WHITE PAPER

A Fresh Approach Toward Document Management With Open Source

Published: August 2008

rivetlogic
ARTISANS OF OPEN SOURCE

INTRODUCTION

Despite the relatively mature state of enterprise software, companies still find it difficult to easily manage the thousands (and sometimes millions) of documents generated throughout the normal course of business. Because of this difficulty, project teams suffer from miscommunication, lost information and disorganized workflow – all of which results in project delays, wasted time, and redundant work activities, not to mention frustrated employees, partners and customers.

The need for improvement is prevalent across enterprises and government organizations of all sizes, but those feeling the pain most acutely are managing multiple documents, a variety of document types, and project teams and enterprise workflows that operate across geographically dispersed locations.

Teams of all types—ranging from cross-functional product development teams to departmental groups that support repeatable business processes—can benefit from a robust yet simplified approach to document management and team collaboration.

PART ONE

Typical Document Management Solutions

Enterprise activities generally consist of either repeatable business processes—such as accounts receivable, purchasing, and employee reviews—or collaborative project teamwork. And for both cases in most every instance, employees engage in a number of common activities that include:

- Sharing information
- Discussing options and alternatives
- Authoring, reviewing, approving, and sometimes formally publishing documents of a variety of types
- Collaborating from remote locations
- Reusing and repurposing content into different formats for publishing to internal portals, external Web sites, mobile devices, among others

How teams manage and perform this work depends on the company and the size of the team. For enterprises large and small, email and shared network drives remain the quick and dirty—but cheap—approach.

The more formal—and expensive—approaches center around traditional document management or enterprise content management software applications from proprietary vendors.

An emerging third approach—which is proving to be both robust and cost effective—uses a new class of open source technologies such as content repositories, portals and wikis. In many respects, the open source alternatives surpass the traditional approaches not only in cost effectiveness, but also in capability.

The following discussion reviews the advantages and disadvantages of the standard methods for managing documents within enterprises and organizations before discussing the emerging trend of open source solutions.

Email and Shared Network Drives

Many project teams and employees today collaborate using email and shared network drives. Email helps teams in two ways:

- Fast and easy team communication
- Document review and approval

Email enables teams to communicate easily—team members simply set an email group for the project team and filter subsequent emails into project folders. Team members can find, discuss, and store various topics catalogued via email threads and subject lines. User can form subgroups to discuss specialized topics.

Sharing, reviewing and editing documents via email is also easy. A team member attaches a document to an email to those in the group; team members than make and track changes until the document is completed or the boss sends out an "Approved!" email.

The problem with team collaboration via email is that the process rarely works smoothly. Email messages proliferate at alarming rates, resulting in important or time sensitive messages or action items getting lost in clogged inboxes, filed in the wrong folder, or deleted altogether. Retrieving this information is often time consuming and frustrating.

Although revising documents via email is easy, preserving the historical trail of previous versions is not. Document versions reside in various team members' personal email folders, making it impossible to roll back to a previous version. And, team members often lose track of milestone dates, resulting in costly delays.

When combined with email, shared network drives do help improve results. Like email, sharing a network drive is easy; the team administrator selects a folder on a network file system, develops the permissions to provide access to the team members, and starts dropping documents into the shared drive. Desktop integration is automatic—users simply open/edit/save documents from any application.

As with email, shared network drives have their downsides as well. To organize documents, the team administrator creates a folder hierarchy that allows teams members to put documents in their proper locations. However, all team members must conform to the same categorization structure or chaos results (usually the latter).

File systems don't provide document version control, so some teams will resort to file naming conventions (file.v3.2.doc, or finance.v.5.6.1.xls)—an easy fix but cumbersome for larger projects or teams. File naming also relies on team members storing their updated versions on the network drive versus their local file system.

Other teams use a software source code control system to control access and track versions, but these tools are difficult to use and were not designed for document management functions.

Finding documents once they have been filed poses problems too. Desktop and enterprise search tools allow team members to easily search for documents. However, searches are limited to phrases within documents, document types, document titles, and the few metadata fields (such as author, description, and subject) supported by desktop applications. And, if searchers don't carefully craft search phrases, they end up sifting through hundreds of non-relevant search results to find what they need.

And finally, because collaborative teams often include outside partners, customers or subcontractors, project teams must burden their corporate IT departments with the task of providing secure access to the server where the network shares reside. Granting access to only certain portions of the folder hierarchy can be hard to maintain and thus error-prone, increasing the risk of security holes.

Traditional Document Management Software Solutions

The second common approach to document sharing and team collaboration includes a proprietary software solution, such as a document management (DM) system or an enterprise content management (ECM) system.

