

WHITE PAPER
October 2005

Discovering the Business Value of Archived Information

ABSTRACT

The value of archived information (and even the definition of “archived” information) eludes many of today’s IT decision makers. Unless driven by specific mandates or by defined applications (and hence clear business needs), most IT shops are missing opportunities to provide competitive advantage through the repurposing of historical information stored in archives. This paper explores the potential business value of information contained in the 30-plus exabytes of data stored in archives today.

1	Executive summary	2
2	Introduction	3
3	A historical perspective on archives	3
4	Intelligent archiving	4
5	Sources of archived data	5
6	The amount of archive information	6
7	Uses of archived information	7
8	The business value of archive information	8
8.1	The amount of archive information	8
8.2	Customers, suppliers and partners	8
8.3	Improved decision making	9
8.4	Cross-functional alignment	9
9	Key applications using archive information	9
10	Moving forward	10
11	Summary	11

1 Executive summary

Archiving is the capture, management and retrieval of fixed-content information for the purpose of its use or reuse. Archives saved without significant attention to retrieval are difficult to access and manage. In addition, large organizations often have many disparate archive repositories. An approach that supports using archived information to drive business value is needed. We call this “intelligent archiving” — archiving optimized for business value.

Intelligent archiving is an integral part of an information lifecycle management (ILM) implementation. Information lifecycle management is a sustainable storage strategy that balances the cost of storing and managing information with its business value. ILM provides a practical methodology for aligning storage costs with business priorities. Key functions of information lifecycle management include archive, advanced data protection and virtualization.

In our interviews with high-level IT executives, we have uncovered the sources and uses of archive data that are common today. Our discussions resulted in a list of key uses that can provide guidance to IT and business decision makers as they work to derive value from archives. The key application areas identified include:

- *Externally facing applications (customers, suppliers and partners) to create revenue and business advantage* — Most examples of effective use of historical and current archive data that have generated value involve providing information to external users. These applications often leverage relationships and affect business results. In addition, several of our discussions raised the subject of an additional benefit, the advantage of reducing cost via customer self-service.
- *Customer, partner and supplier support* — Applications for internal users who provide external support (such as help desks and sales support) can leverage archived information and provide benefits similar to those discussed above.
- *Discovery to serve compliance needs* — Although compliance-driven data is not a significant percentage of archived data, compliance and discovery are certainly in the “front of mind” and are likely serving as a galvanizing force in evolving the management of archived data. Organizations that turn compliance archives into information assets that drive business value will have an advantage over those that see compliance only as onerous.
- *Knowledge worker applications for decision support and collaboration* — These applications were mentioned several times during our interviews. Historical information, the significant fixed content created daily by desktop applications, and extracts of information stored in databases often end up in archives. Unleashing this information to support decision making and to help align cross-functional groups is a key benefit of intelligent archiving.

META Group asserts that research shows today’s records management functions are little more than “document dumping grounds” with limited organizational impact.¹ We believe that there are cultural, operational and technical reasons why archived data is not being used to its fullest potential in driving business value.

¹ META Group, “Effective Information Asset Management Strategies for SOX Compliance and Beyond” (Stamford, CT: META Group, November 2004).

- *Archiving is the capture, management and retrieval of fixed-content information for the purpose of its use or reuse.*
- *Historically, archives were based on paper, film and fiche. Increasingly, archives have been digitized.*
- *The repurposing and reuse of archive data is a relatively new phenomenon.*
- *Archives created without retrieval as a key design point are problematic*

The solution involves strategy, process, systems and infrastructure. Information asset management is a discipline that provides approaches to managing information to drive its business value. Linking intelligent archiving, information asset management and advanced applications strategies, organizations can realize the potential value hidden in their archived data. Historical and current archives can and should be optimized for use, generating revenue and driving competitive advantage and the loyalty of customers, suppliers and partners.

2 Introduction

Discussions of archive seem to start with confusion over what is meant by the term. Merriam-Webster defines the term “archive” as [noun] “a place in which public records or historical documents are preserved; also, the material preserved” and [verb] “to file or collect (as records or documents) in or as if in an archive.”²

Information technologists define archive as “a collection of computer files that have been packaged together for backup, to transport to some other location, for saving away from the computer.”³

No wonder it’s confusing. IT people apparently think archiving is about long-term data protection, and business people think archiving is about preserving information. In this paper, we will use what we consider business-oriented definitions. Later we will provide some insight into why this distinction is important.

