PART I
Introduction to Enterprise Content Management

- CHAPTER 1: What Is Enterprise Content Management?
- CHAPTER 2: The SharePoint 2010 Platform
What Is Enterprise Content Management?

WHAT'S IN THIS CHAPTER?

- Defining ECM as used by this book
- Gaining a historical perspective of ECM
- Defining the components of an ECM system

Considering that this is a book both by and for architects and developers, devoting an entire chapter to talking about the enterprise content management (ECM) industry and trying to define it, rather than just jumping into the bits and bytes that you probably bought the book for, might seem strange. However, by introducing ECM as part of an industry, instead of describing how the SharePoint world perceives it, we hope to provide a perspective that wouldn’t otherwise be possible if you make your living inside the SharePoint ecosystem.

ECM, within or outside of the SharePoint world, seems to be a much-abused abbreviation used to describe a variety of different technologies. Of course, people often adopt new or existing terms, applying their own twist to the original meaning, and this is certainly the case with ECM. The difficult part is determining which meaning is actually correct. Sometimes even the words representing the initials are changed. For example, in the halls of our own company, sometimes “electronic” is used instead of “enterprise.” In other cases, ECM is confused with specific technologies that are part of it, such as DMS (Document Management System), IMS (Image Management System) or WCM (Web Content Management).

Clearly, ECM means a lot of different things to a variety of people. There is no doubt that some readers of this book will think something is missing from the definition, while other readers will find something included that does not fall into their own definition. That being said, this chapter introduces ECM not necessarily from a SharePoint perspective, but from a historical perspective; then it provides an overview of the components of an ECM system.
You can skip this information, but we believe it is important to clarify the problems we are trying to solve, rather than just write code based on our own assumptions.

INTRODUCTION TO ECM

The “content” aspect of enterprise content management can refer to all kinds of sources, including electronic documents, scanned images, e-mail, and web pages.

This book uses the definition of ECM from the Association for Information and Image Management (AIIM) International, which can be found on their website at www.aiim.org:

Enterprise Content Management (ECM) is the strategies, methods, and tools used to capture, manage, store, preserve, and deliver content and documents related to organizational processes. ECM tools and strategies allow the management of an organization’s unstructured information, wherever that information exists.

As this definition states, ECM is not really a noun. That is, it’s not something as simple as an e-mail system or a device like a scanner, but rather an entire industry for capturing and managing just about any type of content. The key to the definition is that this content is related to organizational processes, which discounts information that is simply created but never used.

Moreover, ECM is meaningless without the tools that accompany it. You might say that the tools that solve your content problem also define it. This idea is explored in the next section, and hopefully clarified by a short history of a few of the technologies involved.

A HISTORICAL PERSPECTIVE

Although the term ECM is relatively new, many of the components that make up an ECM system started appearing in the 1970s. The world of information systems was vastly different 30–40 years ago. The Internet as we know it did not exist, the cost to store data was astronomical compared to today, server processing power was a mere fraction of what it is today, and desktop computers didn’t even exist.

The history of ECM can be traced back to several technologies that formed that first stored and managed electronic content: document imaging, electronic document management, computer output to laser disc (COLD), and of course workflow, which formed the business processes.

Document Imaging

As evidenced by the first systems to take the management and processing of documents seriously, paper was one of the first drivers. These systems were often referred to as electronic document management or document imaging systems. By scanning paper and storing it as electronic documents, organizations found a quick return on investment in several ways:

- It reduced the square footage needed to store paper.
- It resulted in faster execution of paper-based processes by electronic routing.
It eliminated the time it took to reproduce lost documents.

It reduced overhead because paper documents could be retrieved electronically.

In addition to a reduction in manpower, there were other benefits to storing paper electronically — namely, security and risk benefits, which preceded regulations such as Health Insurance Portability and Accountability Act (HIPAA) and Sarbanes-Oxley by more than a decade. Some of these included the following:

- Password protection of documents
- Enterprise security restraint brought about by secure networks
- Management of records needed for legal holds
- Management of the document life cycle, such as retention periods
- Audit information about the document life cycle and requests about the document

The first document imaging systems for commercial consumption became available in the early 1980s, and they quickly started to replace the previous technology for removing paper from organizations, which was microfiche. Billions of documents were stored on microfiche, but indexes and location data were often stored in databases. Conversions of these systems to document imaging are surely still being handled today.

