

WHITE PAPER

Gain Business Value from the Disparate Landscape of Corporate Content with Content Archiving

Sponsored by: Hitachi Data Systems

Laura DuBois
February 2008

EXECUTIVE SUMMARY

The growing volume of corporate content and content repositories for enterprises as well as for multinational and domestic firms must be managed to mitigate legal, compliance, and business risk and to reduce costs while serving the needs of discrete business units and content stakeholders. Adequate controls must be applied to mitigate content leakage or unauthorized access, enable compliance and preservation, and prove content integrity. Redundant or duplicate data should be eliminated to reduce risk and cost. Retention periods must be enforced, legal hold orders instantiated, chain of custody preserved, and disposition audited and proven. Firms are also benefiting from the growth in corporate content by reusing it for business advantage such as mining data and using that information to increase revenues, clients, customer satisfaction, and strategic advantage.

To effectively manage the growth in corporate content and leverage the information for business advantage while mitigating content risk, firms must apply disciplined, structured data management principles to this growing volume of unstructured content. Content archiving is a technology solution that meets these legal, regulatory, and business initiatives.

SITUATION OVERVIEW

Growing Landscape of Corporate Content

Capacity growth realized in the structured database world is being seen in unstructured and semistructured content such as email, file systems, video, and Web content and office documents in content management systems. This growth in corporate content varies by industry and geography but averages 52% annually, resulting from new systems coming online, new content types, mergers and acquisitions (M&As), market expansion, globalization, and transition from a paper to a digital workplace and economy. Growth can come from a broad set of drivers, including increases in retail locations, products, customers, component applications, publishing channels, new applications, and use of mobile devices in the corporate office and industrial markets.

Valuable unstructured corporate content can include:

- ☒ **Traditional content:** Microsoft Office files, other file formats such as PDF/A and TIFF, computer output, and enterprise reports
- ☒ **New media content:** Video, audio, Web content, images, blogs, wikis, and RSSs
- ☒ **Collaborative content:** Instant messages, email messages and attachments, and unified communications

These content types are the result of increasingly global communications, additional workloads, and improved business processes and workflow. Applications such as online training, online publishing, video on demand, video on download, video surveillance, online and new distribution commerce, remote medical imaging and diagnostics, distance learning, voice over IP, and specific industry applications are also contributing to the growth in corporate content.

Disparate Content Repositories

Enterprises need to not only direct and make use of newly created content but also manage disparate numbers of existing content repositories. It is not uncommon for a large enterprise to have many different email systems and archives, dozens of CRM applications, hundreds of content management systems, and thousands of file systems distributed across systems, vendors, and locations. A worldwide software company today has over 100,000 SharePoint portals, while a leading financial services firm has over 200 homegrown and commercial content management repositories. Some firms are trying to standardize and reduce the numbers and types of repositories to reduce risk and administrative overhead, legal costs, and complexity. However, the usability, economics, and availability of products such as SharePoint often drive departmental implementations. Broader content management systems provide a uniform workflow to business processes and control the versioning, publication, record retention, and approval of various corporate content and documents.

Today's disparate unstructured and semistructured corporate content repositories include:

- ☒ Distributed user file data on endpoint devices
- ☒ Email and instant messaging systems
- ☒ Enterprise content management systems, computer output, and reports
- ☒ Windows and Unix file systems and file shares
- ☒ Web 2.0 content, including blogs, wikis, and RSS feeds

These corporate content repositories tend to be siloed by geography, business unit, department, or division, which can fuel creation of more new content repositories. Each repository has its own controlled content, operating systems, administration

methods, media, and archive models. The disparate nature of these systems prohibits the unified application of retention and preservation policies and systematic keyword, phrase, or concept search and discovery of material content in response to legal discovery or for content reuse.

Content Reuse to Meet Business Needs

Content in disparate repositories not only presents legal and regulatory risks and increased cost but also inhibits business transformation and the reuse or repurposing of content to service customers, meet supplier demands, and drive new business initiatives. A federated search of disparate repositories can support improved call center initiatives, servicing of patients, access to customer history, integrated project and product information, and the like. Firms can use a federated search not only to satisfy reactive litigation and audit demands but also to leverage content in repositories for business value and content reuse. The concepts of data mining and business analytics applied in the structured database world to large data warehouses are now being sought in the unstructured content world. Mining content for reuse involves extracting relevant content from different systems in different locations for some business function. The first step of locating and retrieving "relevant" content is paramount and is based on specified user criteria or search. Firms seek to mine content for reuse for applications such as the following:

- ☒ **Customer content:** A call center needs to access all content related to a particular customer, case, or project and wants to pull up "relevant" email records and check images and credit card statements.
- ☒ **Project content:** A manufacturing facility needs to manage all content related to a particular project such as contracts, client communications, and associated engineering files.
- ☒ **Patient content:** A healthcare provider needs access to relevant patient diagnostic, treatment, financial, and administrative records over a specified time period.
- ☒ **Retail location content:** A retailer needs to provide location-specific employees with sales, customer, and inventory content during a specific retail period.

