xenos
a division of ACTUATE

Xenos White Paper

Best Practices in ECM Document Archive Migration

Executive Summary

"Enterprises today are often faced with multiple content platforms and repositories without a unifying strategy or technology to make the data actionable.They need the ability to quickly and easily access and repurpose information and make it available through self-service channels."

— Kenneth Chin, Vice President, Gartner Research

Businesses today are forced to deal with a dizzying array of disparate corporate and departmental data sources, including relational databases, document repositories, email stores, and file servers. Compounding the daily challenges of managing this complex environment are the requirements surrounding corporate acquisitions, regulatory legislation, information governance, and mandates to reduce operational cost through vendor and infrastructure consolidation.

Perhaps more than any other corporate data source, enterprise content management (ECM) repositories, also know as archives, IDARS or COLD systems, have felt the full affect of these contemporary business challenges. ECM repositories are used to store millions of statements, policies, images, and other customer-facing documents. They rely on Indexes or Metadata associated with individual documents for discovery, content validation, storage organization, retrieval, distribution and delivery.

Indexes are typically associated with content at the point of creation by composition engines, or during ingestion into an archive. Once in an archive it is very difficult to enhance or augment these indexes to meet changing business needs such as eDiscovery, ePresentment and the creation of a global (360) customer view.

As organizations acquire other corporate entities, consolidate departmental silos of content into corporate archives, and address regulatory compliance, they are faced with two major metadata challenges:

1. De-duplication of key metadata naming standards, for example account numbers, which may be unique within a single organization but duplicated across acquired corporations
2. Creating a global (360) customer view – customers with multiple accounts across organizations are provided a consolidated view of their information.

Often thought of as a "nightmare project", ECM Migration offers the perfect opportunity to address the indexing enhancement and augmentation activities required to address the de-duplication of key indexes and provide the unique identifiers required to facilitate a truly global (360) customer view.

Since the economic downturn, the Xenos ECM Migration team has engaged with a number of major enterprises struggling with corporate acquisitions, regulatory requirements, information governance, and vendor/infrastructure consolidation.

In the past, the document migration process was time-consuming, resource-intensive, and mired with challenges that often far outweighed the benefits. Today this has changed. Organizations can now migrate their documents quickly and efficiently, and during the process realize additional value from their document content. Benefits include providing additional channels of document access, extracting actionable information previously locked away in static document formats, or enriching the customer experience through personalized communication providing the opportunity for competitive differentiation and increased customer retention.

A successful migration demands comprehensive planning to be completed by experienced individuals in advance and involves a number of best practices that can in general be segmented as follows:

Discovery: Analysis and understanding of the current ECM systems, and the types of documents contained within them.

Extraction: Establish the available means to extract document content from the source ECM system, select the best approach, and execute it at the appropriate time.

Transformation: Re-purpose document content and manipulate its format to add additional value and meet business requirements.

Auditing: Throughout the process the documents must be tracked to ensure that all information is accounted for and that reports are available for future audit purposes.

Indexing: Indexes or metadata provide actionable information about your documents. The migration process provides an opportunity to not only move the existing metadata, but also to add or enrich the indexes available.

Loading: Insert, or load documents, metadata and resources into the target ECM system and ensure that fidelity and accessibility are retained.

The end goal of any ECM document migration project is to move the critical business information locked away in legacy systems to a contemporary ECM solution that offers the features demanded by today's businesses. Additional benefits include a reduction in total cost of ownership and the assurance of performance and functional scalability to meet future demands.

In order to achieve these goals an organization must ensure that the process is planned and executed by experienced individuals using a proven methodology and who have at their disposal the appropriate technology to achieve the desired results.

Choosing a vendor that is committed to working with internal technical and business owners and has field experience with multiple ECM, IDARS and COLD systems can significantly lower the risk (and pain) associated with any document migration project as well as reduce the time to return-on-investment. Over the last 30 years, Xenos has garnered the expertise to adapt, recognize and meet these requirements. The award-winning Xenos Enterprise Server™, upon which the ECM Migration solution is based, is uniquely architected to handle semi-structured print stream, structured data, image and office document formats.

