An *IT Briefing* produced by SearchWinIT.com

Compliance and Information Stewardship
Lesson 2: Archiving and Retention

TechTarget
The Most Targeted IT Media

Sponsored By: IBM

Compliance and Information Stewardship Lesson 2: Archiving and Retention

By John Burke

BIO

John Burke, Principal Research Analyst with Nemertes Research, has worked in IT since 1988. He has worked as an end-user support specialist, programmer, system administrator, database programmer and report writer, database administrator, network administrator, network architect, and systems architect. As an analyst, Burke draws on his experiences as a practitioner and director of IT to better understand the needs of IT executives and the challenges facing vendors trying to sell to them.

This *IT Briefing* is based on an IBM/TechTarget Webcast, "Compliance and Information Stewardship Lesson 2: Archiving and Retention."

This TechTarget *IT Briefing* covers the following topics:

About TechTarget *IT Briefings*

TechTarget *IT Briefings* provide the pertinent information that senior-level IT executives and managers need to make educated purchasing decisions. Originating from our industry-leading Vendor Connection and Expert Webcasts, TechTarget-produced *IT Briefings* turn Webcasts into easy-to-follow technical briefs, similar to white papers.

For inquiries and additional information, contact:
Dennis Shiao
Director of Product Management, Webcasts
dshiao@techtarget.com

Compliance and Information Stewardship Lesson 2: Archiving and Retention

Introduction

This document focuses on archiving and retention in relation to compliance and information stewardship. The concept of information stewardship is reviewed, followed by an examination of the issues surrounding records retention, and how archiving and content management tools can be applied to the tasks of records retention and retrieval.

The information in this document has been developed by Nemertes Research. Nemertes is a research advisory firm that focuses on distilling the business value of emerging technology. To understand the trends and best practices around a particular issue, Nemertes draws from a pool of several thousand IT executives to track emerging technologies and to better understand how these executives are organizing themselves, the tools they are using, their favored vendors, and why and how they are spending their time and money.

Information Stewardship

One focus of Nemertes, over the last several years, has been an examination of the core set of disciplines surrounding the management of information in the enterprise. This set of core disciplines, as presented in Figure 1, when approached holistically as an interlocking, intertwining set of mutually supporting endeavors, is called information stewardship.

These disciplines set forth the idea that, for every byte of information in the enterprises, policies will be set,

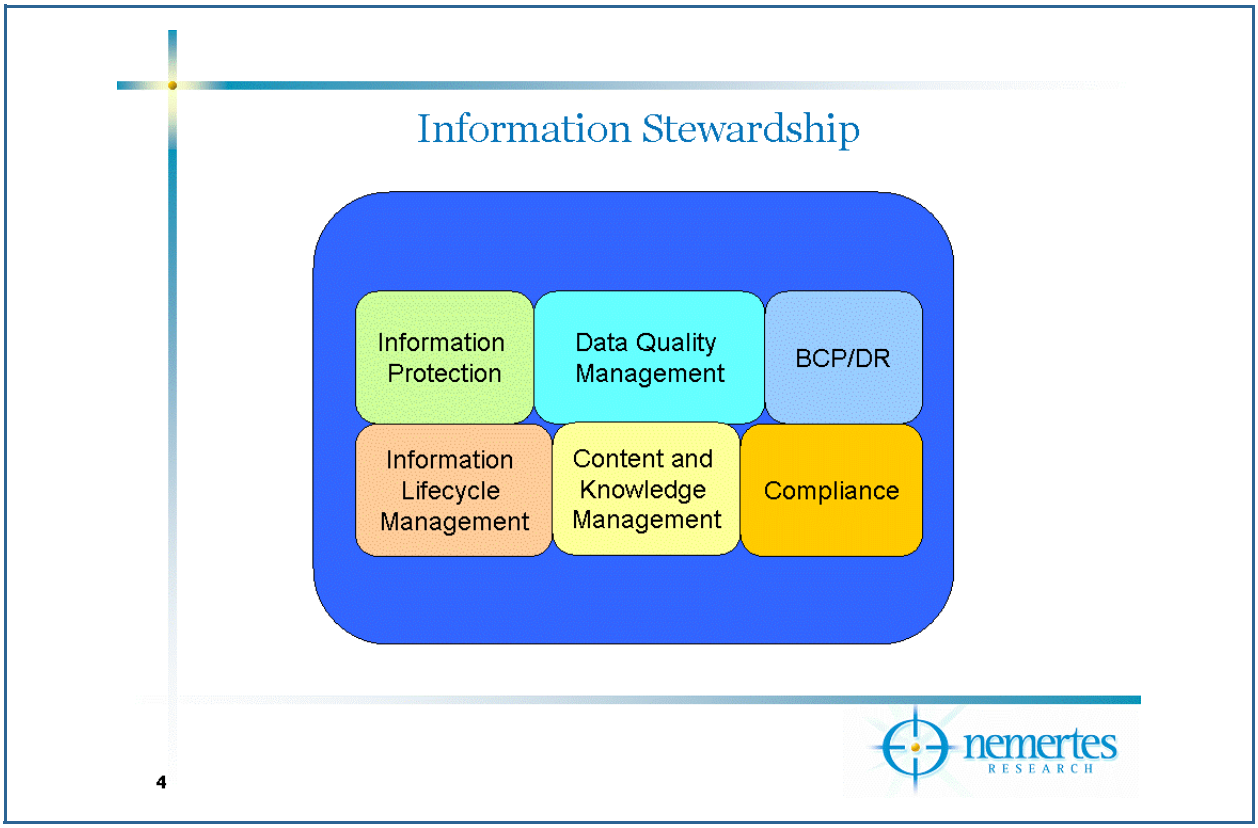

Figure 1

enforced, and audited. These policies govern how data is acquired, handled, and protected and that it is protected. This is not just in the sense of security, but also in the sense that it will be preserved and available in the case of disaster or other kinds of service disruptions.

