents—or even all business records. So does a short retention period, when applied to all types of content.

For example, a company may propose to save all email for seven years. As only a fraction of emails are business records, this save-everything strategy tends to drive significant over-retention of short-term records and transitory information.

Furthermore, a monolithic retention policy will not meet compliance requirements: a small percentage (but material number) of critical electronic records will have retention requirements that exceed the monolithic retention period. These records will be under-retained, resulting in noncompliance.

Be Wary of "Fauxpliance"

Choosing a shorter retention period, in a monolithic retention policy, will reduce the number of emails that are over-retained—but will significantly increase the number that are under-retained. This approach may reduce short-term costs, but expose the company to significantly increased risks and costs in the medium and long term.

CASE STUDY

Fauxpliance in Real Life

- Record retention policy depended on employees to decide which emails and other documents were long-term records
- Employees stated compliance with policy, but the company had no policy audit
- No centralized email storage; policy was for long-term emails to be “saved locally” by the employee on their workstation
- ASUS faced patent infringement litigation (Phillip M. Adams & Assoc., LLC v. Dell, Inc., 621 F. Supp. 2d 1173 (D. Utah 2009))
- Critical emails were not available, Asus argued “good faith deletion”
- Court ruled “While a party may design its information management practices to suit its business purposes, one of those business purposes must be accountability to third parties.”
- “The absence of a coherent document retention policy” is a pertinent factor to consider when evaluating sanctions
- Court ruled spoliation of evidence and defendant was sanctioned

The most obvious flaw of manual classification is the inconsistent behavior of employees, who may not adopt and sustain the required procedures and practices. Some organizations have attempted to address this issue with an automated process for compliance notification and employee response. For example, the system may send each employee an email every month, asking them to click a button to certify that that they are filing, retaining and disposing of documents in compliance with the applicable record classes and retention rules. The problem is that employees may not be complying, in fact, and the system has no way to tell.

Fauxpliance = Faux + Compliance

This approach may provide an illusion of compliance, or “faux compliance”—but it does not directly control or inspect the actual classification, filing and retention of the actual documents. Employees may not be saving the right documents, for the right retention periods—or deleting the documents that should be deleted.

This auto-reminder approach can provide useful reminders for some employees, and it may seem to shift

the burden of compliance from managers to individual employees. However, in the end, the company cannot be sure that it is following its stated retention policy—and taking reasonable steps to ensure compliance. As improved tools become available, expectations will rise as well. Companies will be expected to apply automation, where it can reasonably and effectively enhance or compensate for imperfect manual processes and inconsistent employee behavior.

Failure to follow your policy can create substantial risks in the event of litigation discovery or regulatory investigation. Inappropriate retention can also increase exposure to data privacy risks.

The message here is, don’t settle for fauxpliance. Focus on finding effective ways to actually implement and comply with your records retention policy.

RETENTION POLICY QUESTIONS

What did you say you were going to do?

- Rim assessments
- Complete policies
- Compliant records retention schedules
- Standards and benchmarks
- Court ruled spoliation of evidence and defendant was sanctioned

How did you do it?

- Effective tools
- Documented processes
- Appropriate file plans
- Employee training
- Governance processes
- Change management
- Defensible retention
- Defensible legal holds
- Defensible deletion practices

How did you check it?

- Re-training
- Monitoring / supervision
- Metrics tracking
- Reporting
- Audits
- Periodic refresh

Regulatory Compliance and Auto-Classification: What Regulators Look For

One of the major drivers for records retention is regulatory compliance, so it is useful to understand what regulators look for when they inspect a company's records. Generally, regulatory scrutiny extends beyond the records themselves.

Regulators look at the company's records retention policy, and also examine the way the policy has been implemented in documented procedures and actual practices—whether those involve human filing behavior, automated recordkeeping systems, or both.

The first question may be, "What is your stated retention policy?" Assuming that you have a policy, and it is well formulated, the next several questions will focus on how you have implemented the policy, and what measures you have employed to ensure that your company achieves a satisfactory level of compliance with its requirements.

While regulators may not expect perfection, they do expect to see reasonable measures, diligently applied. Courts may ask similar questions, and may set the bar higher in terms of expectations.