Introduction

The purpose of this white paper is to outline the challenges of migrating documents from ECM systems, educate the reader, and provide best practices that ensure success. The following topics are covered:

- Reasons why your organization might consider embarking on an ECM system document consolidation or migration project.
- An explanation of ECM systems and how documents are stored.
- The various data formats documents are stored in.
- The six stages of the document migration process (Discovery, Extraction, Transformation, Auditing, Indexing, and Loading) and the challenges that each stage presents.
- The document retrieval process, which explains how documents are then extracted from your ECM systems and presented to customers and other users.
- A checklist of traits to look for in a solution or partner to help you solve these challenges.

Drivers for Document Consolidation and Migration

There are numerous reasons why your business would consider moving document data from one or more ECM systems to another.

Cost: Businesses are challenged to reduce the total cost of ownership of their ECM systems due to:

- High annual maintenance charges
- Per seat licensing charges for multiple ECM systems
- Ongoing operational costs for specialized staff due to the specific skill sets required to maintain and upgrade proprietary systems
- Maintenance and support of physical hardware and infrastructure
- Lack of integration between siloed applications resulting in no or inadequate access to information
- Continued risk of incurring penalties for not meeting regulatory compliance requirements or deadlines

Consider that these factors are further multiplied by the number of ECM systems that a business owns.

Mergers and Acquisitions: According to analyst studies, large organizations have accumulated between 6-25 disparate ECM document archives or repositories through numerous mergers and acquisitions. This results in duplicate systems, increasing total cost of ownership.Positioning for Future Growth: It goes without saying that upon increasing your customer base, expect increased volume, demand for functionality and stress on your existing ECM systems.

Improved Customer Experience: When documents relating to a single customer are stored in a number of ECM systems, it is difficult to provide users with a truly seamless experience; providing

self-service channels becomes a challenge. Customer service representatives are challenged by having to deal with numerous systems, and information becomes hard to get at.

Regulatory compliance. Organizations are challenged to meet legal and governance requirements such as Sarbanes-Oxley, BASEL II, MiFID, FSA or others. The ECM solution must enable the organization to meet these standards.

Mainframe "modernization". Organizations often look to reduce the cost of mainframe computing infrastructure, storage and processing cycles, by migrating to distributed platforms. This typically requires a system utilization assessment which can provide the business justification to migrate to fewer or even one ECM system.

Document Content Enabled Vertical Applications. ECM moves from the back office to the front office. Organizations now view ECM solutions as critical to enabling front office business applications. Older repositories, IDARS and COLD systems do not provide the rich feature set required by modern business practices.

ECM Vendor Consolidation: The ECM vendor market has consolidated. IBM acquired FileNet, Oracle purchased Stellent, and OpenText bought Hummingbird. As a result, the vendor choices that were once available are more limited. Expect older ECM systems to move towards becoming unsupported more rapidly and mandatory upgrades to new versions to become commonplace.

What is ECM? "Enterprise Content Management (ECM) is the technologies used to capture, manage, store, preserve, and deliver content and documents related to organizational processes."

— AIIM Web Site www.aiim.org

Enterprise Content Management Systems

Let's review some of the basics of ECM systems, in particular with regards to documents.

AIIM defines Enterprise Content Management (ECM) as the technologies used to capture, manage, store, preserve, and deliver content and documents related to organizational processes. Content can consist of documents, images, structured data, audio or video among others. An Enterprise Content Management System is the amalgamation of technology and methods used to capture, process, persist and deliver large volumes of content to consumers both internal and external to the organization.

Some ECM systems rely on an existing relational database or file system for content storage, while others provide their own, proprietary databases and storage methodologies.

Naturally, each of these systems provides tools for inserting documents into the system (known as "loading"), as well as graphical and programmatic interfaces for getting documents back out ("retrieving"); however, these retrieving interfaces are rarely suited for the mass extraction of content.

For the most part, ECM systems store documents unaltered although some do change the original format during the loading process. In addition, due to space limitations, large document data streams may be split into segments.