The disciplines that make up information stewardship are as follows:

- **Information protection** encompasses not only protections such as access controls or encryption, but also the idea of being able to audit and track things such as who has accessed information and when it was accessed.
- **Data quality management** encompasses controlling data quality by reducing the amount of bad data entering the system. This eliminates the need to factor in bad data through the information life cycle.
- **Business continuity planning (BCP/DR)** ensures that the information is always available when needed.
- **Information life cycle management** ensures that information is where it needs to be, when it needs to be there.
- **Content and knowledge management** encompasses the idea of the necessity of using tools and systems to organize information, so that it can always be accessed in an orderly way.
- **Compliance ensures that information** is handled according with all the laws and regulations that apply to an organization, and that it can be demonstrated when called on to do so.

Why Retain Records?

One key aspect of information stewardship from a compliance perspective is the retention of records. This has implications for all the other parts of information stewardship.

Why retain records? From the perspective of businesses operations, records are retained for the following reasons:

- Operation—Records must be kept at hand or near at hand in order to conduct business. Records are required as proof of interactions with customers or, from a medical perspective, patients.
- Compliance—Records must be retained in accordance with laws or regulations dictated by federal agencies or by practice guidelines from professional organizations. Records are retained in order to demonstrate that an organization has behaved in accordance with the law or the regulation or the guidance.
- Litigation—Records are retained past necessity in order to provide an organization with information to draw upon and respond to in the event of lawsuits or for the purposes of discovery in a legal process.

How Long Should Records Be Retained?

Because there are three different reasons why enterprises hold on to information, there are also different lengths of time for which they will hold on to things. Figure 2 charts information retention rates.

Information that is retained for operational purposes is usually retained for a shorter period. For example, sales records may only be held for each day through the quarter in which the sales were conducted. Those numbers are then rolled up into quarterly figures and then those are rolled up into annual figures. Each kind of information is retained for a different amount of time, but not forever.

For compliance purposes, oftentimes the length of time that records must be retained is defined as a part of the law or the regulation. This period varies according to the law or regulation for which the record is being retained. It may be as little as three years after the close of transaction, for example. It may be as long as forever if the records involved are clinical trials that a pharmaceutical company conducted.

For legal or discovery purposes, all the same records that are typically held for operational or compliance reasons are held, but for a longer period of time. These records are held in case the need arises for them to be produced in a lawsuit. They are held in the event of a lawsuit in which the enterprise is either the defendant and needs to tell its side of the story or is the plaintiff and needs to document what has gone wrong and why the enterprise is seeking redress.

During the most recent benchmarking interviews on security and information protection, the expectation was that enterprises would retain records for as along

