

EMC SourceOne for Microsoft SharePoint Technical Guide

Applied Technology

Abstract

This white paper reviews the basic functionality of EMC SourceOne™ for Microsoft SharePoint. SourceOne for Microsoft SharePoint can reduce the primary storage load on SQL servers and improve SQL Server performance by externalizing active content to tiered storage. Through archiving, it also allows inactive SharePoint content to be managed with consistent retention and disposition policies that support regulatory compliance, eDiscovery, and litigation readiness.

August 2010

Copyright © 2010 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. EMC², EMC, Centera, Documentum, and where information lives are registered trademarks and Infoscapes is a trademark of EMC Corporation. All other trademarks used herein are the property of their respective owners. h4865

Table of Contents

Executive summary	4
Introduction	4
Audience	6
SourceOne for Microsoft SharePoint: Technology overview.....	6
Operational efficiency—reducing the primary storage load on SQL Servers	6
Externalizing active content.....	6
Good information governance—archiving information for compliance, eDiscovery, and litigation readiness.....	7
Archiving inactive content.....	7
Security	10
Searching the SourceOne archive for SharePoint content.....	10
Conclusion	13

Executive summary

Microsoft SharePoint has become an enormously popular tool for aggregating and exchanging enterprise content. Many SharePoint deployments house thousands of sites, millions of documents, and terabytes of data. As SharePoint content grows, it can negatively affect the performance of the SQL servers that store it. The SharePoint user experience suffers, and backup windows get longer and longer.

Moreover, as much as 25 percent of SharePoint content is inactive¹, which means that it's either orphaned or has exceeded its useful "shelf life" according to corporate information policy. Yet, the managed enterprise information infrastructure rarely extends to all SharePoint content. So whatever should happen to inactive SharePoint content often doesn't happen.

EMC SourceOne™ for Microsoft SharePoint is a solution that can address two problems. It can reduce the load on SQL servers by externalizing active content to a more cost-effective tiered storage environment, while leaving the metadata in SQL Server. With metadata in SQL Server, SharePoint retains "ownership" of the content and users are unaware that the content is stored elsewhere. EMC SourceOne for Microsoft SharePoint can also manage inactive content, moving it to an archive and enabling it to be brought under the same control and information governance policy that applies to other enterprise information. The archived content remains searchable by and accessible to users directly through SharePoint, with no impact to the user experience.

Introduction

If there is one thing upon which all information technology analysts agree, it is that businesses large and small are awash in a flood of digital information of their own making. According to the analyst firm IDC, in its white paper *The Digital Universe Decade*, our digital universe reached 0.8 zettabytes in 2009². A zettabyte is 1 trillion gigabytes. To help us grasp that, IDC suggests visualizing a stack of DVDs that reaches to the moon—and back.³ Although individuals *create* the great proportion of digital information, most of it passes "through the servers, network, or routers of an enterprise at some point." When it does, "the enterprise is responsible at that moment for managing that content, protecting user privacy, watching over account information, and protecting copyright." Finally, most of this information is unstructured⁴—the kind that increasingly lives in SharePoint. Much of this information is extremely valuable; some is pretty useful—at least for awhile—and quite a lot is digital junk. All of it, however, needs to be properly managed and stored, at ever increasing expense.

Significant legal and regulatory requirements accompany many types of information, requirements that impose a hefty burden on information management systems. Yet failure to comply can lead to litigation, monetary sanctions, an onslaught of bad publicity, a loss of public trust, and a large dent in brand image.

Clearly, information has a substantial downside risk. Reducing and managing that risk is the realm of information governance. According to The 451 Group, information governance concerns

- what information an organization has,
- where it is,
- how long it should be retained,
- who has access,
- how is it protected, and
- how policies, standards, and regulations are enforced.⁵

¹ "Gathering MOSS? Revealing SharePoint Opportunities and Costs," InfoTrends, August 2009

² Ganz, John, and David Reinsel. "The Digital Universe Decade – Are You Ready?" IDS iView (2010): 1.

³ Ibid, p. 10.

