WHITE
PAPER

Is It Time to Consider Automated Classification?

Part Two: Ensuring Compliance and Defensibility with Automated Classification

Abstract

This white paper is the second of a two-part series. Part one outlined the factors that compel enterprises to consider automated information governance tools, including technologies for automated classification—for effective retention of needed records, defensible disposition of expired records and non-record materials, and control of sensitive and privacy data. Part two provides an overview of the key factors that organizations must consider when evaluating potential strategies and processes for automated classification to ensure compliance and defensibility.

Sponsored by:

opentext™

contoural

READY. COMPLIANT. IN CONTROL.

Introduction

Automated classification products have evolved to a point where many organizations should take a closer look. This is especially pertinent if traditional, manual techniques for information classification are just too painful or if huge volumes of transitory information that could be deleted drive up the costs and risks of eDiscovery and non-compliance. But technology itself can only go so far; there's no magical approach to drive costs down, promote compliance and ensure legal defensibility. Success requires a combination of technology, planning and organizational commitment. There are some specific steps that organizations can take right now to reap the benefits of autoclassification while ensuring compliance and defensibility.

What You Should Expect from Autoclassification

Organizations should expect four things from autoclassification:

- Reduce the manual classification burden on employees for records and long-term business productivity information
- Identify "transitory" or non-records information that can be flagged for immediate deletion
- To "Tag" information for later use by data loss prevention and information rights management tools—a key to privacy compliance
- To be defensible in the event of regulatory or legal challenges

Classification Tools for Tagging of Sensitive Privacy and Security Information

Over the last few years Microsoft in Office 365 and many Enterprise Information Management (EIM) vendors have built auto-classification for privacy and security into their product suites. Many tools now have pre-built configurations for Personally Identifiable Information (PII), which can help organizations to comply with regulations like the European General Data Protection Regulation (GDPR) and the California Consumer Privacy Act. These tools run continuously in the background, "tagging" files and email messages with metadata so that Data Loss Prevention, Information Rights Management and other tools can protect the sensitive information both "at rest" while it is being stored and "in motion" when it is moving around on the internal network or the external Intranet. While the pre-built patterns in these tools can identify PII, companies will need to build the configurations in order to provide security for company trade secrets, financial data and other sensitive information, to accurately classify the unique patterns found in their business. Much of the work necessary to "train" automated classification engines is in building and maintaining these patterns over time.

Once the classifications are built, the second step in using automated classification tools is to set the policies that individual classifications should trigger. Sensitive data tags can be used to trigger various actions, including encrypting emails leaving the company, preventing the copying of financial documents to computers or mobile devices, and preventing the sharing of certain information with individuals outside of the organization. Most of these capabilities are defined by policies built into other tools or systems. Many vendors have emerging technologies that allow the original classification "tags" to be read by a variety of different tools, regardless of how a file moves around the network or between different repositories.

Reduce the Burden of Manual Classification

"Active" information is generated during ongoing operations: content that is generated "starting right now and going forward." "Legacy" information is the accumulation of days, months, and perhaps decades of what was once "active" but may now be difficult to access, may no longer be needed and is consuming disk space (and could be subject to future Discovery requests.) Manual classification of this information would take many hours by individual employees and an extended time frame to complete.

Consider a company that has a core group of records and information management (RIM) professionals and 5000 employees who routinely engage in review and manual classification of active documents. If the chore becomes too burdensome then people either delay the task or ignore it altogether. Assuming that employees are actually engaged in manual classification, courts and regulators can hold an organization accountable for their decisions about classification and retention. Either way, the organization is exposed to considerable compliance and legal risk.

	MANUAL CLASSIFICATION	AUTOCLASSIFICATION
RIM Team Program Development Hours	5 people x 12 weeks = 2,400 hours	5 people x 24 weeks = 4,800 hours
RIM Team Program Monitoring and Execution	3 people x 12 weeks = 1,440 hours	3 people x 24 weeks = 2,880 hours
Total Annual Employee RIM Hours	5,000 people x 1 hour/week = 250,000 hours	5,000 people x .25 hour/week = 60,000 hours
Total Hours	**253,840 hours**	**67,680 hours**

Introducing autoclassification will require additional effort by members of the Security, RIM and IT teams. The nature of such work is described later in this white paper.

