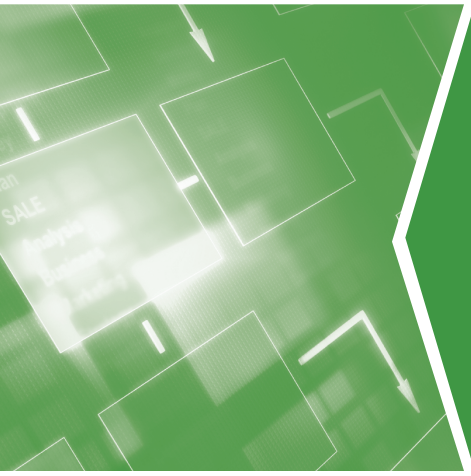
WHITE PAPER



GET YOUR DATA UNDER CONTROL WITH

Automated Content Categorization

Improve speed and accuracy while reducing the cost of information management.

The rise of mobility, cloud computing, social networking and advanced storage capabilities enables more data to be shared in more ways than ever. This surge in connectivity has resulted in an explosion of data for organizations of all sizes and across all industries. Although this data can provide valuable business intelligence, today's enterprises are often challenged to find the information they need and make it useful.

As a result, CIOs face increased IT complexity with the sheer volume of systems and applications required to integrate this data, and the increased cost of doing so. At a granular level, IT cannot manage or apply policy to uncategorized content, users are unable to find relevant information in a timely manner and businesses are struggling to address emerging content types and their locations (social media, cloud computing and the like). These issues create a cascading effect: An unorganized content environment typically leads to difficulties in complying with federal and industry regulations as well as significant time and cost implications for e-discovery inquiries.

Determining what information to keep or retire—whether it's for business, records management, compliance or e-discovery purposes—creates additional challenges for CIOs. And research shows that the problem is only intensifying. According to an IDC report, the amount of enterprise data doubles every 18 months, with 35 percent of the data being subject to regulatory compliance or e-discovery.[1] But the problem doesn't stop there: Data overload also has an impact on productivity levels, with employees wasting on average more than two hours every day looking for information.[2]

RECOMMIND®

To effectively control data overload and make information more accessible, organizations have turned to manual or fully automated solutions. However, both approaches present serious limitations for managing data growth and accessing vital content. What CIOs need is the best of both worlds: a content categorization solution designed to deliver a hybrid approach that leverages the knowledge of an organization's subject matter experts (SMEs) to dramatically improve automated classification processes.

CIO
Custom Solutions Group

[1] Gantz, John and David Reinsel. "As the Economy Contracts, the Digital Universe Expands." IDC, May 2009.

[2] "Wasted Time at Work Costing Companies Billions." America Online and Salary.com, July 2005.

Key Benefits of Decisiv Categorization

- Keep the data you need for doing business while deleting the rest—automatically
- Proactively reduce the costs and risks associated with e-discovery and compliance by reducing the amount of information being stored
- Reduce the time and effort needed to organize, route and distribute information
- Increase the efficacy of investments in enterprise-wide information management systems such as Microsoft SharePoint

Keeping Content Forever Increases Risks and Costs

"Many organizations have drawn a line in the sand when it comes to handling new data and are creating processes to help move them forward, including improved e-mail management. However, the question is, What does the business do with all the legacy data—the 25 terabytes of content it already has sitting on a file server?" says Neil Etheridge, director of product marketing at Recommind Inc., a software company that provides predictive information management solutions.

The data explosion exposes serious flaws in the old retention policy of keeping content forever. Not only does it open the door to unnecessary risks—such as security breaches or privacy rights violations—but the practice is also costly for many organizations.

During the earliest stages of the data explosion, many IT leaders embraced archiving as a logical fix. But CIOs soon discovered that archiving was like using a Band-Aid to stop a hemorrhage. Archiving has failed to address the business issue of information governance and has ultimately become as problematic as the systems it meant to help. Although enterprises need a data management solution, they also need to address the challenges associated with archived legacy data.

As organizations select a sustainable solution to the growing data problem, the logical process is to understand first what can be deleted, what is a business record and what to do with items within the gray area between them. For many CIOs, organizing data within these three buckets represents a valuable step toward controlling the data surge. The most common approaches to controlling data are manual and automated classification:

MANUAL CLASSIFICATION. As the more traditional approach, manual classification and categorization requires employees to devote time looking through mountains of data and making individual decisions regarding content placement, categorization and classification. The key benefit of this process is that accuracy can be extremely high, especially considering an organizational employee's ability to understand documents that fall within the gray area.