⁴ Ibid, p. 2.

⁵ "The Rise of Information Governance: From Reactive e-Discovery to Proactive Information Management," The 451 Group, August 2009

Today many IT projects are driven by information governance and the risk and compliance considerations it addresses. In lieu of a proactive information governance strategy and the tools to implement it, many businesses simply save everything—which brings us right back to unbridled information growth.

And, speaking of growth, there's a new element in the information infrastructure equation: Microsoft SharePoint. Organizations of all sizes are adopting Microsoft SharePoint. Over 60 percent of the organizations in a recent IDC survey either currently use or are planning to use SharePoint.⁶ SharePoint is easy to deploy and it connects seamlessly to the Microsoft Office suite of applications that knowledge workers use every day, so it meets a critical information management and collaboration need.

Many enterprises take on hundreds of new SharePoint sites every month. These sites can be located anywhere within a global organization, contain millions of documents⁷, and account for terabytes of data. Microsoft has released some impressive customer figures to substantiate this trend. For example, the National Oceanic and Atmospheric administration has 12 terabytes of data in Microsoft SharePoint sites.⁸ At Eli Lilly and Company, 110 million documents reside in SharePoint.⁹ And financial services giant Raymond James, one of Microsoft's largest SharePoint users, adds 10,000 new documents a month to its SharePoint trove, which stands at 40 TB and 400 million documents.¹⁰

Yet the aged and outdated content stored in Microsoft SharePoint is rarely archived, considered for long-term preservation, or managed under compliance and even more rarely attached to any existing system that governs content. That means a vast unmanaged wilderness of SharePoint data exists detached from the data center and subject to no information governance whatsoever. Even SharePoint users feel that managing this content is a problem.¹¹

As if the legal and regulatory risks of growing SharePoint content were not enough, it can pose other challenges as well. This glut of SharePoint information can degrade the performance of the very production servers that host the application. That's because the SharePoint SQL Server database not only stores metadata but content as well—as binary large objects (BLOBs). As SharePoint content grows, it is an efficient use of SQL Server resources to move these BLOBs out of SQL Server and focus its role to metadata storage only. Storing metadata in SQL Server ensures that SharePoint retains ownership of the content, which is especially important if there are workflows or business processes attached to it.

So good information governance and smart IT resource allocation demands a solution that enables an organization to fully support the use of Microsoft SharePoint while:

- Mitigating content growth and its burden on the SharePoint production system by moving active SharePoint content from the SQL Server environment to external, lower-cost, tiered storage
- Reducing backup windows
- Archiving inactive content spread across multiple SharePoint Team Sites and governing it according to consistent corporate retention and disposition policies and industry regulations
- Preserving a transparent end-user experience while enabling easy access to archived content
- Ensuring litigation readiness by making content readily accessible for discovery

EMC SourceOne for Microsoft SharePoint meets all these objectives. Archiving is a foundational technology for information governance whether the goal is operational efficiency, compliance, litigation

⁶ *IDC 2009/Microsoft Office and SharePoint Traction: An Updated Look at Customer Adoption and Future Plans* (IDC #220237, October 2009)

⁷ "Gathering MOSS? Revealing SharePoint Opportunities and Costs," InfoTrends, August 2009

⁸ Colligo Networks Webcast: *Colligo Networks Webcast: y For Enterprise Content Management?* March 2, 2009

⁹ Ibid.

¹⁰ Ibid.

¹¹ *IDC 2009/Microsoft Office and SharePoint Traction: An Updated Look at Customer Adoption and Future Plans* (IDC #220237, October 2009)

readiness, or all three. The remainder of this white paper provides an overview of SourceOne for Microsoft SharePoint.

Audience

This white paper gives an overview of EMC SourceOne for Microsoft SharePoint and its functional capabilities. It is intended for CIOs, developers, IT administrators, and line-of-business managers who would benefit from the operational efficiencies and advanced information governance that SourceOne for Microsoft SharePoint provides.

