

# CHAPTER 1

## The State of Information Management

## 2 Transforming Infoglut!



What is infoglut?

*Infoglut* is information overload. It is the inability to find what you need. It is skyrocketing storage costs, even when storage is getting cheaper. It is not knowing where to find authoritative information assets. It is the risk of not being able to find those assets at the instant they are needed by customers, employees, and auditors. It is too much data, and too few tools to manage it.

Infoglut is the inevitable byproduct of our digital world, compounded by Internet applications and Web 2.0 technologies. In a world where anybody can create e-mail, Web sites, digital photos, instant messages, or Microsoft Office documents, it's only a matter of time before everyone is overwhelmed. As Bruce Schneier says, "data is the pollution of the information age."

Ultimately, how can you transform your infoglut into strategic assets? How can you turn this massive amount of information into a competitive edge? How can you reduce the costs associated with maintaining this data? How can you secure your information? How should you wade through all this information to find what you need, when you need it, in the format you need?

The analyst firm IDC estimated that in the year 2006 the world created about 161 billion gigabytes of information. By 2010, they predict that the world will create 988 billion gigabytes of information *in one year*. That is a six-fold increase in only four years. How much of that total is because of your organization? How much data did you create in 2006? Are you prepared to manage six times as much by 2010?

The analyst firm Gartner believes that the first organizations who learn to transform infoglut into assets will gain a significant competitive edge over those who don't:

\*Gartner DL, Newman D. Spotlight on Enterprise Information Management. June 2, 2006. "Effective information management will be critical in the next decade, differentiating those enterprises that will implode under the infoglut from those that will use it to dominate the global economy."

This goes beyond simply managing the storage of data, but also determining the value of the data. What content is relevant? To whom? Why? When? And in what format?

These trends are also affecting the role of the CIO in the enterprise. In July 2008, the IBM Governance council predicted that within four years the value of data will be a line item on a company's balance sheet. The CIO will have to quantify its worth to the CFO, and guarantee that the proper governance is in place to increase value and minimize risk. The purpose of this book is to set you on the path towards transforming your infoglut. It presents a plan for the vital first step: getting a handle on your unstructured content with a pragmatic enterprise content management strategy. Unstructured content is 80 percent of the infoglut problem, and if you can properly align your strategies, you could accomplish even more.

## Information Management

Every organization has problems with information management. Enterprise information management is about one thing: *organizational knowledge*. In other words, it is about the *current state and direction of your organization*. It may seem like a strange question, but where does that information currently exist? Is it stored in highly structured database tables? Is it strewn about multiple employee laptops as Microsoft Office documents? Is it locked away, undocumented, in the minds of your employees?

Is this information entirely under your control, or is some of it in the hands of your customers, partners, and competitors? In an ideal world, where would you put it? How should it be secured? What is important to keep, and when should it be destroyed? How will people find it when they need to know it? And perhaps most important, how will people find it even if they *don't know* that they need it? If your unstructured content storage will increase by a factor of six in the next four years, will your current systems be able to manage it?

Enterprise content management (ECM) helps solve this problem. ECM is not just a line-of-business application, nor a framework, nor middleware, nor infrastructure. It is all these things, plus an *initiative* about creating a culture of information sharing. It's about people first, context second, content third, and technology last.

## 4 Transforming Infoglut!

Every ECM vendor has a slightly different offering, but overall they agree on the following definition of ECM, created by the Association for Information and Image Management (AIIM) in 2006:

“Enterprise Content Management is the technologies 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.”

This book describes a set of ECM tools and techniques that you can use to solve the fastest growing and hardest to manage part of the information management problem: unstructured content.

### **Structured Content vs. Unstructured Content**

Information management experts usually talk about placing content into two types of systems: structured content repositories, or unstructured content repositories.

The term *structured content repositories* applies to highly organized data that is typically used by enterprise applications. This includes lists of employees, customers, products, orders, inventory, and purchases. The structured data is usually stored in a relational database with a rigidly defined structure. The purpose behind them is usually to help automate well-defined business processes: make sure all invoices are paid on time, keep track of inventory, or ensure all customers are contacted once per quarter. These repositories include common enterprise applications like enterprise resource planning (ERP), customer relationship management (CRM) systems, human resource management applications, as well as a plethora of home-grown database applications.

