

Assured Federated Records Management: Achieving Compliance Across the Enterprise with EMC Documentum

Technology Concepts and Business Considerations

Abstract

As the word “compliance” continues to gain mindshare in the executive suite, companies are beginning to rethink their approach to records management. This white paper details how the EMC[®] Documentum[®] platform provides enterprise records management that can consistently declare and classify records in a folder tree or file plan, and apply records-related metadata, retention and disposition policies, and legal holds, to a wide variety of document types throughout the enterprise.

October 2009

Copyright © 2007, 2009 EMC Corporation. All rights reserved.

EMC believes the information in this publication is accurate as of its publication date. The information is subject to change without notice.

THE INFORMATION IN THIS PUBLICATION IS PROVIDED “AS IS.” EMC CORPORATION MAKES NO REPRESENTATIONS OR WARRANTIES OF ANY KIND WITH RESPECT TO THE INFORMATION IN THIS PUBLICATION, AND SPECIFICALLY DISCLAIMS IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.

Use, copying, and distribution of any EMC software described in this publication requires an applicable software license.

For the most up-to-date listing of EMC product names, see EMC Corporation Trademarks on EMC.com

All other trademarks used herein are the property of their respective owners.

H3499.1

Table of Contents

Executive summary	4
Records management at the enterprise level	4
Introduction	4
Audience	5
What is federated records management?	5
Benefits of federated records management	6
Features of federated records management	6
Technical challenges	7
The diversity of enterprise records	8
The structured data problem	9
The immutability problem	11
EMC Documentum Assured Federated Records Services	12
Conclusion	13
The bottom line: Demonstrable risk management	13

Executive summary

Records management at the enterprise level

As the word “compliance” continues to gain mindshare in the executive suite, companies are beginning to rethink their approach to records management. It used to be that records referred to special isolated document sets closely controlled and protected by professional records administrators. Today the term “record” has expanded to cover all information assets subject to any regulatory retention, privacy, or audit requirement, or discoverable in civil litigation. Most of those documents today live not only outside the purview of records administrators, but are not even stored in a managed content repository. Instead they live in network file shares, employee desktops, collaboration portals, and— most notoriously—e-mail systems. The risk that a ticking bomb lurks somewhere in there is making legal departments increasingly anxious.

Moreover, as the new U.S. Federal Rules of Civil Procedure (FRCP) have made clear, such records are not limited to documents. FRCP changes that went into effect in December 2006 now make all forms of “electronically stored information” subject to e-discovery. That includes data in addition to documents. The inability to disclose in an “early meeting” all relevant information, where it is, what it contains, and how it will be produced, can invalidate a company’s claim of privilege over that information, and risks heavy damages in the litigation itself.

An electronic record is more than just a document in a content repository. That stored document is just the content component of the record. The record also contains metadata describing the content, its relationship to other records in the company record structure or “file plan,” its retention policy, security, and other properties. That metadata absolutely controls access to the record content, does not allow it to be deleted or modified during the retention period, and even strongly controls changes to record metadata.

Records are subject to retention policies, dependent on their classification, which determine how long they must be retained in unaltered form as well as their ultimate disposition at the end of the retention period. During the retention period, events such as an impending investigation or litigation can place a legal hold on a set of records, preventing their disposition even if their normal retention period has expired.

The EMC® Documentum® platform, which provides robust infrastructure for storing and managing any kind of enterprise content, has extended that enterprise approach to records management. Enterprise records management (ERM) requires a platform that can consistently apply records-related metadata, retention and disposition policies, and legal holds, to a wide variety of document types throughout the enterprise.

EMC Documentum Retention Policy Services (RPS) does exactly that, supporting not only typical office documents but e-mail, threaded discussions, and physical records such as paper. The RPS modular architecture allows different parts of the enterprise to implement records management at its own pace and degree of formality. Retention policies, security policies, and the choice of formal records management versus simple retention management can all be tailored to the needs of individual departments and record types without losing the benefit of a single set of retention policies for the enterprise. In addition, EMC Documentum RPS is built on top of the Documentum platform and user applications, so content does not have to be moved in order to be managed as records, and Documentum users can access records through a familiar interface.