- We define backup (or, more broadly, data protection) as any activity that copies, replicates, logs or moves data for the purpose of safeguarding active, changing information.
- We define archive as the capture, management and retrieval of fixed-content information for the purpose of its use or reuse.

3 An historical perspective on archives

Archiving has been evolving for many years. In an historical context, archives were paper copies of letters, invoices, statements, drawings and so on. The material, originally filed in cabinets, was moved to boxes and shipped to warehouses for long-term storage. The boxes were labeled, usually by content or topic (e.g., customer) and date.

Along the way someone got the brilliant idea that archived information could be stored in digital form. The rapid adoption of personal computing, image applications and electronic mail added to the growth of digital archive data.

Today, most organizations have more fixed-content data than dynamic-database data. The metaphoric boxes we have been storing it in are still labeled by file names and dates. Most organizations admit that their approach to storing historical information has left a good portion of it inaccessible.

² Merriam-Webster Online, <http://www.m-w.com> (June 2005).

³ TechTarget, Whatis.com, <http://whatis.techtarget.com> (June 2005).

Some organizations are beginning to recognize the value of archived information as a tool to serve customers and knowledge workers. What was once stored with the fervent hope that it would never need to be retrieved is now recognized as an untapped information asset. Unfortunately, archives have a problem; they were stored as a great haystack of information. And now someone is asking for the proverbial needle. Oops.

4 Intelligent archiving

A new approach to archiving is needed to serve the business needs for use (and reuse/repurposing) of archived data. Archives saved without significant attention to retrieval capabilities are often difficult to access. An approach that supports using the archived information to drive business value is needed. We call this “intelligent archiving” – archiving optimized for business value. The basic requirements of intelligent archiving include:

- **Capture** — The sheer amount of data being captured implies that capture operations must be managed cost-effectively.
 - Concurrently, metadata, content access and classification of information must be generated to support policy-based management and topical access to archived information.
 - Capturing data at its source and assuring that content/context meta-data is immediately associated with it is key to creating archives with business value.
 - Backup applications are optimized for restore operations, generally file- or volume-driven. Archive data must be accessible by content or topic.
- **Manage** — Data volumes and IT cost-containment efforts drive a focus on cost-effective and automated management of archives. The following points are among management aspects of an intelligent archive that drive business value of archived information:
 - *Compliance and governance requirements must be served.* The archive must be capable of data provenance via write once, read many (WORM), audit controls and serialization capabilities. Compliance data repurposing may provide significant positive benefit to requirements often associated with cost only.
 - *Retention management is key to achieving business value from archived data.* Keeping data that is actively referenced, maintaining data that is being held for governance reasons, and disposing of data in an intelligent manner once it has lost its business value and expired from regulation represent significant complexity and potential conflict. Intelligent retention management must address this complex environment to ensure that business value and business risk are managed concurrently.
 - *The archive must have capabilities that ensure data protection of archived data.* By its very nature, archived data is static. Therefore, data protection for archive data does not involve the same “time or window pressure” as dynamic data protection. However, there must be provisions to keep remote copies of archived data. Valuable data should be well protected.

Intelligent archiving

- *Optimization of archiving for business use of archive information*

Capture

- *At creation time, capture of information, metadata and context*

Manage

- *Support for compliance and governance requirements*
- *Protection of archive data*

Use

- *Provision of topical access from multiple applications*

Archive data comes from five primary sources:

- *Artifacts of back-end applications*
- *Documents, presentations and spreadsheets*
- *Digitization*
- *Image capture*
- *Communications*

- Policy-based, automated data management and intelligent data movement across a hierarchy of storage can help to ensure that storage and storage management costs are contained. Cost containment is important to achieving business value, as it allows storage of a broader spectrum of information for longer periods of time, within governance standards.
- **Use** — Enabling access to archive data via content-based access is critical to achieving business value. It is in fact, the only real source of business value from archived data. Simply put, archive data has business value upon use (or reuse/repurposing). Capturing it and managing it without using it are somewhat useless tasks and expenditures.
 - An open architecture with multiple application support ensures that access will be available as front-end applications technologies change.
 - Intelligently managed tiered infrastructure helps to ensure that access requirements can be met while maintaining the lowest possible cost. Policy-based staging of information to high-performance tiers makes archived data available to business applications with appropriate performance characteristics.

