Content Import Integration with

Prepared By
Lee Dallas
Principal Consultant
Armedia, LLC
October, 2008

Abstract

This white paper identifies the high-level patterns for system integration with unstructured content and discusses the use of Armedia Caliente! to solve this pervasive technology need. Content management systems are rarely deployed as closed systems. More often than not, the system is put into place to manage content that is produced by applications and processes outside the control of the platform. Similarly, the repository often is required to supply content to other systems with little or no influence on the downstream applications. Armedia Caliente! offers a low-cost framework to develop and support file system-based content integration for EMC Documentum® and Microsoft® SharePoint.

Introduction

Seldom deployed as closed systems, content management systems most often are used to manage content produced by applications and processes outside the control of the platform. However, this content must be managed by the platform at some point during its lifecycle.

Many organizations require manual intervention due to the complexities of the import requirements. Some try to develop and maintain different interfaces over time. Both tactics are time consuming and costly, as well as a needless repetitive drain on limited IT resources. As companies grow, either organically or by acquisition, there is a need to support an ever-expanding list of content sources. IT departments often will create an import solution for an immediate, specific requirement, and there is seldom time built into a given project to deal with the myriad possibilities that might result in future projects. This leads to different solutions for different projects.

Today's off-the-shelf import products are tailored toward either one-time migration efforts or user-driven import tasks. Both are solutions for specific scenarios in which the functionality fits the requirement but is probably not providing the full solution for various import projects. A more reasonable approach is to establish a protocol and integration/import framework that provides a simple and familiar development process for multi-system integration. The import/integration framework should maintain a high-level *upgrade tolerance* so that changes in either the source or the target can occur independently, so long as content requirements do not change.

Armedia Caliente! is an import framework, providing a flexible environment for the development and unattended batch processing of content and metadata to be loaded into a content repository. This paper will demonstrate how Caliente! can be deployed successfully in common CMS design patterns to provide a low-cost, flexible platform for current and future content integrations.

Design Patterns and Content Management Integration

There are many techniques and software packages that support import and export requirements of content systems. Independent of programming language, however, there are a limited number of forms that the integrations actually take. These forms, or patterns, represent the various roles performed by middleware to integrate a CMS with other applications in the enterprise.

One of the primary reasons for defining patterns is to provide a language to describe components for reuse in different scenarios. Once defined, analysis of an integration problem quickly qualifies which patterns are required. This suggests which previously developed components might be used. In content management integration, there are seven common patterns that are typically combined to

© 2008, Armedia LLC

solve specific integration problems.

Assembler

The assembler pattern is characterized by the need to bring disparate components together in a single document. This pattern could be triggered by a change or the receipt of one or more of the components. The integration engine (Caliente!) would, in some cases, have to retrieve components from elsewhere in the environment to complete the package before sending it to a target system, such as EMC Documentum® or Microsoft® SharePoint.

Cache

The cache pattern of integration is used when it is necessary to provide, for whatever reason, a replication of a content store for access by the target without requiring connectivity or state awareness of the source. A web content management system that pushes files to a web server where they are accessed through a browser is a common cache integration pattern.

Disassembler

Disassembler processes decompose content from one source into a set of components that are then expected to be managed independently by the target. For example, a company may receive a ZIP archive file and then extract the individual content files. This pattern also may include metadata extraction from the content for use in content categorization on the target.

Loader

The loader pattern is the automatic process that transfers content to the control of a target system. The loader performs the application of business rules required by the target with respect to content lifecycle and metadata. Frequently, this pattern is dependent upon preceding patterns that have appropriately prepared the source for ingestion.

Notifier

When content is received or is modified on a source, there are often triggers and notifications that need to be sent through various mechanisms, such as SMTP. The notifier pattern is the set of functions that represent this sort of state change messaging.

Transformer

At times, when content published by one system is to be consumed by another, it is preferable to import in a format or structure different from that of the original. Transformation, in this context, represents both structural changes to conform to a new schema and basic format conversion from editing to view only formats, such as Word to PDF.

Validator

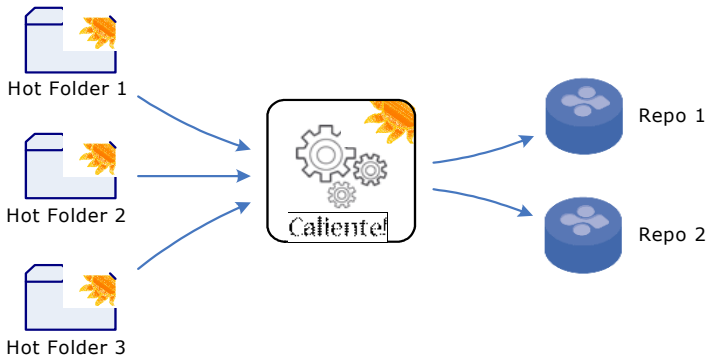
The validator pattern applies when it is necessary to ensure that the content or associated metadata conforms to particular business rules prior to subsequent processing. Absence of required fields, format types, naming conventions, and adherence to a particular schema are all conditions that

© 2008, Armedia LLC

would fall under the domain of the validator functions.

