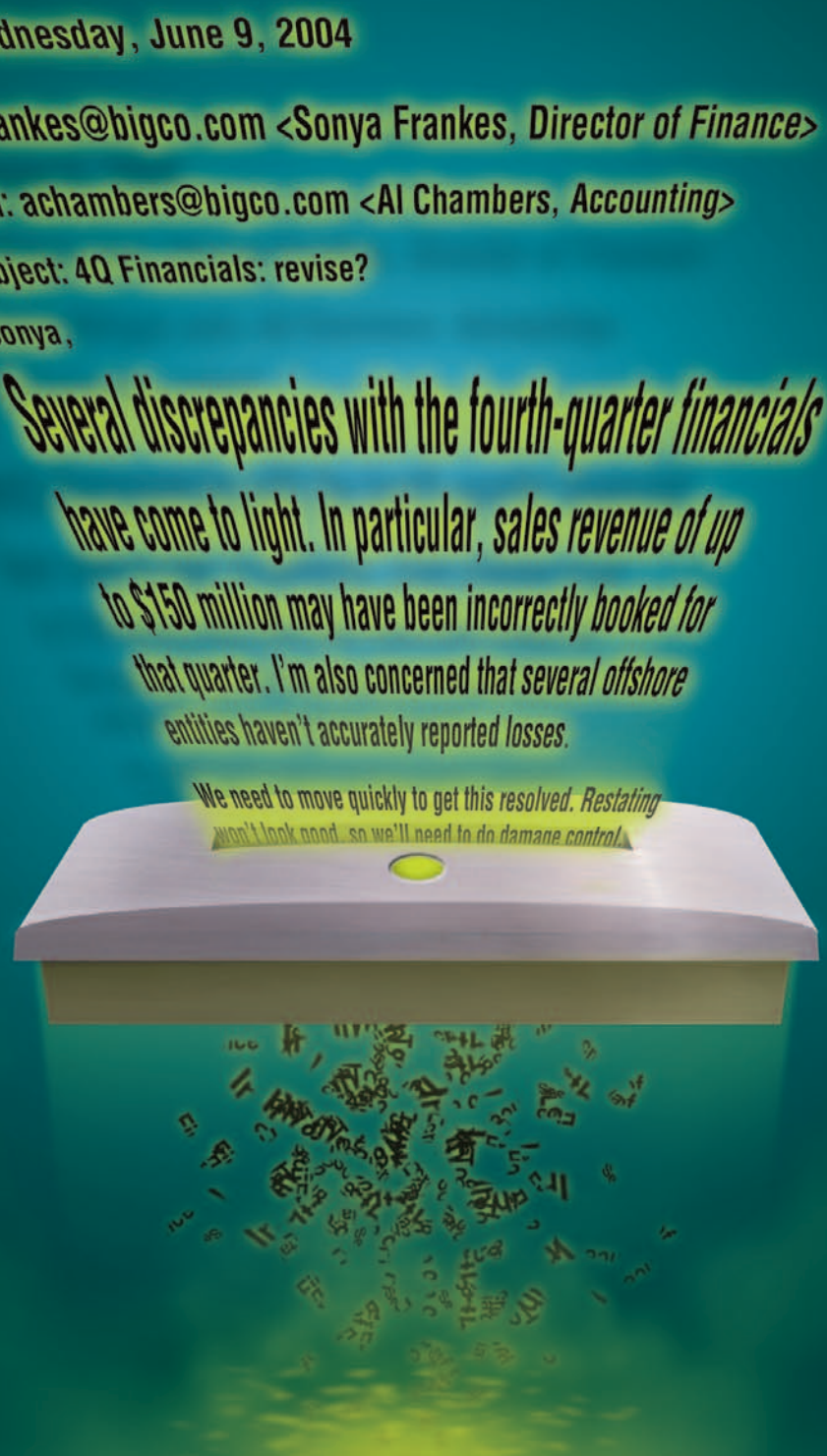
InformationWeek
Research & Reports

Appetite For Destruction

CIOs rethinking the 'save everything forever' approach must ensure that their disposition strategies are razor sharp to cut through retention regulations **By Andrew Conry-Murray**



Sek Leung

WHILE THE OIL AND gas refined by CVR Energy will someday run out, the company generates a seemingly inexhaustible supply of data: 3 to 5 TB of information in 2008 alone, says CIO and senior VP Mike Brooks. He expects that load to double every year for the foreseeable future.