SourceOne for Microsoft SharePoint: Technology overview

EMC SourceOne is a family of next-generation information governance products for archiving, eDiscovery, and compliance. Currently, the family includes SourceOne Email Management, SourceOne Discovery Manager, SourceOne eDiscovery — Kazeon, and the most recent addition, SourceOne for Microsoft SharePoint.

EMC SourceOne for Microsoft SharePoint supports a proactive information governance strategy, which features improved operational efficiency, centralized content archiving, and consistent application of retention, disposition, and overall lifecycle management of corporate information. SourceOne for Microsoft SharePoint ensures that the right data is retained and managed according to industry and corporate regulations—while providing accessibility to native and archived SharePoint content.

Operational efficiency—reducing the primary storage load on SQL Servers

EMC SourceOne for Microsoft SharePoint can help manage the explosive growth of SharePoint content and reduce the performance impact of this growth on the production environment. SourceOne for Microsoft SharePoint accomplishes this in two ways. First, it can externalize “active” content, content that is being accessed regularly in the SharePoint environment. Second, it can archive “inactive” content, content that is orphaned, no longer in use, or has aged past its defined end-of-life date. Both capabilities can reduce the load on the SharePoint production servers.

Externalizing active content

Microsoft discusses the issue of active content storage on its developer network website. The company estimates that *“as much as 80 percent of data for an enterprise-scale deployment of Windows SharePoint Server consists of file-based data streams that are stored as BLOB data. However, maintaining large quantities of BLOB data in a Microsoft SQL Server database is a suboptimal use of SQL Server resources. You can achieve equal benefit at lower cost with equivalent efficiency by using an external data store to contain BLOB data.”*¹²

EMC SourceOne for Microsoft SharePoint follows Microsoft best practices for externalizing SharePoint BLOB content, which enables SourceOne to solve the three most pressing operational issues related to content growth:

- **Cost:** Large volumes of content stored in SharePoint demand increased expenditures for high-performance storage. Eighty percent of current MOSS users purchased additional IT infrastructure when they implemented SharePoint.¹³
- **Data protection:** The larger the SharePoint farm, the longer it takes to back up.

¹² "External Storing of Binary Large Objects (BLOBs) in Windows SharePoint Services." 2009. <http://msdn.microsoft.com/en-us/library/bb802976.aspx> (accessed September 25, 2009).

¹³ ESG, “Getting the right data into SharePoint,” April 2009

-
- **Scalability and performance:** As databases hit their object count/document count limits, performance degrades. The larger the database, the slower the application performs.

EMC SourceOne for Microsoft SharePoint helps resolve these issues with no impact to the user experience; it is 100 percent transparent. Active content metadata remains in the SQL Server database. Ultimately, SharePoint still “owns” the content.

How the solution works

To support external data stores, Microsoft released an Application Programming Interface (API) called external BLOB storage or EBS. EBS is a low-level API that intercepts the reads and writes directed at the SQL server and dictates whether the data is stored in the database or is redirected to an external file share. The API was included in the Service Pack 2 release for Microsoft Office SharePoint Server MOSS 2007 and Microsoft Windows SharePoint Server (WSS) 3.0.

EMC SourceOne for Microsoft SharePoint supports the Microsoft EBS API. Besides reducing the data management demands on SharePoint SQL servers, it enables tiered storage management for redirected BLOB content. SourceOne for Microsoft SharePoint can redirect BLOB content to different levels of storage, reducing cost while optimizing SQL Server. The SourceOne for Microsoft SharePoint EBS provider runs below the SharePoint application stack and will not break major Microsoft Office applications, ensuring complete transparency to SharePoint users.

EMC SourceOne for Microsoft SharePoint can also deduplicate content at the storage level. In terms of return on investment (ROI), the combination of tiered storage and deduplication can deliver significant savings. EMC estimates the cost differential between tier one and archive-level storage is in the range of \$50,000 per terabyte per year. Tiered storage also decreases SQL Server backup windows and restore times.

Keep in mind that the Microsoft EBS API operates at the farm level of the SharePoint containment model, which is the information hierarchy that SharePoint follows. The farm is the top level of the model, while everything else — web applications, site collections, sites, lists, and items — live below the farm. Therefore there are two externalization choices: externalize everything or nothing.

