

Electronic Communications Retention/Retrieval under Sarbanes-Oxley

Arthur J. Riel

Introduction

The Sarbanes-Oxley Act passed by the US Congress in 2002 requires that all public companies retain their business records, including electronic communications, in an easy to access location. Many public companies are only now beginning to analyze and/or acquire electronic communication archives to satisfy their compliance obligations. Too few of these companies are armed with the necessary knowledge to examine the various technologies in this field to determine which solution best meets their needs. In this article we examine the myriad issues that revolve around electronic communications retention and retrieval with respect to business records compliance.

The Whats of Electronic Communication's Archiving

There are a number of electronic communications channels that need to be reviewed by a team including the legal/compliance, internal audit, and IT departments. In general, electronic communications include internal email, external email, third party email (e.g. Bloomberg, Reuters, Yahoo, AOL, Hotmail, etc.), instant messages, electronic faxes, and cell phone text messages. Voicemail recordings are frequently considered out of scope, but management needs to be wary of technologically advanced voice over IP phone systems that email WAV files to their users. These are frequently considered in scope if they are business related.

Electronic communication archive products typically focus on the capturing and processing of both internal and external email from the company's email plant. Third party email systems are often problematic to capture/process for archiving. Web-based mail applications such as Yahoo and Hotmail are notoriously difficult to capture and are very frequently prohibited in organizations that fall under the SEC 17a-4 or Sarbanes-Oxley rules. Bloomberg and Reuters mail logs are often obtained as daily feeds from the vendors for script based processing and insertion into the archive. Instant messages are typically logged and captured as individual datum that should be threaded into more meaningful conversations before being captured in an archive.

Companies examining electronic communication archives should ask the following questions when reviewing potential acquisitions: 1) Does the archive support multiple types of communications?; 2) Does the archive allow for the automatic threading of instant messages into conversations?; 3) Does the archive have built in support for Bloomberg/Reuters email feed processing, or is this something that needs to be done externally?; 4) How easily will the archive integrate with existing technologies/applications (e.g. an enterprise content management system)?; 5) Does the archive require specific infrastructure (e.g. operating system, database, storage device), or can it work on the current IT infrastructure base of the company?; 6) Does the archive support capture from the company's choice of mail server? (Note: Many products are MS Exchange only).

The Hows of Electronic Communication's Archiving

There are two main methods that electronic communications are captured by electronic communication archives: near real-time and batch modes. Most companies will rely on both methods depending on the type of communications. A typical scenario sees an archive capturing email from journal files in near real-time (i.e. draining the queues as quickly as the infrastructure will support) while batch processing daily feeds of Bloomberg mail, instant messages in need of threading, and electronic faxes from the firm's fax plant. For most companies, electronic communications archiving need not be concerned with the handling of peak loads (unlike your

electric company). Most firms size the hardware to handle average daily load, relying on the infrastructure or application to queue additional data for later processing. Most target a maximum window of 12 hours for the final processing of captured email data.

Once the data is captured, processing should produce metadata for use in the later retrievals from the archive. A relational database is often used to capture interesting header information such as message ID, sender(s), receiver(s), date, time, size, and subject. In all but the smallest companies, the full text indexing of messages and their attachments is required to achieve reasonable retrieval times for the inevitable keyword/phrase/regular expression inquiries. Given the large volumes of electronic communications, archives that allow for front end filtering of incoming data have an advantage. Companies can determine the criteria for a subset of email that can be discarded even before it reaches the archive.

Once the data reaches the archive, single instance storage algorithms should be used to minimize the amount of storage required by the system. While most commercially available archives claim to employ single instance storage algorithms, the techniques come in several flavors with varying results. A typical algorithm computes a checksum on the entire email and compares it to other emails in the archive. This is often useless in removing large quantities of storage, e.g. if two copies of the same Exchange email passes through two different Exchange databases then the checksums on the two emails will differ. A better solution is to break the attachments away from the message body and perform a single instance storage algorithm on each component. Not only does this eliminate the problem of checksum variance related to email databases, but it also weeds out duplicate attachments detected when an existing email is forwarded to a third party or even when an attachment is saved locally, attached to an unrelated email and re-sent. The latter algorithm is a prerequisite to creating separate full text indices on each component of the email. This partitioning of text indices will be very useful in the retrieval portion of our discussion.

Another improvement on single instance storage is the use of block-level analysis of the attachments. This reclaims storage when a large document is modified minimally and re-sent. Given the error prone nature of block-level single instance storage, it is recommended that users push this algorithm into an intelligent storage device to accomplish these goals. Solid compression algorithms should be used regardless of the single-instance storage algorithm employed by the archive.