Our example shows that the extra work amounts to about 120 person-weeks before production roll-out, and 72 person-weeks for tuning, monitoring and adding new content types once autoclassification has been proven and accepted. Assuming that autoclassification accuracy is 75%—which we believe is reasonable with current technology—the burden of manual classification can be reduced from one hour to 15 minutes per employee per week.

Of course, your results may vary, but autoclassification offers the potential for significantly less effort to be expended across the organization while achieving better compliance with internal standards and external regulations.

Identify Transitory Content that can be Safely Deleted

Now, suppose that the same organization has gathered similar legacy information stored on file servers, archives, content management systems, email repositories and other sources over the last ten years or more. For both active and legacy information, manually classifying and separating what should be properly retained and protected from the far greater amount of "transitory content"—largely, everything else—is a task that few organizations undertake willingly. However, avoiding eDiscovery on such unneeded information justifies the effort. Autoclassification can make this work far less daunting.

Defensibility

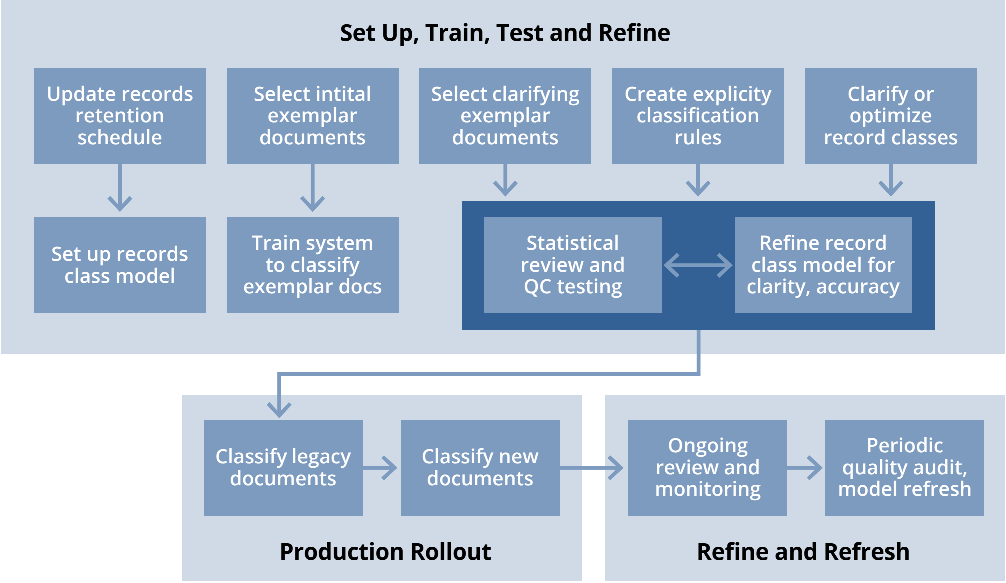
For an organization's processes to be legally defensible, they must withstand legal scrutiny. Reasonable steps to protect the organization itself and its information must be demonstrated. At this time, we know of no case law that validates the sufficiency and defensibility of automated classification for records management. But to be consistent with our definition, we believe that autoclassification must:

- Be part of a transparent process so that the basis for decisions can be readily understood, tuned and explained to non-IT and non-records professionals
- Facilitate adequate sampling to demonstrate both precision and completeness of the results

There Is No "Easy Button"

Why not just point autoclassification technology at a collection of electronic documents and let the system itself figure out what records are there and how long they should be retained? Advances in data analytics are making this more feasible, but haven't yet reached that level of capability. Up-front efforts by a team of RIM professionals, the legal group, business units and IT are needed before the system is enabled and while it is in production.