Different Content Stakeholders

Compounding challenges associated with disparate content repositories and overall content growth is the management of different content stakeholders' objectives. These content stakeholders include functions such as compliance, legal (inside and outside counsel), security, risk management, records management, technology, human resources, finance, and various business units. This broad stakeholder community tends to be concerned with the content itself, the policies under which the content is managed, or the underlying infrastructure that houses the content. Collectively, these stakeholders must mitigate corporate, legal, and regulatory risk; reduce the total costs of corporate content infrastructure; and drive new business initiatives and revenue streams.

Managing Content Risks, Timely Access, and Costs

Mitigating content risks and providing timely access to content while managing costs is a primary concern for most firms. This concern compounds as the number of repositories expands and the amount of content grows annually. Risk results from the lack of or untimely access to content during litigation or audit, the discovery of and ineffective handling of content fraud, the inappropriate use of content by employees or those seeking to compromise a firm, or even the mishandling of corporate content by the firm itself in violation of corporate policy. Firms seek to mitigate these risks to avoid court sanctions; regulatory fines; corporate litigation time and costs; damage to corporate brand, partner, and customer trust; and loss of shareholder value. Managing corporate content to ensure compliance with regulatory requirements spans many geographies, cultures, and legal environments. U.S. businesses are faced with an onslaught of over 20,000 regulations with record-keeping requirements, but this is not only a domestic problem. Regulatory compliance is mandated on a country basis with international regulations in the United Kingdom, Germany, and Japan receiving increasing focus from governmental bodies. Countries under the European Union are increasingly being affected by regulatory requirements. The following is a sample list of regulations outside the United States with record-keeping requirements that are affecting both domestic and multinational firms:

- ☒ **J-SOX:** Japan's 164th Diet approved and passed the Financial Instruments and Exchange Law (FIEL) — commonly referred to outside Japan as J-SOX. The FIEL reforms, which affect publicly listed companies, are intended to enhance investor protection, promote financial innovation, and ensure the competitiveness of Japan's financial markets relative to global financial markets.
- ☒ **The EU Data Retention Directive:** This directive applies to all EU member states and aims for retention of telecommunications data to ensure that the data, which can identify the caller, the time, and the means of communication, is available for the purpose of the investigation, detection, and prosecution of serious crime. Under the directive, the data retained will be made available only to competent national authorities in specific cases and in accordance with national law. The data will be retained for periods of not less than six months and not more than two years from the date of communication.

In the event of a regulatory audit, legal discovery, or compromise to corporate customer or employee content, the time and manner in which a firm addresses the content crisis is paramount. A firm does not have weeks to locate the data in question, to understand the scope of content exposure, or to perform manually intensive processes to locate the point of content compromise. Discovery costs during litigation can scale to hundreds of millions of dollars based on the scope of the discovery, the media the content is on, and the format in which it is stored. Moving from a manual to an automated and unified discovery and collection model of a litigation event can cut down the discovery costs, better prepare the firm for legal review, reduce the time to respond to discovery requests, and allow stakeholders to return to their normal operational roles.

BUSINESS STAKEHOLDER OBJECTIVES

Business stakeholders have an interest in the risk associated with managing or handling corporate content. These stakeholders concern themselves with an application- and content-oriented view. This community is tasked with the establishment of policies of how corporate content is managed, and these policies then inform technology and infrastructure decisions made by technology stakeholders.

Business objectives relative to corporate content are at once both reactive and proactive. Business stakeholders want to reduce risk and cost associated with content while using content repositories to support new business initiatives, drive new business value, increase customer satisfaction, and improve competitive advantage. Business stakeholder objectives include those discussed in the following sections.

Reduce eDiscovery Risks and Costs

In an increasingly litigious corporate environment, it is not uncommon for a firm to face hundreds of legal matters at a single point in time. Each single legal matter represents potential risk and cost to a firm. Risks can come in the form of a firm's inability to locate, access, or retrieve discoverable content as well as the lack of adequate controls to manage preservation orders. Corporate ramifications could include court rulings of spoliation, inverse inference, or obstruction of justice and could have a material effect on the final ruling in the case or result in court sanctions. A unified search of disparate content repositories can reduce discovery and collection times and costs and allow for quick export of content into legal review tools such as Concordance, Attenex, and others. Costs in the ediscovery collection as well as review phases to electronic discovery can be minimized by proactively placing indexed content on accessible media where data can be searched and culled, duplicates eliminated, and content exported to legal review systems. The culling and duplicate elimination tasks are vital because they can reduce the amount of data that a legal professional must review for privilege or relevance.

Consolidate Search for Litigation Readiness

Historically, electronic discovery has often been targeted at email and instant messaging systems. However, the December 2006 amendments to the Federal Rules of Civil Procedure (FRCP) represent an effort to update the discovery rules in light of differences between paper and electronic discovery. Specifically, Rule 33(d) was amended to recognize the importance of electronically stored information and Rule 34 distinguishes between electronically stored information and "documents." The rule goes on to outline that the requesting party can stipulate the form in which the electronically stored information is produced, such as in a TIFF format. However, absent a court order, agreement, or specific request for production, a party can produce electronically stored information in a manner consistent with how a responding party maintains such information, such as in native format.