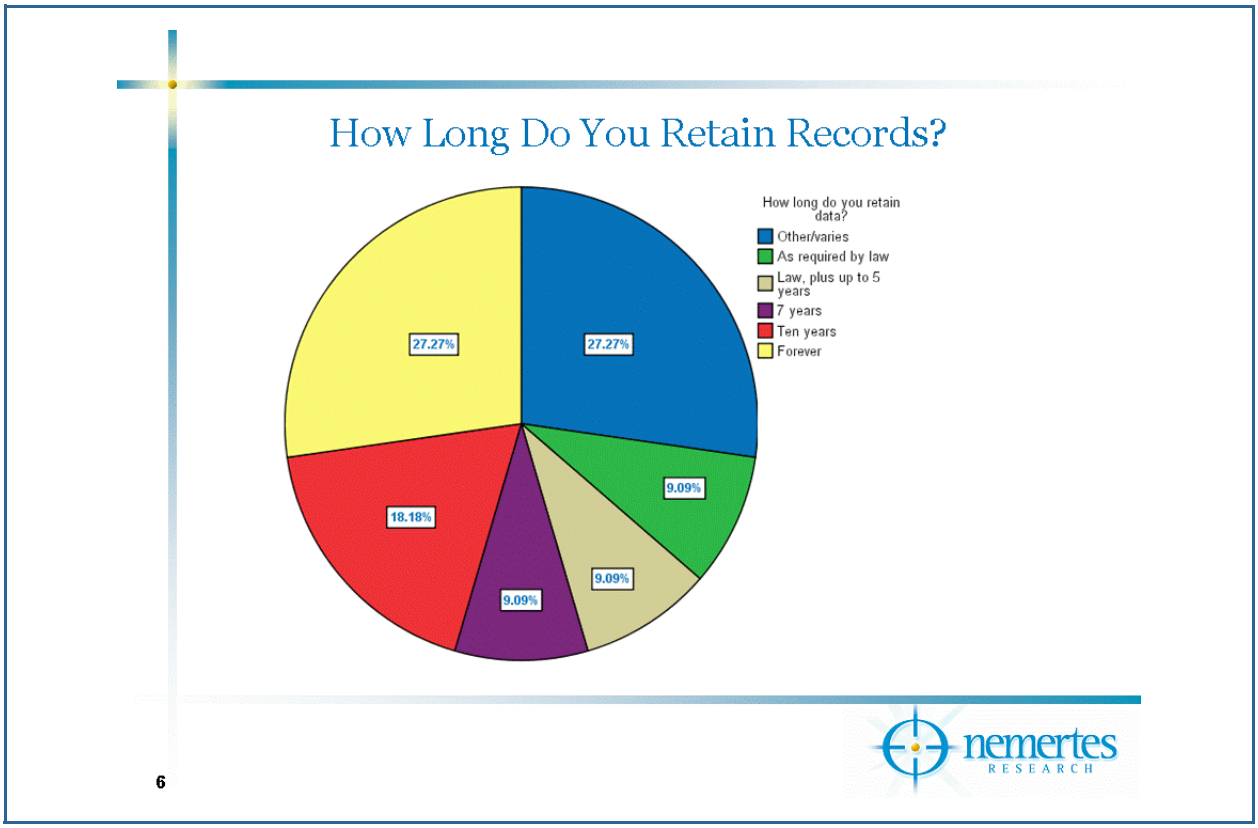

Figure 2

as required by law. However, it was found that the practice was in the distinct minority, making up about 10% of the responding pool. Surprisingly, three-quarters of the responding businesses either defined a fixed period for retaining records, if that was a superset of any of the regulatory requirements placed upon them, or they committed to a broadly varying set of retention practices, none of which was limited to what the law required.

Approximately one quarter of the respondents retain records forever; no matter what the requirements of the law, these enterprises decided to keep records permanently. This was somewhat of a surprise, not that enterprises do this, which was already understood, but that so many are doing it this way.

Records Retention as a Response to Lawsuits

Enterprises that retain records forever do so out of a fear of lawsuits. The worry is the inability to produce information during the course of a discovery process in a lawsuit. Retaining records forever is a response to lawsuits over the last three to five years in which companies were called upon to produce information that they were not required to retain or that had gone past the end of its required retention period, and they were unable to produce it. In many of these cases, it was seen that the companies were penalized for this inability to produce records in court, either by setting up a bad perspective on the company for the jury or by causing judgment against it during the production process and pre-trial motion by the judge.

The eventual results of these lawsuits in terms of either a settlement or a fine imposed by the jury ran into not just millions, but tens of millions of dollars. Therefore, these companies, which are typically large companies in particularly litigious branches of industry, have looked at the landscape in the courtroom and changed their practices. Rather than fall behind the uncertain shield of doing exactly what is required by law, these would take a more proactive approach and retain records if there was a good chance that they would ever be needed in court. These companies have weighed the risk of essentially unlimited damages being imposed upon them against the cost of retaining this information forever, and they have decided the costs were far outweighed by the risk and so are holding onto things forever.

What Records Should Be Retained?

With companies increasingly keeping records forever, the question that can be asked is, what records are being retained? The answer is that a company retains what it needs to operate for as long as it needs it to operate, and then retains what it needs to stay in compliance for as long as it needs to stay in compliance.

These companies have decided that retention periods need to be essentially forever. In some cases these companies have also decided that they could no longer rely on being in good standing with respect to the law, by retaining records as long as they needed to retain them under the law. They also could no longer depend on staying in compliance with the law with respect to what to retain for that fixed period of time.

These companies have seen the risk of being unable to produce information in court even if it was not something that was required to be retained. They understand that the risk of being unable to produce that information is greater than the cost of retaining it, and so are essentially retaining everything that they use in their operation. That means that everything, not just those things required by law, but everything required operationally gets preserved forever.

Data Classification

For enterprises that choose not to retain everything, the question of what to retain becomes a critical decision, and data classification is key to the decision making. Classification encompasses data in structured data systems such as databases, in semi-structured systems like e-mail, and in unstructured systems like Word and Excel documents. Each piece of information must be identified as to whether it is needed for operational, compliance, or possible discovery purposes. Data can be classified according to the following guidelines:

- Operational data is classified as data that is needed during the year and must be accessed quickly.
- Compliance data is classified as information that must be retained for some finite period.
- Discovery data is classified as data that may be asked for in a lawsuit. This type of data must be identified as being either protected or discoverable.

Having a rich system of classification is a key to effectively managing information through its life cycle, managing information for compliance, and managing information for discovery.

The Place of Information Technology

IT can lead the efforts in the enterprise to classify information, but it cannot do it alone. Typically, IT can make a very good first pass at classifying information for operational purposes. Some information needs to be kept online and fast, some can be off-line, and some can be near-line, for example, online but slow. IT can make these determinations because it wrestles with these issues on a daily basis on behalf of the enterprise. However, when it comes to classifying information for compliance or discovery purposes, IT really needs the involvement of the legal staff, the Human Resources Department, and the business line. IT is an excellent place to begin the effort and to bring in the other lines of the business and the other departments as quickly possible.

Retaining Data Is More than Just Storing It

Retaining data is not just about placing it on a disk somewhere, although this is often done. Retaining data means securing it, not just in operational use but in whatever form it will be permanently retained. Also it must be shown that the data is secured properly for compliance purposes. For example, it will be necessary to have logs of that activity and of the accesses to that information, and if logs are available, they must be auditable, which means that the log data itself also has to be stored and secured properly.