These questions may be raised by other stakeholders as well, including outside auditors and insurance firms that are engaged to render an opinion regarding the potential risks and costs.

To the extent that autoclassification can improve the consistency and accuracy of records classification and retention, it can become a useful tool for demonstrating diligent implementation of the stated retention policy.

The Impact of Poor Record Compliance on eDiscovery and Data Privacy

The impacts of inaccurate classification, and improper retention or deletion, extend beyond regulatory compliance. Needed documents may be missing, due to misclassification and/or early deletion. Those missing documents (which should have been retained under your retention policy) may hamper early case assessment, increase the costs of litigation discovery, or even lead to adverse judgments and unfavorable settlement terms.

Conversely, inaccurate classification and retention may cause unnecessary documents to be retained too long. Once they are subject to legal hold, such documents cannot be deleted. They can multiply the costs of legal review, and the ongoing costs of data storage and retrieval. Early case assessment can be hampered by missing documents that should still be available under the company's retention policy; it can also be impeded by the retention of large numbers of unclassified or misclassified documents.

One often-overlooked impact is the opportunity cost of legitimate claims that the company does not pursue, after assessing the documents that are available—or attempting to find them. If the documents are needed to support a claim, the practical difficulties of supporting the claim with needed evidence may exceed the amount that could be recovered.

Privacy laws in many industries now require the protection of personally identifiable information (PII) regarding customers, employees and other individuals. Poor recordkeeping policies and practices may cause a company to retain PII longer than it is needed or indeed legal to do so, or to fail to apply appropriate privacy and security controls. Such failures can expose the company to increased risk of privacy breaches, and consequent legal actions or reputational harm.

Incidentally, information security classifications and records retention classifications are generally organized in separate taxonomies, and applied separately. Thus a single document may be classified in multiple ways. While this discussion focuses on classification for records retention, it is important to recognize that employee- driven or automated approaches can also be applied to information security classifications for data protection and data loss prevention. Privacy and security is becoming a rapidly adopted part of the classification landscape and your program needs to take this into account along with record retention.

When considering potential autoclassification approaches and requirements, companies may find that some tools are applicable to multiple types of classification, and can thus help the organization meet a variety of information governance needs.

Is It Time to Consider Autoclassification?

As electronic documents continue to multiply, traditional approaches based on manual classification are becoming less effective, exposing companies to increased costs and risks. The old approaches are not working well. It may be time to consider a different approach.

Meanwhile, autoclassification tools are improving. Vendors are developing capabilities that will make these tools useful for additional purposes, and apply them to a broader range of documents and repositories. Many of these tools feature fairly complex out of the box taxonomies that can identify privacy, financial, health and other types of information without extra work from your IT staff.

While the technology is not universally applicable, it may prove suitable for classification of many types of documents and files—especially when they are already housed in supported document management systems and repositories. Effective autoclassification may make information easier to search and retrieve.

Since autoclassification will inevitably play an increasing role in document classification and management, companies need to determine when and how it will make sense for them to deploy such tools and techniques.

Is Autoclassification Accurate?

If it is to become a reliable basis for document retention and destruction, an autoclassification process must be trustworthy. A trustworthy process will be able to demonstrate the accuracy of its results, and the defensibility of its processes.

Part two of this white paper will focus on specific processes that are needed to make autoclassification accurate and defensible, and will outline related strategies for successful adoption. But it is appropriate to set the stage here, with a brief introduction to the meaning of “accuracy” in the document classification context.

In the evaluation testing of an automated classification tool, a large number of documents are presented to the classification process, which classifies the documents according to its particular set of rules and pattern-recognition tools. The results are then evaluated in terms of two metrics: precision and recall.

Precision is the primary metric for classification accuracy: For any given document retention category, the precision of the tool's output is the percentage of the classified documents that a human expert would also place into that record category. At 100% precision, all of the classified documents do belong in the category—there are no "false positives."

Recall is a distinct but related metric, often framed in terms of completeness. For a given document category, recall is the percentage of all the valid documents that the tool was able to classify correctly. At 100% recall, all of the sample documents that meet the criteria for inclusion are found in the classification result—there are no "false negatives."