SourceOne for Microsoft SharePoint externalization is also a day forward solution from the point when EBS is enabled. But most organizations will have many active SharePoint sites prior to this point. So, to get around this drawback, Microsoft recommends backing up the target farm, enabling EBS, and restoring the backup, which will externalize the content.

Good information governance—archiving information for compliance, eDiscovery, and litigation readiness

While EBS (externalization) is not archiving it does provide significant operational efficiencies. Archiving is applied intentionally and *selectively* to inactive content that needs to be managed with consistent retention and disposition policies.

Archiving inactive content

Archiving does not externalize inactive content from production servers to external storage. It copies or moves the SharePoint content to an archive repository. So while archiving can provide operational value by improving SQL Server performance in the same way externalization does—by removing content from the SQL Server database—its true sweet spot is in information governance and regulatory compliance.

Many global organizations apply stringent information policies to business-critical content assets such as, SOPs, price lists, contracts, NDA submissions, and so forth. With EMC SourceOne for SharePoint, they can consistently apply and enforce those same policies to SharePoint content, without affecting the end-user experience.

How the solution works

From the SourceOne Management Console, an administrator selects a SharePoint archiving activity.

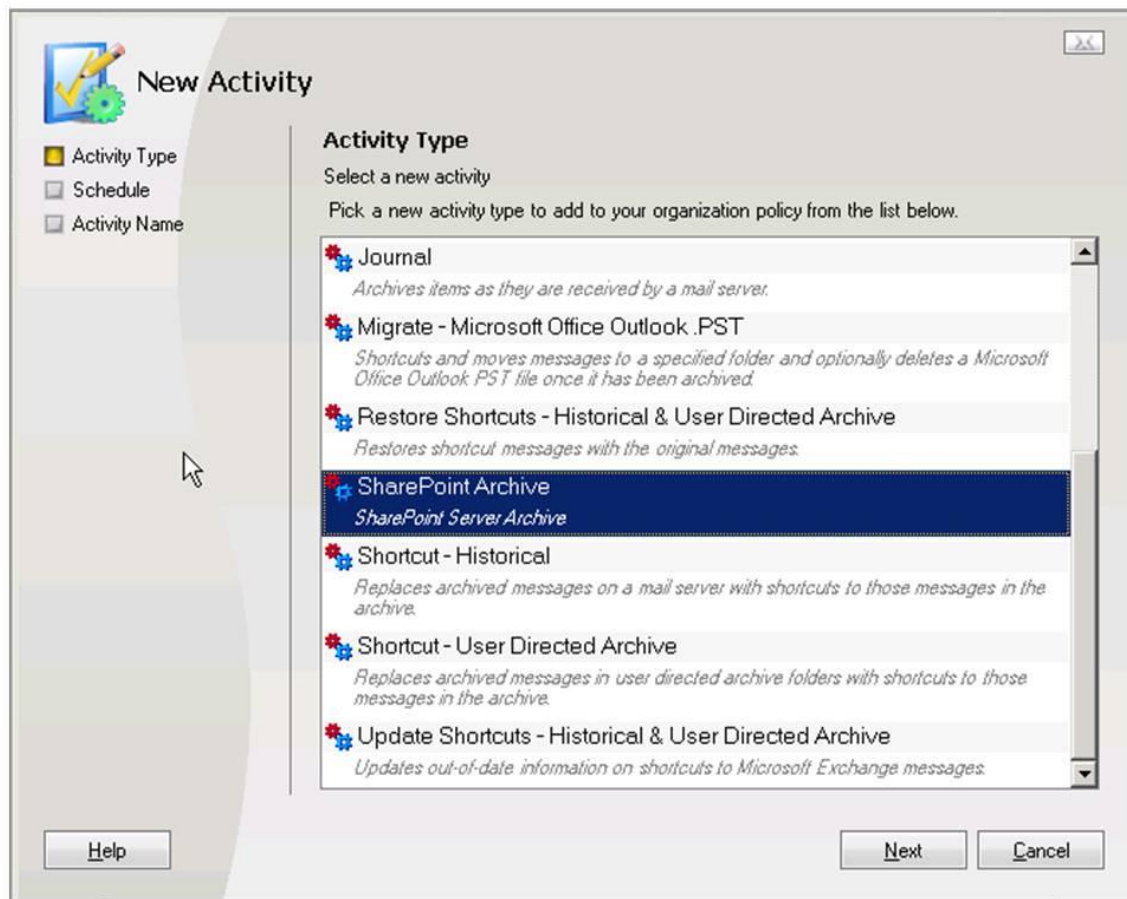


Figure 1. The SourceOne administrator console displays a list of activities, including SharePoint Archive

Next the administrator identifies data sources—from an entire SharePoint farm, site collection, or site to files, discussions, images, and calendar entries at the item level. These sources are called “scopes.” The following sections of the white paper discuss data sources and destinations and content types in more detail.

