USING DATA PROFILING TO MITIGATE 7 'RED FLAG' INFORMATION RISKS

Data profiling technology can help an organization identify what electronic information it has and where it is located, which is the first step to ensuring that information governance policies are applied to it, reducing the organization's costs and mitigating its seven greatest information risks.

Jim McGann

Today's records and information management (RIM) professionals are tasked with mitigating the risks that come along with ever-changing regulations and escalating threats to information security, all while controlling exponentially growing data costs.

More like detectives than RIM professionals, they are charged with uncovering sensitive records, such as Outlook personal storage table (PST) e-mail archives and aged e-mails created by former employees, unencrypted files containing personally identifiable information (PII), and copies of contracts and research data lost within network file shares.

But before they can find and manage this data, they must determine what information exists across the enter-

prise, where it exists, and who created it. To address the risks that come with not knowing this, some RIM professionals are using data profiling.

Data Profiling Defined

Data profiling examines data from all sources and collects metadata-level information on the content to create a searchable and reportable repository about the information, identifying such things as the information's owner, age, type of file, location, date last accessed or modified, and whether it is a duplicate.

Data profiling allows RIM professionals to create a map of what data exists and where, so they can actively enforce and audit compliance with the information governance policies that dictate the use, disposition, retention, and management of corporate data, protect the firm's assets, and manage long-term risk.

This helps organizations control costs and risks, as well as address the seven greatest issues – red flags– that have a critical impact on their electronic records systems today: PII, hidden PSTs, user shares, departmental servers, legacy backup tapes, aged data, and large multimedia files.

The Information Governance Landscape

The volume of information is exploding. More and more business information is digitized, and users create data 24/7 as they carry their office in their pockets. It is estimated that data growth is reaching unprecedented levels, with Gartner stating in its June 2012 research report "Organizational Collaboration and the Right Retention Policies Can Minimize Archived Data and Storage Demands" that the volume is growing 40% to 60% annually. The numbers are nearly impossible to comprehend and rising each year. Technology teams are keeping up with this data growth by increasing storage capacity.

The data environment also has become more complex; for example, copies of user content are typically replicated and archived many times to ensure it is never lost, leaving records managers with multiple versions of the same mess to clean up.

Archives that were created to store specific important data, such as content preserved for legal holds and documents required for compliance, have become bloated unstructured repositories where other data is put in and never seen again. Highly sensitive information becomes lost in the shuffle and largely unmanageable.

The Legal, Regulatory Climate

Over the past decade, e-discovery has provided an expensive learning lesson for some organizations. They have spent significant time and resources identifying sensitive user data and collecting it to support active litigation. It is not unusual for a single litigation to cost hundreds of thousands or even millions of dollars. The Searle Civil Justice Institute's 2010 "Litigation Cost Survey of Major Companies" found that for the period 2006-2008, the average *Fortune* 200 company that responded to the survey paid average discovery costs of $621,880 to $2,993,567 per case.

Ten years ago, it was fairly easy to claim that specific content was not easily found within the complex corporate infrastructure and, therefore, it was "inaccessible," or it placed an "undue burden" on the party being asked to produce it for litigation.Today, this argument is less successfully used.

Judges, opposing counsel, and expert witnesses have all been educated on and have a better understanding of the corporate data environment. Today, it is a high-risk proposition to enter a court without the requested data, as there is a good chance for the judge to admonish or even issue fines and sanctions for not producing it.

***Data profiling* examines data from all sources and collects metadata-level information on the content**

Beyond the growing demand to produce information for e-discovery, compliance and regulatory requirements have been increasing, resulting in renewed emphasis on RIM strategies. These growing and evolving regulations – such as requirements to encrypt sensitive records or archive specific classes of correspondence – require significant updates to corporate policies and new strategies for data managers.

As a result of these issues, organizations are taking a fresh look at policies and implementing information governance strategies, such as using data profiling, that are aimed at protecting them from risk and liability by providing them knowledge they need to make proper decisions.

Data Profiling and Policy

Corporate data policies are complex.For many organizations, the complexity stems from their attempts to define policies without having adequate knowledge about what data exists. Policies are created, but they can't be enforced or monitored.

By default, then, many organizations are leaving it up to end users to implement the policies. But users tend to neglect making decisions about data, opting to keep it forever "just in case" they might need it. Very few know such things as how PST files are made, where they go, or what the consequences of sending a client's credit card number through an e-mail could be.