Electronic communication archives typically digitally sign each record to ensure that material has not been modified since capture. Digital signatures can also be used between rows of any relational metadata database to ensure that entire records have not gone missing. For organizations that need to guarantee that material does not get modified (as opposed to just ensuring that such modifications are detected) then Write Once Read Many (WORM) storage mechanisms can be used (e.g. WORM tape, WORM optical drives).

The Basics of Electronic Communications Retrieval

All electronic communication archives should handle the basics of electronic communications retrieval. This requires that the archive be able to quickly sort through large quantities of data to retrieve the responsive communications for queries that include any number of senders/receivers, a date/time range, and specified keyword(s), phrase(s) and/or regular expression(s). There are some common pitfalls even among these basic requirements. In order to retrieve data in a timely fashion, archives must have an efficient method for finding the required emails from their metadata. A prerequisite for this is the use of a relational database (or similar home grown architecture) to capture the header data of each email and index it appropriately. Finding archives that have the proper index structure can be surprisingly complex. For example, should an archive

index the metadata based on the message ID? If one is thinking only in terms of regulatory support you might not create an index on this field, but if you expand your thinking to include mailbox management (i.e. the offloading of messages from the email plant to an archive for later retrieval by a mail client on the user's desktop) it is absolutely essential.

Another major area of system deficiency is in the text indexing of the product. Only the tiniest companies with the smallest amount of data in their archive can hope to retrieve keyword/phrase/regular expression queries in a reasonable period of time without a full text index. Today, many products have full text indexing built-in to the archive process but are deficient in their creation. Known pitfalls to watch for are products that do not build the indices using all attachment types, or in some cases no attachments at all. Such systems are clearly not appropriate for regulatory compliance or civil litigation support. A great test to make of a candidate system is to examine what it does with composite/compressed files such as .zip or .rar files. If it does not recursively unpack such files and build the appropriate indices then the system is not sufficient to fulfill regulatory or litigation inquiries. Try explaining to a regulator or judge that material was not produced to them because the smoking gun attachment must have been in a compressed format!

The Finer Points of Electronic Communications Retrieval

When it became clear to the electronic communication archive marketplace that the Sarbanes-Oxley Act would greatly expand the market for products in this space, many companies rushed to fill the perceived void. Many of these companies did not have previous experience with regulators and/or civil litigations, at least with respect to electronic communications retrieval. Their product's lack of certain helpful, if not essential, features acts as testimony to their naivety. We explore each of these additional criteria below, providing an explanation of the usefulness of each.

Department Level Searching

Electronic Communication Archives should allow users to execute queries on aggregations of data, not just the raw data itself. By allowing for enhanced metadata such as department codes for senders/receivers it becomes possible to execute queries that look for email between the marketing and sales departments, or Chinese Wall violations such as emails between Investment Bankers and Research Analysts (for just two examples). The enhancement to the metadata can be executed at archive time, or by batch processing hours, days or weeks after the message is captured. The enhancement requires that the company has some source of data for mapping employees to departments. This is typically captured in a third party or homegrown application so companies need to be sure that the archive can take a standard feed of data from its system.

Email Group Support

Many electronic communication archives ignore the notion of email groups in their retrieval processing. This is a serious error from a regulatory or litigation viewpoint. When a regulator or judge requests all email received by an employee they are not just asking for emails in which the target was one of the recipients. The inquiry should include any email group of which the target was a member. This requires the archive to capture knowledge about which employees belong to which email groups on a given day. This data is typically stored in a third party or homegrown application so be sure that the archive can take a standard feed of data from your system.

Multiple Email Address Support

In large organizations employees will typically have several email addresses. This is most often the result of a long term employee working through several domain changes due to mergers/acquisitions, corporate re-branding exercises, or the support of long and short email

addresses. For example, in many organizations the author has had both Arthur.Riel@company.com and riela@company.com email addresses. Electronic communication archives should support this facet of corporate reality by accepting a feed of supporting data. The alternative is the manual tracking of the addresses, for use by retrieval personnel on each inquiry. This is very error prone and sure to result in future problems.

Bcc Injection

This feature of an electronic communication archive is only developed by those companies who have actively worked in the regulatory or litigation areas extensively. An interesting problem arises during email production and delivery to regulators or litigating parties in organizations that make extensive use of “bcc:” in their emails. Let’s assume a retrieval request is made for all email sent by Arthur Riel to John Doe for a regulatory or civil litigation request. A production of the email is made and the party or regulator suddenly sees emails sent by Arthur Riel but received by other parties besides John Doe. The assumption is made that the company is placing irrelevant material in the production to bury the regulatory or litigation party in extraneous material. In reality, the production is accurate. John Doe is a blind carbon copy (bcc) on the email message and therefore does not show up in the header. Optional bcc injection allows the inquiry to inject a “bcc: John.Doe@company.com” into the header so it is clear to the reader that the material is relevant. (A similar example actually came up between the SEC and a large investment bank).