Over time as the volume of documents stored within the system increases, some documents may be moved to lower cost external "off-line" storage mechanisms such as tape.

Alongside the documents stored within an ECM system is “index” data (or “metadata”) about the document itself or its contents. This index data is used by users searching or querying for required information stored within the system. Indexes for a customer statement, for example, may include the date the statement was produced, the customer’s name and associated account numbers.

Depending on the system, other elements that are used to display or print the document—such as fonts and images (known as print resources)—may also be stored within the ECM document archives.

Document Data Formats

Documents that are stored in an ECM system are represented in a number of formats, depending on how they were created, and how they are to be used. High volume documents that are generated for customer correspondence are typically created in a print stream format, capable of being printed on high-volume production printers. Other formats include image, PDF or proprietary desktop formats.

Overview of Common Document Data Formats

Images

Documents can be represented as a plain image, typically in the Tagged Image File Format (TIFF). The most common source of these documents is from scanning applications.

As an image, document text and images alike are represented strictly as dots on a page.

The benefit is that images provide a full fidelity representation of the document, and are guarantee that documents are displayed and printed consistently from computer to computer and printer to printer.

However, because parts of the document are indistinguishable from one another, text cannot be easily extracted or searched upon (without the use of character recognition software). This limits the use of these documents beyond simply displaying them. Documents in image formats can also take up a considerable amount of disk space, requiring additional hardware to store them and more bandwidth to deliver them to consumers.

Adobe Portable Document Format (PDF)

Adobe PDF has become the de-facto standard for representing printed documents for viewing electronically.

In this format, text is represented as strings, separated from images and other objects, making documents in this format searchable. The PDF standard also allows for features, such as bookmarks and links, which allow for easy document navigation. In general, PDF provides an opportunity to make static documents more useful, and dynamic. There has also been a new

standard introduced for the long term archiving of PDF documents called PDF/A. This standard defines the structure of a PDF document that will ensure it remains supported and viewable in the future.

Print Data Formats (Print Streams)

Your business likely prints customer-facing documents such as statements and policies on large, high-volume printers. These printers, from vendors such as IBM , Xerox, and HP work with documents in specific data formats including IBM Advanced Function Presentation (AFP), Xerox Metacode, and HP Printer Control Language (PCL).

Documents in these formats are not designed for viewing except after being printed. They must be transformed into another format before being presented to users electronically. Document print streams tend to be very large, multi-gigabyte files that contain hundreds, thousands or even greater amounts of documents.

Any given page on a document makes use of print resources such as fonts, images, overlays, and forms. Resources can be stored “in-line”, meaning that they are included in the document data file. However, the majority of print resources are stored externally, typically on the printer, and are brought in when required.

Choosing an Appropriate Format

Of these options, which format should your business use? This depends on which users require access to documents, and how documents will be used:

- Users over the web: This potentially implies a requirement for low-bandwidth formats, depending on the average size of your documents.
- Internal customer support: Customer service representatives often require access to an identical version of the statement that was sent to the customer on the other end of the phone, in a format that can be rapidly navigated. Features such as bookmarks and full-text search become important here.
- Reprints: This implies that the original print format or image should be stored, in order to accurately reproduce the same printed statement that was sent to your customer in past periods.

Complaince: Regulations may require that the original document format be stored without transformation to ensure compliance or that specific long term archive requirements be met using such format standards such as PDF/A.

The Document Migration Process

Successful document migration projects require a good deal of planning, make use of very specific technology components and require experienced individuals familiar with ECM solutions, document formats and project management disciplines. A proven methodology is critical. The Xenos DETAIL Methodology™ outlined below includes best practices for document archive migration:

- Discovery: Analysis and understanding of the current ECM systems, and the types of documents contained within them.
- Extraction: Establish the available means to extract document content from the source ECM system, select the best approach and execute it at the appropriate time.
- Transformation: Re-purpose document content and manipulate its format to add additional value and meet business requirements.
- Auditing: Throughout the process the documents must be tracked to ensure that all information is accounted for and reports must be available for future audit purposes.
- Indexing: Indexes or metadata provide actionable information about your documents. The migration process provides an opportunity to not only move the existing metadata but to also add or enrich the indexes available.
- Loading: Insert, or load documents, metadata and resources into the target ECM system and ensure that fidelity and accessibility are retained.