Introduction

This white paper introduces federated records management, a means to manage records in place, including the benefits, features, and technical challenges such as data structure and immutability. The paper also

explains how EMC Documentum Assured Federated Records Services works to assure record integrity, and how it provides demonstrable, documented, and auditable risk management.

Audience

This white paper is targeted toward records managers and IT managers who are responsible for enterprise content management systems and records management systems. Others who may benefit from this paper include business managers or compliance officers who need to understand what role federated records management can play in their information management strategy.

What is federated records management?

The simplest way to implement ERM would require all the records in the enterprise to be copied or moved to a single repository—such as a single Documentum repository. That, however, does not meet the needs of many organizations, which might already have multiple Documentum repositories and other enterprise content management repositories as well. Moreover, they also have voluminous content not currently stored in any managed repository, but instead reside in the file system, e-mail systems, SharePoint portals, and legacy systems. And those are just the documents! The new FRCP rules cover all forms of electronically stored information, including structured data, which today is stored in databases and enterprise applications.

To bring it all under the ERM umbrella, you could copy all of the electronic stored information to a single repository, and keep doing it for each new document added to those systems. At the enterprise level, this is not practical.

What users want instead is federation. That means the ability to manage the content in place—wherever it currently resides—as if it were a single ERM repository, that is, with a single set of classification and retention policies, consistently applied and enforced throughout the enterprise. Many of the native content stores—file systems, e-mail systems, and legacy systems—however, don't provide either the necessary metadata or the control methods to support retention and other records management requirements on their own.

Realistically, federated records management means leaving the record content in place, in disparate stores throughout the enterprise, and securing that content by centralizing the record metadata in a master repository. In other words, there is a single master database of records for the enterprise, but the content of each record does not have to be moved. The ERM repository controls access to record content stored remotely, ensuring that it is not deleted, altered, or moved during the retention period.

EMC Documentum Federated Records Services (FRS) provides federated records management (FRM), but recognizes its inherent limitation, which is that some content stores cannot absolutely delegate full control over retention to ERM. They still allow local system administrators to delete or alter content through the content store's native administration interface. While problems can be minimized through strong management policies and procedures, some residual risk remains. To eliminate that risk, EMC Documentum is introducing an enhanced form of federated records management—EMC Documentum Assured Federated Records Services.

Assured Federated Records Services allows the verification and provides audited proof that content stored in various disparate repositories across the enterprise continues to be centrally managed as enterprise records under a single set of policies. To overcome the risk inherent in federated records management, Documentum provides a patented *assurance engine* that uses statistical sampling techniques to systematically compare records with the federated content they point to, generating alerts on any missing or altered records detected, along with a verifiable audit trail. The comparison and sampling rules are tunable

to each remote system, or even folders within each system, to match the business and technical risks of each.

Benefits of federated records management

Federated records management is usually not a trivial undertaking to implement enterprisewide; however, it offers significant benefits.