5 Sources of archived data

Understanding archives by tracing their sources and uses may provide some insights into the usefulness of archived data. The following is a taxonomy of sources:

- *Artifacts of back-office applications* — Transactional workflow management applications — enterprise resource processing (ERP), customer relationship management (CRM), and software configuration management SCM, among others — produce artifacts that are different from raw data. They produce descriptions of transactions, such as orders, that contain the complete information associated with the transaction (e.g., the buyer's name and address) and are a composite of specific data elements in the transactional application databases.
- *Documents, presentations and spreadsheets* — The ubiquitous Microsoft Office products (even without Outlook) may create more data every year than any other source.
- *Images* — Digital image capture applications (X ray, magnetic resonance imaging [MRI], digital video, check image and so on) produce medical and other images. This source is generally industry-specific and creates a significant quantity of data annually.
- *Digitization* — Historical data in paper, image or analog formats is an additional source with a unique impact. There is a “digitization penalty” associated with capturing analog/ image data in digital format. The process is (in general) inefficient: A 35-mm photo digitizes at 5 megabytes, whereas a similar size (and quality) digitally captured 1,600 x 1,200 pixels per inch (ppi) JPEG is about 750 kilobytes. Nevertheless, a significant number of large firms have digitization initiatives under way.⁴

⁴ Enterprise Strategy Group, “Reference Information: The Next Wave” (Milford, Mass.: Enterprise Strategy Group, 2002).

- *Communications* — Applications such as e-mail, instant messaging, conference and voice applications produce text, attachments, messages, digitized voice messages, conference recordings and so on. This may be the fastest-growing source of fixed-content archive data. Some of the estimates of the potential impact of voice digitization are truly impressive, suggesting a potential doubling (or more) of archive data volumes in a very short time span.

6 The amount of archive information

- *Today, archives may account for as much as 30 exabytes of data worldwide.*
- *The portion of that data with high business value is unclear.*
 - *We estimate that about 10% of archived information has been promoted to disk solutions.*
- *Intelligent archiving will obviate the “either-or” (disk versus tape versus optical) decision by managing data across tiers of storage based on policies.*

Archived data is a significant majority of the data stored by most organizations. Some estimates place archive data at five to eight times active working data (including the data protection copies of both data types).⁵ Several different assessments of “archived” data put the range of archive data from 20 to 30 exabytes, mostly depending on what’s being counted. In any event, the amount of archive data is huge and is growing fairly fast, in the range of 30% to 50% annually.

Certainly archives contain copies. It seems fairly safe to assume that the 20 to 30 exabytes of data represent no more than 10 to 20 exabytes of information and are probably on the lower range of those numbers. The interesting question is, “What portion of that has business value?”

Some organizations have a good handle on their archives, having built them with use in mind. Others have been keeping data without a clear vision of its importance or potential use.

Our interviews with large IT organizations indicated a broad range of estimates of data versus usable information. IT executives estimated that one third to two thirds of their data in historical archives was valuable. One notable comment: “It’s hard to tell...occasionally we have found a surprising use of historical data, in a way that no one had previously conceived.” The same respondent went on to aver that “it’s likely that the surprises will yield the most value.” We are not astonished that the “surprises” tend to yield “disruptive value” and to engender new ways of doing business and competitive advantage.

For the past several years, IT organizations have faced an either-or decision about where to store archive information. Without intelligent archiving, the choices — disk, tape or optical — have been either-or. One of the clear indicators of information value is promoting data from a lower-performing tier to a higher-performing tier, generally from tape or optical to disk. We certainly do not mean to imply that there are not many high-value archives on automated tape libraries; there are many examples of high-performance tape automation serving active reference information repositories. Good indications of the volume of archives having high business value are the current volume of archive data on disk systems and the migration rate from tape to disk. In 2005, almost a full exabyte of reference information will be stored on “server-attached storage.”⁶ Based on total archive volumes of 10 exabytes of *information*, this means that at present, about 10% of archived data has proven to be so valuable that it has forced a platform decision (and likely resides concurrently on multiple platforms). It is likely that, in an intelligent-archive approach, far more information would be “promoted” (at least on a temporary basis).