The ability to scan existing paper documents in order to create electronic documents, as discussed in the next section, led to the vision of a “paperless office,” a commonly used phrase by the end of the century. Of course, this lofty and often pursued goal of a paperless office has yet to materialize, and paper is still the original driver behind many business processes. As shown in Figure 1-1, focusing on paper is a good starting point to quickly begin realizing the benefits of an ECM system.

With the average cost of each wasted page being about six cents, a company with 500 employees could be spending $42,000 per year on wasted prints.
Electronic Documents

The invention of computer-based word processors (in the 1970s) created the need for a way to store and quickly retrieve these documents. Electronic documents share similarities with document imaging systems, yet they are unique in that they are typically dynamic; that is, they often require ongoing modification, whereas scanning paper was typically performed for archiving purposes.

The first electronic documents were created through word processing software, driven in the late 1970s by WordStar and Word Perfect. Although the former has been abandoned, WordPerfect still exists today and is part of an office suite from Corel.

Soon after personal computers and electronic word processors hit the market, electronic spreadsheets became available, beginning with VisiCalc, followed by Lotus 1-2-3 and eventually Microsoft Excel. Spreadsheet documents are now as commonplace as word processing documents.

Today, electronic documents exist in countless types and formats, ranging from simple ASCII text files to complex binary structures.

COLD/Enterprise Report Management

The widespread use of computers, beginning with the large mainframes, resulted in an unprecedented use of paper. Early computers all over the world started producing reports, typically on what is known as green bar paper. As the need for information from both mainframe and mini computers grew, so did the need for computer-generated reports. Necessary at first because structured methods for viewing data electronically did not exist, this excessive use of paper continued to plague organizations into the 1990s and even into this millennium.

Out of this problem grew a solution coined computer output to laser disc (COLD). Instead of generating paper, these reports could be handled in a type of electronic content management system, typically storing the ASCII data and rendering it onto monitors. These systems enabled not only search and retrieval of the reports, but the addition of annotations, and of course printing of the documents when using a monitor is not adequate.

The term COLD was eventually replaced by enterprise report management when magnetic storage replaced the early optical storage systems.

Business Process Management/Workflow

Storing content electronically was a great step forward, but moving content to a digital medium quickly put the content in front of the right user at the right time.

The first business process management (BPM) systems, launched in the mid 1980s and called workflow systems, were created by the same companies that brought document imaging systems to market. These early systems were far less complex than the BPM systems used today, however, as they primarily enabled content to be put into queues to be processed by the same workers that processed the paper.

It was almost 10 years later when the first graphical components became available for creating complex workflow maps. This was the beginning of BPM as we know it today, which enables organizations to create, store, and modify business processes.
ECM COMPONENTS

In order to understand ECM, it is necessary to understand the common components that comprise such as system. The following sections provide a brief overview of these components, which, like the definition for ECM provided earlier, have been defined by AIIM.

Capture

Capture is the process of gathering the data, regardless of the source, including classification, indexing (sometimes called tagging), and rendition. These tasks are required before storage is possible, in order to understand what type of content is being managed, which keywords will be used to search for the content, and to ensure that the content is in a form that can be easily retrieved and viewed later.

Paper

Paper is still the primary driver of ECM. The reason is simple: Because of the volume of paper that most companies need to handle, efficiently managing that paper can provide the greatest return on investment. Figure 1-2 gives you some idea of the cost of paper in an enterprise.

**FIGURE 1-2**

Paper capture is primarily done by using document scanners specifically built for the purpose. These scanners can capture both small and large batches of paper. Paper documents are typically divided into three categories: structured, semi-structured, and unstructured.

Structured documents typically represent forms such as tax documents, applications, or other preexisting forms. Because they are always the same, these documents are usually the easiest to automatically extract information from. Technology such as Zonal OCR can be easily applied to structured forms because the key data always exists in the same place.
Semi-structured documents are similar to structured documents, but they are different enough that structured zones can no longer be used to extract data. A common type of semi-structured document is the invoice. Most invoices are similar enough to be recognized as such, but each company designs its invoice with enough nuances to distinguish it from others, so these documents require either manual manipulation of the data or more intelligent automation.