The traditional DM/ECM approach, however, has its own drawbacks. First, these software tools can be quite expensive. Even the cost of basic document management solutions priced on a per-user basis can escalate quickly. Second, the corporate knowledge base becomes locked into a particular vendor's system—making extraction or information migration difficult, time-consuming and expensive. Third, many of these systems were built over a decade ago, making them difficult to use and integrate with existing software applications and costly to maintain or customize.

The Emerging Trend: Commercial Open Source Applications

Given the limitations of email, network drives and traditional software approaches, enterprises have looked to a new wave of emerging alternatives: commercially supported, open source enterprise software applications.

Open source software has proven itself at both the operating system level (e.g., Linux) and the middleware level (e.g., application servers, email servers, databases). Just recently, viable open source alternatives have appeared at the application level.

The benefits of a commercial open source approach include:

- Zero up-front license fee
- Free community support
- Complete visibility of product capabilities, flaws and roadmap
- Easier customization
- Pay-as-you-go commercial support and training
- No vendor lock-in, with standards-based approaches typically
- High level of innovation

While the benefits of open source development are well documented, there is an important distinction between open source and commercial open source. For an open source technology to be considered commercial open source it must be supported by a financially viable commercial entity. Further, the company should have a track record of successful implementations in mission-critical environments, which can guarantee technical performance and availability, along with global 24x7 support services. In the open source world, this is what distinguishes a commercial open source company from an open source project. Enterprises require commercial open source solutions from companies such as Alfresco, Red Hat and Ingres.

The types of commercially backed open source enterprise applications that support document management include enterprise content management systems and repositories, enterprise portals, and wikis.

Before jumping head first into the open source waters, enterprises must first evaluate the viability of an open source system because only a handful exist that are suitable for enterprise deployment. Teams considering open source tools must also consider the fundamental requirements of enterprise content management in general, and document management in particular.

For example, most open source content management systems primarily support only Web site content, and therefore aren't suited for general-purpose enterprise content such as project or line-of-business documents and files. Standalone portals and wikis provide team spaces that enable collaboration, but generally do not provide full support for the structured workflows and the extensive document management capabilities needed by all except the smallest of organizations.

Additionally, enterprise environments that are global in terms of customers or operations typically require the systems to be available 24x7. Other enterprises have strict compliance requirements for how content or data are managed, from the end user to the data center. Open source technology shouldn't force customers to sacrifice the availability or security of their content management systems,

Part Two of this paper discusses the capabilities one should expect when considering an open source solution. A solution that includes the key functions outlined below can help enterprises improve productivity through better document management and enhanced team collaboration. For those familiar with traditional software solutions, some of these keys will be familiar, while others will be new.

Such is the advanced stage of open source application software. In many respects, open source has equaled and even surpassed the capabilities of proprietary, traditional software alternatives.

PART TWO

Document Management Requirements and How Open Source Can Help

When considering a document management solution, look for the following capabilities and how they are supported with potential open source solutions.

1. A rules-based document repository that replaces the shared network drive

A document repository should be as easy to use as a network share. For example, users should have access to the repository via desktop applications, such as Microsoft Office, and it should support the creation of hierarchical folders, graphical browsing and hidden folders where needed.

More importantly, however, it must overcome the limitations of a network share. In particular, the repository should process documents according to configurable rules developed by the project administrator. Examples of rules that boost productivity include:

- Automatic versioning of documents
- Enforcing the checking in and out of documents to prevent collisions
- Assigning workflow to documents that proceed through well-defined phases
- Automatically extracting metadata from files or scanned documents
- Sending an email alert when certain types of documents are modified
- Automatically transforming documents to other formats, such as converting documents to PDF upon approval or for archiving purposes

In addition, the repository should allow users to create templates that contain a hierarchical folder structure with pre-defined documents, rules, workflows and permission structures for reuse on other projects, or for similar business processes. Repository folder templates enable teams to define and use best practices, and eliminate duplicate efforts.

2. Effective search and retrieval

Project information and documentation is worthless if it can't be accessed when and where it's needed. Some researchers estimate that most employees spend more than 20 percent of their time searching for information, so quick and effective access to the right content is vital to team productivity.

The basis to efficient information access is:

1. **Content modeling**—Applying metadata to documents that accurately characterize important properties based on document type
2. **Search**—Retrieving relevant information quickly and accurately, based on full-text indexing of content as well as file metadata.
3. **Federated and cross-repository search**—Searching

across multiple document repositories, and across other repositories that hold enterprise content such as blogs and wikis.

A basic file search capability—where searchers use phrases within a file's name or content—is necessary but woefully insufficient. Basic metadata (the information about the document such as author, description, etc.) is separate from the file's actual content and can improve information retrieval. Significant improvements can be realized, however, with more elaborate metadata—that is, information that precisely captures document properties and makes them available for search.