Retaining information is one thing, but being able to find it and produce it when it is needed is another. While data is only useful once it is retained, it is also only useful if it can be found. Therefore, it is necessary to have search and indexing tools available. And as more data is stored, it takes longer to search through it, so the searches become slower and it takes longer for staff to fulfill a request for information. So, it is imperative to have strong indexing and classification schemes, in order to lessen that lengthening of latencies and that slowing of searches.

Storage Issues

Storing all the information is still a major concern, and it must be remembered that the growth in storage is

explosive in many sectors of industry. Figure 3 shows that growth.

Consider e-mail, for example, as a driver for that kind of growth. Users are sending more e-mail every day, both for personal purposes as regular means of interpersonal communication with co-workers and as a part of defined business processes. E-mail is being built into the workflow and used as a tool to get information from person to person within a workflow at the same time, and in part as a consequence of this, e-mails are larger per user than they were in the previous year and larger on average than the previous year.

Both of these things, more messages and larger messages, mean the amount of e-mail information is growing extremely rapidly. This means there is more data to manage in the information store, more data to retain in the backup system, and more data to protect and replicate.

Ever increasing amounts of data mean slower retrieval times, which mean the information will be more difficult to find and retrieve quickly. This has an impact on retrieving information either to demonstrate compliance to a regulatory body or during discovery in a court case. It increases the likelihood of errors and it increases the likelihood that responses to requests will be slowed, which can increase the risk of fines, for example. All these issues are just in reference to e-mail and e-mail is at least semi-structured.

Looking at how storage overall is increasing, research has shown anywhere from 15% year-over-year growth at the bottom end of the curve for telecommunications type companies to over 130% year-over-year annual growth for financial services companies. And that is the average. There are examples of 230% growth within financial services, and more than 300% within healthcare, although those were unusual. However, the problem is that a company cannot keep increasing disk capacity; that is what many enterprises have been doing up to now, but they are hitting the limits of that practice. These enterprises never throw anything away or they do not throw anything away in an orderly and process-driven way. Therefore, they just throw disk capacity at the problem of having more data to store. This in known as the curse of cheap disk, because it helps hide from these enterprises the underlying problem, which is the need to manage the information as responsible stewards and not just warehouse it.

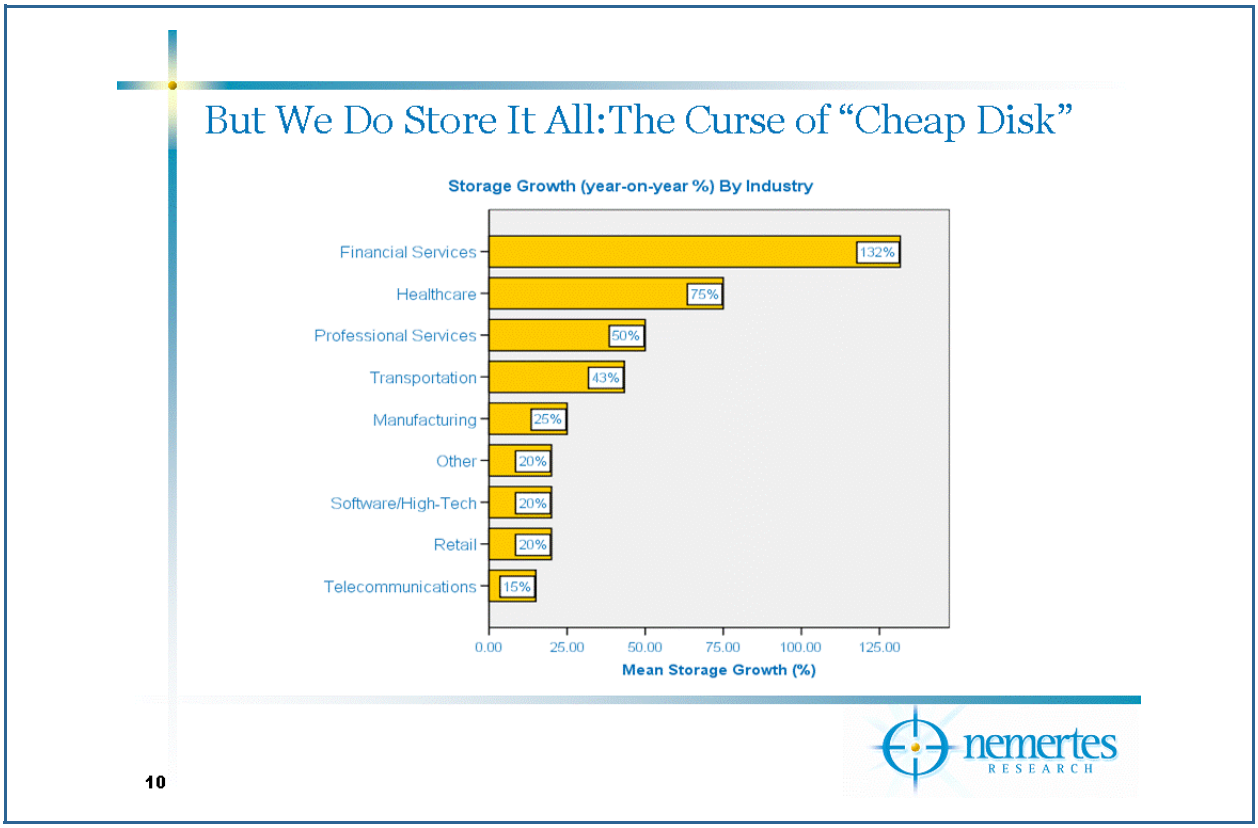

Figure 3

Information Retrieval Is Still Manual

In addition to just increasing disk capacity to solve retention issues, most companies, as illustrated in Figure 4, still do not use technology to help classify, store, manage, and retrieve information when it is necessary.