This amendment is significant because it substantially broadens the universe of discoverable information to include digitized voicemail, digitized video, or instant messages. As there is not a substantive definition for "electronically stored information" in the amendment, it allows for a flexible and inclusive approach to what is "electronically stored information." This opens up the door for electronic discovery to include information sources such as file shares, local PST files, applications and databases, voicemail, instant messages, and even remote devices and USBs. For a large firm with hundreds to thousands of content repositories with potentially relevant information, performing electronic discovery tasks on a serial, one-by-one basis is not a financially viable or legally prudent approach. Firms need a means to perform a federated search of disparate content repositories.

Avoid Risk with Amendments to the FRCP

The previously mentioned amendments to the Federal Rules of Civil Procedure (FRCP) stipulate the following:

Parties must "meet and confer" early to address issues relating to electronic discovery, including the form and preservation of electronically stored information, the problems of reviewing the electronically stored information, and assertion of privilege. Firms facing high volumes of legal matters must prepare themselves for electronic discovery by putting in place retention, preservation, and legal hold order policies. Additionally, firms should conduct inventory assessments of and document information repositories, specifying what content exists which will position a firm for early readiness in a "meet and confer" meeting. Electronically stored information that is not "reasonably accessible" does not have to be produced, unless the requesting party can show good cause. Firms should include in their inventory documentation what information sources are considered accessible and which are considered inaccessible and the time and cost associated with retrieving content from respective "inaccessible" information sources.

Note: To read more about the Supreme Court's action on the amendments, refer to www.uscourts.gov/rules/#supreme0406.

While the amendments to the FRCP are specific to the United States, with the expansion of multinationals and global business transactions, there is a high probability that large enterprises will increasingly be involved in cross-border ediscovery investigations. However, the discovery rules in different countries vary significantly, thus making ediscovery even more complex. Knowing where to find all the rules and regulations for each country is the first step in successful overseas litigation.

Note: To read more about general discovery rules in Europe, refer to the Hague Evidence Convention and The International Organization on Computer Evidence (www.ioce.org).

Satisfy Regulatory Record Retention

Most firms face some type of record retention requirement stipulated by a governing body in the industry in which the firms operate. Regulatory bodies in the United States such as the DHHS, SEC, FDA, EPA, and OSHA oversee regulations that stipulate how companies as well as healthcare providers, payers, and business partners manage and retain content found in records. *Regulations such as those defined by the Sarbanes-Oxley Act, Securities and Exchange Commission rules 17a-3 and 17a-4, and the Health Insurance Portability and Accountability Act (HIPAA)* subject companies to requirements that specific business content be captured and retained in a way that ensures information integrity, security, and accessibility. This legislation is concerned with the content and not the format in which it is stored.

Regulations may stipulate that related content contained in electronic records be retained for specified time periods or outline rules to secure the privacy, security, and lack of compromise to sensitive information. Central to satisfying regulatory compliance requirements is the fast provisioning of electronic records in response to a regulatory audit, typically within a 24- to 48-hour time period. As a result, electronic records subject to regulatory compliance should be indexed for easy search, retrieval, and production. Equally as important as retention is the proper disposition and destruction of electronic records, absent any legal requirements to preserve, once they no longer have any regulatory or business value.

Comply with Corporate Policies

Firms need to comply with corporate policies around how content is managed, retained, preserved, migrated, secured, and discovered. As discussed, lack of compliance can present regulatory and litigation risk. If firms have policies, they must implement, enforce, and audit their policies and processes for compliance and have proof of their compliance. Corporate policies should be endorsed by legal. These policies should take into account not only content policies but also infrastructure issues such as change and configuration management, data protection, and recovery processes. Corporate policies should also include in their scope user, desktop, endpoint, PST, and removable storage policies as well as auditing and training on compliance with documented policies. Table 1 provides some examples of what corporate content policies should address.

TABLE 1**Examples of What Corporate Content Policies Should Address**

Number	Policies should address and answer the following questions:
1	Which data sources, based on their content, constitute official electronic records for a firm?
2	What regulations stipulate this as an electronic record, and how long should it be retained?
3	Which applications, across multiple locations, have electronic records that must be retained?
4	In what format, and with what accessibility and security, should this electronic record be retained?
5	Is this electronic record subject to multiple regulatory requirements for retention?
6	Is a second copy of an electronic record required?
7	With what level of authenticity, integrity, or permanence does the record need to be retained?
8	How quickly does this record need to be retrieved in response to audit or discovery?
9	With what categorization or taxonomy should an electronic record be classified?
10	How should an electronic record be indexed (full content or metadata only), and why?
11	How are electronic records to be searched as part of a discovery or an audit?
12	How are electronic records to be produced as part of a discovery or an audit?
13	During litigation, how is a records hold order implemented and by what systems, people, and processes?
14	How can a records hold order be authenticated, audited, and proven?
15	How are multiple hold orders on the same record managed and enforced?
16	How are records released from a records hold order?
17	How is controlled deletion of electronic records ensured, absent any legal or regulatory retention?
18	How is unauthorized duplication or distribution of sensitive corporate information prevented?
19	How do you ensure that role-based access controls are being applied consistently?
20	How do you audit, monitor for, and verify sensitive corporate information leaks?
21	Is the proper metadata associated with electronic records preserved and tamperproof?
22	What electronic records can be deleted? When, or after how long?
23	Will deletions be automatic or manual? If manual, have users been trained on this policy?
24	Is destruction of an electronic record to a standard such as DOD 5015 required?
25	How is destruction of electronic records on removable media handled?
26	How are assets containing electronic records retired or decommissioned?
27	How are old employee accounts and records handled?
28	Where is sensitive content located? In which applications? How is access controlled?
29	Are distributed PSTs or NSTs with corporate content able to be located and managed, or is the PST function disabled?
30	Are distributed storage devices such as USBs located, managed, or disabled?
31	Are policies able to be audited and verified by a third party?