Simplified examples of additional metadata that can enhance team productivity include:

- An invoice document with "vendor ID" property, allowing easy retrieval of all invoices from a particular vendor
- An insurance application that has a "currently assigned to" property, enabling quick retrieval of all applications in-process within a particular department, such as the sales department or the underwriting department
- A project specification document with a "project type" attribute, enabling a product manager to instantly access all specifications for a particular product line

When combined with a rules-based repository, metadata may be automatically populated and updated as a document proceeds through a workflow—among a wide variety of other events. And, automatic metadata extraction can dramatically improve productivity by populating metadata fields based on content within the document, with no manual intervention.

3. Seamless repository interfaces based on standards

The document repository should also provide an easy user interface, the most common being that of a network share drive. If the repository looks like a network share, then any application can access and save documents into it. This is easily accomplished via the Common Internet File System (CIFS) protocol, the same protocol Microsoft uses for its shared network drive implementation. If a repository supports the CIFS protocol interface, then using the "open" or "save" button within a desktop application is seamless—and requires zero integration or training.

Another important interface, especially for seamless remote access to the repository, is WebDAV support (Web-based Distributed Authoring and Versioning protocol), which allows reading and writing of documents as if they were on a local drive. WebDAV provides the same functionality as CIFS, but is better suited over wide area networks as it runs on top of the HTTP protocol. In contrast, CIFS is better for access over local area networks.

Other interfaces to look for include a Web browser interface, for both power users and administrators, and standards-based content repository interface support, such as the JSR-170 standard for accessing content repositories in Java. Repository access through the File Transfer Protocol (FTP) is a must for remote access and/or bulk transfers of documents. A seamless integration with portals is important, too, for enabling easy access to personalized content throughout the team or enterprise.

4. Library services—version control and file check-in/check-out

Typically referred to as library services, functions such as automatic versioning and check-in/check-out of documents are fundamental to any type of content management system and mandatory for teams of any reasonable size collaborating on documents. Document versioning should automatically store previous versions while document check-in/out ensures that only one person is modifying a file at any given time—with both functions eliminating the chance of collisions and lost data.

5. Discussion forums

As a complement—indeed, alternative to—email discussions, a forum or bulletin board that consolidates project discussions will dramatically improve productivity, ease search and retrieval of project content/ideas/debates, and provide a single source of project historical information. Open source portals, when properly integrated with open source document management systems, generally provide discussion forum capabilities. And some systems provide the ability to attach discussions to individual documents so that a focused discussion can carry along specific to the content under review.

6. Workflow

Simply put, workflow functionality helps ensure the right person is working on the right thing at the right time. But to be effective, teamwork that follows a particular flow of activities should be as easy as receiving an email request from a co-worker. A good workflow engine will automatically send out individualized notifications or alerts when a document needs review, editing, and/or approval. Workflow definitions should be flexible but easy to configure, and should accommodate a wide variety of business documents and organizational processes.

7. Document imaging

Documents within enterprises generally fall into two categories: 1) documents generated from desktop applications, and 2) physical documents that are scanned as images and then typically converted to text (via optical character recognition). A document repository should support integration with document scanners and scanning platforms. Important capabilities include managing the associations between scanned images and OCR'd text files, automatic metadata extraction (via forms recognition or keyword recognition) to populate the repository accordingly, and support for enterprise workflows associated with scanned documents.

8. Simple records management

While only a fraction of enterprises require sophisticated records management solutions to meet legal and regulatory compliance requirements, almost every company can benefit from basic support for document lifecycle management and scheduling of document events, such as document effective and archival dates, and management of cutoff, hold, and disposition events.

9. Security

A project can be compromised if the wrong people obtain access to sensitive information or, worse, a malicious intruder corrupts or deletes critical files. Security starts with strong authentication of users who have access to the repository, and through the definition and use of individualized user roles within the repository. For user-level security, any enterprise system should integrate with the corporate LDAP or Active Directory infrastructure.

User roles can be specified as collections of permissions that grant specific access rights, such as Write access to content or Read access to folders. Roles are generally defined to permit various activities. For example, a role for a team member who only authors content may be called Author. Other examples of user roles include:

- Reviewer, for those who review and approve files
- Consumer, for those who only need to read certain files
- Editor, for those who only may edit certain files

To accommodate real-world project teams, user roles should vary based on location within the repository. A Contributor in one folder may only be a Consumer of content in another folder. Role definitions should extend beyond folders to individual documents for very fine-grained access control when needed.

Administrative control of the repository should not be confused with the administration/management of access to a document or record. A repository system capable of delineating roles for administration ensures compliance of security best practices commonly found in regulated environments. The operating system access to the content should be protected for those who are authorized including physical files being encrypted on the disk and handled for consumption by users who need to update, review or read it using tools such as Microsoft Word or Open Office.