An Automated Classification Workflow for Retention


Set Up, Train, Test and Refine
Update records retention schedule
Select intital exemplar documents
Select clarifying exemplar documents
Create explicity classification rules
Clarify or optimize record classes
Set up records class model
Train system to classify exemplar docs
Statistical review and QC testing
Refine record class model for clarity, accuracy
Classify legacy documents
Classify new documents
Ongoing review and monitoring
Periodic quality audit, model refresh
Production Rollout
Refine and Refresh

Our experience suggests that customers should adopt a programmatic—and common-sense—workflow approach for introducing the technology. There are three phases to the workflow: setup-train-test-refine, production rollout, maintain and refresh. By following distinct worksteps and involving key stakeholders in the process and decisions at each phase, organizations can be sure that their records program with autoclassification produces results that are defensible and promote compliance.

Start with the Basics

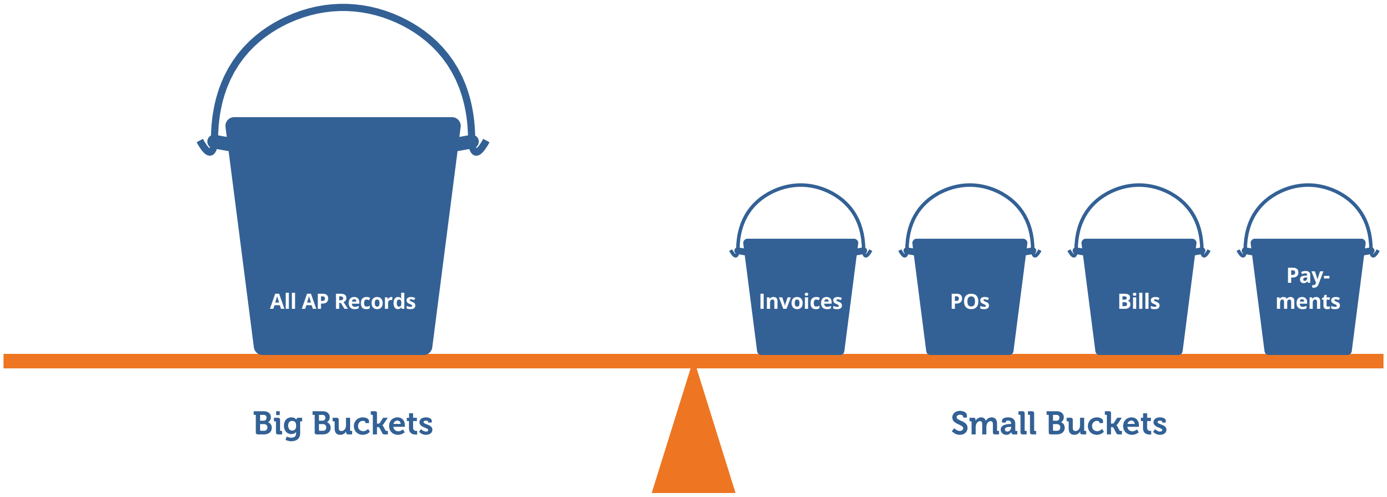
Before autoclassification is introduced, make sure that a stable records management program is in place with well-defined records classes and retention periods. You must also inventory the data you wish to protect for security and privacy regulation compliance and build a starting point set of sensitive information. While not always feasible, combining these two requirements into a single set of classifications will make classification easier and simpler for both manual and automated classification.

Up-To-Date Retention Schedules

How long should information be retained and which rules apply? Such decisions are embodied in retention schedules that are up-to-date. It makes no sense to map new technology against old schedules as the classification will fail to meet current legal, regulatory and business requirements. Early in the process, strive for alignment among the compliance office, legal department, IT, business units and RIM staff about record definition and retention. Remember that current retention schedules must drive autoclassification program design, and not the other way around. Don't let technology define the requirements.

How Big are Your Buckets?

A crucial decision must be made about the number and scope (granularity) of record classes, or "buckets." To illustrate the issue, consider a requirement to retain certain accounts payable (AP) documents as formal records. Separate, small buckets for each subtype of AP record (e.g., invoices, purchase orders, etc.) might make sense. But most, if not all, AP documents are related to each other in some way and may share common retention requirements. Perhaps a large bucket for all AP records would be more appropriate.