⁵ Freeman Reports, “The Growing Importance of Archive” (Ojai, CA: Freeman Reports, 2005).

⁶ Enterprise Strategy Group, “Reference Information: The Next Wave” (Milford, Mass.: Enterprise Strategy Group, 2002).

Archive data is used in five major areas:

- *Discovery to serve compliance needs*
- *Externally facing applications (customers, suppliers and partners) to create revenue and business advantage*
- *Customer, partner and supplier support applications with internal users*
- *Knowledge worker applications for decision support*
- *Collaboration applications*

7 Uses of archived information

The following current uses of archived information all center on being able to access the data by topic via content or context rather than by file name(or across file names).

- *Discovery to serve compliance needs* — Picture responding to a request from your legal department to “supply all transaction records, communications and documents pertaining to customer XYZ from June 15, 2002, to October 15, 2002.” Our interview respondents indicated that a fair amount of work would be involved. The data would be spread across many files and tapes and would require multiple restores and queries, as well as considerable human resource. It’s possible, but cumbersome and expensive. Generally, infrastructure-independent, content/context access is required to make such applications feasible.
- *External user applications* — Customer, partner and supplier applications all have high potential for using previously archived data (e.g., statement histories held by consumer banks) to create competitive advantage. Some of these applications will also involve “customer self-service,” driving down administrative costs.
- *Service applications* — Customer, partner and supplier support applications (e.g., help desks) may leverage historical information to provide higher levels of service and improved satisfaction and retention.
- *Knowledge worker applications* — These include using information for improved decision making (e.g., an energy company using historical seismic data to model oil reservoirs; a consumer goods company researching historical point-of-sale records to determine purchasing patterns).
- *Real-time collaboration* — Combined with conferencing technologies, current fixed-content and historical archived data can be key to ensuring that organizations have cross-functional alignment across various teams and broad access to information.

There are several problems associated with reformatting historical data to make it consistent across time and usable by current (or planned) applications. Data integration efforts are often involved to resolve different formats and data points. Generally, we have been advised that the “reconstruction” of historical data into a usable reference archive can be extremely difficult.

New applications and newly created fixed content being archived are less of a problem. In these instances, data formats are current, and there is a reasonable likelihood that the data is being stored with retrieval in mind.

8 The business value of archive information

Table 1 summarizes the high-level business benefits of the use of archived information

- *Using archived information to support revenue-generating activities will change the current view of compliance from “compliance equals cost” and “compliance is good medicine” to “compliance information can drive revenue.”*
- *Externally facing applications — self-service or supporting customers, partners and suppliers — can make use of archive information to leverage loyalty, reduce cost and drive revenue.*

Application area (archive data use/reuse/repurposing)	Business benefits
Discovery	<ul style="list-style-type: none"> . Reduced business risk . Compliance cost management
External user applications	<ul style="list-style-type: none"> . Increased loyalty and retention driving increased revenue and market share . Cost reduction via the transfer of internally staffed tasks to customers, suppliers and partners — e.g., self-service applications that replace help desks
Service applications	<ul style="list-style-type: none"> . Improved service to customers, suppliers and partners
Knowledge worker applications	<ul style="list-style-type: none"> . Improved decision making — additional information
Real-time collaboration	<ul style="list-style-type: none"> . Alignment of cross-functional teams and improved decision making

8.1 Reduced business risk and compliance cost management

Using an intelligent archive approach to storing compliance data for legal purposes provides the business benefits of risk reduction and cost containment. Intelligent archiving provides the functionality required to meet regulatory requirements, automatically manages tiered storage, and can provide opportunities for reuse of compliance data. We have all heard the theme that “compliance is good medicine for business,” implying that doing compliance well forces better practices and policies.