These terms are also used in the evaluation of the document search tools that are applied during electronic document discovery. The object of a document search is to produce an output set with high recall and high precision, in a cost-effective way.

Similarly, to be considered accurate, an autoclassification process must demonstrate its ability to classify sample documents with high recall and high precision. Strategies and processes for achieving this result are outlined in Part Two of this whitepaper.

Conclusion

This white paper outlines the problems inherent in traditional approaches that rely on manual classification of documents. Simplistic attempts to avoid manual classification, such as monolithic retention policies and fauxpliance, can actually make the problems worse.

Our conclusion is that autoclassification may prove to be a useful tool that enterprises can use to solve these problems—enabling them to more effectively comply with records retention, security and privacy requirements while reducing the associated costs and risks. A combination of autoclassification, assisted classification and manual classification may hit the sweet spot of policy compliance and information governance.

Of course, organizations will need to approach the problem intelligently. They need to take effective steps to ensure that the classification decisions are sufficiently accurate to meet retention policy requirements, and that the classification process itself is well-documented, supported and defensible.

Part Two outlines implementation approaches that can help ensure that autoclassification provides valuable benefits for records management, in ways that are both realistic and defensible.

~~About OpenText~~

~~OpenText, an enterprise software company and leader in enterprise content management, helps organizations manage and gain the true value of their business content. OpenText brings two decades of expertise supporting millions of users in 114 countries. Working with our customers and partners, we bring together leading Content Experts to help organizations capture and preserve corporate memory, increase brand equity, automate processes, mitigate risk, manage compliance and improve competitiveness.~~

~~As a publicly traded company, OpenText manages and maximizes its resources and relationships to ensure the success of great minds working together.~~

Additional Materials Available

Please visit www.contoural.com for the following materials.

Complimentary Webinars

- Records Retention Policy and Schedule Development
- Records Retention Policy and Schedule Refresh
- Records Schedule Citation Development and Legal Review
- Records Management and Information Governance Maturity Assessments, and Strategic Roadmap Development
- Enterprise Behavior Change Management
- Legal Hold and Discovery
- Technology Requirements and Adoption
- Legacy Paper and Data Disposition
- Email and Unstructured Data Placement
- Records Management and Information Governance Organizational Development and Governance

White Papers

- Stop Hoarding Electronic Documents
- Metrics Based Information Governance
- Email Classification Strategies That Work
- Is It Time For Auto-Classification? Parts 1 and 2
- Ten Elements of Electronic Records Retention
- Seven Essential Storage Strategies
- Six Steps to Controlling eDiscovery for Email
- Ensuring Compliance and Reducing Risk
- Archiving Approaches
- What Do We Do With Legacy Data?

About Contoural

Contoural is the largest independent provider of information governance consulting services focused on Records and Information Management (RIM), litigation and regulatory inquiry readiness and control of privacy and other sensitive information. We do not sell any products or take referral fees, store any documents or provide any lawsuit-specific "reactive" e-discovery services, serving as a trusted advisor to our clients providing unbiased advice. We have more than 30% of the Fortune 500 as clients, across all industries, as well as federal agencies and local governments. Contoural offers a range of record management and information governance services:

- Records retention policy and schedule development
- Records retention policy and schedule refresh
- Records schedule citation development and legal review
- Records management and information governance maturity assessments, and strategic roadmap development
- Enterprise behavior change management
- Legal hold and discovery
- Technology requirements and adoption
- Legacy paper and data disposition
- Email and unstructured data placement
- Records management and information governance organizational development and governance

Disclaimer
Contoural provides information regarding business, compliance and litigation trends and issues for educational and planning purposes. However, legal information is not the same as legal advice—the application of law to an individual's or organization's specific circumstances. Contoural and its consultants do not provide legal advice. Organizations should consult with competent legal counsel for professional assurance that our information, and any interpretation of it, is appropriate to each organization's particular situation.

contoural
READY. COMPLIANT. IN CONTROL.

335 Main Street, Suite B, Los Altos, CA 94022
650.390.0800 | info@contoural.com | www.contoural.com