Though disk may still be cheap, Brooks says, it just doesn't make financial sense for CVR to store every bit of electronic information indefinitely. Besides raising hardware, software, and utilities costs, outsized data stores make backups and enterprise search less efficient, and legal e-discovery more burdensome. When you're paying lawyers hundreds of dollars an hour to review e-mail and documents, a smaller pile means a smaller bill.

That's why CVR, a $3 billion-a-year refinery based in Sugar Land, Texas, is undertaking a massive data disposition project, hammering out policies that will govern how long the company stores its information and when it can be disposed. Between

deletions based on the new rules and other technology approaches, such as deduplication, Brooks hopes to cut CVR Energy's disk use in half.

He isn't alone. More organizations are evaluating—if not yet implementing—data disposition strategies. By 2013, half of all Global 2000 companies will have formal records management systems to shepherd data through its life cycle, Gartner estimates.

But this is one area CIOs must approach with caution. There are significant technological, regulatory, and organizational hurdles to clear before organizations can eliminate data with confidence. At the top of the list are compliance and legal. Every industry has government-mandated retention requirements. On the legal side, general counsel and human resources may worry that critical pieces of information that could support their positions—in case of employment discrimination or harassment claims, for example—may be destroyed.

Technological and organizational challenges are just as daunting. Before you can dispose of information, you must identify it and know every place it resides—not a simple task. And users aren't quick to give up the mail and documents they produce. As with NRA members, you may have to pry PST files and PowerPoint decks from their cold, dead hands.

Do You *Really* Want To Save That?

90% Drop in access rate of some older data, such as e-mail, within 60 days

$72 Cost per gigabyte for Tier 1 storage

40% Respondents who gained high or very high benefits in meeting retention policies through information life-cycle management

Data: Gartner, Oracle, and 2006 *InformationWeek* reader survey of 291 respondents

TEAR IT UP

Getting rid of data generally goes against the corporate grain. Much time and effort is devoted to producing, protecting, and preserving information, and now you want to shred it?

But if there's one thing that can focus executive attention, it's litigation. An evolving legal landscape is encouraging enterprises to reconsider this preservation instinct. In December 2006, the Federal Rules of Civil Procedure, which set litigation guidelines at the federal level, were updated to include electronically stored information in discovery requests, in which one party asks the opposition for records relevant to a lawsuit. This means parties in litigation can request