- **Leverage existing IT assets.** As compliance and risk management become “enterprise” concerns, the ability to continue using existing IT systems, rather than to rip them out and replace them, is a significant benefit. Federated records management acknowledges the huge investment companies have in existing content management systems, file systems, report management systems, and other sources of record content. The key benefit of federated records management is leverage of those existing systems and allowing content to remain in place without losing the assurance of secured retention. This does not mean that repositories that have no further use other than for retention of content cannot be de-commissioned. Content from these repositories can be migrated to the master repository for management and retention at any time. It does, however, provide the enterprise with the ability to plan, prepare for, and execute these migrations when it is the right time to undertake such a venture.
- **Improved compliance and records capture.** A nagging fear in the general counsel’s office is that discoverable information is lurking unnoticed in the e-mail system, network file system, and other unmanaged stores. Federated records management allows companies to apply retention policies automatically to “controlled” folders without having to move record content to a centralized store. By making enterprise records management practical, federation helps to ensure that a larger portion of discoverable information assets is captured and retained.
- **Lowered cost of discovery.** When litigation, investigation, or audit occurs, the cost of finding and producing relevant documents and other electronically stored information can be staggering. Federated records management provides an efficient cost-effective way to search across all content stores, meet FRCP early meeting requirements, apply legal holds across all of the records, and produce them on demand. In addition to lowering the cost of discovery and record production, federated records management lowers the risks of not being able to produce all requested records. Situations such as these are notorious for adverse consequences costing hundreds of millions of dollars.
- **Improved policy management.** When records are distributed across multiple systems, the policies governing them are usually distributed as well. That means as those policies change, or as new regulations take effect, the changes must be implemented separately in each system. Doing this consistently and cost-effectively is difficult and a nagging management headache. Federated records management allows retention policies, disposition policies, and file plan classification policies to be maintained centrally, assuring consistency across all information assets and agility in the face of ever-changing requirements.
- **Improved policy enforcement.** If the ability to maintain policies centrally is valuable, the ability to enforce those policies consistently across all records in the enterprise is priceless. In today’s world of enterprise-level compliance and risk management, it is impossible to secure access to every record through the mediation of trained records administrators. New records are being created all the time, in many cases invisibly to their authors. The application of retention policies has to be invisible and automatic as well. Federated records management does not rely on simply publishing changes to corporate retention policies, but it automates their enforcement on all federated content stores even those without native retention management features.

Features of federated records management

Some vendors use the term “federated records management” loosely. For example, the ability to perform a single query to locate records stored in multiple repositories across the enterprise is a valuable thing, but it falls short of being federated records management. For example, content integration technology that supports mapping of metadata between disparate repositories, such as Documentum Content Integration

Assured Federated Records Management: Achieving Compliance Across the Enterprise
with EMC Documentum

Technology Concepts and Business Considerations

Services, can provide this. Or if you are just using text queries, even standard search engines like Google can do it. But that does not fulfill the requirements of federated records management. So how would you recognize FRM if you saw it?

- **Centralized policy management.** FRM must provide centralized policy management for the enterprise, that is, a single place to maintain retention policies, disposition policies, access control and security policies for all records in the enterprise, no matter where the content is stored. Centralized policy management is an essential element of compliance and risk management at the enterprise level. However, centralized policy management could be implemented without achieving true FRM. For example, a master policy repository could advise remote records management systems on changes to policies, but it would have to trust that the advice was implemented uniformly throughout the enterprise. So centralized policy management is a necessary component but not sufficient for a complete FRM solution.
- **Centralized policy enforcement.** Centralized policy enforcement is the key to true FRM. In particular, the retention policy applied via the central file plan is not just a rule on paper but the system actually prevents deletion or alteration of record content on the remote system. This requires the FRM system to provide adapters to each type of supported content store through which it can disable deletion or alteration on the remote system. The difficulty in ensuring this 100 percent is the motivation behind EMC Documentum Assured Federated Records Services.
- **Assured Federated Records Services** allows the verification and provides audited proof that content stored in various disparate repositories across the enterprise continues to be centrally managed as enterprise records under a single set of policies. To overcome the risk inherent in federated records management, Documentum provides a patented *assurance engine* that uses statistical sampling techniques to systematically compare records with the federated content they point to, generating alerts on any missing or altered records detected, along with a verifiable audit trail. The comparison and sampling rules are tunable to each remote system, or even folders within each system, to match the business and technical risks of each.

Technical challenges

The FRM promise to secure all enterprise content in place seems too good to be true and in many cases, it is. Federation is not a cure-all to the enterprise records problem. Sometimes there is no effective alternative to copying or moving record content to a centralized repository. With some kinds of information, such as structured data, it may not even be obvious how to do this in a records management context.

EMC Documentum Assured Federated Records Services represents a patented and innovative approach to the technical challenges of record federation. Before discussing the approach, let's take a closer look at the challenges involved.