Identifying data sources

The primary archive data source or scope can be an entire farm or any site or site collection within a farm. Once the scope is selected, the range of content within that scope is defined. For example, the primary scope (parent) could be site collections within a particular farm. That scope might narrow that source to a series of “child” scopes—only one site collection and only specific sites within that collection. EMC SourceOne for Microsoft SharePoint delivers a fine degree of granularity in choosing content to be archived, which ultimately extends all the way to the item level of the SharePoint hierarchy.



Figure 2. Selecting EMC SourceOne for Microsoft SharePoint archive data sources or scopes

Using this level of granularity is optional. All the data that a scope includes (that is, lies beneath it in the hierarchy) will become part of the archive unless certain categories are excluded. In other words, the default range of a scope is everything that that scope contains. Any new content added to the scope is included the next time the content is archived.

Choosing content types

After identifying and defining data sources, content types are selected for the archive. EMC SourceOne for Microsoft SharePoint can ingest all SharePoint content types, making them searchable. The default setting for SourceOne for Microsoft SharePoint includes all content types. Content types can be selected based on criteria such as:

- Last modified date
- Date created
- Created before or after
- Aged older than
- Owner

Once content types have been chosen a series of filters are applied, which can further refine the archive contents. Content can be filtered by:

- Version—choose all versions or the latest version
- Attachments (file types)—choose file types to include and exclude
- Item size—choose items above or below a size threshold

Selecting destinations

Once data sources and content types are chosen, a destination folder in the EMC SourceOne for Microsoft SharePoint archive is selected. Figure 3 shows an archive with three possible destinations.