The Xenos DETAIL™ Methodology for Document Migration

Discovery
Extraction
Transformation
Auditing
Indexing
Loading

Each of these stages is defined in more depth in the following sections.

1. Discovery

The first task in any data migration project is to study the source of the data. In this case, the goals are:

1. to ensure the documents themselves are well understood
2. to analyze the ECM system(s) in which the documents are currently stored
3. to understand the business environment in which the ECM system operates

The questions below can help to frame your understanding of these aspects.

Understanding your Documents

What are the various types of documents stored in the system?

For example, an annual financial statement might present a summary of information across a given customer's accounts, while a monthly account statement may present activity in a single account, over a shorter time period. While both of these documents share a similar purpose—each presents financial information for a given customer—these documents are laid out differently and are used and accessed in different ways.

What metadata or indexes are used to describe each document type?

Individual document types typically possess a unique set of indexes that describe their contents. In the above example an annual statement might be uniquely identified by searching for the customer's name and a given year, while a monthly statement would be associated with a given account number and month of the year. Understanding these differences is necessary to recreate these relationships in the target ECM system.

How do documents relate to one another?

These relationships need to be understood and maintained during the migration.

Understanding your ECM System

How is document data physically stored?

Are documents stored in a database, file system or third-party storage solutions. Are documents stored offline, for example on tape? What storage formats are used? There is a need to understand the source system in order to determine the most efficient means of accessing the data for migration.

How is metadata stored?

Indexes and other information that describes the documents stored in your system can be located in databases, control files or actually appended to the document contents themselves within the source system. This metadata is crucial for the retrieval process and can be maintained and migrated to the target system. If additional metadata is required in the target system it is possible to "mine" the contents during the migration process extracting index information to meet the target system requirements and business use.

What techniques are available to extract documents from the ECM system?

Having the answers to the above questions dictates the best approach to extracting documents from the system. This is covered in more detail in the next section.

Understanding your Environment

How can the migration be performed without compromising the quality of service currently provided to your customers and other users?

Unfortunately, it is not realistic to shut your business down in order to perform a document migration. Day-to-day operations involving these systems must be supported while migration takes place behind the scenes—ideally, the migration process should appear to be seamless for customers and other end-users.

Since extraction requires putting a large strain on the ECM system, care must be taken as to when the procedure occurs. Techniques include: copying or cloning the system, scheduling the migration during off-peak hours, or migrating documents from the old system in low volumes in parallel with a new system. Without professional experience, some of these methods are time-consuming and can literally take years. With the proper professional experience and tools, the time spent implementing these methods can be greatly reduced.

What are your business requirements?

Do industry regulations require your business to maintain documents in a long-term archive?

2. Extraction

Extracting all of your document data and associated metadata from an ECM system is challenging.

Methods

Typically, ECM systems provide a mechanism by which to pull out individual documents, one at a time for presentment. Pulling out every document stored, however, is a completely different matter. Your options may include the following:

- Batch Tool: Some ECM systems provide tools to both load and extract large volumes of document data directly to and from the ECM system. Consider yourself lucky if an extraction tool is available to you.
- Programming Interface (API): ECM systems generally provide an application programming interface (API) with which programs can be written to retrieve individual or multiple documents for presentation or other purposes. This interface can also be used to extract data for the purposes of migration. However, this technique can prove to be impractical due to slow performance, or due to an inability to iterate over all data stored in the ECM system since these interfaces are designed for speedy retrieval of “hit lists” and individual documents.
- Direct Database/Storage Access: Some ECM systems rely on an existing relational database or file system, making it a possibility to directly extract binary data and potentially other metadata from a data source. Understanding how data is stored and interrelated becomes vitally important, in this case.
- Third-Party Expertise: For the reasons previously stated, your business is not alone since document migration projects are becoming increasingly common. Take advantage of this experience by engaging a vendor with a set of proven best practices to help manage and facilitate your move between ECM systems.