Unstructured content represents the majority of the information in the average corporate enterprise. Almost all human correspondence is a good example of unstructured content. Although this content ends up being stored as the same content type in each company, on a per-page basis they have very little in common. Manual indexing is typically required for this type of data.

Developing applications to handle all the different types of paper that may come into an enterprise is a daunting task. Fortunately, toolkits are available to drive most document scanners on the market today, and these are normally compliant with either (or both) the TWAIN and ISIS driver standards. Although the physical process of scanning a piece of paper is simple, building a good process for either manually entering information or automatically extracting it is difficult, and probably not cost effective for custom applications.

**Office Documents**

Because this book is about ECM and (Microsoft) SharePoint, it focuses on the most common type of electronic documents that need to be managed: (Microsoft) Office documents. These documents are created using word processing software, spreadsheet software, presentation software, and so on. These documents are typically pre-classified on creation, as they frequently start from a template; therefore, extracting data for searching is often overlooked. Although each word from the document can be added to search indexers, it makes more sense to use specific keywords to identify the document. Sometimes pre-identified form fields are used, but often the most important data is keyed by hand before sending the document to storage.

**E-mail**

Capturing e-mails into an ECM system is becoming an increasingly common scenario. E-mail is often used to drive a business process, as it is becoming an acceptable form of correspondence in most organizations. Although some data can be indexed automatically, such as the sender, receiver, and subject, as shown in Figure 1-3, it can be difficult to extract the useful information contained in the body of the message, and manual intervention is usually required.
E-mail attachments are often more important than the e-mail message itself. Although Microsoft Outlook and other e-mail clients are improving ECM integration, extracting the attachment and exporting it to an ECM system usually requires specialized software in order to properly tag the content with searchable data.

Reports
As mentioned earlier, technology originally known as computer output to laser disc (COLD) and later as enterprise report management enables computer reports to be parsed into electronic files. Classifying and indexing these documents is typically automatic because they are in a form that is very structured.

Like the handling of paper, building a system for handling enterprise reports is most likely not cost effective when you compare the needed functionality versus the difficulty of obtaining it. Consider being able to read EDI or other electronic streams and extract the necessary data from them. Also, although data storage may not be an issue, you must consider how it will be displayed to users in a readable format.

Electronic Forms
It was initially believed that the goal of the paperless office would be achieved with the help of electronic forms. After all, the form templates provided in many applications, such as InfoPath, enable users to fill out preexisting fields and submit them directly to a content management system. With the type of content already known and the data being put into electronic form as it was gathered, it stood to reason that the paper forms could gradually disappear. However, human habits die hard. It may take another generation of computer users, who have been raised from birth with computers and who use them for everything from social networking to bill-paying, to realize the truly paperless office.

Other Sources
Although the most common types of capture have been identified here, there are many other possible data sources and data types. Multimedia, XML, and EDI are other well-known data formats that can arrive from many different sources. Indeed, just about any type of data can be consumed in an ECM system.

Store and Preserve
The store and preserve components of an ECM system are very similar; storage traditionally pertains to the temporary location of content, whereas preservation refers to long-term storage. In the past these were separated because online storage was costly. Content was usually stored in the temporary location only during its active life cycle, when it was frequently accessed as part of the business process. Once the active life cycle was complete, content would move to long-term storage, known as offline storage or nearline storage, which was much less expensive. The term “offline” reflects the fact that the content wasn’t accessible without human intervention; the term “nearline” typically refers to optical discs that were brought online automatically, such as in the case of a jukebox. The following sections describe both the software and hardware components of storage and preservation of content.
SOFTWARE

The software required to store and preserve data varies widely according to the needs of an organization. Sometimes the software required is part of the operating system, and other times there are specific services used to store, deliver, and allow transactions on the content itself.

Repositories

Repositories refer to the actual software that defines the taxonomy of a storage system. In the case of SharePoint, the repository refers to the site, libraries, and content types that define the taxonomy of the system.

The term library services is also used to describe not just the repository but the services built around repositories.