Instead, they rely on technicians to pull up the indexes or backup tapes to review the content, look for things that might be relevant, pull them off the tapes and put them in front of the legal staff, for example, for review. Nemertes has determined that 80% of companies do not have any kind of technological assistance, specifically, for example, for the problem of information retrieval.

E-mail and Instant Messaging Retention

A more disturbing situation is that half of these companies, as shown in Figure 5, do not procedurally or officially include e-mail and instant messaging traffic within compliance efforts.

These records are not being retained the way they are supposed to be and the companies are not logging retention the way they are supposed to. This is despite the ubiquity of e-mail and the growing ubiquity of IM as communications media. The best implementation of retention and logging of e-mail and instant messaging is within financial services and healthcare. Outside those industries, though, it gets very spotty very quickly.

Archiving, Classification, and Records Retention

What do companies need to do? One of the keys to handling records retention properly is to again address it, not just as a problem of holding information and dumping things on disk, but as a problem of managing information for long-term retention. One of the key technologies here is archiving. Figure 6 provides an overview of the archiving and classification mechanism.

This includes using tiered storage, which is fast and more reliable and therefore more expensive, and some storage that is slower and possibly more prone to failure, and therefore much cheaper. Then, using

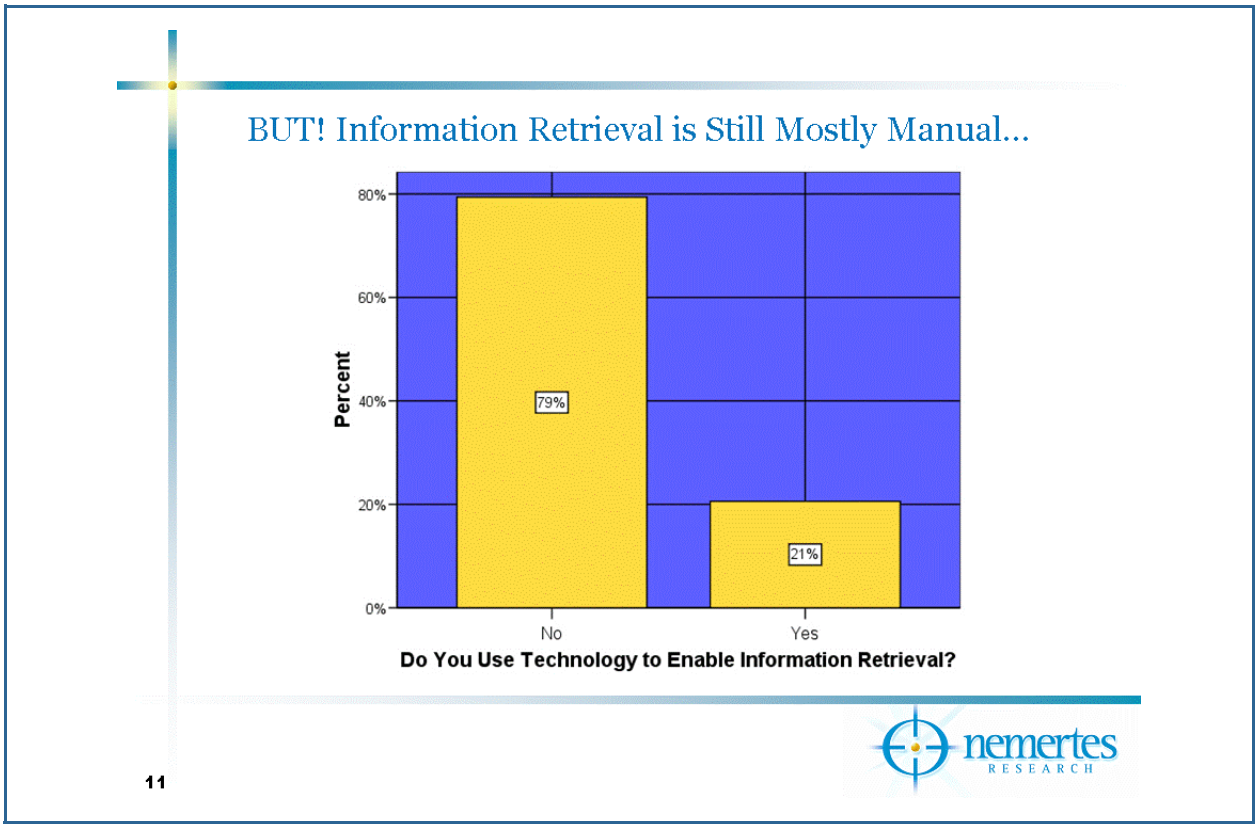

Figure 4

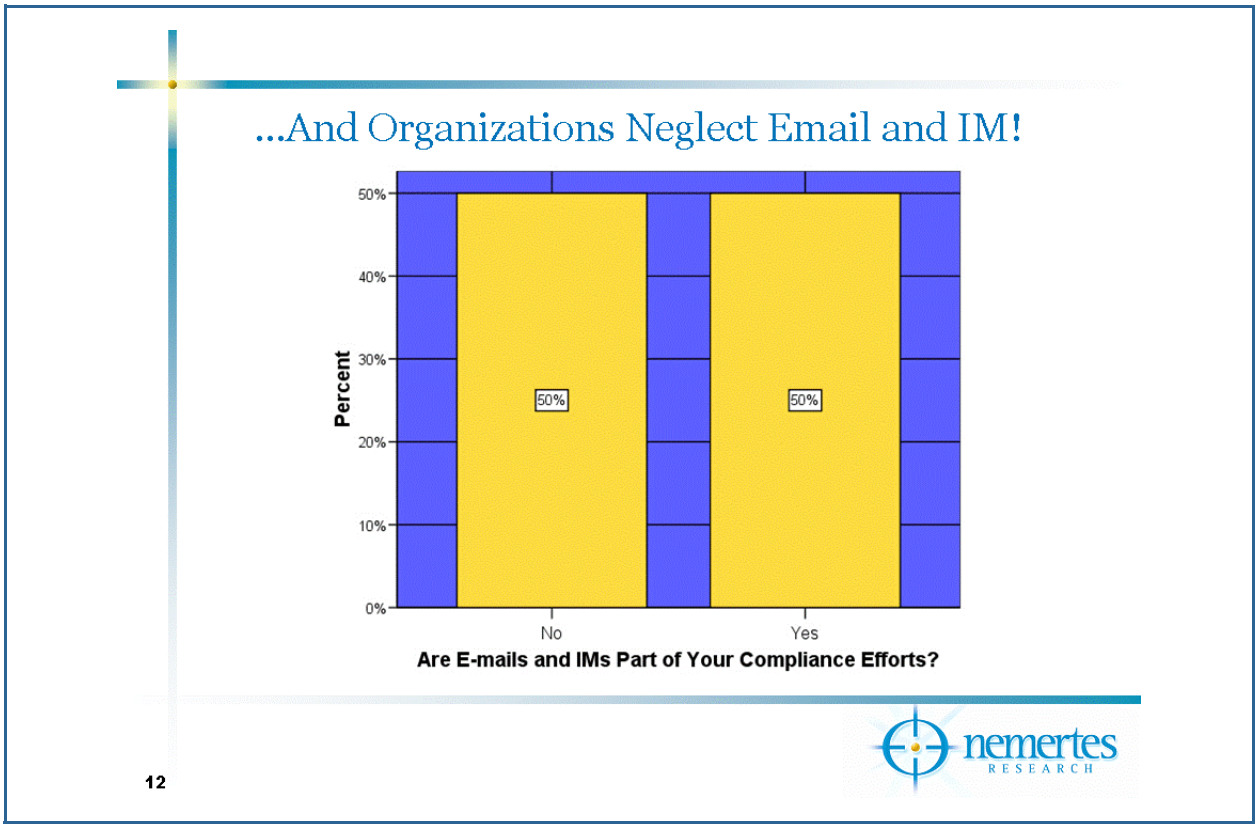

Figure 5

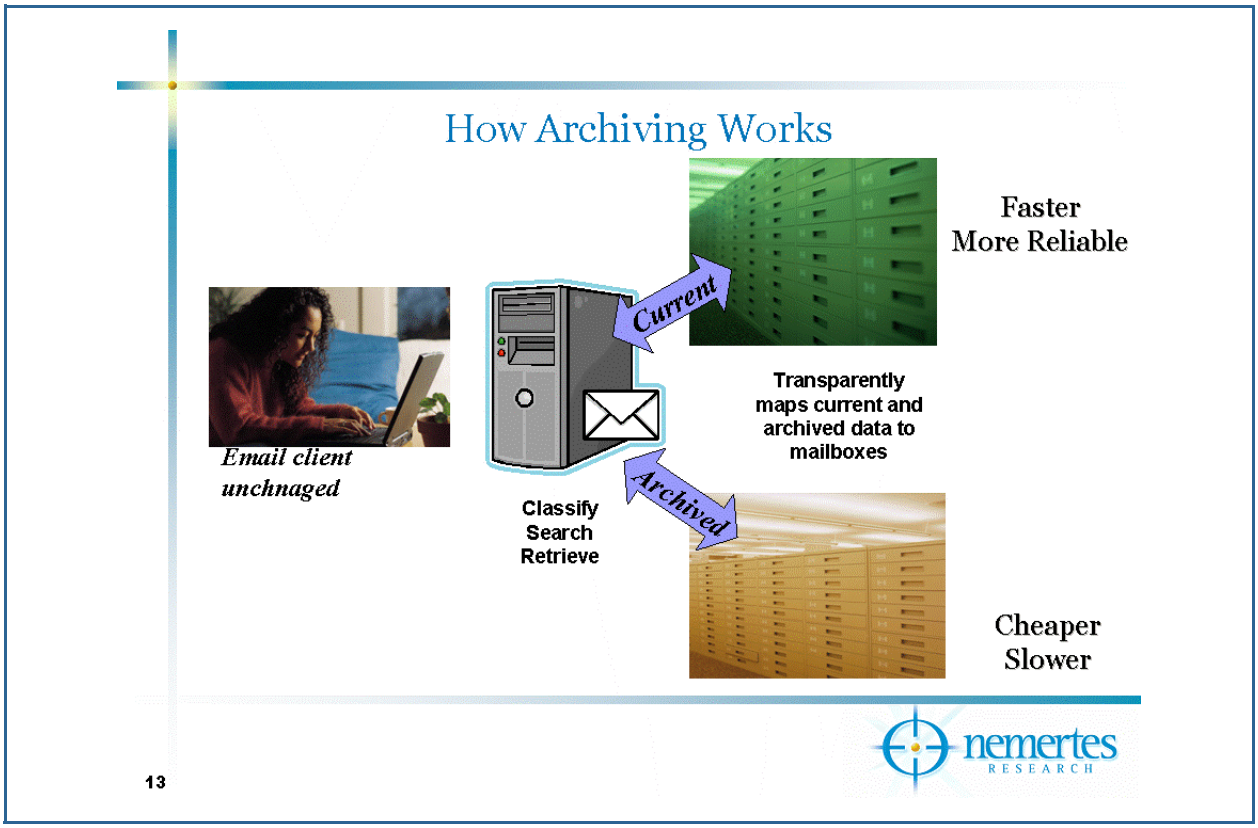

Figure 6

classification schemes, decide where the information can live. Does it have to be faster, does it have to be super-reliable, can it be slower, is it okay to have it go to backup tape if necessary? Finally, using information life cycle management-type tools, keep the information where it belongs.