Source: IDC, 2008

Document Chain of Custody

For the purposes of litigation, a party seeking to introduce an item of electronic evidence must prove that the item is taken from a particular person or place, which makes the item relevant as evidence in the trial. Such proof is provided by testimony identifying the item as having been taken from that person or place and by evidence tracing custody of the item from the time it was taken until it is offered in evidence. As a result, firms need to have accurate written chain-of-custody records tracking the possession, handling, and location of evidence from collection to presentation in court. This documentation is necessary to avoid claims of substitution or tampering by opposing counsel. Firms should have the ability to use logs and audit trails to document chain of custody and use write once, read many (WORM) media to satisfy that evidence has not been compromised or tampered with and is considered tamperproof.

Support "Duty to Preserve" Obligations and Legal Hold Orders

A firm that reasonably anticipates litigation has a "duty to preserve" electronically stored information that may be relevant to the potential case. The scope of a company's duty to preserve evidence has been outlined in a series of five court opinions in the employment discrimination case of *Zubulake v. UBS Warburg LLC*. In the fourth written *Zubulake* opinion, the Court explained the nature of a litigation party's duty to preserve relevant electronic information:

Once a party reasonably anticipates litigation, it must suspend its routine document retention/destruction policy and put in place a 'litigation hold' to ensure the preservation of relevant documents. As a general rule, that litigation hold does not apply to inaccessible backup tapes (e.g., those typically maintained solely for the purpose of disaster recovery), which may continue to be recycled on the schedule set forth in the company's policy. On the other hand, if backup tapes are accessible (i.e., actively used for information retrieval), then such tapes would likely be subject to the litigation hold.

The impact of failure to preserve relevant electronic information can be monetary sanctions or court rulings of inverse inference.

Mine Content and Gain Business Value

Historically, through data reuse and business analytics, organizations sought to capitalize on the value of electronic information resident in structured databases for revenue generation, customer satisfaction, and strategic advantage. However, increased levels of advantage can be realized in the reuse of unstructured data. Because firms need to comply with regulations that stipulate the retention of content contained in unstructured repositories for specific periods of time, new usage patterns for that data emerge that can yield a competitive advantage.

For example, at a credit card processing firm, historical customer account information is retained. The disk-based archive and retention of these records can be done in a manner that provides fast access in minutes versus days. Because the records are accessible, they can be made available, through information portals, to internal and external customers for self-service use. The provision of these records in an online and self-service manner saves on employee costs and allows the credit card processing firm to charge for portal services. These extra services create additional opportunities for revenue generation and higher customer satisfaction. This manner of data reuse is difficult to achieve with older, traditional archive technologies because the information has not been indexed, is not easily accessible, or if stored on removable media, runs the risk of being lost.

Avoid Compromise to Sensitive Content

As a result of regulatory and public exposure pressures, firms increasingly need to mitigate the risks of sensitive content leaks within unstructured data. To alleviate potential exposures of sensitive information, the discovery and indexing of content within distributed content sources is ideal. Once content within the data is understood, the risk of exposure can be mitigated by the proper tagging of data based on corporate classification or role-based access, or in some cases, data can go through disposition. Step one is to understand any content exposures and then to apply information management policies such as retention, access restriction, or disposition.

Consider the risks for a firm in the exposure and visibility of sensitive information such as company information (e.g., financials, intellectual property, engineering data, sales data, or business strategy), and customer information (e.g., name, social security number, account number, patient health information, medical history, or credit history) by the wrong people. Content-aware discovery and classification can help a firm address the challenges with litigation, regulatory compliance, and optimization of storage capacity while protecting sensitive information from unwanted leaks or exposures.

TECHNOLOGY STAKEHOLDER OBJECTIVES

Technology stakeholders want to reduce capital and operating costs and manage infrastructure more effectively while serving the business needs outlined previously. This group of technology stakeholders cares about the management of the content as well as the management of the infrastructure and seeks to leverage investments in existing infrastructure to the extent possible. Technology stakeholder objectives include those discussed in the following sections.