In contrast, *unstructured content repositories* are more free-form. They include anything that doesn’t fit, or wasn’t put, into your structured repository. They usually contain information needed by emerging processes, projects that are rapidly evolving, or initiatives that are difficult to define. This content is frequently used for communication, and is therefore frequently delivered through multiple channels. These may contain training documents, scanned paper documents, newsletters, research reports, e-mails, spreadsheets, contracts, specifications, Web content, audio and video assets, and the like.

The terms “structured” and “unstructured” are somewhat misleading. All information has structure of one form or another; otherwise, it would be impossible to understand. So shouldn't we only need structured content repositories? No.

This distinction is important not because of the *what*, but because of the *how*. What matters is not *what* the information is, but *how* it is created and consumed. In order to benefit from a structured repository, you need to define all possible data structures beforehand, and force your users to create information with a very strict process. These usually involve Web pages or desktop clients that require a great deal of highly specific data entry. In contrast, an unstructured repository would allow a contributor to insert information in any way that is convenient to the contributor; the repository transforms the content, and extracts useful information behind the scenes to add structure.

The unstructured content repositories exist because of the inevitable gaps in larger, more rigid enterprise applications. Even if you have an enterprise application for storing financial information, your employees are likely to store vital content in loosely structured Excel spreadsheets. They will also need to refer to scanned paper documents from these systems to verify that the data entry was correct in your financial application.

Why not store this unstructured content inside your enterprise application? Simple: structured repositories take time to properly design. It is easy to add a new table in a spreadsheet, but not to your database. Doing so can cause side effects to your financial application, and as a result, your IT department has complex rules about who can change it. The necessary change-management process is usually so long and complex, that your employees are likely to prefer doing their job by sending spreadsheets around via e-mail.

However, now you have another problem. How can you secure that important data in the spreadsheet? How will people find it again? How will they be able to change it, and ensure everybody works on the most recent revision? Solving that problem is the realm of ECM.

## Process Workers vs. Knowledge Workers

In general, highly structured information repositories are usually designed for *process workers*. These people follow a well-defined process in order to accomplish their jobs. Their roles are highly *tactical*: accounts receivable, accounts payable, invoicing, shipping, HR forms processing, and the like. A structured content repository helps you streamline this process, and makes it more efficient.

## 6 Transforming Infoglut!

However, this efficiency is a paradox: once a process can be automated, soon everybody will be doing it, and it ceases to be a competitive advantage! There is some value in being an early adopter, because then you are automating processes before anybody else. There is also value in being a late adopter, because then you can benefit from the commoditization of industry-specific enterprise applications.

In general, the reliance on process instead of people is a rapid race to the bottom. Competitive advantage comes from *creating* novel processes that give you an edge, or in being able to accomplish tasks where no clear process can yet be defined. These are sometimes called a *barely repeatable process*, and it is the role of *knowledge workers* to design them, or to get things done without them.

Naturally, knowledge workers both produce and consume a great deal of unstructured content. They need to find both specific and general information on a range of topics. They need instant access to determine what action to take, whether it is a documented process or not.

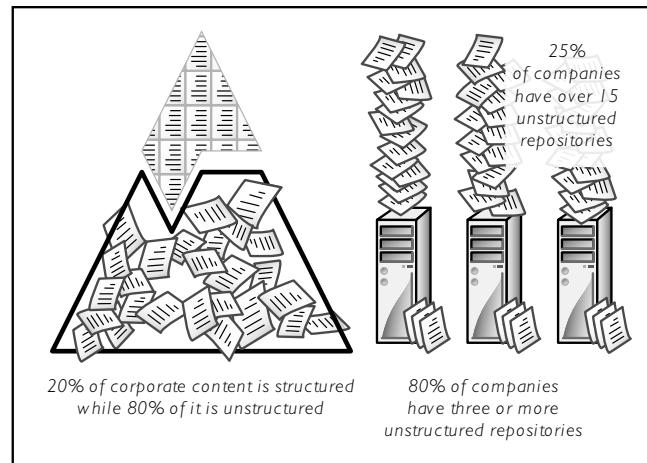
There is no clear line between the process worker and the knowledge worker: many people fill both roles at different times of the day. In fact, the best way to improve your process is frequently to empower your process workers to spend more time as knowledge workers.

As a result of this blurred line, a coherent information management strategy needs to understand where the automated process ends, where knowledge begins, and how they overlap. It is also vital to ensure that both process workers and knowledge workers get the tools they need to boost their productivity. Some of this is the role of ECM applications; some is the role of enterprise applications.