Archiving and E-mail

Building into an e-mail system the ability to spread information across these different tiers of infrastructure makes it possible to shrink the size, for example, of the e-mail operational store. This allows more and more information to be saved into the archived store, so the e-mail system does not have to manage it actively but can go and get it when needed. Moreover, this allows for better use of the classification tools within a system and the use of those for search and retrieval, for discovery, and for compliance reporting, but it hides the fact that information is living in different places from end users. End-user e-mail clients can remain unchanged, but if they make a request for something that has been archived, the archiving system will retrieve it and supply it to the e-mail system transparently. At most, the end user will see a "one moment please" kind of a message while the information is being retrieved.

Archiving Techniques for Other Systems

As illustrated in Figure 7, these archiving techniques can be replicated in other systems, when using content management and archiving for the files in collaboration tools such as a Groove or SharePoint systems, for unstructured files such as Word documents or Excel spreadsheets located in a user's personal space, and for instant messaging logs that a company should be retaining, whether or not the law currently requires it.

If all these kinds of information are being stored properly, they can be archived properly using tiered storage and made available more easily for compliance applications to find the information that is needed to help demonstrate and generate reports on compliance initiatives or needed by the legal staff for discovery purposes. Again, this makes it easier for them to find information and to omit from those searches information that is known to be protected but to otherwise reach broadly so everything that is relevant is found.