The diversity of enterprise records

The most obvious challenge to FRM is the vast diversity of content that needs to be federated, and by implication the diversity of native content stores that would need to be brought under the FRM umbrella. Figure 1 illustrates the diverse sources of record content in the enterprise and their native stores, to which we would like to apply through federation, a common set of records management services: classification, search, policy enforcement, and disposition.

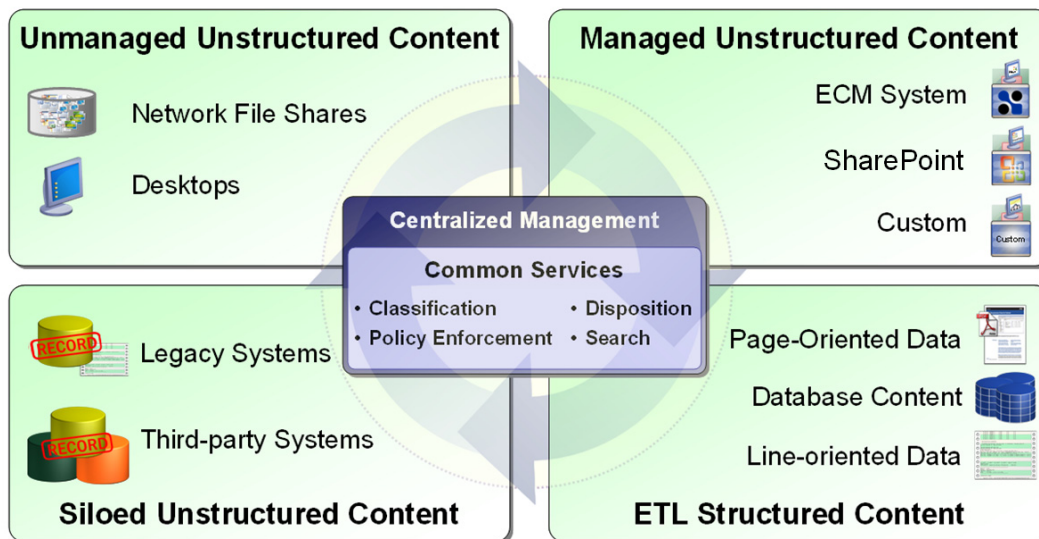


Figure 1. Enterprise content includes diverse information types and stores that were never designed for records retention.

The managed unstructured quadrant at the top right is the most familiar, represented by enterprise content management systems such as EMC Documentum, collaboration portals such as Microsoft SharePoint, and others. We call it unstructured because the content is in the form of documents, typically formatted text, without any fixed or computer-readable structure. We call it managed because access to the content is controlled by an application, with metadata applied to each content object, which also allows the management of security and controls access.

The top left quadrant, represented by files on network servers and user desktops, is called unstructured because, again, the record content is typically in the form of documents. But this content is unmanaged because access is provided directly by the operating system, which provides minimal metadata useful for search, access control, or version control.

Structured content, better known as data, has become an important record category under the new e-discovery rules. It includes information stored in “live” form in databases and enterprise applications, as well as snapshots of that data captured in host-generated reports, statements and invoices, and log files. The “batch” nature of structured content creates special problems for securing individual records.

“Siloed” unstructured content, shown in the lower left, includes both structured and unstructured information created by legacy systems, home-grown applications, and other proprietary systems. It is called out separately here because that information is not easily transformed into standard content objects, and can only be viewed and managed from within the native application.

The important thing to note here is that the native content stores differ not only in the format of the content but in the availability of metadata, methods of securing access control, and the scope of an individual information object. Thus, FRM has to apply different technologies and procedures to bring them all under a common management umbrella.

The structured data problem

The structured data quadrant presents special problems for records management. It is not uncommon that critical information needed in audit or e-discovery is some element of structured data, and the ability to retain it and produce it on demand is vital to any corporate records management strategy.