Databases

Databases have been the standard for storing transactional data for decades; however, they were typically not a popular place for content storage until SharePoint began to use SQL Server for storage. As discussed later in this book, the rapidly growing number of SharePoint databases created for storage led third-party vendors to find new ways to move content into other storage mechanisms.

Hierarchical Storage Management

Hierarchical storage management (HSM) has dwindled in recent years due to the lower cost of online storage. The purpose of an HSM is to automatically transfer content between online and offline storage. These systems have enough intelligence to bring content offline when it was no longer actively accessed and back online when requested.

Often, the storage of content during its process life cycle and its permanent archival is the same; they are sometimes separated when the cost of having a large amount of documents for instant retrieval is too high.

File Systems

File systems are the most common location for content storage, but they are not the most common location in ECM systems. If fact, they are probably the worst possible place your content could be stored, especially if your system lacks a good backup strategy.

Hardware and Media Technologies

Software would not be very valuable if content were limited to computer memory. Very simplistic to advanced hardware and media technologies are used in combination with software to provide online, nearline, or offline storage locations for content.

Magnetic/RAID

At one time very expensive, the cost per gigabyte of magnetic storage has plummeted over the past 10 years. Storing content online has historically used magnetic disk; but now that the cost is so much lower, many systems continue to use magnetic storage for long-term preservation as well.
Redundant array of independent disks (RAID) allows several disks to be used together to add redundancy, which enables systems using magnetic technologies be more fault tolerant.

Optical

In the earliest days of ECM, specifically document imaging systems, large optical disks provided the least expensive long-term storage method for data. Although magnetic storage has become much less expensive, smaller optical media, such as DVDs, still offer a valid and often used method for permanent storage of content.

The obvious disadvantage of optical storage is that keeping it online is next to impossible; therefore, it is necessary to use nearline storage methods, using jukebox-type devices that bring the disks online as needed.

SAN/NAS

With the reduced cost of magnetic storage, both the storage area network (SAN) and network area storage (NAS) have become very popular in the world of ECM. While both can work for file-level and block-level data, a SAN is often optimized for the block-level access required by software such as databases. A SAN is typically several disk arrays that connect through fiber channels and that appear as local disks to a server. They all allow the fastest possible read times in order to meet the needs of relational database systems. NAS devices are detached storage arrays on less expensive hardware than a SAN, which makes them more cost effective in scenarios such as document archival. A NAS also works well with file-level access, making it a great solution for server-based file systems.

Cloud

- The cloud is the likely future location for content storage and quite possibly the entire ECM landscape. Cloud technologies combine both software and hardware technologies to provide centralized storage over the Internet (and sometimes intranets). One advantage of cloud technologies is that cloud data centers purchase hardware in bulk, drastically reducing the costs to customers.
- Cloud data centers are environmentally conservative, offering greatly reduced energy consumption.
- Software and hardware are optimized to work together.
- Commercial cloud providers offer a level of fault tolerance not otherwise possible for the same cost.

Of course, there are disadvantages too. Some are technical, such as the latency for users who need rapid access to large files. Others are more business related, such as security issues and loss of data control.

Cloud storage is rapidly evolving. In terms of total services provided, the current leaders seem to be Amazon S3 and Microsoft Azure. By the time this book is published, however, many other competitors will emerge to offer storage and management capabilities in the cloud.
CHAPTER 1  WHAT IS ENTERPRISE CONTENT MANAGEMENT?

Management Components

All the components described in the following subsections have their own chapters in this book, so here you are just briefly introduced to each of these management components of ECM.

Document Management

Document management can be a complete solution itself for many organizations, or it may represent a component of a larger platform. Chapter 4 focuses on the complete solution; but from a component standpoint, document management is an underlying technology for document imaging, records management, and web content management.

As a component, the most important functionality provided is the taxonomy. In terms of SharePoint, this means libraries, content types, columns, and even perhaps higher-level items like site collections and sites. When users refer to the taxonomy, they are typically referring to the assembly of these features within their organization.

Other functionality provided by document management systems are security, check in/out, versioning, and of course document retrieval. Document management systems can contain many more features, but these are the major components related specifically to documents.