We believe that the real gem in compliance applications has yet to be exposed. The information being stored, while being driven by regulation, has value in serving customers, suppliers and partners more effectively and generating loyalty and revenue.

8.2 Customers, suppliers and partners

Applications using archived information to improve customer, supplier and partner services will provide very strong and easily justifiable business benefits. Loyalty, retention and revenue will be directly influenced by leveraging archive information. Our interviews indicated that the easiest way to find value in archive information is simply to ask the question, “What information would customers like to have that they currently don’t have?”

One CIO from the automotive industry suggested that if he could get historical service information out of the archives, he could leverage customer retention by measurable percentages. Since his company was generating profit measured in the billions of dollars, his comment that a quarter point of market share represented “real money” was taken as a serious indication of business value. An operations executive from a public-sector company described an opportunity her organization has to leverage archived information for a self-service application that would significantly lower administrative costs and improve service for citizens.

Improved decision-making

- *Applications using archive information to leverage decision making, while difficult to justify with “hard” benefits, often have strong value*
- *Real-time collaboration will leverage the information in archives within and between companies, customers and suppliers, creating improved alignment.*

In both of these cases, the use of archived information as “reference information” drove strong business benefits. The key to defining the opportunity was to ask the question, “What’s in the archive that our external users would find valuable?”

The key benefits of using archive information for external applications — and for internal applications that support customers, suppliers and partners — are increased loyalty and retention driving increased revenue and market share, and cost reduction via the transfer of internally staffed tasks to customers, suppliers and partners.

8.3 Improved decision making

Generally, this type of business justification is very difficult to quantify. In our interviews, the approach to valuing new information ran from the simple (it depends on who demands it) to some very complex calculations based on the value of information and financial options-valuation theory.

Financial options are valued by using techniques such as Black-Scholes (the “simple” approach) or a binomial pricing model (the complex approach). Using intelligent archiving to mine previously inaccessible data gives you the option, but not the obligation, to act on the information you have discovered. Just like a financial option, this information option has value.

8.4 Cross-functional alignment

The benefits of “real-time collaboration” within and between organizations and businesses are derived from innovation, adaptability and speed to market. Business performance and growth are highly dependent on knowledge and knowledge-sharing activities. Archived data is an important element of collaboration applications. Historical reference material is critical to maintaining a high level of institutionalized knowledge in service industries and in the development and distribution activities of most organizations. Intelligent archiving can serve collaboration applications by providing an effective repository of knowledge that can be shared by multiple applications and users. Today “collaboration” is mostly served by tactical solutions that focus on the delivery layer. In the future, “contextual collaboration” will span businesses and external users to leverage a knowledge-based approach to business.

9 Key applications using archive information

Some of the application areas discussed in the previous section are applicable across all industries, and some are likely to be industry-specific. A review of current application development forecasts and input from interviews suggests that the following applications will be key in the next few years and will make extensive use of fixed content archive information.

- *Single customer view (customer data integration)* — An accurate and timely repository of all customer information. The term customer data integration (CDI), which started as a value proposition from marketing service providers Acxiom and Experian, is being touted as the next “killer” application.
- *Clinical data repository* — A repository for clinical information, organized in a patient-centric manner, supporting computer-based patient record systems (CPR) driven by healthcare and insurance companies.

Key business applications using archived information:

- *Single customer view — customer data integration*
- *Clinical data repositories*
- *Collaborative development*
- *Content and rich media distribution*
- *Supply chain management*
- *Product lifecycle management*
- *Product information management*
- *Financial value chain management*

- *Collaborative development* — Integration of workflow with co-workers, suppliers and customers “as if they were in the same room.” Generally, this includes technologies for instant messaging, group chat, whiteboard collaboration, application sharing, desktop sharing and co-browsing.
- *Content and rich media distribution* — Commercial and government distribution systems that support knowledge workers with multiple types of integrated content. It includes collection, aggregation, and distribution/presentation of content to businesses, organizations and consumers.
- *Supply chain management* — Oversight of materials, information and finances as they move in a process from supplier to manufacturer to wholesaler to retailer to consumer. Supply chain management involves coordinating and integrating these flows both within and among companies.
- *Product lifecycle management (PLM)* — The management and control of all product-related information over the entire product and asset lifecycle.
- *Product information management (PIM)* — A single source for distribution of product information and content to multiple channels, internal users and customers.
- *Financial value chain management* — The monetary part of the overall value chain for inter-enterprise trade, which also includes goods/services and content movement components — intra- and intercompany financial transaction-based functions and processes linking suppliers, customers and the financial community.