## The Specific ECM Problem

The challenge with unstructured repositories is in managing the content after it is created. How should the contributor find it again? How should other people find it and reuse it in the future? When should it be archived or deleted? Should it be repurposed and published to other systems? How do you make sure that this content doesn't fall into the wrong hands?

According to most technology analysts, 80 percent of the useful information in your organization is unstructured content, as shown in Figure 1-1. Unfortunately, on average only 10 percent of this content is being properly managed in a repository. The rest resides in department-specific collaboration



**FIGURE 1-1.** *The vast majority of useful content is unstructured and insufficiently managed.*

project spaces, shared disk drives, FTP servers, e-mail servers, home-grown applications, hosted applications, or desktops. All together, these systems have dozens of duplicate items; it is difficult to find content, let alone understand what's recent, and what's accurate. These systems are also not properly secured, and it's difficult for auditors or other departments to find and use this information.

In the ECM world, we call these systems *content silos*. According to industry analyst Gartner, nearly 80 percent of organizations have more than two unstructured content repositories, and 25 percent of organizations have over 15 such systems. This is why nearly 30 percent of an employee's time is spent just looking for information. If that weren't expensive enough, on average, 40 percent of your IT budget will be spent on integrations. Also, analyst firm IDC estimates that there are on average eight copies of every content item in your enterprise. Do you have the storage space for that? How about four years from now, when you need six times that?

Also, how risky is it to have so many copies of digital content sprawled around your enterprise? What happens if you are sued, and a judge demands that you produce relevant documents? The expense of finding relevant digital content—called eDiscovery—can be extraordinarily

## 8 Transforming Infoglut!

expensive: on average, one gigabyte of data will cost \$2500 for a lawyer to review. Analyst firm Forrester believes that the market for eDiscovery alone will grow from \$1.4 billion in 2006 to more than \$4.8 billion in 2011.

If you had one unified system instead of 15, then your employees would be able to find what they need. You would have a “single source of truth,” well integrated into your enterprise, allowing everyone to find the single authoritative version. You would have one system dictating who has access rights to the item and what the life cycle of the item needs to be in order to follow industry regulations. It would also properly destroy content once it is no longer relevant, minimizing legal risk. This yields significant cost saving. It means fewer IT resources spent on storage and integrations. It means knowledge workers can find what they need, when they need it, to become more efficient. And finally, it means a reduction in risk, because your content is secure, and it is retained according to the applicable business practices and industry regulations.

However, is it feasible to place all content into one single repository? Can one single system satisfy all of your content management needs? Will it continue to grow to meet all future needs? Historically speaking, the answer to this is no. Therefore, a good information management strategy needs to be dedicated to federation as well as consolidation in order to be effective in the future.

Instead of focusing on content management, your strategy should focus on *making content manageable*, no matter where it exists.

### **The Unstructured Content Problem**

As stated, the main problem is infoglut. For many years information management centered on the goal of “information at your fingertips”—the idea that access to content and data was the key to knowledge management. Vast enterprise-wide visions were provided at leadership forums and IT conferences focusing on users that would have instant access to information from wherever they happen to be working. The state of information management 20, or even 10, years ago was such that the vision of instant access to information seemed to solve the majority of knowledge problems in the workplace.

This is understandable. Twenty years ago, most content went through an information broker of some kind, and often had a physical location. If you needed a report you went to a person who had the report in a drawer. An old memo might be in a nearby filing cabinet or perhaps down in the basement archives. The fact that information was not readily available was, by most accounts, the major barrier to collaboration and knowledge sharing.

Technology has made massive strides forward in information access. In most cases, users do have the ability to get to a massive amount of content directly and easily from their desktop, if they know where to look. Often the information that was once hidden away by information brokers and scattered across the enterprise in physical repositories is now available immediately. Provided that the information brokers are motivated to share information, and allow others to access their once hidden data, we can move beyond this problem of content “silos” that hoard information. We can instead place all of this information online in a readily accessible format.

However, this solution has led to a new set of challenges. Organizations must now contend with providing the *right* access to information. That is, more specifically, they must secure the information that is now so readily available—a problem previously solved in many cases by the existence of an information broker. They must also enable consumers to find their content quickly, and ensure that it’s accurate.