Impact Assessment: Data Disposition

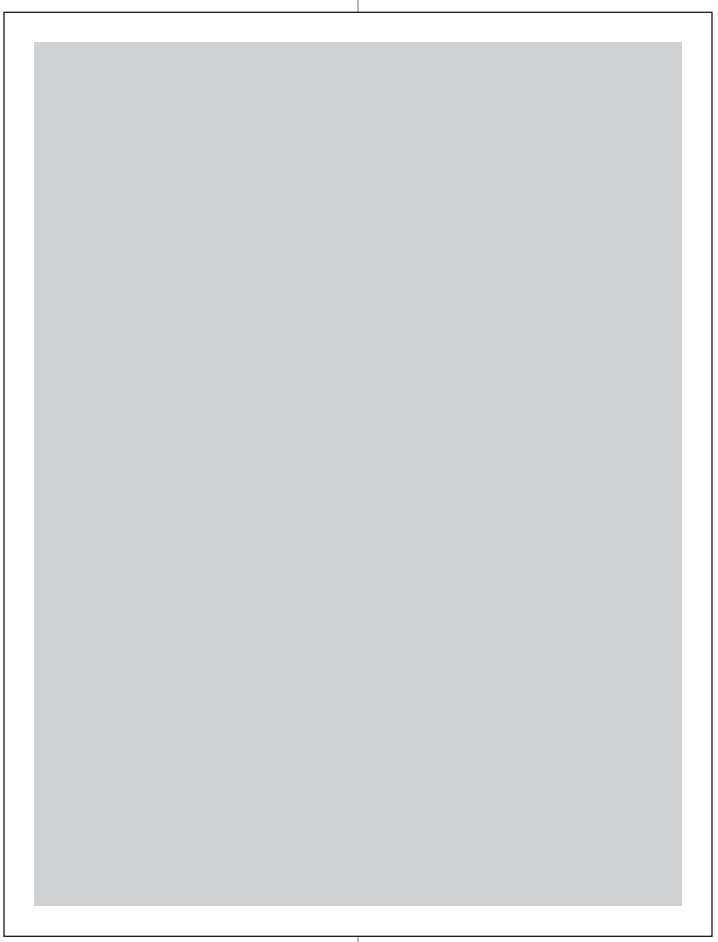
	Benefit	Risk
IT organization	Data disposition yields a leaner, less costly storage infrastructure. Systems required to execute disposition, including information discovery and classification, make the storage architecture more efficient.	Disposition is hard to do well and easy to do poorly if IT fails to take a strategic approach founded on policy and supported by key constituents, including executives, general counsel, and compliance officers.
Business organization	There's no value in preserving data beyond its retention period. Disposition technologies may be leveraged by enterprise search initiatives, enabling users to find and retrieve information more effectively, improving productivity.	Companies walk a fine line between deleting data that should've been preserved and keeping data that should've been eliminated. Failure to demonstrate adherence to a disposition policy is a significant risk.
Business competitiveness	Disposition is the ultimate test of an efficient information management system. For litigious enterprises, an efficient system can hold down discovery costs.	Improperly destroying data, especially that on litigation hold, can be fatal in court. In addition to fines and sanctions, companies risk damage to their reputations when disposition goes wrong.

Bottom Line

It doesn't make financial or legal sense to store information indefinitely. Disposition is key to managing growing volumes of unstructured and semistructured data, and the technologies required for a proper disposition system underpin an efficient information management ecosystem. A sound disposition policy can reduce the cost of legal discovery.

both physical documents and electronic information, and organizations have a legal obligation to produce all relevant material. Most discovery requests focus on e-mail, but the scope of the rule is broad enough to include Office documents, instant messages, text messages, .wav files, and so on.

Companies spend jaw-dropping amounts of money on e-discovery. Fiona Schrader, principal product manager of EMC's compliance division, says DuPont estimates that one legal discovery bill came to $11 million. Let's be clear: DuPont didn't spend $11 million total on a lawsuit; it spent that amount on the discovery portion. In that same discovery effort, DuPont found that $4 million to $6 million worth of records had already met their retention deadlines and should have been destroyed.



"Companies aren't getting the connection between what they are keeping and what it means for time and expense when litigation hits and you have to pay lawyers to look through everything," says Michael Sands, a partner at law firm Fenwick & West and chairman of the firm's electronic information management group. Schrader agrees and estimates that less than 10% of her customers have active, automated disposition practices.

There are three main reasons for this foot dragging. First, some companies aren't sure it's legal to get rid of data. It is. The Supreme Court has ruled that it's permissible—though usually under very specific circumstances. A large constellation of rules and regulations governs how long various types of information must be stored: 17 years for patient health records, six years for dealer/broker records, the lifetime of a building for construction and architectural documents, and indefinitely for certain kinds of environmental records and reports. But once mandated compliance periods are met, information should be destroyed.

Companies also must be aware of another stipulation to legal data destruction: the litigation hold. This is a procedure in which information that may be relevant to a case is preserved, even if it's nearing or has reached the end of its retention period.

"Litigation hold and disposition are intimately related," says Sands. "Any automatic system to purge is fine, as long as there's a way to turn it off so you aren't destroying documents you have an obligation to preserve."

Companies also are reluctant to dispose of data because they think they'll find information to help them prove their case during litigation. They probably won't. "As a litigator," says Sands, "the number of instances we find a document we wished wasn't there far outweighs the times we find something where we say 'Whew! Glad we saved this!'"