Integrate Archiving Technology

IT professionals need to manage and integrate several different technology layers: the content-generating application, the archiving software, and the archiving storage solutions. Each component plays a critical role in managing corporate content as follows:

- ☒ **Content-generating applications** such as Exchange, Notes, SAP, SharePoint, and Microsoft Office allow users to generate content that may require long-term archiving as a result of legal, regulatory, or business drivers.
- ☒ **Archiving software**, which is either available standalone or integrated with enterprise content management (ECM) applications, manages the retention and disposition of content. Collectively, this software provides an automated and efficient way of interfacing with the content-generating application to store, index, retrieve, migrate, copy, retain, expire, delete, and shred content, including automatically migrating or recalling data between primary and secondary systems based on policy. Archiving software often maintains links to the archived content supporting user access via normal interfaces and paths. Unlike other processes, archived content is accessible without a restoration process and data is indexed, either at the full-content level or at the file-attribute level. Based on the application, the integration between content and the archiving software layer will vary.
- ☒ **Archiving storage solutions** integrate with the archiving software solutions via file system, Web, or API interfaces to consistently support and enforce policies set by the archiving software. Moreover, archiving storage solutions provide an additional layer of integrity checking and protection required for long-term retention, preservation, and content reuse. For example, WORM storage is used to ensure data cannot be changed or deleted until expiration has been reached and/or a legally defensible chain of custody is maintained. The hash-based naming schema for each piece of content also ensures that data sent to the storage media is not corrupt. Lastly, archive storage solutions can provide a federated search capability to run a single search criterion or a series of search criteria across all content, including from different applications, thus saving time and enforcing consistency in policy.

Leverage Investments in Tiered Storage Infrastructure

Tiered storage has been accepted and adopted by firms today as a means to control costs and minimize tier 1 storage purchases by placing data on a tier of storage with the right levels of performance, reliability, and access based on the business value of the content. A firm's investment in a particular storage architecture or platform should not be overlooked when selecting a tier of storage for content repositories and archiving. Making use of existing infrastructure not only drives savings in capital expenditure but also supports a common administrative model, integration with broader storage management software, and minimal to no domain training and improves interoperability. As static tiered storage continues to be adopted and

virtualized storage environments continue to proliferate, the use of dynamic tiered storage will increase. As this schema for data to storage placement is adopted, the relationship between primary and archive storage will blur and data will be able to be dynamically placed on the right primary, secondary, or archive tier of disk storage based on usage patterns, metadata, and relevance of the content to the business.

Reduce Administration Costs

Firms want to standardize on infrastructure as a means to reduce administration costs and implement a single management framework that can be used to more effectively administer the storage infrastructure. It is not uncommon for IT professionals to be pulled away from important operational and infrastructure tasks to respond to discovery or audit requests. Some firms remedy this by assigning dedicated personnel to support discovery requests. Others form liaisons with forensics, discovery, and compliance teams to serve as an interface between business and IT stakeholders and mitigate disruption to normal IT staff. A unified management and discovery model across repositories allows for repurposing of IT personnel, minimizes the size of liaison departments, and improves institutional knowledge. Benefits of a common physical architecture versus archive silos include:

Improve Retrieval Performance

Some firms still use tape for archiving; others use optical devices. Each approach has its own challenges. Tape can be unreliable and damage prone. It requires media management and tape tracking and, due to its serial nature, requires extensive amounts of time and administrative cycles to locate content. Media degradation of tape volumes makes tape an impractical media in the event of an audit or a lawsuit. Tape failure rates can be high, introducing risk, and often many multiple recoveries are required, which increases administrative overhead. For retention, tape management is required and old tapes have to be collected and content transferred to new media and new technology in a manual process. Optical has its limitations in the amount of content that can be stored on an optical platter and in that it also requires media management. Due to the removable nature of optical devices, recovering content can be as time intensive as locating relevant content. Because business stakeholders increasingly are fighting external pressures to locate and make use of content, fast retrieval times are critical. Disk is a natural complement to this requirement. Moreover, given the amount of data and the number of repositories, ingestion performance is also of concern. More firms are using disk as the media of choice for archiving of content repositories, and this media allows firms to take a proactive approach to discovery and compliance. In fact, migrating content (including tape and optically stored content) from legacy archives to disk-based content archive architectures enables consistency in policy enforcement and avoids archiving silos.

Reduce Backup Windows

Ask datacenter managers what their challenges are, and they will tell you about their backup window issues. Due to the rate of growth in unstructured content, and limited or shrinking backup windows where clean, consistent backups of production data can be performed without impacting users, applications, or networks, firms are looking at archiving infrequently accessed or fixed content in secondary or archive tiers of disk storage to help reduce backup window pressures. Moving older, nonchanging data to a secondary or disk archive tier can significantly reduce backup window challenges and improve performance and operational efficiencies with primary applications and systems.

Execute on Documented Business Policies

Technology stakeholders often seek establishment and endorsement of business policies by legal or compliance functions. In some firms, this can result in two strategies or mental models, neither of which is often ideal. The first strategy is for a firm to "save nothing" and to apply aggressive deletion policies of 30- to 90-day retention. This obviates the ability of a firm to derive business value from its content and may present a risk during litigation if policies around destruction are not established and documented or if relevant content that could be used as evidence is destroyed. Moreover, destruction policies that are not consistent and applicable to items such as PSTs, remote devices, or removable tape storage could also raise questions around inconsistent retention and disposition practices. The second strategy firms often endorse is one of keeping everything as long as possible, which presents practical limitations in terms of content management and gain from what is often called a "tsunami of information." While a firm may never lose data, storage and electronic discovery costs skyrocket.