Web Content Management

Although many users argue that web content management (WCM) is the primary driver of ECM, this component came along much later than the others mentioned earlier, with perhaps the exception of collaboration. WCM is an underutilized technology that can greatly reduce the costs of managing websites by introducing document management, publishing, and even workflow to the process.

WCM, unlike other components such as document management and BPM, is a technology used for the specific purpose of managing website authoring.

Business Process Management and Workflow

Business process management (BPM) and workflow enable electronic content to be delivered to the right people at the right time — and much more inexpensively than managing the content through manual processes.

These two terms are often used interchangeably but one is really part of the other. Workflow, in the context of ECM, refers to the process of automating a business process, while BPM is the process of not just automating but redefining, improving, and managing the business process. Workflow is part of BPM. The important thing to know is that BPM, or just workflow, can not only accelerate the return on investment of an ECM system, but also be the primary reason to adopt an ECM system.

Records Management

Records management, covered in detail in Chapter 8, is similar to document management and document imaging but has some key functional requirements that differentiate it. Although records
management often refers to both paper and digital records, this book focuses on records management from a digital perspective.

Records management is defined by ISO 15489-1:2001 as “the field of management responsible for the efficient and systematic control of the creation, receipt, maintenance, use, and disposition of records, including processes for capturing and maintaining evidence of and information about business activities and transactions in the form of records.”

Records, in the context of records management, can then be understood as any evidence and information about business activities and transactions. The best source for more information on records management is ARMA International. Their public website is www.arma.org.

**Collaboration**

Collaboration is a relatively new management component of ECM, and it probably wouldn’t be even considered today if not for SharePoint. The ability to create online content and have a group of users act on that content has been around since forums first started appearing on Internet newsgroups. These newsgroups thrived throughout the late 1980s and 1990s; and although they still exist, they have given way to website forums, blogs, and wikis. Collaborative information is the ultimate unstructured content.

**Social Media**

Social media, such as Twitter, YouTube, and LinkedIn, are not typically considered a part of ECM, but these popular sites have vast amounts of content, and it is managed.

Although organizations largely ignored social media in its early years, they are starting to understand the value of mining the information it contains, and they are looking for the best ways to manage it.

**Document Collaboration**

Document collaboration enables a set of users, typically authorized, to save documents to repositories, where they can be shared and manipulated. These documents are often a crucial part of the business process.

Although previous document management systems allowed some collaboration, SharePoint initially focused on collaboration, rather than ECM, until version 2 (SharePoint Portal Server and Windows SharePoint Services 2.0).

**Delivery**

The final component of ECM is delivery. Without a way to deliver the content, there is no reason to capture and store it in first place, and management would be impossible.

**Search**

The capability to search content is a key step to delivering information to users. How do users even know it exists unless it is easily identified?
In the earliest days of ECM, table indices on databases were used to find content with a structured taxonomy. Using this taxonomy to define the exact data needed made databases a great technology for searching content. This technology is still heavily used in ECM systems. Even SharePoint makes it possible with its new Metadata navigation capabilities on list column data as well as being able to make this column data available to the SharePoint and FAST search crawlers.

In the past decade or so, the industry has been switching to a more unstructured approach to searching. Whereas the first search systems for ECM were simply based on database tables and indices, now content is crawled by complex engines that can store seemingly unlimited amounts of data. These crawlers interrogate text and media on websites and in documents, as well as inside databases. Modern search technologies need to be able to aggregate this information for users while still being able to perform simple structured searching based on taxonomy.

**Viewing**

Of course, the viewing of content is a critical output point for an ECM system. Delivery of content is said to be completed when it is available to a viewer. The delivery method used varies according to the type of content and the purpose for which it is intended.

**Document Readers**

Document readers are used to display content, generally text, to users in an easily readable manner. Adobe Acrobat Reader is probably the most well-known example of a document reader. Although the term “reader” is being used, these applications often allow for much more than simply reading content and may also be the applications that were used to create the content to begin with. Applications such as Microsoft Word and Microsoft Excel are the best-known examples.

**Web Browsers**

Web browsers are rapidly becoming the most common delivery method for content, as much of the world’s unstructured information is now delivered within websites. Even most document viewers today have plug-ins for the most common browsers, or advanced viewers that are delivered directly within the browsing experience, as shown in Figure 1-4.

