OpenText™

Best Practices in Enterprise Content Management Migration

Six stages of ECM migration can help organizations reduce costs, improve customer experiences, and position for future growth

Businesses today are forced to deal with a dizzying array of disparate corporate and departmental data sources—all compounded by the requirements surrounding corporate acquisitions, regulatory legislation, information governance, and mandates to reduce operational cost through vendor and infrastructure consolidation. Enterprise Content Management (ECM) repositories have felt the full effect of these contemporary business challenges, but ECM migration offers the perfect opportunity to move critical business information locked away in multiple legacy systems to a more cost-effective ECM solution that offers the features demanded by today's businesses.

OpenText™

Table of Contents

Introduction

There are numerous reasons why businesses might consider migrating output from one or more ECM systems to another, and these factors are further multiplied by the number of ECM systems that a business owns.

- **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 access or inadequate access to information
    - Continued risk of incurring penalties for not meeting regulatory compliance requirements or deadlines
- **Mergers and acquisitions**. According to analyst studies, large organizations have accumulated between 6 to 25 disparate ECM output 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 when you increase your customer base you can expect increased volume, demand for functionality, and stress on your existing ECM systems.
- **Improved customer experience**. When output relating to a single customer is 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 that can provide the business justification to migrate to fewer or even one ECM system.
- **Output-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. As a result, the vendor choices that were once available are now limited. Expect older ECM systems to become unsupported more rapidly and mandatory upgrades to new versions to become commonplace.

In the past, the ECM migration process has been time-consuming, resource-intensive, and mired with challenges that often outweighed the benefits, but organizations are now realizing they can migrate print stream outputs quickly and efficiently while also gaining more value from their content.

Enterprise Content Management System Basics

To understand the ECM migration process, it's best to review some of the basics of ECM systems, particularly with regard to high-volume documents.

The Association for Information and Image Management (AIIM) defines 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, video, and more. An ECM system is the amalgamation of technology and methods used to capture, process, store, 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 output storage, while others provide their own proprietary databases and storage methodologies.

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

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 output data streams may be split into segments.

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

Alongside the output stored within an ECM system is "index" data (or "metadata") about the high-volume document itself or its content. This index data is used by those 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.

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 that ECM migration can help address:

1 De-duplication of key metadata naming standards such as account numbers, which may be unique within a single organization but duplicated across acquired corporations.

2 Creation of a global (360 degree) customer view—customers with multiple accounts across organizations are provided with a consolidated view of their information.

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

High-Volume Document Formats

High-volume 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.

Images

Output can be represented as a plain image, typically in the Tagged Image File Format (TIFF). The most common source of this output comes from scanning applications.

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

The benefit is that images provide a full-fidelity representation of the output, and are guaranteed to display and print consistently from computer to computer and printer to printer.

However, because parts of the high-volume 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 high-volume documents beyond simply

"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

displaying them. Output 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.

Portable document format (PDF)

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)

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

Output in these formats is 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 in a high-volume 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 output 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

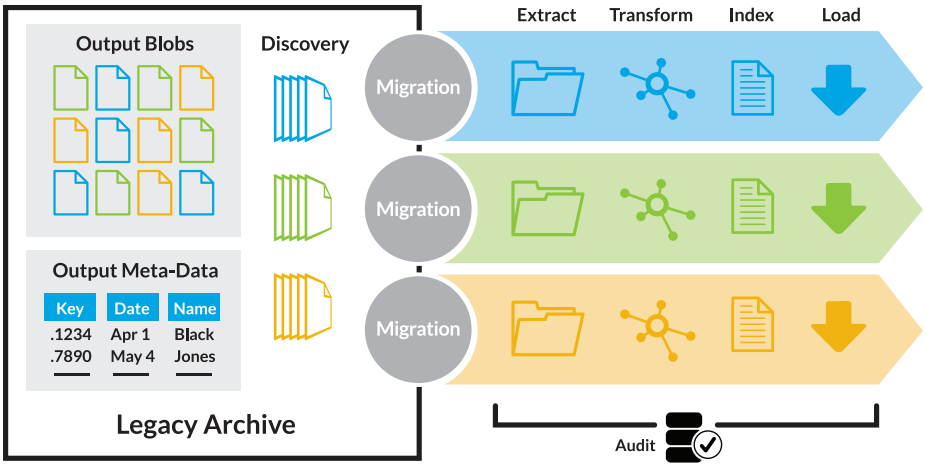
Of these options, which format should your business use? This depends on which users require access to output, and how the output will be used:

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

The High-Volume Document Migration Process

Successful migration projects require a good deal of planning, make use of very specific technology components, and require experienced individuals familiar with ECM solutions, output formats, and project management disciplines. A proven methodology is critical. The OpenText™ DETAIL Methodology™, outlined here and delved into deeper later in this white paper, includes best practices for archive migration:

- **Discovery**. Analyze and understand the current ECM systems, and the types of output and metadata contained within them.
- **Extraction**. Establish the available means to extract output from the source ECM system, select the best approach, and execute it at the appropriate time.
- **Transformation**. Repurpose output and manipulate its format to add more value and meet business requirements.
- **Auditing**. Track output throughout the entire process to ensure that all output has been migrated to the target system and create reports to confirm the status for each high-volume document to satisfy audit requirements.
- **Indexing**. Indexes or metadata provide actionable information about your output. 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 output metadata and resources into the target ECM system and ensure that fidelity and accessibility are retained.