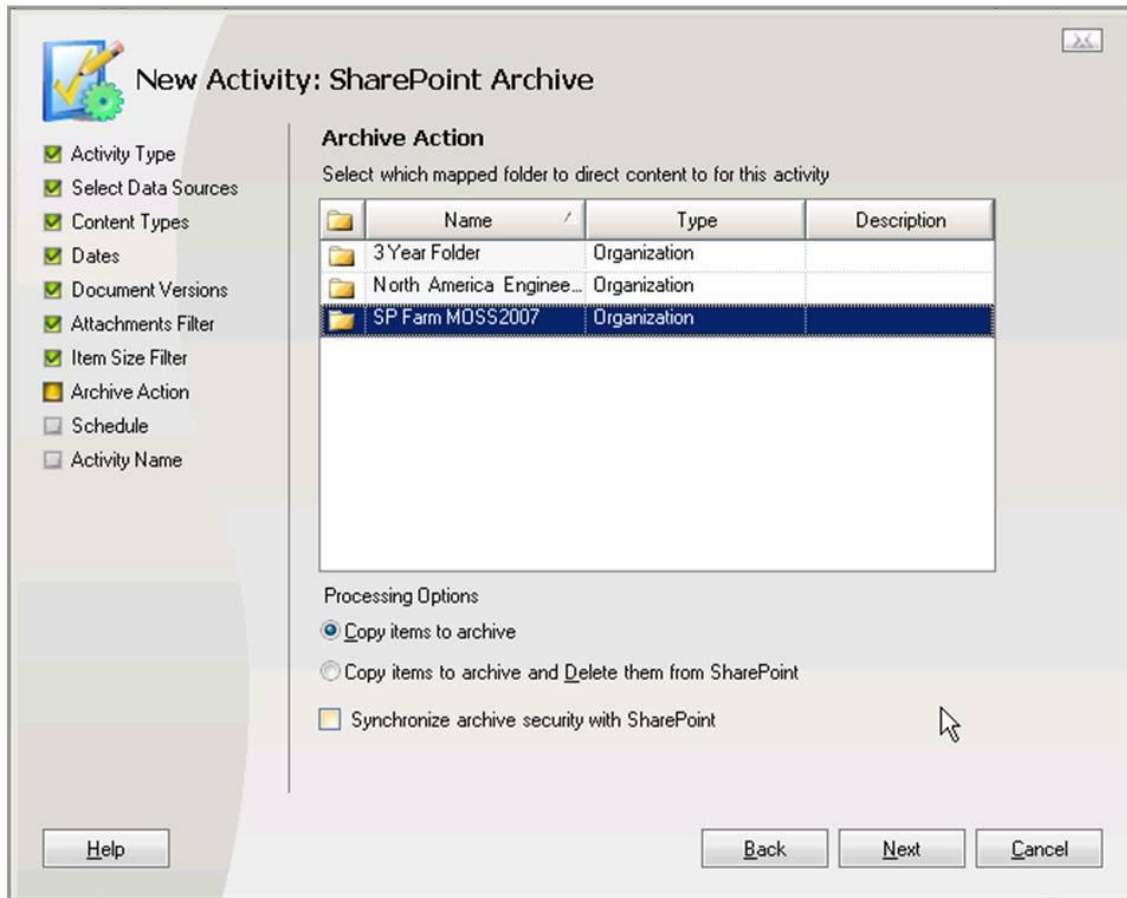


Figure 3. Directing content to a mapped folder

Each folder can have different retention and disposition policies. The 3 Year Folder may be governed by policies that are appropriate for compliance requirements or eDiscovery activities.

This screen also presents different processing options. In this instance, the archive administrator has chosen to copy content to the archive but leave it in SharePoint as well. The copy and delete option enables administrators to use archiving to improve operational efficiency in the production environment.

Security

EMC SourceOne for Microsoft SharePoint supports Microsoft Active Directory for user authentication. Typically, access control to SharePoint content is applied at the site collection level through user groups. Sites inherit their access controls from the parent collection. EMC SourceOne for Microsoft SharePoint stores user groups for authenticating access to SharePoint content. Access control to lists, sites, and collections can use SharePoint groups, Active Directory groups, or individual entries.

Searching the SourceOne archive for SharePoint content

As mentioned previously, EMC SourceOne for Microsoft SharePoint can ingest all SharePoint content types, making them searchable. The EMC SourceOne for Microsoft SharePoint search application sits on top of the EMC SourceOne Search Services platform. The application is a collection of web parts, a site

template, services that run on SharePoint, and an administrative site for configuration that also runs on SharePoint.

EMC SourceOne for Microsoft SharePoint search enables end users to access archive content that no longer resides in SharePoint. It provides a transparent search tool with a nearly identical look and feel to SharePoint's native search environment. EMC SourceOne for Microsoft SharePoint search uses the same Microsoft search metaphors with which SharePoint users are familiar. The same metaphors are also used in EMC SourceOne for Microsoft SharePoint Archive Web Search.