Additionally, despite the fact that digital information no longer has a fixed location, people still organize it as they would organize files in a giant folder. This mentality has led to its own kind of content silos. Only those with the proper training can navigate the arcane taxonomies that were useful to information brokers, but incomprehensible to everyone else.

In short, by breaking down these content silos, you have created a flood of information that overwhelms your users. Some of it is very useful, and you would never have found it if you had to contend with an information broker. However, most of this content is not relevant, leading to infoglut. How can your users find what they need and make sure it’s authoritative, recent, and secure? How can they find what they need when they don’t even know if it exists?

## **Adding to the Problem**

In almost any organization, information is scattered between hard drives on user computers, shared network drives, e-mail inboxes, databases, content management systems, miscellaneous Web and FTP servers, and within individual applications. Users have a need to store this information within the context of the places where they work. This need explains the myriad of locations across which content is scattered for any given user.

As an example, think of the way the average business document is created, such as a knowledge worker creating a presentation in Microsoft PowerPoint. They begin by creating a presentation file on their laptop computer. That file is stored on the local file system, perhaps in their

## 10 Transforming Infoglut!

Documents folder as the user works on the document locally as the sole owner of the presentation. Multiple versions of the document may be saved as different revisions, commonly using file naming standards such as “preso\_V1.ppt” and “preso\_V2.ppt.” This is done manually by the user since most operating systems do not currently provide standard versioning capabilities in their core file storage system. As the document nears completion others are asked to contribute to the document and also to review its contents, often by e-mail.

Now a collection of additional users collaborate on the presentation using the native application to view and edit the content, again, often saving their edits and changes as unique new files called something like “preso\_V1\_apm.ppt” or “preso\_V1\_bex.ppt” to reflect the changes made by a specific user. Those changes are then sent out to the group for review. Now, in addition to the multiple copies on the original author’s local file system, we have copies of various revisions on local file systems of the different users as well as throughout the e-mail inboxes of all the participants. If the file is large, they sometimes participate via shared file systems or FTP servers instead of e-mail.

Once the document completes this review process, it may be published and distributed. This may put additional copies of the finished document in various locations within the network. It may also extend the number of copies outside the firewall if the document is e-mailed or uploaded somewhere external. Other users may store the document in locations to share with their peers. Very quickly we have different versions of the same document in different states stored across e-mail, network shares, e-mail inboxes, and even at third-party locations. In essence, the way in which we work with content today perpetuates the root problem of storing documents in decentralized, inconsistent ways.

There is also the core issue of access and process. Consider how the goals of knowledge workers conflict with the goals of process workers. A process worker needs a robust enterprise application to store highly structured data about the state of the business. These systems contain vast amounts of useful information, locked away so only a select few can access it. The information is highly structured and typically accessed via the enterprise application interface. The process workers need that level of control to secure the structured information and ensure its integrity.

Now consider the knowledge worker, who needs reports from those systems in order to plan for the future. In many cases, knowledge workers are not granted sufficient access to those systems, but they absolutely need

access in order to properly do their job. This creates a strong incentive to make them violate the security policy through cajoling or favors to get access to the sensitive data. They then promptly analyze the data, save their results in a spreadsheet, and share it with their team members via e-mail or shared drives.

The result? Your highly sensitive raw data is now stored, along with an even more sensitive analysis of the data, in an unsecured spreadsheet. One misdirected e-mail, one lost laptop, or one disgruntled employee is the only thing between you and a critical security leak.

In order to make sure that your employees can find the information they need, when they need it, in the right format, and with the right security,

### **What's in Your Digital Landfill?**

According to the Department of Commerce, US companies spend \$91.9 billion in non-capitalized expenditures and \$141.6 billion in capital expenditures on IT each year. What is troubling is that most of that is spent on “structured” information. But that leaves an entire world of information untouched: the world that I like to call the “digital landfill.” This is the place where all the application files, images, web pages, text messages, e-mails, and a host of other bits and pieces of electronic information wind up. For the most part, these are unmanaged and out of control.

There is a growing awareness in organizations that the digital landfill is a problem. A recent IDC analysis concluded that by 2011, the “digital universe” will be ten times as big as it was only two years ago. A lot of this information is business critical, and yet poorly managed.

Our user surveys tell us that on a scale of 1 (terrible) to 10 (excellent) 54 percent would give themselves a grade of 5 or less in terms of the effectiveness of their organization in managing information. 90 percent of organizations view their ability to manage electronic information as critical to their future, yet 52 percent of organizations have “little or no confidence” that their electronic information is “accurate, accessible, and trustworthy.”