The algorithm for bcc injection is more complicated than simply inserting the statement into the header. The injection should not modify the context of what is being searched. Consider the case where person A emails person B and blind carbon copies two persons C and D. If a party requests all emails sent by person A then two bcc statements should be injected, but if they request all email received by person C, then only the bcc for person C should be injected since person C would have no knowledge of person D receiving the email as well. Since litigation and regulatory investigations often focus on who knew certain facts and when they knew them, we do not want to imply more knowledge about the email than would have been conveyed when it was sent (e.g. person D and person C knowing that they were both blind carbon copied on the email in question).

Scope Filtering

Electronic communication archives should maximize the filtering possibilities for the user. Some obvious filtering is the support for wildcard characters in fields, case sensitivity/insensitivity and internal/external/both communication type selection (which will require the system to accept configuration data specifying which domains should be considered “internal”). More complex schemes include the construction of separate text indices for each part of the email message so that keyword/phrase/regular expression searches can confine the contextual search. Ideally, users should have the capability to specify searches only on the subject and/or header and/or body and/or attachment file names and/or the attachment file contents. Advanced text based searching will allow for fuzzy searches, synonym searches, frequent misspellings and possibly context sensitive searching (e.g. find all emails (or specified parts of emails) that contain the key phrase “Arthur Riel” within 20 words of “SEC” but not within 10 words of “Company X”).

Total Cost of Ownership

The features and functionality of electronic communication archives is a primary driver in the selection process. However, when it comes to total cost of ownership, there are a number of factors that a company must take into consideration before making a final determination. These other factors should be considered against the backdrop of your current and the future uses for your electronic communications data.

The choice of hardware and software infrastructure can greatly impact the cost of ownership over time. Decision makers must weigh any additional capital and operational costs they will incur if they select an archive product that does not fit into their current IT infrastructure base. The cost of supporting foreign servers, databases, operating systems, and storage devices should not be underestimated. This fact underscores the value of infrastructure agnostic products that will migrate easily with changes to a company's infrastructure blueprint. For example, having to purchase expensive exotic storage devices will quickly erode any license cost savings between two products.

One area of electronic communication retention/retrieval that does not fall under Sarbanes-Oxley considerations can actually underwrite the costs of regulatory compliance and litigation support. This facet is often called mailbox management. Mailbox management is the aggressive migration of email away from the email server plant to an electronic communication archive. This migration greatly reduces the management costs of older email as it reduces the amount of expensive, high-availability storage; additional servers and their associated email licenses. Email older than a certain number of days, or from mailboxes that exceed aggressive quotas are moved to cheap and deep storage solutions. Mailbox management requires an archive solution that seamlessly integrates with your email server (typically Exchange) and your mail client of choice (often Outlook). For those firms using these platforms, the cost of the archive can be recouped in as little as several months just in email plant upgrade savings. The "insurance" of regulatory compliance and/or litigation support are essential free.

A predicted future use of electronic communication archives is the creation of a corporate wide knowledge center. As more companies create and maintain electronic communication archives for regulatory compliance, litigation support and/or mailbox management, they will realize that their archives contain numerous information assets. The majority of communications with clients, strategic decisions, and other important corporate knowledge assets are captured in the sea of unstructured data that is the typical archive. Electronic communication archives that have built-in data mining tools/modules will have the advantage of allowing their users to easily capture the embedded knowledge with minimal effort. Without such tools, companies will not be able to create useful knowledge bases out of their archives without additional tool acquisitions and potentially extensive integration exercises.

Summary

Over the next several years the majority of public (and many private) companies will be acquiring electronic communication archives for regulatory compliance, litigation support (e-discovery), mailbox management, knowledge management, and internal policy compliance. We have examined a large number of criteria that companies can use to determine which electronic communication archive products best suits their needs. These criteria cover common items published in various forms in the relevant literature as well as those that are not easily detected except by those individuals who have managed extensive electronic communication retrieval projects for regulators and/or civil litigants.

While features and functionality play a major role in selecting an archive product, total cost of ownership encompasses a number of other factors that should be considered during the purchasing decision. These factors include a product's infrastructure requirements, mailbox management features and data mining capabilities. Armed with these criteria, companies can make educated choices based on their current and future requirements of their electronic communication archive.

Published in the November 2006 issue of "Sarbanes-Oxley Compliance Journal".