It is useful to separate data into three types of information objects. First there is the live data stored in relational database tables in a Database Management System (DBMS) or enterprise application, for example, a customer's account register. For recordkeeping purposes, more often we want to secure not the live database but periodic snapshots of that database, and secure each record within that database snapshot separately, for example, for each customer account or transaction.

The two basic types of database snapshots are page-oriented and line-oriented, representing two types of content objects in FRM.

Page-oriented content renders the data, along with formatted text, in printed form for human consumption. Typical examples are bank statements, invoices, and insurance policies. The content is typically produced from standard relational data but capture and retention of the information with full print fidelity (that is, as presented to the customer) are important for compliance.

Line-oriented content, on the other hand, is not typically a snapshot of a relational database but a log of events or transactions written one at a time as they occur, such as a bank transaction audit file. This information is not designed for human consumption but represents an important record of the state of the business at some point in time.

Many systems that produce and maintain structured data can also produce page-oriented or line-oriented content. However, that is not always the case. In those cases where page- or line-oriented content cannot be produced, special technology called Extract, Transform, and Load (ETL) can be employed to create content snapshots of the data.

Whether starting from page-oriented content, line-oriented content, or ETL-generated content, managing structured data as enterprise records requires the same three steps: archiving snapshots of the data in some standard form; bursting the archive so that record metadata can be applied at the proper granularity; and then securing bursted pieces with records management policies.

The archiving step is straightforward, as there are numerous tools, including ETL, available to capture data snapshots and render them in a standard recordkeeping format. Where print fidelity in the rendition is important, Adobe PDF—or, more specifically, PDF/A, a version of PDF intended for long-term archiving—is the obvious choice. Otherwise, XML could also serve as a standard archival format.

However, these reporting or archiving tools are batch-oriented. A single PDF or XML file created from a daily or monthly run is going to contain thousands of customer statements, invoices, or transaction reports. Given the huge volumes involved, this is the only way to do it for performance reasons. But while it solves an archiving problem these massive content objects create a records management nightmare, since you can't be assured that the same policies apply throughout.

Even if they all are subject to the same nominal retention period, say seven years, what happens if litigation places a legal hold on one customer's statement? Does that mean that all other statements in the batch need

to be held as well? Ideally, you'd like to leave the massive content object—the single PDF file—alone, but create separate record objects with each pointing to an individual customer statement.

EMC Documentum Archive Services for Reports is an intelligent report archiving solution that allows this. Like other tools, it normally creates a large PDF file from a batch print run, say, of customer statements. But unlike other tools, it allows a user to check out a single logical section, say a single customer statement, and treat it as a single object separate from the batch PDF. Archive Services for Reports updates its internal index so that it knows that the particular customer statement is now a separate object, but the PDF pages are left in the original file.

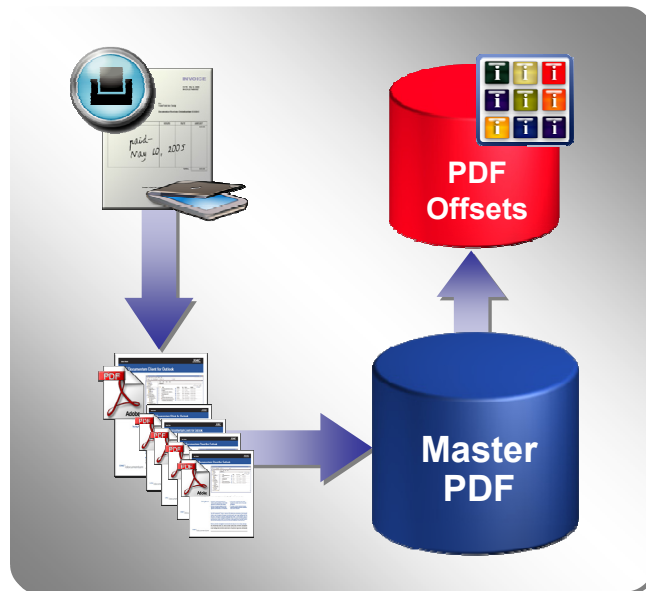


Figure 2. EMC Documentum Archive Services for Reports supports managed retention of individual records within massive batch report PDF files.