It's no coincidence that companies that have been through litigation at least once are more amenable to implementing data disposition policies, Sands says.

The third reason organizations are slow to get rid of data is technological. Before information can be destroyed, IT has to know where it is, what it is, and which retention rules must be followed. Records management, content management, and e-mail archiving systems play a role in retention and disposition. But they're often deployed tactically rather than as part of an enterprise-wide strategy. These products also have limits, which we'll discuss.

IDENTIFY YOURSELF

Before you can chuck a piece of information, you have to know what it is. Thus, index and classification technologies are key. That's where CVR's Brooks is

starting. The company bought Autonomy's Intelligent Data Operating Layer, or Idol, a software platform for enterprise search and classification, and centralized its storage around 10 geographically dispersed storage area networks. The platform uses connectors to tap into the SANs to index the content stored there.

Brooks started with a backlog of unindexed information stored in the SANs, including 1.9 million e-mails and 600,000 documents. It took about 10 days to create a searchable index of those data stores, and now the Idol engine keeps up with new data that gets moved into the storage networks.

It sounds great, but the dark side of indexing is that it adds to your overall data store. In fact, Brooks' team initially failed to properly size the database for the index because the team didn't anticipate just how large it would be. Autonomy says a typical Idol index runs 20% to 25% of the total data store, depending on the level of indexing, from basic metadata to cataloging the full contents of a file.

The next step is to categorize all this information for retention and disposition. CVR is still working through its disposition policy, though Brooks expects it to be in place by the first quarter of 2009. "Our objective is to take out the human element," he says. "Two people can look at the same document and cate-

DIG DEEPER

WEB 2.0 FACTOR New collaboration tools make it difficult to track corporate data that must be managed, but it isn't impossible. Download this *InformationWeek* Report at:

informationweek.com/1182/report_web2.htm

See all our Reports at informationweekreports.com

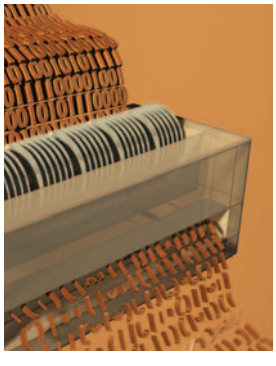
Off The Record: Enforce Retention Policy

IF YOUR ORGANIZATION IS going to claim in court that records aren't available because they've been destroyed, be prepared to back up that assertion with a retention and disposition policy—in writing. You also should have records that demonstrate how the policy is implemented and how employees are trained in retention and disposition. When creating the policy, IT, legal, and compliance officers need to be involved, as do line-of-business managers. But don't forget the people who actually create the content.

"Users are in favor [of disposition] unless it's their data," says Mike Brooks, CIO and senior VP of CVR Energy. A disposition policy has to be cognizant of users' desire to have some information that lives forever. Overly strict policies will prompt users to find ways to thwart the rules—and that could have harmful compliance or legal repercussions.

You also need to have an audit trail, such as electronic log files, to show that disposition is applied regularly and uniformly in accordance with the written policy. Be sure the process includes legal-hold capabilities, and expect to have an IT or storage director be deposed, or even appear in court, to explain the policy and how it's implemented.

Whatever you do, don't write a policy and then fail to follow it. "That's arguably worse than not having a policy at all," says Michael Sands, partner at law firm Fenwick & West. Many in-house lawyers see retention and disposition as a checklist item, he says. They get a sample policy, slap the company logo on it, put it in a drawer, and forget about it. That's dangerous.



"When a company has a 'policy' that they aren't following, they have defined their own standard of care that they have then failed to meet," Sands says, adding that if your opponent can show that you say one thing but do another, you've already lost.