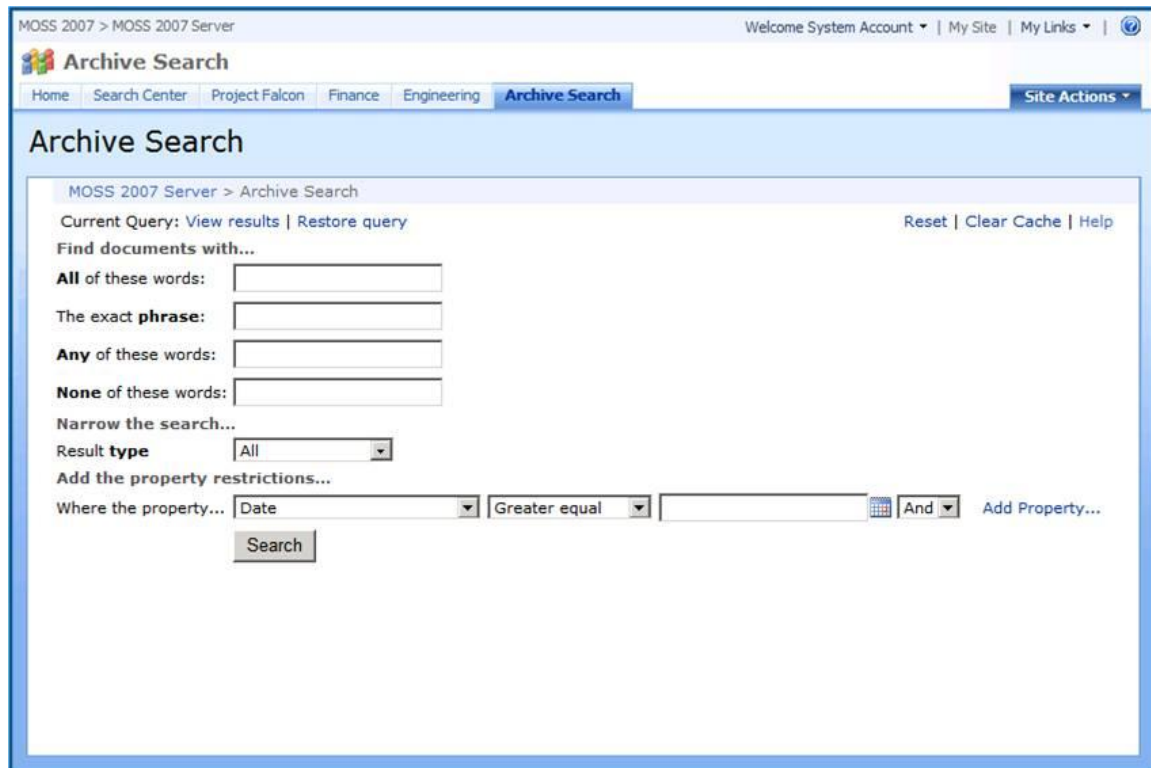


Figure 4. Archive Search added to the SharePoint search model

The list of types exposed as “first class citizens” in SharePoint Archive Search include:

- Document Library
- Contact
- Discussion Board
- Wiki
- Picture Library
- Calendar
- Tasks
- Issue Tracking
- Generic Item

A search results list follows a simple paging model that uses EMC SourceOne Search Service server-side paging.

Archive Search Results

MOSS 2007 Server > Archive Search

[Show Diagnostics](#) | [View Query](#)

Results 1-6 of 6. Your search is **complete** (25.66 Seconds)

ES1 SharePoint Requirements

Version: 1.0 Last Modified By: 8;#Don Mace Last Modified: 2010-02-23T20:29:28-05:00
 Created By: 8;#Don Mace Created: 2010-02-23T20:29:28-05:00
 Check-in Comments:

[Preview](#) · Size: 91KB · Folder: sp farm moss2007

Office UI Guide FAQs

Version: 1.0 Last Modified By: 8;#Don Mace Last Modified: 2010-02-23T20:29:08-05:00
 Created By: 8;#Don Mace Created: 2010-02-23T20:29:08-05:00
 Check-in Comments:

[Preview](#) · Size: 82KB · Folder: sp farm moss2007

ES1 Policy and Rules

Version: 1.0 Last Modified By: 8;#Don Mace Last Modified: 2010-02-23T20:28:48-05:00
 Created By: 8;#Don Mace Created: 2010-02-23T20:28:48-05:00
 Check-in Comments:

[Preview](#) · Size: 1021KB · Folder: sp farm moss2007

EX6 Operations Node

Version: 1.0 Last Modified By: 8;#Don Mace Last Modified: 2010-02-23T20:18:45-05:00
 Created By: 8;#Don Mace Created: 2010-02-23T20:18:45-05:00
 Check-in Comments:

[Preview](#) · Size: 655KB · Folder: sp farm moss2007

Figure 5. A SourceOne for Microsoft SharePoint search results list

Each item listing contains a preview link. Figure 6 shows the preview for an archived contact.