Within the master PDF, Archive Services for Reports maintains these objects as offsets (Figure 2), so that the record points only to a section of interest within the file. In the Archive Services for Reports variable object storage model, users can elect either no bursting, bursting specific documents on demand, or full bursting on ingestion. If a legal hold is applied to that statement, the batch PDF can be deleted at the end of its normal retention period, but Archive Services for Reports ensures that the held statement is retained. This lets you apply records management on massive volumes of structured data by exception only.

Archive Services for Reports is also effective for archiving and retaining line-oriented data. As in the previous case, a print run archives a large volume of line data to PDF. For example, a single PDF file representing one day's transactions could contain hundreds of thousands of lines. The file is stored as a single object in the Documentum repository until it is required for discovery or analysis.

At this point its contents are loaded in to a temporary database. A query of the temporary database returns data as a standalone XML file, which is then turned into a record. Since the XML extract file is generated automatically by the system, it can be shown to be a faithful representation of the original data, establishing a reliable chain of custody. Thus, Archive Services for Reports combined with Documentum RPS provides an effective way to bring both page- and line-oriented structured data under the FRM umbrella.

The immutability problem

Besides finding a common basis for managing all types of content, FRM must deal with the problem of guaranteeing enforcement of retention policies on disparate remote systems. Even content stores with available APIs securing their objects against deletion cannot prevent a local administrator, using that content store's native administration tools, from overriding those protections, either inadvertently or maliciously. Overcoming this inherent obstacle is the motivation for the EMC assured federated approach.

Figure 3 illustrates the general procedure for federating records management. The first step is to discover content on the remote system, typically using an agent or scheduled job on the remote system.

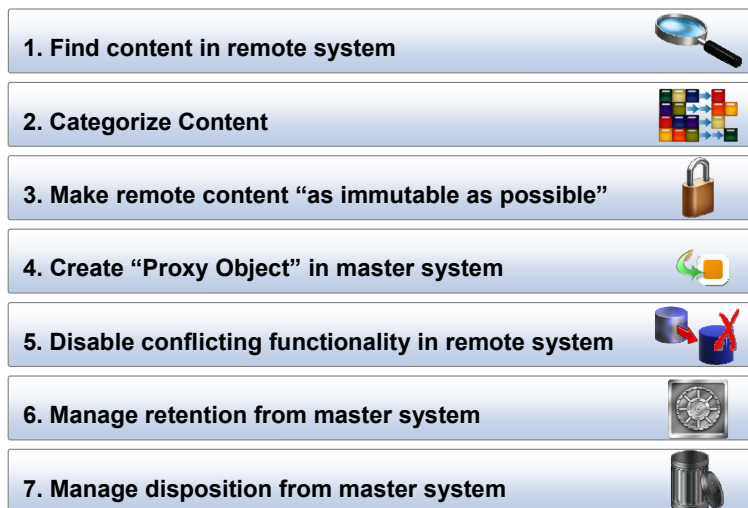


Figure 3. Federated records management requires making remote content "as immutable as possible."

The discovered content must then be classified in the enterprise file plan, based on metadata or content. Next, you need to make the remote content immutable—meaning it cannot be deleted or altered—using the APIs on the remote system, and disabling any conflicting functionality on the remote system. If the remote system lacks such APIs, it cannot be federated but must be copied or moved to the master system.

You then create a record in the master database with the classification metadata and a proxy object in the master system representing the content. The proxy object is a pointer to the original content on the remote system plus a convenience copy of metadata used in a record search result set, plus a hashed value representing the record content. The hash is used to assure the integrity of record, such that, if the record content changes even in the slightest the hash value will no longer match. Since the original content, left on the remote system, is immutable, the proxy object stays in sync with the original object. The master system manages retention and disposition centrally, controlling deletion on the remote system. For example, after deleting the record in Documentum, FRM deletes the content on the remote system, and then deletes the proxy object.