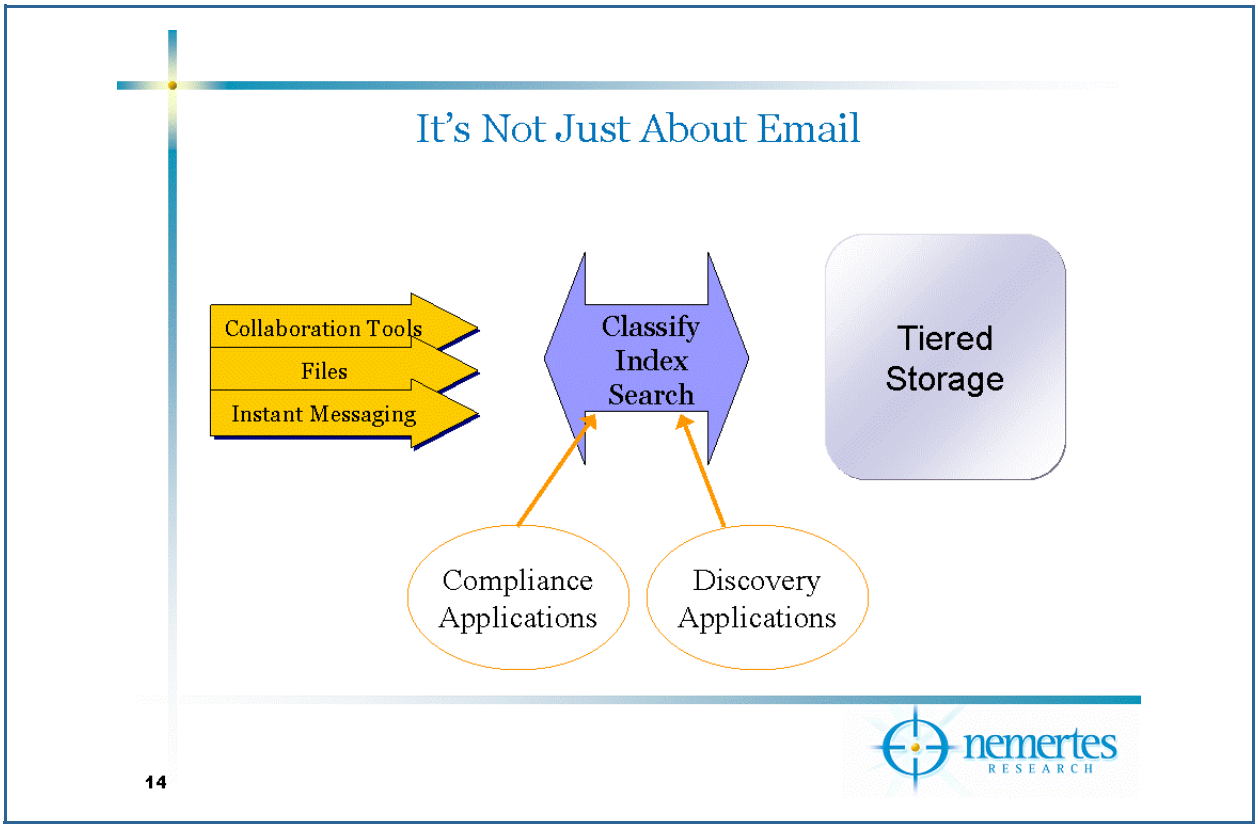

Figure 7

Archiving and Content Management

Archiving and content management generally are key enabling technologies for retention efforts. They can help a company adhere to applicable laws and regulations and they can lower the cost of discovery. Almost every enterprise has to go through discovery from time to time; lawsuits are common. However, these tools can save an enterprise tens, hundreds, even thousands of minutes of expensive legal time reviewing documents by speeding the process of sorting out the protected from the unprotected information and by using search tools to find relevant information more quickly.

Using such systems also allows the enterprise to retain more of the content it wants. For example, with a content management system, a single copy of a document can be kept and accessed by all users rather than storing the copy and revisions for each of the people who have looked at it. This reduces the amount of space taken up by useless copies of information and makes economical copies available to all. Reducing the amount of duplicated information on the disk leaves more room for required information. This in turn reduces the need to purge content as often and, since content is not being purged as often, reduces the risk of accidental loss of information. Finally, through the use of content management and archiving systems and the features within them that provide for the classification and indexing of the information, it is easier and more productive to search for information.

Archiving and Cost Reduction

For some of these same reasons, using such tools can help reduce cost and improve services to end users. If a significant fraction of a company's information—and in many cases the majority of its information—can be placed into slower, cheaper storage, the company can spend a lot less money on its storage infrastructure. If a company is still in the process of converting its paper records to electronic media, it will also be able to save on things as mundane as physical storage space. If it is necessary to store box after box in a climate-controlled warehouse somewhere, an enterprise can convert that to storing a few of tapes in a climate-controlled warehouse, and realize some savings in that way as well.

Archiving and Disaster Protection

It is also possible to enhance the ability to continue operations in the event of disaster when content is classified and tiered by importance as well as compliance needs. Companies are able to more easily restore what is needed first and to know the order in which to restore everything else in the event of an emergency. If it is the practice of a corporation to replicate storage among different locations so that it can operate continuously, then it also has the ability (again, with classification) to know which things it is important to replicate synchronously and which things asynchronously.

Archiving and File Storage Reduction

By archiving and content-managing information, it is possible to do things like consolidate or migrate systems, e-mail systems for example, more easily. By reducing the size of the operational information stored within e-mail, for example, it is possible to improve application availability. The less information in the store, the more reliable the operation of the store and the quicker its response to a request.

Information End-of-Life Considerations