Note also that there are different levels of "deleted." The most common method is to overwrite data with other information. However, data erased by a simple overwrite often can be recovered using forensic software. Other methods, such as overwriting multiple times with ones and zeroes, encrypting without preserving the decryption key, using a strong magnetic field to wipe a disk, or physically destroying storage media, are used for high-security deletion to defeat forensic data recovery. When it comes to data disposition and federal civil court cases, a simple overwrite is sufficient. Courts are less interested in the method used to delete data than in a litigant's ability to demonstrate a comprehensive and repeatable disposition system.

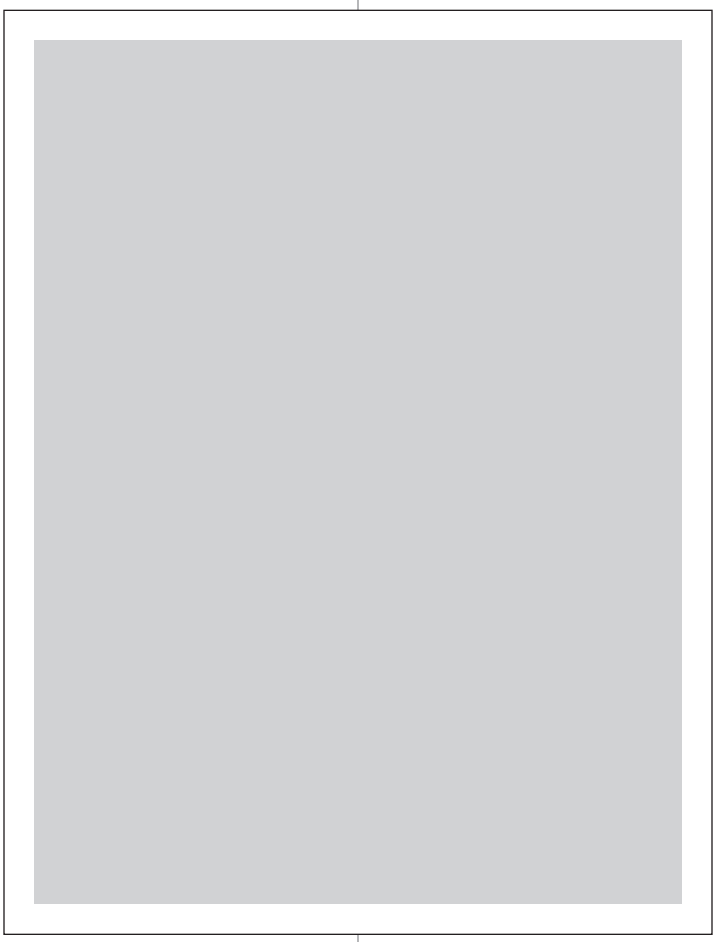
"If you say, 'We don't have e-mail from the following people prior to March 2006,' courts are generally going to accept that as long as it's supported by a declaration from someone in IT explaining retention and disposition policies and practices," Sands says. "A court isn't going to make you prove a negative."

However, it's useful to maintain clear disposition records, such as audit trails. Many archiving products, such as Hewlett-Packard's Integrated Archive and Autonomy's Idol software platform, create date and time stamps for each object, such as an e-mail or document, stored in the archive. Tracking the archive date is crucial for time-based disposition policies.

—ANDREW CONRY-MURRAY

gorize it differently. Any time there's human intervention, courts can question your consistency." By automating the process, he hopes to avoid dispute on the final disposition of a file.

Brooks' team is working with various company departments, including legal and accounting, as well as business units on a policy that will designate different information categories to meet all the requirements for retention. Once the policy is in place, the Idol engine will assign data to the most appropriate category. "If it goes into a folder that has policies for financial documents, in seven years it will get disposed of," Brooks says. "If a document is environmental, that's lifetime storage."



Because CVR's policy isn't finalized, the company hasn't gotten rid of any data. Brooks also says that once information reaches its retention limit, the company will start with a manual review to ensure the data should be destroyed. But his ultimate goal is to automate the destruction. "The manual intervention is where you get in trouble—everything becomes a judgment call," he says. "If the machine is doing it based on algorithms and parameters, at least your company can be consistent."