Generally, big buckets make it easier to classify records and assign retention periods by reducing ambiguity and complexity. Less time is required to "teach" humans or an automated system to classify records properly. There are two potential disadvantages:

1. Because retention periods default to the longest retention requirement of the stored content, big buckets might result in over retention of some information. The risk of over retention should be balanced with the ease of classification.
2. Suppose legal and regulatory requirements change, affecting only certain record types. If those records are classified in big buckets along with others that are not subject to the new rules, it may be difficult to apply the new requirements.

We observe that most organizations take a hybrid approach, with some records in large buckets, some in smaller ones. If retention across sub-types is the same, our advice is to not spend inordinate effort to define numerous small buckets.

Socialize the Plan

The decisions about retention and granularity will ripple through the records program and affect the ultimate success of autoclassification. They should be made up-front and with the agreement of major stakeholders.

So far, we've only alluded to this important and sometimes overlooked aspect to the overall records program: gaining early organizational agreement and maintaining it throughout the process. We advise the formation of a cross-functional team to oversee the project. Such a team would include representatives from key stakeholders such as the legal, records management compliance, security and IT departments, as well as executives and end-users. Make sure that the plan is well-understood so that there are no surprises. Don't wait for the production rollout to discover objections that can delay or derail the whole project. Compare expected results to what is achieved with current, manual processes.

Train the System

A common stakeholder expectation is that autoclassification will be 100% accurate. While this is a noble goal, it's unrealistic. Training the system properly—and documenting the procedures and tactics used—will help you to prove reasonable and good faith efforts. This builds your case for defensibility, and, after all, such efforts are what courts and regulators expect.

Accuracy

Borrowing terms from the engineering world, accuracy can be measured in two dimensions: completeness (also known as "recall") and precision. Autoclassification is complete if the system can correctly and consistently identify documents that should be records while excluding transitory, non-record information. It is precise if the records thus identified are placed in the right buckets and proper retention rules are applied.

Well-trained subject matter experts can achieve manual classification accuracy between 90 and 100 percent. On the other hand, studies show that average employee manual classification varies tremendously—from 20 to 80 percent. We're convinced that good autoclassification technology, supplemented by proper monitoring, testing and tuning, can achieve 60 to 90 percent accuracy rates.

Measure success by achieving results that are better than what you experience today, while reducing costs and risk for the organization. Don't let perfect be the enemy of good—or the enemy of done!

Identify Exemplar Documents

Start "small"—pick out one or two records categories as initial targets, and identify exemplar documents that are typical of those record classes. The technology will use the content of the exemplars to learn the rules that define each record category.

Leverage the experience of subject matter experts and records management professionals to find exemplars in your current records management system. Such documents have already been classified and retention rules have been applied.

Accuracy will improve with a greater variety of exemplars, so the set should include documents with different authors, titles and formats.

Equally important, the system should recognize what is not a record. Identifying transitory content is potentially the greatest potential benefit that can be gained using autoclassification. Removing such known "transitory" content results in less to retain and reduces the risk and potential cost of eDiscovery.

Test and Tune the Process

Having established autoclassification rules based on exemplars, the next step is to identify a separate set of documents against which the rules can be tested. Using these documents, the first test run will produce an initial set of classification results.

At this stage, human oversight is essential to determine if the test documents are classified properly. The system should be able to call out any mistakes and identify the exemplars that were used to cause the misclassification. Remove documents from the exemplar set or add different ones to improve classification accuracy.

Some documents may simply be unclassifiable. Perhaps information from a particular business unit or discipline simply does not fit the records model. Are the buckets too small? Should a new class of records be specified? Are additional training iterations needed?

It's important to document this "quality assurance" process for autoclassification, calling out exceptional situations and describing remedial procedures. Doing so will help build a solid case for defensibility. Socialize the process so that stakeholders are confident about how the autoclassification system should perform and what results can be expected.