10 Moving forward

Driven by e-mail volume, hyper-competition and compliance, many organizations are examining the ways they are storing and using archives. What they are finding is that their archive repositories are fragmented physically and logically. The most usable ones are functional or departmental, and not a corporate wide information asset. Their deep archives are often stored in backup formats and are difficult or (close to) impossible to access on a topical or content basis.

Progress is being made; almost 10 percent of these archives have been (recently) promoted to disk or were stored on disk initially. Recently stored archives (in particular e-mail archives) are being designed for recall and use, as opposed to being a dumping ground for infrequently used data.

We recommend several steps in discovering the value of archived information:

1. The obvious first step is to conduct an inventory. As one of our interview participants said, “It’s pretty hard to make much progress until you know what you have.”
 - 1.1 *Inventory data stored in so-called “deep archives.”* This includes inactive data offline and in libraries, juke boxes and vaults. Include the physical information assets (paper, film, fiche) that also represent potential digitization targets.
 - 1.2 *Inventory data artifacts currently being stored in application, departmental and functional silos.* Analysts’ studies over the past few years have indicated that in organizations of more than 1,000 employees, it’s not unusual to have more than 40 disparate repositories of archived data.

Remember that this is how true archive (information stored for the preservation of records) evolved. There are probably more of these than appear obvious at the surface. One of our discussions indicated that the participant's business had more than 200 disparate repositories, most of which were not known to IT prior to the firm's inventory project.

2. Identify opportunities to enhance projects in the current development portfolio with current fixed-content and historical data. This is probably the most interesting part of discovering business value in archives. In this paper, we have included a list of application areas for reuse (external applications, support applications, decision support, collaboration and, of course, compliance). We have also explored some industry trends in applications that will make use of archive data. For many organizations, an assessment of these opportunities by an outside firm would be useful in adding some external creativity to this part of the process.
3. Integrate into information asset management (IAM) plans. As you formalize "managing information as a strategic asset" and implement your IAM function or organization, ensure that the charter extends to all information assets — online, nearline, offline and physical. As your organization seeks to derive the business value extracted from information and optimize information use, ensure that archives are viewed as a key resource. Integrate IAM plans into your applications portfolio plans.
4. Implement infrastructure to support your plan. Because applications drive demands for storage services, the appropriate and flexible supporting infrastructure is critical. Intelligent archiving is nascent. As the market matures, expect several innovations, new entrants and competition. Don't allow your organization to get caught in the either-or solutions that vendors have been presenting for the last several years. If additional infrastructure is necessary to address a current need, install infrastructure that can later be embedded into a single-instance intelligent archive.
5. Execute. Enjoy the competitive advantage, improvements in customer loyalty, and cost and risk reductions associated with the effective use of archive information to serve customers, suppliers, partners and your own knowledge workers.

11 Summary

Information lifecycle management is a sustainable storage strategy that balances the cost of storing and managing information with its business value. It provides a practical methodology for aligning storage costs with business priorities.

Information lifecycle management requires a hierarchy of infrastructure to provide cost benefit and the lowest possible total cost of ownership while meeting service level requirements. Its set of services includes data protection, archive management and virtualization. The integration of retention, protection and virtualization for data (outside the primary data path) is fundamental to the ILM vision.⁷ The services can be presented as different types of storage-related applications based on customer needs. The linkage of archiving within information lifecycle management to other key services optimizes storage management efficiency and effectiveness.

⁷ *StorageTek, www.StorageTek.com, "Information lifecycle management vision; White Paper" (Louisville, Colo.: StorageTek, 2005).*

Obtaining a high level of business benefit from archiving requires that the capture, management and use of archive information be optimized for information access. Both historical archives and currently captured archives represent opportunities to leverage customer, partner and supplier relationships and to improve the quality of your knowledge workers' decision making.

There is indeed "gold in them there hills." Mine it.