Once extracted, the document data itself may not yet be in a usable state. To save space content management systems employ schemes such as data compression using common or even proprietary algorithms. Data is also often encrypted as a security measure. Document data may also be specially structured to accommodate secondary artifacts such as print resources.

Metadata

It may be necessary to maintain metadata associations with content. However, this can prove to be a challenge upon extraction:

- Metadata may not be able to be extracted from the existing ECM system.
- There are no standards for defining metadata. Most legacy systems were built before the days of XML.
- Different document types infer different metadata rules, and indexing requirements.

One practical solution to this problem, as discussed later, is to rebuild indexes during migration through a technique called “re-indexing”

3. Transformation

Once documents have been unlocked from your existing ECM systems, it may be necessary to convert or re-purpose your documents from one format into another prior to loading them into a second ECM system. Here’s why:

- Preparation for loading: ECM systems generally require document data streams, associated resources, and metadata to be in a specific format, prior to loading. For example, some ECM systems require documents to be in a stacked file (all individual documents concatenated together in a single file) with associated indexes structured in a specific format (discussed in more detail in a later section). In other systems, where document transformation is not possible upon retrieval, documents may need to be converted and stored in PDF format, ready for presentation.
- Resource extraction and versioning: The sheer abundance of print resources such as fonts and images used by various document types can lead to problems down the road. Resources with duplicate names, for example, can result in the incorrect logo or signature showing up on the wrong document. The transformation process allows for embedded resources to be extracted and essentially catalogued, so that the retrieval process can accurately recreate the original document.

In addition, the transformation process has other important benefits:

- Eliminate redundant printers: Consider a business that has standardized on one vendor’s printing solution, say IBM’s, and the business acquires or merges with another company that had previously standardized on Xerox technology. A transformation step would allow Xerox Metacode to be transformed into IBM AFP, thereby eliminating a dependence on Xerox technology.
- Document proofing: Transforming documents allows for easier validation that what came out of an ECM system is the same as what goes into another.

In general however, document transformation allows for other value to be realized from your documents:

- Rapid Online presentment: Store documents in a web-viewable format for rapid extraction and presentation, with no additional overhead required in your customer-facing applications.
- Long-term storage: Convert to standard data formats such as PDF/A (PDF for Archiving), to comply with industry regulations.
- Reverse composition: Extract all elements of the statement into formats such as XML, for use in other applications. Consider this an alternative approach to data integration, as all of the relevant information has already been gathered in a single document.
- Enrichment: Add color, images, and other creative touches to boring legacy documents, to enhance the customer experience.
- High-volume printing: Convert non-print formats such as PDF into print formats to take advantage of high-volume printers.
- Print proofing: Create a web pre-flight application to provide a means of efficiently producing and approving new types of printed documents.

4. Auditing

Migration Statistics and Reporting

When documents are migrated, detailed information must be readily available to ensure that each and every document has been migrated. This is achieved by the recording of an audit trail throughout the conversion process and applying a series of checks and balances during each phase of the conversion. This gives the client a higher level of confidence that the entire archive was migrated successfully.

In order to determine if a successful migration process has occurred it is necessary to track and provide an audit trail documenting the following:

- Where did the document originate?
- What metadata (indexes) were associated with the document?
- What processes touched the document, and how was it changed? Perform a validation step to ensure that any changes to the document are consistent with the original version and that the original document fidelity is retained.
- The document can be identified and located within the new system, in a similar manner as in the original system?
- All of the documents from the source system destined to be migrated are accounted for within the process.

Unfortunately, due to the high volume of data in play, it is not practical to perform a manual verification of each and every document. Instead, manual spot checks should be performed, at the very least for a batch of documents for each given type.

5. Indexing

Indexes include key information about a document, such as an account number, statement date, or a customer name. Using a combination of this metadata, a search can be performed to access a particular document within an ECM system.

The indexes that are available depend on the type of document, the information that can be found on each page, and on the choices your business made years ago. If you were lucky, you were able to extract this metadata from the ECM system as-is and preserve it in some form; if not, at this point you will have to re-index your documents.