Technology stakeholders can help inform a firm's decision to take a midpoint approach to these two extremes whereby critical information is retained and nonbusiness relevant content such as spam, personnel content, and other low-value/no-value content is eliminated. However, central to this midpoint approach is the application of retention policies for regulated content and the enforcement of a firm's "duty to preserve" during litigation. IT plays an important role in helping business stakeholders to understand what is feasible from technical, time, and cost perspectives. IT has a major voice in current and future business policies as related to content: IT should let legal make the decision based on the risk level of the firm and then help the firm understand the ramifications of that business policy. Once there is a common framework and a set of shared expectations, IT can confidently execute on documented business policies.

Ensure a Scalable, Performing, Reliable, and Secure Infrastructure

For technology stakeholders, the infrastructure for content repositories must provide resiliency in the event of physical and logical failures and must be able to sustain multiple component failures simultaneously as well as provide for disaster recovery. The architecture must scale in terms of capacity, not only for what is required for today but also for future archive capacity, and be designed to support petabytes of data storage. The infrastructure must support different content repositories under a single architecture and ensure content is secure and compliant with controls for authentication and immutability to ensure compliance and integrity. Technology stakeholders expect the infrastructure to be resilient while also providing no technical compromise. Additionally, the infrastructure must provide economies of scale and enable, over a long period of time, a lower total cost of ownership (TCO). The reduced TCO is measured in terms of unit costs, software, services, and administration.

Avoid Platform and Vendor Lock-In

From a technology perspective, IT professionals and executives want to avoid platform and vendor lock-in, in particular for content that will persist for a long period of time. With technology changes and supplier M&A trends, API and media lock-in can present risk in terms of long-term accessibility to corporate content. To the extent possible, technology stakeholders seek open, published standards-based formats and interfaces. This increases the likelihood of long-term interoperability among applications, file formats, and storage media. Content written and stored in an open format can be more easily and painlessly migrated to next-generation application versions and media migrations and helps with data, application, and infrastructure longevity. Ideally, technology stakeholders want the ability to easily migrate — in an online fashion — applications and data written 20 years ago to new hardware technology so that the leading application of the day can open, access, and read the data. Today, firms are forced to conduct application migrations every 12–18 months and hardware migrations every 3–5 years. These migrations are disruptive and time intensive, and they require planned downtime.

THE VALUE OF CONTENT ARCHIVING

What is content archiving, and why is it important? Content archiving is a process that involves policies and technologies to support how a firm intends to manage its information for the purposes of meeting regulatory and legal needs and supporting business and operational objectives. Content archiving solutions consist of the archiving software layer and the archiving storage, which both provide critical functions to enable a firm to formalize policies and processes around the management, disposition, storage, access, search, retrieval, and movement of fixed content. Table 2 illustrates the components of content archiving and the functions provided.

TABLE 2**Integration and Functions of Content Archiving Components**

Component	Examples	Functions
Archiving Software (database, email, file, ECM) – Interfaces to Applications	Email archiving: Symantec Enterprise Vault, CA Message Manager, Zantaz EAS	<ul style="list-style-type: none"> • Archive/index – move mail to archive and retain index for access • Mailbox management – reduce load on primary mail storage • Watchdog – review keywords and quarantine questionable language • Legal discovery support – search and case management tools for legal discovery • Records retention – set retention on emails to meet compliance standards
	Database archiving: Princeton Softech (IBM) Optim, Solix	<ul style="list-style-type: none"> • Ability to archive "business objects" – collect relationships and schema information • Rules engine – apply rules to archive, store and access enterprise application data • Report archive – extract commonly used reports for archiving
	File archiving: Symantec Enterprise Vault, Arkivio, CommVault Data Archiver, Enigma Data Systems	<ul style="list-style-type: none"> • Rules engine – automate file migration to other storage tiers • Migration agents – move data across various tiers of storage • Stubbing – pointers or "stubs" are left behind to ease retrieval of data • Archive – sets retention policies when moving fixed content to archive
	ECM/ERM: Open Text, FileNet, CA Records Manager	<ul style="list-style-type: none"> • Document management – check in/check out, version control, security, and library service for business documents • Records management – long-term archiving, automates retention and compliance policies • Document capture and imaging – captures paper documents • Document-centric collaboration – threaded discussions, project teams • Workflow – supports business policies, routes content, assigns work tasks, creates audit trails
Archiving Storage Solutions (disk-based) – Interfaces to Archiving Software	Hitachi Content Archive Platform (HCAP)	<ul style="list-style-type: none"> • Consolidated archive of different content • Object metadata model – stores content with metadata in open format • WORM ensures immutability of data, once written • Proactive authentication ensures content does not change • Open file system (NAS) and Web interfaces (HTTP) to ensure easy archive application integration • Enforces retention policies and litigation hold at storage layer • Federated search capabilities across all content/archive types • At-rest encryption of all content in archive • Deduplication for storage optimization • Detailed logging of administrative actions • Nondisruptive, secure content migration across generations of storage hardware

Note: This list is not meant to be exhaustive.