MOSS 2007 Server > Archive Search > Archive Search Results >

Contact

Date	2/23/2010 8:58:34 PM
Source Type	Contact
DateLastModified	2/26/2010 5:45:10 PM
Creation Date	2/23/2010 8:58:34 PM
Last Modified By	1073741823;#System Account
Created By	8;#Don Mace
Version	1.0
Last Name	Mace
First Name	Don
Full Name	Don Mace
Person Title	Architect

Figure 6. An archived contact preview

The top of the results page can also display a summary of the query as shown in Figure 7.

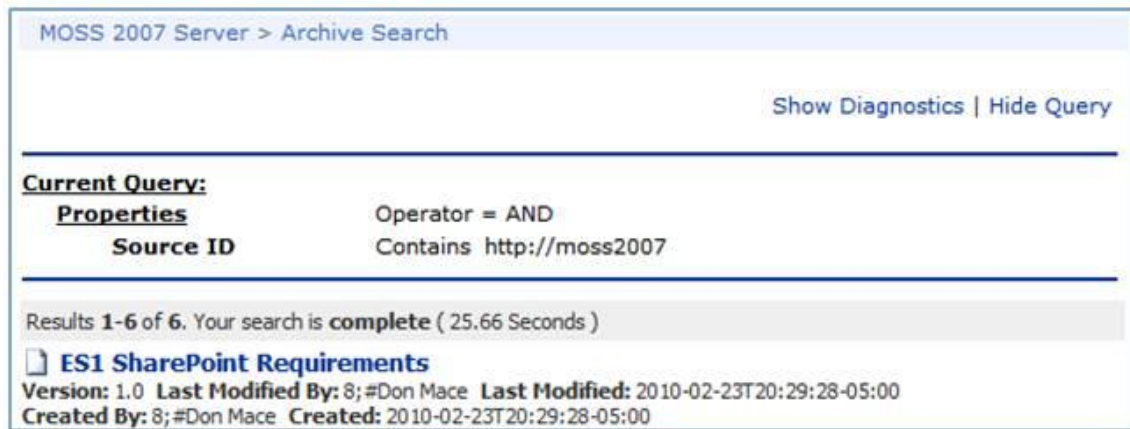


Figure 7. Search results with an optional query summary

Conclusion

EMC SourceOne for Microsoft SharePoint helps IT administrators, and the organizations they serve, cope with the rapid growth of SharePoint content. Through its intelligent, tiered storage management capabilities and support of the Microsoft EBS API, EMC SourceOne for Microsoft SharePoint can improve performance in the production environment and reduce the operational and management costs associated with active content. It can shorten backup windows and protect information through low-cost recovery, restoration, and data protection without disrupting the transparent, single point of access to which SharePoint users are accustomed.

For records managers, compliance officers, and legal staff who are concerned about eDiscovery, litigation preparedness, regulatory compliance, and risk mitigation, EMC SourceOne for Microsoft SharePoint can do much more than that. It can apply full lifecycle management including retention and disposition to inactive SharePoint content, while it resides outside the production environment yet remains easily searchable through the SharePoint user interface.

As the regulatory and legal environment for businesses grows more complex, and the number of Microsoft SharePoint deployments and the information they contain steadily increases, EMC SourceOne for Microsoft SharePoint will become an indispensable tool for managing content across the enterprise information infrastructure.

To learn more about EMC SourceOne for Microsoft SharePoint, please visit EMC online at <http://www.emc.com/products/detail/software2/sourceone-microsoft-sharepoint.htm> or call us at 1.800.607.9546 (outside the U.S.: 1.925.600.5802).