It is not possible to consider managing content and archiving it without also considering the end of life of the information in an enterprise. If a company is not retaining everything forever, it should have policies in place dictating how and when information is removed from the system. If a company retains information for as long as required plus five years, its policy should state that and a process should be put in place to annually purge the information that is past that compliance-plus-five-years deadline.

In order to do that effectively, content management systems can really help because they can help to identify where all the copies of information are and where all the revisions of the information are kept. If all the content is just sitting in old file shares, there is no way to be really sure, without using comprehensive full-text searching and the like, that somebody does not have a copy of a document that needs purging or is stored under another name somewhere else. It is information like this, turned up during the course of hospital discovery processes, that has led to many negative judgments in the recent past and many large fines or settlements. However, if content management is used, and used comprehensively, it is possible to know where all the copies of something are and to ensure that they are all destroyed when it is time to remove the record.

Likewise, a company can make sure that the removal of records from its system is an event that gets logged and therefore it can be audited later both for content management and archiving systems. This is a great benefit because if a document cannot be produced, it really helps to be able to state that it cannot be produced because it and all copies of it were removed two years ago "according to our policy." And again, if a company is still using paper records, it must recognize that, for example, control over copying that paper is not possible, but it is possible using electronic records. It is easier to find rogue copies of things with electronic media than it is with paper. If somebody has made a photocopy of a document that should have been destroyed because it is five years past its useful life, there is no way to know about that photocopy. However, software systems make it possible to know about electronic copies.

Conclusion

The important points to take away from the information presented in this document are:

- Information life cycle and content management are part of holistic data management in the enterprise information stewardship.
- Content management and archiving systems can be crucial tools in helping companies survive the data explosion they are all experiencing. These systems present a more collective solution than just throwing disk capacity at the problem.
- Classification is the key to content management, information life cycle management, archiving, discovery, and compliance.
- The judicious use of archiving systems can help improve retention efforts and reduce operating costs, and potentially even improve the services offered to clientele within the enterprise.
- The decision to retain or destroy records, the method by which a company decides information needs to be retained, and whether information is discoverable or is protected—all of that is a matter of enterprises risk management and a question for business line and legal staff, not for IT. These are not technology questions at their root and so more of the enterprise must be engaged in addressing these subjects.