ECM is emerging as a mainstream industry to help solve these problems, but it has barely begun to tap the massive opportunities available.

—John Mancini,  
President, Association for Information  
and Image Management (AIIM)

you need *in-context enterprise content management*. This means being able to slide content management into the normal business process of your knowledge workers. You need to ensure that they do not need to significantly alter their habits, in order to take advantage of ECM. Ensuring that your documents are up to date should be as easy as opening them up on your desktop. Securing your documents should be effortless, and they should remain secure even when taken out of the repository. An enterprise application should be able to present unstructured content to the end user, without that user needing to leave the content of their application.

# Why Do I Need ECM for Unstructured Content?

Almost any database application can store unstructured content. If you treat it like binary data, you can just store it in a table and have simple rules for managing revisions. That's content management, right? Not quite. There's a huge difference between revision control and enterprise scale management of unstructured content.

Does your system encourage findability and reuse, or is it just another content silo? Does your system encourage good information architecture, or is it merely allowing content to proliferate in chaotic new ways? Do you need workflows or subscriptions to control notifications about changes in the content? Do you need to store descriptive information alongside this document—commonly called *metadata*—such as title, keywords, or department to help people find and process this content? How about allowing people to make comments or annotate the content? Can you transform the content from one format to another so people get the content however they need? What parts of your business process can you automate with unstructured content services like routing, notifications, or conversion? How easy is it to integrate your system with an externally defined security policy? Are you confident that your system can scale to a billion content items? Are you confident that you have the tools you need to properly administer a repository of that size?

By way of comparison, every enterprise application needs to manage users, passwords, and access rights. When network applications were new, it was common for every application to maintain their own user database. When new systems came online, administrators might migrate the entire user repository to the new system, or they might just re-create the users they needed.

However, as the number of systems grew, architects found that they could no longer ensure that all systems had the correct user information, nor were they confident that all systems were properly updated when employees entered or left the company. It was simply too difficult to manage user data in every application, even if every application needed user data. This ultimately led to centralized user repositories and identity management systems. Modern enterprise applications integrate with these systems, instead of trying to replicate the functionality. This ensures best-of-breed technology and easier administration.

Enterprise content management solves a parallel problem. Your enterprise has dozens of applications that need unstructured content. They need to find it, display it, and change it. They also need assurance that the content follows the proper compliance and security policies. They need content services to repurpose assets for the Web, for handheld devices, or for marketing brochures. They also need a guarantee that the item is authoritative: once one application changes the item, all should view the new revision. There should be no proliferation of copies. And finally, at the end of the content item's life cycle, these applications need assurance that the item will be properly archived or destroyed.

In short, your enterprise applications need enterprise content management.

## **ECM Is Empowerment, Not Control**

The primary driver behind old document management solutions was to ease the burden for a small team that produced large quantities of complex documents. People needed systems that guaranteed to have the most recent version of items, along with authoring tools to assist in the creation of very large documents such as books, legal contracts, or product documentation. These tools had value in helping experts create documents, but not for average users to create or find useful information. These old systems catered to the expert creator, not the average creator, nor the consumer.

One of the primary goals behind modern content management is to empower contributors. Sometimes this takes the form of giving power users extremely complex tools that help them author complex documents. In other cases, it means allowing people to contribute content to a knowledge base or a Web site without having to be retrained to use a new tool or a new technology.

## 14 Transforming Infoglut!

Web content management changed this ecosystem in the early part of this decade. As people placed content on Web sites for easy access, suddenly the number of Web sites proliferated. Some organizations had thousands of Web sites, and you needed to know HTML in order to update them. This turned the webmaster into a bottleneck for updating the information on those sites, and stale content proliferated.

The solution was to empower both the webmaster and the content creators. The webmaster needed powerful tools to control the overall look and feel of the site. Contributors needed simple tools that enabled them to quickly and easily modify Web content. Together, this creates the best of both worlds: Web sites that conformed to branding guidelines and IT standards, but allowed untrained users to keep the content updated.

A content management system needs to stay focused on the needs of all users of the system: contributors, consumers, managers, administrators, and developers. It needs tools of varying degrees of complexity in order to help each of them create, find, share, and manage content.

It should empower all contributors, not just the experts. It should empower consumers, so they can find information regardless of how the contributor chose to organize it. It should empower developers, by giving them the tools they need to integrate with ECM, and never again be the bottleneck in content authoring.