He's also aware of the need for legal holds. In the event of litigation, the plan is to use the Idol technology to search for relevant data and then move that information to a separate repository. Brooks' IT team also wrote agent software that moves data off corporate laptops and into the SANs whenever the laptops attach to the corporate network. When data is destroyed on the SANs, the agent also will erase it from the laptops.

Data disposition is a crowded vendor field. For instance, vendors of enterprise content management (ECM) systems—including EMC, Open Text, and IBM (via its FileNet software)—are adding classification, retention, and disposition capabilities to their portfolios. ECM products focus on records management to maintain strict control over official paper and electronic records, such as business contracts and legal documents, while providing content repositories, mechanisms for end users to check documents in and out of those repositories, and version control enforcement.

EMC's Documentum content management system offers the Retention Policy Services module, which lets IT create folders that will enforce specific retention policies. Administrators can choose between automated and manual disposition when information reaches the end of its retention period, and the module supports legal holds to suspend disposition. Documentum licenses the Fast enterprise search engine (recently acquired by Microsoft) to index and search information.

Open Text's Enterprise Library Services, rolled out in October 2007, provides a retention and disposition policy layer across a variety of content repositories, such as archives, file systems, Microsoft SharePoint, and SAP. In December 2007, IBM announced a SOA-based connection between FileNet and the IBM Classification Module. The module automates the classification of unstructured content, including e-mail, through full-text analysis. In March, Hewlett-Packard announced it would acquire Tower Software, an Australian document and records management vendor, to expand its legal discovery and regulatory compliance capabilities.

Before the purchase, HP had included Tower's software in its Integrated Archive Platform, an

archive appliance that serves as a central repository for a variety of data, including e-mail, Office documents, and SharePoint and Web content. Once inside the Integrated Archive Platform, the Tower software indexes and categorizes content so administrators can set up retention schedules. At the end of the retention period, the appliance destroys the data, essentially by writing over it in the repository.

BLIND SPOTS

While classifying information is a challenge, finding it often proves an even higher hurdle.

Major data stores, such as network-attached storage filers or e-mail archives, are the low-hanging fruit. Storage administrators generally know where they are. But other data stores are trickier. SharePoint servers, for example, are relatively easy to deploy, which means line-of-business managers can set up one or two on their own, without IT's permission or knowledge. After a recent audit, one of HP's bank customers found more than 5,000 SharePoint implementations it wasn't aware of, says Jonathan Martin, chief marketing officer for HP's information management software group. Those servers likely hold information that falls under a retention and disposition policy.

Online collaboration tools—such as Socialtext, PBwiki, and Google Docs—are another area of concern. Users can upload business content to these sites in seconds with IT none the wiser, and the data moves beyond the reach of classification and disposition systems. Proactive IT organizations will provide sanctioned collaboration tools that blend administrative controls, such as provisioning, deprovisioning, and authorization, with the ease of use of Web 2.0 apps. This way, you can ensure that content created in these collaborative environments can be discovered—and destroyed—in accordance with policy.

Just as significant are user desktops and laptops. User hard drives are chock-full of corporate data, as are portable flash drives and other removable storage media.

So what's to be done? For user devices, agents are a good answer. EMC talks about using its RSA Data Loss Prevention agents, which are deployed on endpoints and can find and identify content, for information management. These agents are focused mainly on enforcing use policies, such as preventing certain kinds of information from being attached to an e-mail or saved to a removable drive. But the classification capability may be repurposed to also ensure that information on user endpoints meets retention policies. Backup agents could play a similar role. These agents already are copying data from local machines to be stored on backup servers, so they're naturals for legal discovery and retention and disposition purposes.

No vendor has yet made product or road map announcements to this effect, but as HP's Martin says, "It's a natural evolution that organizations want to leverage the investment they've made in backup for more than just simple recovery."

Data disposition has clear benefits for IT and for the business. A sound disposition policy will help enterprises reduce storage costs and reclaim disk space. The tools needed to find and classify data can be leveraged as part of an information management strategy. Regular purging also will reduce discovery costs in the event of litigation. It's shredding time.

Write to Andrew Conry-Murray at acmurray@techweb.com.