PART THREE

First Steps in Considering an Open Source Document Management Solution

Rivet Logic offers deep experience with open source document management solutions. Based on our experience, enterprises of all sizes—from small and mid-size businesses to the Fortune 500—benefit most by considering the following open source software technologies:

- Alfresco, an open source ECM platform that provides a strong foundation for document management, among other enterprise content management needs. See www.alfresco.com
- Ingres, an open source database platform that enables a robust, highly available and secure Alfresco repository. See www.ingres.com
- Open source portals, such as Liferay, for team collaboration. Portals provide support for personalization and aggregation of enterprise content and applications. See www.liferay.com
- Open source wikis, from which there are many to choose, for team collaboration. See www.mediawiki.org (PHP-based) and www.jspwiki.org (Java-based) for examples. The Liferay portal also provides a wiki portlet as part of its bundle.
- Open source workflow engines, such as JBoss jBPM and Intalio BPMS. Alfresco natively incorporates the jBPM engine, and third-party integrations with Intalio are underway. See www.jboss.com and www.intalio.com

Unlike most other open source content management systems, Alfresco has been designed from the ground up to address the gamut of enterprise content management, in contrast to many others that focus solely on Web content management. This comes as no surprise considering the Alfresco team includes the co-founder of Documentum (John Newton), as well as developers from Documentum and Interwoven. Alfresco, as a result, provides strong capabilities for document management. Alfresco's out-of-the-box capabilities include a rules-based content repository, available on Ingres technology, with built-in library services and an extensible content model. In addition, Alfresco supports a wide variety of interfaces, including CIFS (for shared drive emulation), FTP, WebDAV, and JSR-170.

Alfresco also provides support for workflow, discussion forums, scanning platform integration, basic records management, and a strong security model. When Alfresco is used with an Ingres database, the content repository will support highly available and security requirements, aspects that are commonly found in enterprise data centers. For search and retrieval, Alfresco provides two fundamental capabilities: 1) Alfresco incorporates the Lucene open source search library and its native support for full-text indexing and 2) provides support for the OpenSearch standard for federated and cross-repository search, including searching in parallel across multiple Alfresco and even non-Alfresco repositories, which may include legacy document repositories as well as existing enterprise wikis and blogs.

Rivet Logic has learned from experience that Alfresco users get maximum benefit when the content model is extended to include metadata and document types that precisely characterize the documents and files that are being managed.

More benefits accrue when metadata is automatically populated based on document content or other contextual information. For example, Rivet Logic developed – for a professional services organization—a custom content model and an automatic metadata insertion capability to support management of a wide variety of project documents. The result has been improved project delivery and higher customer satisfaction. In other cases, repeatable business processes such as vendor timesheet approvals and customer application processing have been streamlined with extended content models and customized functionality.

Open source portals and wikis should be considered when an organization would like to supplement Alfresco's basic team collaboration capabilities, which are document and content centric. Portals, for example, enable project teams to create workspaces for sharing task lists and calendars. In addition, a portal will provide capabilities for personalization and aggregation of other enterprise content and applications. Rivet Logic has integrated the collaboration and personalization features of the Liferay Portal with the strong document management and workflow features of Alfresco for several clients.

Wikis generally provide a very user-friendly mechanism for teams to share information and collaborate on any type of project. They are also useful for developing documentation or a knowledgebase about a project, product, or topic. Wikis may be standalone, or they may be integrated into a portal.

Based on Rivet Logic's experience, an organization will benefit most from an integrated solution. Starting with Alfresco as the content repository, combined with Ingres as a scalable and secure database, a complete solution would integrate a portal, wiki or a custom user interface to fully meet an enterprise's particular needs. In all cases, dramatic productivity gains may be realized when an enterprise leverages all of the fundamental document management capabilities that are now supported by best of breed open source technologies.

Call Rivet Logic to Learn More

For more information about open source document management and team collaboration solutions and how one can benefit your business, call Rivet Logic Corporation at **703.955.3480** or visit **rivetlogic.com**

About Rivet Logic

Rivet Logic is an award-winning consulting and systems integration firm that helps organizations better engage with customers, improve collaboration and streamline business operations. Through a full suite of solutions for content management, collaboration and community, Rivet Logic enables organizations to fully leverage the power of industry-leading open source software. With deep expertise in the Alfresco Enterprise Content Management platform, Liferay Portal, and JBoss Enterprise Middleware, Rivet Logic crafts content-rich solutions that power next-generation Web properties, Enterprise 2.0 applications and collaborative communities. With offices in Reston, Virginia and Cambridge, Massachusetts, Rivet Logic serves clients across a wide range of industries. Rivet Logic—Artisans of Open Source. **Visit rivetlogic.com**