Either way, the migration process provides an opportunity to enrich, or add new indexes, creating new ways of accessing your documents and meeting business and regulatory needs.

Consider a scenario in which a customer finds a charge on their credit card bill for a service that your company provided. Adding a new index value, such as the amount due, could make it possible to locate and retrieve the appropriate statement in record time.

Some ECM systems do not allow new indexes to be added once documents have been loaded, so this may be a unique opportunity to enhance the value of the documents stored in your ECM system.

6. Loading

Once documents have been extracted, transformed, re-purposed and indexed they are ready to move into their new home.

Preparing Inputs

For the most part, the inputs that you will feed to your new ECM system consist of the document data itself, metadata/indexes, and print resources.

If documents were extracted as individual files from an existing ECM system, file system performance can become a problem. To allow documents to be loaded efficiently, some ECM systems allow for document files to be combined into a single stacked file.

A stacked file is a single file that contains hundreds, or thousands of document files, appended to one another. Metadata that describes these documents are contained in a separate index file that contains byte offsets and lengths, pointing to the location of an individual document in the stacked file.

Loading Methods

Enterprise content management systems provide one or more mechanisms by which to load documents in large batches:

- Loading utility: Most ECM systems provide an application that, provided one or more document data files and index information, will load the given database.

- Programming interface (API): Some systems provide an application programming interface with which custom applications can be written to load documents.

The Document Retrieval Process

Now that the migration process is complete, users can conduct searches on your ECM system to locate and then retrieve documents. Document transformation ensures that documents are delivered in the appropriate formats, as well as provide other benefits. How you choose to deliver and present documents at this point depends on your business objectives.

Retrieval Methods

From a technical standpoint, ECM systems provide a couple standard mechanisms with which to query and retrieve individual documents:

- ECM client: Most ECM systems provide a graphical client front-end in the form of a desktop or web application. This client is generally not designed as a customer-facing interface and thus is suitable for internal use only.
- Programming interface (API): To integrate with your web application, modern ECM systems provide a means by which to write a program to search upon and extract documents in a raw binary form. Depending on the format they are stored in, documents can then be converted for presentation, or delivered directly to a user's web browser.

Document Transformation

Once a document has been retrieved, it likely needs to be transformed before it can be presented to a user.

In addition to providing the document in the appropriate format for the user, transformation provides a number of benefits. At this point, direct marketing messages can be applied, color and images can be added, and sensitive information can be removed, all without affecting the original document. More importantly, an intelligent transformation process can import the appropriate print resources based on revision control information added during the migration transformation stage.

To minimize the average time your customer waits for a given document, any transformation solution that your business considers must be highly efficient. Re-purposing, by nature, is memory- and processor-intensive, due to the "richness" of most documents. To mitigate these issues, a high-performance transformation solution utilizes multi-threading techniques, and caching of commonly-used resources to deliver documents in the fastest way possible.

Meeting your Objectives

No matter what methods you choose for retrieving and presenting your documents, being flexible will allow you to meet the changing needs of your users, and provide the best customer experience possible.

Conclusion

While a large-scale ECM document archive migration project may seem daunting at first, a good amount of planning guided by best practices methodology is a critical success factor.

Choosing a vendor who has met the following criteria can significantly lower the risk (and pain) associated with any document migration project as well as reduces the time to return-on-investment. Ask questions of the vendors you are dealing with to ensure they have:

- Experienced migration consultants with deep technical knowledge regarding the extraction and and loading of ECM, IDARS and COLD systems.
- A proven detailed migration methodology.
- Configurable migration components designed to automate the migration process.
- Specialized applications designed to transform and extract information from a variety of print stream formats.
- A track record of success, speed, efficiency and reliability.
- Reference accounts you can speak with.

Copyright © 2011 Actuate Corporation. All rights reserved. Actuate and the Actuate logo are registered trademarks of Actuate Corporation and/or its affiliates in the U.S. and certain other countries. All other brands, names or trademarks mentioned may be trademarks of their respective owners.