Along the way, needed information gets lost in the shuffle. Over time, research and intellectual property data ages and becomes difficult to leverage. This hidden data, which has value but cannot be found within the infrastructure, is not tapped to help support current users.

As knowledge workers leave the organization, their data typically remains scattered about the network infrastructure. Since the owner is no longer around to manage this content, it quickly gets lost; current users don't even know it exists.

Data profiling helps both sides of the information quandary. Using it, content can be searched, found, and leveraged to support business needs, and it can be purged, encrypted, or secured to mitigate risks.

... personal media files have no business value but are being managed and backed up daily

The 7 Red Flags

The profiling process begins by becoming aware of the greatest threats to data breaches, compliance issues, and bloated storage budgets. As described below, these seven red flags will have the most immediate and critical impact on an organization's management of electronic records.

1 *PII*

PII includes credit card and Social Security numbers that, if lost, could put customers at risk. Organizations usually have privacy policies that prohibit sending PII through e-mail and mandate the encrypting of all files containing PII, but that doesn't mean they are being followed. Organizations are responsible for safeguarding this information, and any violations – such as of the requirements of the U.S. Health Insurance Portability and Accountability Act or the Fair and Accurate Credit Transactions Act – make them financially and legally liable.

2 *Hidden PSTs*

Microsoft Outlook allows users to create PSTs and store messages, contacts, appointments, and other information on their hard drives or a network server. PSTs, much like PII, usually have a policy surrounding them that isn't readily enacted or enforced. The truth is, most non-IT personnel probably don't know what PSTs are, let alone that they are creating them. Managing PSTs can be challenging; even from an IT standpoint, finding PSTs is difficult. Legal and compliance teams are often surprised to find this hidden e-mail during high-profile litigation. Auditing for PSTs helps legal teams evaluate these highly sensitive e-mail archives and determine disposition.

3 *User Shares*

Organizations typically set up network shares where users can store files and other content. These shares often grow to the point where managing the content is impossible. With data profiling, this mystery content can be analyzed and an action plan defined. For example, data profiling can find all data that has not been accessed for a specific length of time, allowing it to be moved off this environment or even purged if it has no business value.

4 *Departmental Servers*

Profiling the content on shared servers and desktops within a network that was created by a department that frequently works with valuable content, such as intellectual property, consumer information, or research data, is a logical place to begin. This will provide the knowledge required to determine its disposition; much of the data may be archived for long-term retention or secured to protect sensitive intellectual property.

5 *Legacy Backup Tapes*

Backup tapes contain copies of all existing user files and e-mails amassed over time. In the past, these were often considered burdensome to access, so courts did not always demand for it to be produced for litigation. However, technology now makes this content reasonably accessible and, thus, a corporate liability if not managed according to policy. Organizations are now cleaning up this content based on a data profile and preserving only what has value.

6 *Aged Data*

Profiling aged data – that which is more than seven years old and has not been accessed for a long period – will reveal what and where it is, allowing it to be moved to less expensive storage platforms, migrated to the cloud, archived with related documents, or purged if its retention is no longer required.

7 *Large Multimedia Files*

Chances are that audio, video, and photo files are not putting organizations in legal jeopardy unless there is a possibility of copyright infringement, but they could be affecting storage budgets. User shares likely are packed with personal files from employees watching movies or searching through vacation albums during breaks or downloading music libraries to listen to while working. These personal media files have no business value but are being managed and backed up daily. Locating them enables organizations to audit how

much space these files are using, purge them from storage, and develop policies governing the use of company resources to play or view and store them.

Data Profiling and Disposition

Data profiling helps organizations understand and manage the needs of their information governance policies by ensuring the policies are followed with the proper action and data dispositions. It finds the red flags before breaches or litigation can. It often prevents problems and saves valuable storage costs.

Data profiling is one of the best resources that RIM professionals have to support information governance policies. It provides a deeper understanding of assets and is the key to being able to control risk and liability. Without a data profile, it's nearly impossible to manage mystery content and enforce policy.

Common dispositions include moving essential content to an archive, preserving data for legal holds, removing duplicate content, encrypting sensitive data, migrating to less expensive storage, and purging data that has no business value.

No longer should data remain on networks unmonitored and unmanaged with uncontrolled growth. Using policy as the foundation and data profiling as the support, organizations can leverage and manage data more effectively. **END**

Jim McGann can be reached at Jim.McGann@IndexEngines.com. *See his bio on page 47.*