Accuracy will increase with more testing, so expect to repeat the process several times with more and possibly different exemplars and a larger body of test documents.

Production Rollout

Eventually, autoclassification results and manual review will sufficiently coincide. At that point, the system will have achieved statistical stability and can be used to classify a larger corpus of content.

Our experience shows that new technology like autoclassification is best introduced incrementally, first as a proof-of-concept (POC), before going enterprise-wide for defensibility purposes. Think of the POC as a final test run—the last chance to fine- tune the rules and document the results before a wide-ranging deployment. The POC will help identify issues not caught during testing, such as:

- when and how manual intervention might be necessary
- where record classes should be modified
- where additional patterns are required for security and privacy beyond what is built into the tool or added during the initial phase
- where continued monitoring and review is required to ensure accuracy
- what additional repositories might need to be classified

Identify a business unit whose content can be classified into a small set of buckets, or repositories that contain relatively small amount of content types. Good repositories to start with are individual department areas in electronic content management (ECM) systems, email, Microsoft SharePoint, or file shares. Such repositories are comprised of many document types and are usually well-controlled by IT.

Start with business processes and document categories where autoclassification is likely to be successful and relatively easy and quick to implement, based on what has been learned in the testing process. It is often best to start with legacy documents; autoclassification results can be reviewed and shared with stakeholders without affecting the work habits and productivity of many employees.

Move to departments and record classes that require more effort to achieve reliable autoclassification, and those for which can autoclassification is best approached as a starting point or "assistant" for user-driven classification. Finally, it might be that some departments, users, or repositories are less likely to achieve acceptable results from autoclassification. In those cases, alternative approaches—including manual drag-and-drop methods—may continue to be appropriate. Accept the results and move on!

Change Management

Don't underestimate the importance of helping employees accept and embrace changes to their current working environment.

Autoclassification is likely to change the work of most employees in a positive manner. For some, however, particularly when autoclassification results in migration or files to a more controlled system, disruption may occur. Most people will be happy to be free—well, at least "freer"—from the burden of manual classification and from maintaining "private" classification systems such as personal archives of electronic mail messages divided by content type. Most organizations should expect employees to identify minor issues. For example, certain types of records may have been captured but "...oops...we missed one" or "...this document is not classified properly" could happen. Educate employees about their responsibilities and give them mechanisms to flag issues. Let them know that their information is more accessible and that the organization is better protected from regulatory and legal challenges.

Through the cross-functional team, keep senior management informed about progress. They'll want to know—does autoclassification work? Is it effective?

When will its use be expanded across the organization? The more that success can be reported, the more buy-in and perhaps more funding can be made available to expand the program. Such training and expectation-setting is all part of a good change management process.

Review, Monitor and Audit

At some point, a regulator or litigant might offer this challenge: "...you had a computer classify content and make decisions that a skilled RIM professional would normally make; how do you know the results are accurate, defensible and correct?" By following the workflow approach we have outlined, you will have built an audit trail that provides a solid case for defensibility of autoclassification.

Along the way, everything has been well-documented and stakeholders are satisfied with progress. Most importantly, compare the state of things prior to autoclassification and articulate the project goals. Explain the uses of exemplar documents, test sets, sampling, manual oversight, test results, and the proof of concept trial. Review end-user training and change management procedures.

Just because autoclassification has reduced employee workload, flagged transitory content for deletion, and is defensible, the process is ongoing! New rules and regulations will be written, business requirements will change and case law will evolve.

For these reasons, we recommend that you keep the cross-functional team together and review progress on a quarterly basis. There will be additional work for IT staff and records management professionals as they review classification results and issues raised by employees. As a result, it may be that retention schedules and record classifications will be updated periodically.

Conclusion

By adopting autoclassification as part of a robust records management and security program, organizations can expect to lower cost, reduce employee grumbling and promote compliance. This white paper outlines a workflow approach that can help organizations achieve such valuable benefits 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

® 2018 All rights reserved, Contoural. 101118