**Others**

Just as ECM systems can manage countless content formats, there are numerous ways to view that data. Although web browsers are becoming a standard mechanism for viewing all types of data, some types of content require a specialized viewer.

**Transformation**

The content within an ECM system is not useful unless it exists in a form that can be easily consumed. Often, part of the capture process includes transformation technologies that enable content to be stored in specific (or multiple) formats for easy consumption. Table 1-1 describes several common format types into which content may be transformed.
TABLE 1-1: Common Transformation Formats

<table>
<thead>
<tr>
<th>FORMAT</th>
<th>DESCRIPTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>PDF (Portable Document Format)</td>
<td>Possibly the most common format for archived documents,</td>
</tr>
<tr>
<td></td>
<td>PDF is a cross-platform format for storing almost any type of</td>
</tr>
<tr>
<td></td>
<td>document. PDF documents typically display documents exactly as they will</td>
</tr>
<tr>
<td></td>
<td>print.</td>
</tr>
<tr>
<td>XPS (XML Paper Specification)</td>
<td>Very similar to PDF but developed by Microsoft before given to the</td>
</tr>
<tr>
<td></td>
<td>community.</td>
</tr>
<tr>
<td>XML (Extensible Markup Language)</td>
<td>An open format for describing just about any type of data.</td>
</tr>
<tr>
<td>HTML (Hypertext Markup Language)</td>
<td>The output format of most websites worldwide.</td>
</tr>
<tr>
<td>TIFF (Tagged Image File Format)</td>
<td>An image-only format often used when the original document’s fidelity is</td>
</tr>
<tr>
<td></td>
<td>no longer needed.</td>
</tr>
<tr>
<td>COLD/Enterprise Reports</td>
<td>Typically only report data is captured. However, when displayed, the</td>
</tr>
<tr>
<td></td>
<td>data must be transformed to look like the original paper report that would</td>
</tr>
<tr>
<td></td>
<td>have been generated.</td>
</tr>
</tbody>
</table>
Security

Security ultimately applies to all the components of an ECM system. Good security systems enable system administrators to exercise complete control over how content should be secured. Administrators should be able to apply security in various ways: to each piece of content individually, automatically applied by capture systems, or controlled by other aspects of the system. Table 1-2 describes the three main security levels.

**TABLE 1-2: Security Levels**

<table>
<thead>
<tr>
<th>LEVEL</th>
<th>DESCRIPTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parent</td>
<td>Security applied at the parent level forces content stored in the specific library or other repository item to take the same security level as the parent. Prior to Microsoft SharePoint Server 2007 (Windows SharePoint Service 3.0), this was the only option available.</td>
</tr>
<tr>
<td>Document</td>
<td>Document-level security allows specific security to be applied to a piece of content regardless of where it is located.</td>
</tr>
<tr>
<td>Type</td>
<td>Type-level security refers to security associated with a specific type of content. This set of permissions associated with an object is also known as an access control list (ACL). This type of security does not exist in SharePoint today.</td>
</tr>
</tbody>
</table>

Digital/Electronic Signatures

Digital (or electronic) signatures guarantee the authenticity of the documents with which they are associated. Technologies that handle these signatures prevent a piece of content from being modified without the signature being compromised. Once the content is modified in any way, the signatures become invalid until reassigned.

Information/Digital Rights Management

Documents often need to be secured separately from the ECM security model itself. Because these documents can be transported outside the confines of the system, they need another layer of security to prevent unauthorized access to either part or all of their contents.

Within the industry, this technology is typically known as digital rights management; however, Microsoft has proprietary technology called Information Rights Management (IRM), designed specifically for content created by Microsoft software.

Digital rights management can control all aspects of a document, from preventing access to the document itself to how the document may be used. Examples include not allowing a document to be printed, saved, or even viewed.

SUMMARY

Without using a lot of technical jargon, this chapter provided a quick look at the ECM industry. Many developers understand ECM only as it applies to the SharePoint ecosystem, so the information provided here fills in some of the blanks by describing how the rest of world views ECM. Highlights of the chapter included a working definition of ECM from AIIM International, followed by a historical perspective, and finally the common components that make up an ECM system.