Source: IDC, 2008

In using archiving technology to satisfy corporate information management policies, content archiving addresses both business and technology stakeholder objectives:

☒ **Compliance and records executives.** Time-based, event-based, and infinite retention policies can be executed. Retention policies set at the application level can be enforced and marked for retention and expiration, or optionally, retention policies can also be set at the content archiving layer. Unified auditing of security, access, and retention can be done in a consistent manner across many content repositories, and audit records can provide proof of compliance. During a regulatory audit, records can be located and retrieved in a unified manner, even across different content repositories. This streamlines audits and demonstrates a firm's consistent approach to compliance requirements.

Data integrity and permanence can also be satisfied with content archiving, which can verify that data hasn't been manipulated and chain of custody can be preserved. Behaviors taken with electronic systems and content can be monitored, supervised, and audited for compliance with internal and external policies.

☒ **Legal executives.** Preservation orders can be automated and active on electronic content. Hold orders and multiple holds on the same electronic record can be instrumented and managed. Duplicate records can be culled to reduce legal review cost. Retrieval times are faster with archiving technology than with manual tape processes. Unified search can be done in a consistent manner across many content repositories, and chain of custody can be proven and documented. This streamlines the electronic discovery process, reduces the amount of content required for review and production, and can reduce the firm's overall legal risk.

☒ **Human resources executives.** Sensitive employee, customer, or corporate information can be protected from compromise and inappropriate use. A firm's internal controls around content can be proven and verified. Corporate assets such as intellectual property or financial information can be secured. Unified search of multiple corporate content repositories can be conducted to ensure that no human resources policies have been violated or to support claims of violation.

☒ **Risk management executives.** In addition to meeting the legal and compliance stakeholder requirements, risk management professionals, who commonly take a holistic view of corporate content, are better able to manage corporate assets according to larger corporate governance and key performance indicator metrics, which allow the firm to highlight its role as a responsible corporate citizen and reduce overall financial, business, regulatory, international, and legal risk.

☒ **IT executives.** IT stakeholders can proactively meet the needs of the business stakeholders and avoid costly administrative technical cycles to support reactive legal and regulatory requirements. Additionally, technology stakeholders can reduce capital and administrative costs, leverage existing investments in infrastructure, and provide a federated search across many repositories, thus improving visibility and minimizing IT risks.

WHAT IS NEEDED TO SUPPORT CONTENT ARCHIVING?

Common Data Management Principles

Sheer content growth and the numbers of corporate content repositories have stifled the application of disciplined data management principles to unstructured content repositories. Distributed, disparate corporate content repositories have lacked the management discipline traditionally applied to structured databases. There has been no instrumentation of a common set of controls based on policy establishment and documentation. There has been no unified approach to policy, policy execution, monitoring, and auditing. Each content repository has been managed as a discrete business entity. Therefore, applying structured data management to unstructured content — including data-related policy; data ownership and responsibilities for ensuring legislative compliance; data documentation and metadata compilation; data quality and standardization; and data access and dissemination — would serve both business and technology stakeholder objectives.

Access and Policy Services

Firms are moving to a content world and are disaggregating the application from the content via Web services and component- or composite-oriented architectures. As a result of these dynamics, the role of applications is changing. Content archiving solutions should provide access and policy services that allow for next-generation applications and agile IT environments. Access and policy services should:

- ☒ **Work with existing applications.** Both commercially available and internally developed content applications must be supported. This should be done in an open, standards-based manner supporting methods such as NFS, CIFS, HTTP, WebDAV, and XAM.
- ☒ **Operate in federated fashion.** Content archiving implementations should not force migrations as a rip-and-replace approach. Content archiving should support the ingestion and policy execution of many different systems under a common architecture. However, federated content mining, search, and discovery should not require migrations.
- ☒ **Support longevity of the application and data.** Content archiving should provide an online, nondisruptive migration of both the application and the content format to next-generation application and file formats. This should be transparent to users and their expected access patterns.
- ☒ **Enable infrastructure services.** Content archiving should be able to ingest and manage content from multiple, disparate content repositories; execute on application (or infrastructure set) policies; and then enforce, audit, and verify these policies were executed over the life of the content.
- ☒ **Provide centralized policy execution.** Policies passed down from the application layer — such as retention, preservation, disposition, security, and access — should be enforced and executed by the infrastructure services.

Infrastructure Services

As with access and policy services, firms are responding to changing market dynamics, including:

- ☒ **Dynamic tiering.** With the proliferation of virtualization, pools of storage are created that allow for more dynamic tiers of content based on its business, legal, and regulatory value. Intelligent, dynamic tiering allows for automated, transparent movement of content within the storage cloud, based on algorithms around the content, its history, and its usage patterns.
- ☒ **Physical migrations.** In addition to data and application migration, migration of content from older physical media and storage infrastructure to next-generation technologies is required. This migration, like its logical counterpart, should be done in an automated, nondisruptive, and online fashion.
- ☒ **RAID.** Proven, established RAID technologies allow for redundancy and resiliency from storage hardware and disk drive failures. Supporting investments in existing technologies, content archiving should work with state-of-the-art RAID services up and down tiers of RAID storage devices.
- ☒ **Virtualization.** In the next 10 years, storage environments will be increasingly virtualized. This move to virtualization will be driven by increasingly dynamic and mobile IT environments, compressing recovery and business continuity objectives and the need to reduce complexity while managing increasing storage capacities.
- ☒ **Copy and replication services.** The infrastructure must provide control and management for not only a single copy of content but also the copies that might be created for business continuity and disaster recovery objectives. The controls should be applied to all copies of a piece of content and track the number, locations, and policies associated with the content as well as the master copy. Replication services serve to provide a copy at an offsite location for disaster recovery.
- ☒ **Immutability.** Increasingly, firms require that data not be tampered with or deleted without proper authorization or reaching of a record's end of life. Otherwise, the consequence can be claims of spoliation during ediscovery or raised questions about the authenticity of a piece of information. WORM technology ensures that records are not able to be modified or deleted by an unauthorized user, application, or administrator.
- ☒ **Retention.** Retention at the storage layer provides a physical guarantee and enforcement that data cannot be deleted or modified until the retention period has been reached.
- ☒ **Proactive authentication.** Infrastructure services can provide for proactive authentication using a naming schema whereby each unique piece of data is named with a distinct hash value. This hash calculation is performed before a piece of data is sent to a device and after it has reached its target location and periodically while it is stored. This process ensures the data does not become corrupt in transit or over time as it is stored.

- ☒ **Encryption services.** With sensitive corporate data moving across networks, devices, and locations, security is a paramount requirement in most large enterprises. With archive data, which holds sensitive content about customers, clients, patients, or employees, or even corporate intellectual property, encryption must be available and transparent. Infrastructure services must provide this level of protection of data and content from unauthorized access.
- ☒ **Deduplication services.** Infrastructure services can enable the identification of and eliminate the need to store duplicate data. A deduplication process that replaces duplicate data with references to a shared copy in order to save storage space is mainstream in today's archiving environments. This process also serves to speed discovery and review processes.
- ☒ **Destruction and shredding.** The infrastructure must provide for tracking, retention, and expiration of content and support disposition and destruction of content based on the business rules. This should include, if applicable, the overwriting for the content to specific government standards such as DoD 5015, which requires shredding of content by overwriting it up to seven times.
- ☒ **Policy enforcement.** Policies set at the application level, such as retention, preservation, disposition, security, and access, must be enforced by the infrastructure services, not only across a single copy of content but also across all copies created by the infrastructure.
- ☒ **Scalability.** With data growing between 60% and 100% year over year and content repositories growing in number of files, size, and quantity, scalability requirements for a decade from now must be designed into today's products. First-generation archiving platforms were not able to meet current and future object number, storage capacity, ingestion performance, and rebuild requirements.

ONGOING CHALLENGES

The industry at large still faces a series of ongoing challenges from technology, standards, and business process perspectives.

Backup Versus Archive, Different Workloads

Many firms are still confused about the differences between archive and backup. Backup serves as a process to provide for restoration and recovery in the event of corruption of the primary system or data. Backups are typically done periodically, the media used is rotated, and older backups are destroyed. Conversely, archive is the explicit function of preserving and ensuring retention of specified records for the purpose of regulatory compliance, discovery, and general business use. The hardware and software used to archive and discover information is very different from the hardware and software used for traditional backup and to provide policies for retention, preservation, and disposition.

Legacy Data, Old Formats, Old Media

Many organizations have infinite or permanent retention of critical business or financial records. However, the application that was used to create the data may not be around to support recall of the data if it is required 100 years from now. Moreover, the media on which the data lives 100 years from now may no longer be interoperable with the application on which the data was originally written. The industry is still in a state of standards development, and standards are needed to ensure adequate levels of interoperability between the technology of the data and the data created 100 years prior.

In-Place Retention Approach

Records and content management applications allow users to set a retention policy that can be enforced with the native application. This can be managed for content within the repository; however, as previously described, most firms have many different content management systems as well as email, instant messaging, file systems, and so forth. Thus, the question becomes, How can firms apply a consistent set of retention policies across all these disparate systems? As content becomes fixed, firms can migrate or move content from the primary location to a secondary location but still within the control of the content management application.

Lack of Centralized Policies

Business and technology stakeholders in leading-edge firms are coming together to form steering committees and cross-functional teams. These teams are putting together policies for how firms manage their content. However, the risk is that the nature of these policies can still be specific to an application, a division, a geography, or a department. For large companies that have made investments in compliance, governance, or risk management, this situation is changing, but it will take time for centralized policies to be adopted across all large firms.

SUMMARY

Corporate content is vital to a firm's long-term viability. Corporate content needs to be managed to avoid risk, but it is also being mined and analyzed to derive business advantage for a firm. Content archiving is at once a business and technology approach to meet these reactive and proactive objectives. Applying structured data management principles to a firm's content is a means to derive business advantage from unstructured corporate content.

Copyright Notice

External Publication of IDC Information and Data — Any IDC information that is to be used in advertising, press releases, or promotional materials requires prior written approval from the appropriate IDC Vice President or Country Manager. A draft of the proposed document should accompany any such request. IDC reserves the right to deny approval of external usage for any reason.

Copyright 2008 IDC. Reproduction without written permission is completely forbidden.