However, manual classification introduces an increased level of complexity, which can result in a high number of false negatives. And with individuals processing documents one at a time, scalability suffers. After all, data exists in different places and it's difficult and often time-consuming to access. The manual process also results in increased costs, ranging from excess storage and unnecessary data to delays in meeting e-discovery and compliance deadlines.

"As organizations try to clear up content by using the manual approach, they realize they are very rarely wrong in the classification decisions but are hardly ever able to handle a significant level of data," says Etheridge.

AUTOMATED CLASSIFICATION. As the data explosion has intensified, the fully automated, rules-based categorization approach has gained favor, for its unique ability to rapidly scale to address projects of any size. Typically, using this approach means generating and applying blanket rules such as "delete any document that hasn't been edited during the past 10 years" or "with any document sitting in a folder called XYZ, do X." It's time- and labor-saving.

However, the fully automated approach comes with a few challenges of its own. For instance, since the approach relies on broad, sweeping rules for judging content, the results often contain numerous false positives.

"If you have a lot of false positives, defensibility becomes the key concern," says Etheridge. "How can you prove that you made the right call if the data has been moved into a category or deleted?" In addition, the fully automated approach eliminates people from the equation. In other words, after a rule is created, no avenue exists for expert input, interpretation or intervention.

Leveraging the Best of Both Worlds: A Hybrid Approach to Data Control

Today CIOs have choices beyond the manual and automated approaches, such as Recommind Decisiv™ Categorization. As part of Recommind's suite of data management tools, Decisiv Categorization offers automated classification through supervised learning. The technology effectively leverages the knowledge of human beings to teach technology to better automate classification.

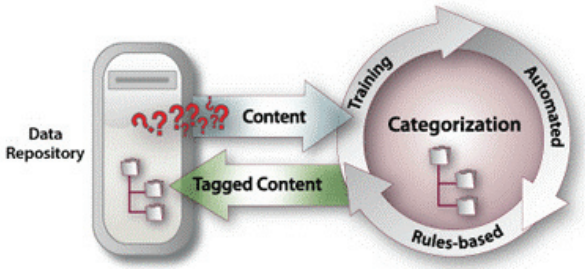
Representing a true hybrid approach, Decisiv Categorization effectively combines the capabilities of auto-categorization, information governance, e-discovery and enterprise search. This enables enterprises to address legacy content, enforce governance policies, mitigate legal risk and reduce IT costs. The

driving force behind Recommind's solution is a methodology and workflow known as predictive sampling, which helps provide defensibility to an organization's categorization decisions.

"Although we do not think manual and automated approaches are wrong, most CIOs would agree that they are both incomplete," Etheridge says. "Our approach understands that people are instrumental to closing the loop. Our technology enables you to merge the two approaches together with very powerful statistical technology. You are able to automate the categorization of information at a high scale while maintaining a very high level of accuracy. You're significantly reducing the number of false positives and false negatives."

Drilling Down: How Decisiv Categorization Works

The predictive sampling process starts with the creation of simple rules. For instance, any document older than 10 years that has not been edited in the past five years can be deleted. Using this rule as a starting point, Decisiv Categorization will search selected databases and find everything that meets the preselected criteria. However, before you delete anything, Decisiv Categorization will take a completely random sample from the data and push it to a subject matter expert (SME) within the organization. This step is a crucial differentiator, since it enables an actual person to review the documents and make the final decision on whether the rule is accurate. Depending on the information under review, SMEs may represent different areas within the organization. For instance, individuals in the legal department may serve as SMEs when the review material deals with industry regulation compliance.



Predictive sampling is just the start. Where Decisiv Categorization becomes really useful is in its incorporation of the supervised learning process. "This is where we truly merge the manual and rule-based by putting together a seed set for each category," says Etheridge.

"As organizations try to clear up content by using the manual approach, they realize they are very rarely wrong in the classification decisions but are hardly ever able to handle a significant level of data."