## **Centralized Policies for Decentralized Information**

The goal of a pragmatic ECM strategy is not to force you to place all content into a single repository. Although this is desirable, it is not always practical. Instead, a pragmatic ECM strategy forces you to centralize your policy, and then has tools that allow you to enforce that policy wherever content may exist.

In the past, IT centralization went through a curious series of boom-bust cycles. In order to reduce costs and enforce consistency, IT laid down the rules about what information management systems to use, what languages to use, and what frameworks to use. Unfortunately, over-centralized IT usually led to insulation from the needs of the business. This meant that IT and business strategies failed to properly align themselves.

Unfortunately, non-aligned strategies inevitably lead to application proliferation. Departments install line-of-business solutions without IT approval to help them become more efficient or collaborate more effectively. These include hosted solutions, open source applications, or easy-to-install collaboration tools. This may even include applications that

are difficult to secure and maintain, but are still valuable because they perform a specialized function that helps one business unit meet their objectives. Because these systems were set up without IT approval, the odds are low that the business units put sufficient thought into how IT should manage or integrate with these systems. Unfortunately, as soon as these systems become business-critical, IT will be forced to maintain them and integrate with them.

In practice, the more centralized your IT policy is, the more likely it is to not satisfy the needs of all users. As Web 2.0 collaboration tools become more commoditized, individual departments will inevitably set up systems outside of central control.

Even if your IT department has 100 percent control over all your business units, there is always the looming problem with mergers and acquisitions. What if your organization merges with another: what should you do with their content management applications? How long will it take to consolidate all the information? Is 100 percent centralization always cost-effective?

A pragmatic ECM system needs to be aware of these problems, and needs to solve it through innovation, not control. A pragmatic ECM strategy needs several pieces. It needs a repository that *could* satisfy the needs of all users and all formats, but doesn't require you to configure it this way. It needs to be an infrastructure that all your applications can use. It needs to support middleware, and act as middleware, to easily integrate with multiple systems across your enterprise. Finally, it needs simple federation tools that extend the reach of content management services to legacy applications that are difficult or cost-prohibitive to change. And finally, it needs to ensure the content is secure, no matter where it exists.

That is the core of the pragmatic ECM strategy, which is the subject of the remainder of this book.

## How You Should Use This Book

The purpose of this book is to help you change your business. However, that will not happen unless you take specific action. To get the most out of this book, I would advise you to follow Seth Godin's sage advice on how to read any book about business strategy.

First, decide that you will change three things about how you work today. At least three. We will present hundreds of options in this book, but if you only change the three things that you need the most, your organization will benefit greatly from the time you spent reading. The purpose of this book is

## 16 Transforming Infoglut!

not to force you to implement a grand 10,000-point plan that you must follow to the letter in order to achieve ECM, nor is it to sell you any specific product. The purpose is to teach you how ECM is an initiative, not an application. It's about creating a culture that shares information, so it's reusable, secure, and findable. A pragmatic ECM strategy requires a good deal of initial planning, but will always be a moving target. Dedicate yourself to making three important changes, decide what those three things are while reading this book, then use the book as guidance for how to change.

Second, write on this book. Your time is far more valuable than the paper you hold in your hands, so take notes! Use a highlighter and mark up the sections that you find most important to you. At the very least, take notes on paper, and leave them in the book. What three changes seem like they would have the best return on investment? What actions should you take to get there? What might be an easy way to improve on how you are doing business today?

Third, share this information with your team. This book was designed to be primarily an ECM strategy book, but it also contains information useful for implementation. You should share this book with your ECM team to make sure that everybody is speaking the same language. This book should make sense to every member of your ECM Center of Excellence, which is a cross-departmental team that brings your ECM visions to reality (covered in Chapter 3). Your Center of Excellence is vital to keeping IT and business aligned, and this book helps you communicate and realize that value.

## Takeaways

- Easy access to online information has not solved the knowledge management problem; instead, it has created *infoglut*.
- Infoglut is information overload: the inability to find what you need in the ocean of information.
- Enterprise content management (ECM) solves 80 percent of the infoglut problem: finding content, securing content, ensuring content is authoritative, destroying it properly, and reducing storage costs.
- ECM is about empowerment, not control.
- Centralized policies, along with strategic integrations and federated tools, help extend content management services beyond the ECM repository.