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 output is well understood.
2. To analyze the ECM system(s) in which the output is currently stored.
3. To understand the business environment in which the ECM system operates.

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

Understand Your Output

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

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

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

Individual document types typically possess a unique set of indexes that describe their content. In the previous 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 an account number and month of the year. Understanding these differences is necessary to recreate these relationships in the target ECM system.

How do high-volume documents relate to one another?

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

Understand Your ECM System

How are high-volume documents physically stored?

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

How is metadata stored?

Indexes and other information that describe the output stored in your system can be located in databases, control files, or appended to the output 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 output during the migration process—extracting index information to meet the target system requirements and business use.

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

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

Understand Your Environment

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

It is not realistic to shut your business down to perform a high-volume 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 high-volume 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 take several years. With the proper professional experience and tools, the time spent implementing these methods can be greatly reduced.

What are your business requirements?

Think about requirements, such as whether industry regulations require your business to maintain high-volume documents in a long-term archive.

2. Extraction

Extracting all of your output data and associated metadata from an ECM system can be challenging.

Methods

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

- **Batch tool**. Some ECM systems provide tools to load and extract large volumes of output data directly to and from the ECM system. Consider yourself lucky if an extraction tool is available to you.
- **Application programming interface (API)**. ECM systems generally provide an API with which programs can be written to retrieve individual or multiple high-volume 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 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 high-volume 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; 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 output 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. Output 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 output. 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 output 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 output has been unlocked from your existing ECM systems, it may be necessary to convert or repurpose your high-volume documents from one format into another prior to loading them into a second ECM system for a couple of reasons:

- **Preparation for loading**. ECM systems generally require output data streams, associated resources, and metadata to be in a specific format, prior to loading. For example, some ECM systems require output to be in a stacked file (all individual high-volume 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 output transformation is not possible upon retrieval, high-volume 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 high-volume 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, 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.
- **High-volume document proofing**. Transforming high-volume documents allows for easier validation that the output data that came out of an ECM system is the same as what goes into another.

In general, however, output transformation allows for value to be realized from your output in other forms:

- **Rapid online presentment**. Store output 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 statement.
- **Enrichment**. Add color, images, and other creative touches to boring legacy high-volume 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 output.

4. Auditing

When output is migrated, detailed information must be readily available to ensure that each and every high-volume document has been migrated. This is achieved by the recording 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 migration process was successful, it is necessary to track and provide an audit trail documenting the following:

- Where did the output originate?
- What metadata (indexes) were associated with the output?
- What processes touched the output, and how was it changed? Perform a validation step to ensure that any changes to the output are consistent with the original version and that the original output fidelity is retained.
- Can the output be identified and located within the new system, in a similar manner as in the original system?
- Are all of the output from the source system destined to be migrated 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 high-volume 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 statement such as an account number, statement date, and/or a customer name. Using a combination of this metadata, a search can be performed to access a particular high-volume document within an ECM system.

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

Either way, the migration process provides an opportunity to enrich or add new indexes, creating new ways of accessing your output 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 high-volume documents have been loaded, so this may be a unique opportunity to enhance the value of the output stored in your ECM system.

6. Loading

Once high-volume documents have been extracted, transformed, repurposed, 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 output data itself, metadata/indexes, and print resources.

If output is extracted as individual files from an existing ECM system, file system performance can become a problem. To allow output 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 high-volume document files, appended to one another. Metadata that describes this output is contained in a separate index file that contains byte offsets and lengths, pointing to the location of an individual high-volume document in the stacked file.

Loading Methods

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

Loading utility. Most ECM systems provide an application that, provided with one or more output data file and index information, will load the output.

API. Some systems provide an API with which custom applications can be written to load output.

The Retrieval Process

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

Retrieval Methods

From a technical standpoint, ECM systems provide standard mechanisms with which to query and retrieve individual high-volume 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.
- **API**. To integrate with your web application, modern ECM systems provide a means by which to write a program to search upon and extract output in a raw binary form. Depending on the format in which they are stored, output can then be converted for presentation, or delivered directly to a user's web browser.

Output Transformation

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

In addition to providing the output 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 high-volume 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 high-volume document, any transformation solution that your business considers must be highly efficient. Repurposing, by nature, is memory-and-processor-intensive, due to the "richness" of most high-volume documents. To achieve these desired results, a high-performance transformation solution uses multi-threading techniques and caching of commonly used resources to deliver output in the fastest way possible.

Meeting Your Objectives

No matter what methods you choose for retrieving and presenting your output, 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 archive migration project may seem daunting at first, a good amount of planning guided by best practices methodology can ensure success.

Choosing a vendor who meets the following criteria can significantly lower the risk (and pain) associated with any output migration project as well as reduce 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 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 output formats
- A track record of success, speed, efficiency, and reliability
- Reference accounts you can speak with

www.opentext.com

NORTH AMERICA +800 499 6544 • UNITED STATES +1 847 267 9330 • GERMANY +49 89 4629-0
UNITED KINGDOM +44 (0) 1189 848 000 • AUSTRALIA +61 2 9026 3400