— Neil Etheridge, Director of Product Marketing, Recommind

The process of categorizing employment contracts is a good example. First employees manually locate and categorize a small selection of employment contracts to serve as a model. Once this has happened, Decisiv Categorization runs its proprietary learning process, which uses the model to understand "key identifiers" such as phrases or terms within the model documents and determines which items fall within the employment contracts classification. The system then automatically goes out and finds other documents it determines should have the same categorization.

After the system has found all the documents that meet the model criteria, the predictive sampling application forwards a random sample of the categorized documents and pushes it out to the previously identified SMEs. This step gives company personnel the opportunity to review a small sample and judge whether the system is accurate. As organizations go through the sample, Decisiv Categorization has the ability to accept any corrections the SME makes and to relearn.

"This is an adaptive technology that can learn and grow in step with the organization. It enables you to build the accuracy of categorization to a very high level, actually higher than human categorization," says Etheridge. "The key to success is to identify people within the company who can serve as SMEs and make judgment calls on categorization activity."

Delivering Immediate Results and Long-Term Benefits

Having data properly classified not only yields immediate benefits but also affords the business an ongoing streamlined process for organizing, routing and distributing data, including:

LEGACY DATA REMEDIATION. Classification yields a reduction in risk and lower costs for data collection and preservation. It also frees the IT department from some associated storage costs when it is able to delete unnecessary data.

IMPROVED E-DISCOVERY AND COMPLIANCE. When a company is subpoenaed to produce all data on its dealings pertaining to a specific topic, it must go through the e-discovery process. This event is stressful as it is, but risks and costs are associated with it whenever an organization's data is unorganized. The result is either under-collection or over-collection of data, both of which are costly. But an enterprise with fully classified and categorized data need not labor through unnecessary documents. Instead, a few keyword searches will garner all the pertinent materials—a significant benefit, considering the costs associated with legal reviews.

ONGOING INFORMATION GOVERNANCE. As a key step toward delivering on the promise of seamless records management, having a well-designed classification system makes it possible to proactively govern all new content. CIOs can take control of data and reduce reliance on data creators to accurately categorize information and data. Adaptive technology empowers the right people within the organization to develop classifications that comply with enterprise standards. This level of control also allows for the creation of an easy-to-maintain enterprise search model, empowering the IT department as well as client-facing employees.

"The benefits are huge, whether the organization is looking back or moving forward," says Etheridge.

"Decisiv Categorization classifies large volumes of data much faster and even more accurately than human experts. We can now automatically classify thousands of newly created documents every day. Simply put, Decisiv Categorization has enabled us to greatly improve the speed and the quality of our information management."

— Martin Steinbach, Head of Archiving, Handelsblatt

"Categorization enables you to cost-effectively get your house in order in a timely manner and keep it that way. You do not have terabytes of over-collection, so the costs of legal reviews are far more reasonable than they would be otherwise."

With more data to control than ever, today's CIOs are challenged to discover cost-effective, accurate solutions that enable employees to access the information they need, when they need it. Although manual and fully automated solutions may provide some relief, the Decisiv Categorization hybrid approach combines the accuracy and insights of human expertise with the speed and flexibility of an automated process. ■

RECOMMIND

About Recommind Inc.

Recommind is the leader in predictive information management software, delivering search-powered business applications that transform the way enterprises, government entities and law firms conduct e-discovery, enterprise search and information governance. Recommind's solutions are all built on the CORE (Context Optimized Relevancy Engine) platform, which automatically accesses, organizes and analyzes large volumes of information in context from myriad sources. With greater control over and more accurate access to information, organizations can lower risk, heighten productivity, increase the value of information assets and improve competitiveness and profitability. Recommind customers include AstraZeneca, BMW, Cisco, Clifford Chance, Marathon Oil, Morgan Lewis, the U.S. Department of Energy (DOE), White & Case and Wilmer Hale. Recommind is headquartered in San Francisco and has offices in Boston, London, Sydney and Bonn, Germany. For more information, go to **www.recommind.com**

Copyright © Recommind, Inc. 2000-2011.
Recommind, Inc.'s name logo are registered trademarks of Recommind, Inc. Decisiv, Decisiv Email, Auto-File, SmartFiltering, Axcelerate, Insite Legal Hold, QwikFind, One-Click Coding and Predictive Coding are trademarks or registered trademarks of Recommind, Inc. or its subsidiaries in the United States and other countries. Other brands and names are the property of their respective owners.