While the remote content is called immutable, it is better termed "as immutable as possible," for the reasons mentioned above. An administrator using the native tools of the remote system or some malware running "as root" can sometimes override the API securing immutability. So while immutability can be promised, it cannot always be guaranteed. This is the source of the problem, since if the original object is modified or deleted, the proxy object is no longer a faithful representation. You may have lost a record without even knowing it!

EMC Documentum Assured Federated Records Services

EMC starts from the recognition that it is impossible to guarantee that the remote system will “behave itself” or in other words that the proxy object will always faithfully represent the remote record content. In spite of this, EMC’s patented approach provides assurance that the federation is reliable. The assurance comes from implementing a monitoring process that tests the integrity of the proxy/remote object relationship on an ongoing basis.

EMC Documentum Assured Federated Records Services works as follows: Scheduled jobs periodically test the integrity of the master records repository. The user selects the relationships and the percentage of records to test. The job then retrieves the remote record content and compares it to the Documentum proxy object. If the attributes fail to match, the test fails. If the attributes match and the remote timestamp indicates the original object has not been modified, the test succeeds. If the timestamp is unavailable or unreliable, the retrieved remote object can be hashed and compared with the recorded value.

The monitoring is configurable to match the business risk. The sampling percentage and matched attributes can be dialed up or down as warranted by the risk represented by a particular set of records.

Assured Federated Records Management allows EMC to provide a unified enterprise records management solution for electronic, physical, and federated records across the enterprise, centralizing policy management without requiring record content to be copied and moved from its original location. The solution is fully integrated with Documentum Retention Policy Services, which means it can leverage advanced features like unlimited event- and time-based retention policies, legal holds, permanent records, and periodic review. The solution supports records management, both manual and automated record declaration and classification, all on a DoD 501 5.2 certified environment. As much or as little records management can be applied in each segment of the enterprise, and the degree of federated integrity assurance can be tailored to the business risk associated with each class of records.

Conclusion

The bottom line: Demonstrable risk management

Enterprise records management starts from the need to manage all of the company's records under a consistent framework of retention and disposition policies. Federation makes that practical by allowing record content to be managed in place leveraging existing IT systems where possible, but it cannot always assure record integrity. EMC Documentum Assured Federated Records Services restores that assurance, and adds the stamp of demonstrable, documented, and auditable risk management.

In today's world, simply capturing and retaining records is only part of the story. At crunch time, you need to be able to prove that you have all of the records, that none have been lost or changed, based on auditable procedures going back over time, and with documented auditable results. EMC Documentum Assured Federated Records Services delivers that. It allows record content to be managed in place, while providing assurance through continuous rolling validation of system integrity and a detailed audit trail. It does not automatically test the integrity of every remote record, but supports a flexible program of statistical sampling that balances computational effort against the business risk, and allows the company to recalculate that balance independently for each record set.

A second key feature of the Documentum Assured Federated Records Services offering is the balance of manual and automated processes. Like other federated solutions, Documentum FRM provides a set of technology adapters to other content stores that allows the master record repository to control retention on the remote store. But unlike other offerings, Documentum recognizes that the remote system may not be able to guarantee integrity with a fully automated solution. Instead, the Documentum assurance engine combines automation with manual procedures, such as tracking of accidental deletion of records. This not only makes the audit trail more reliable but significantly expands the range of systems that can be federated.

Finally, the Documentum solution recognizes that not all records are equal in terms of business risk. Thus the assurance monitoring is tunable to that risk—highly monitored for changes to, high-value records, light for stable, low-value records—allowing the cost of records retention and risk management to be tailored to the business value. Moreover, the integrity and effectiveness of procedures can be concretely demonstrated through monitoring and audit reports.

To learn how EMC Documentum can solve your enterprise records management challenges, contact your EMC sales representative, visit www.EMC.com, or call **800.607.9546**.