The Caliente! Integration/Import Engine

Content can come from many sources, both internal and external. Every company, however, has some ability to accept, route, and deliver content to a file system. Processes as simple as users dragging content into a shared folder from the desktop demonstrate this ability. This file system, therefore, represents the lowest common denominator for content integration. Caliente! extends the idea and introduces the idea of a *hot folder.*



The hot folder is a particular location monitored for content change or delivery. The framework provides a mechanism to apply tasks against the content. The tasks are chained together to perform more complex operations prior to actually handing the files off to the content management system. In this manner, Caliente! is able to support any of the seven integration patterns previously noted. Caliente! supports loading of content to either EMC Documentum® or Microsoft® SharePoint.

The Caliente! Import Engine will install and run as either a Windows® service or as a daemon process in Linux and UNIX platforms. The engine is configured to monitor directories for changes and fire workflow tasks against files as defined in the configuration. A straightforward scheduling mechanism allows the installer to define the intervals between directory scans and other default behavior of the process.

With Caliente!, it is possible to implement all seven of the common integration patterns with a single framework. By creating tasks specific to the company's needs, a library of reusable components can be maintained enabling rapid response to new projects where transactional and file-based content integration is required. Tasks in Caliente! are written using the commonly accepted scripting tool ANT and easily can be extended using a broad range of techniques. In this way, the framework itself is adapted to the methods and processes accepted within a given company's developer community.

A Business Problem

Businesses are faced with a changing landscape of content sources that need to be integrated by importing content into their common content management platform. For example, a company may have contracted with a firm to scan documents from multiple branches of the company and convert them to PDF. The files need to be stored in a particular location in a repository that corresponds to the path on the media delivered from the vendor.

The Caliente! Solution

The problem represents two of the seven content management integration patterns (page 3). The most obvious is the loader pattern for storing the content. Before loading however, a validator needs to verify that the metadata encoded in the path is supplied correctly. The Caliente! solution involves a simple two-step workflow.

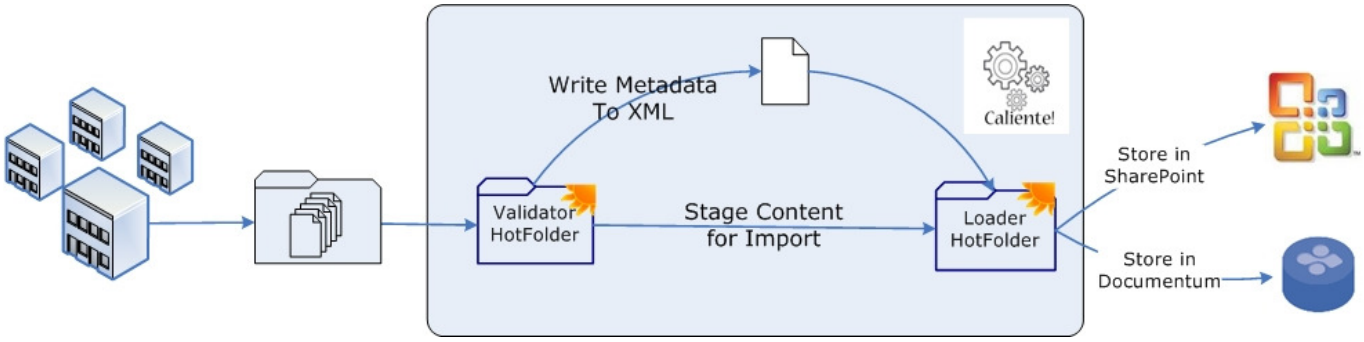
© 2008, Armedia LLC

Step 1: Validation

Once the media is received, the process begins and content is loaded into a hot folder. The engine checks for new files, and, when one is delivered, the task is fired to check whether the path and file name contain correct information. If everything is present, the task generates an XML file containing the metadata and places both the XML and the content in the hot folder for the next task. Error conditions move the file to another folder for investigation later. A repository folder path is derived from the relative folder path in the hot folder and written to the XML metadata file.

Step 2: Loading

The XML file is processed by the loader task, and the folder location is created if it does not already exist. Next, the content is moved into the repository. The loader task also is configured with default values for the task if that information was not supplied.



The Value of Reuse

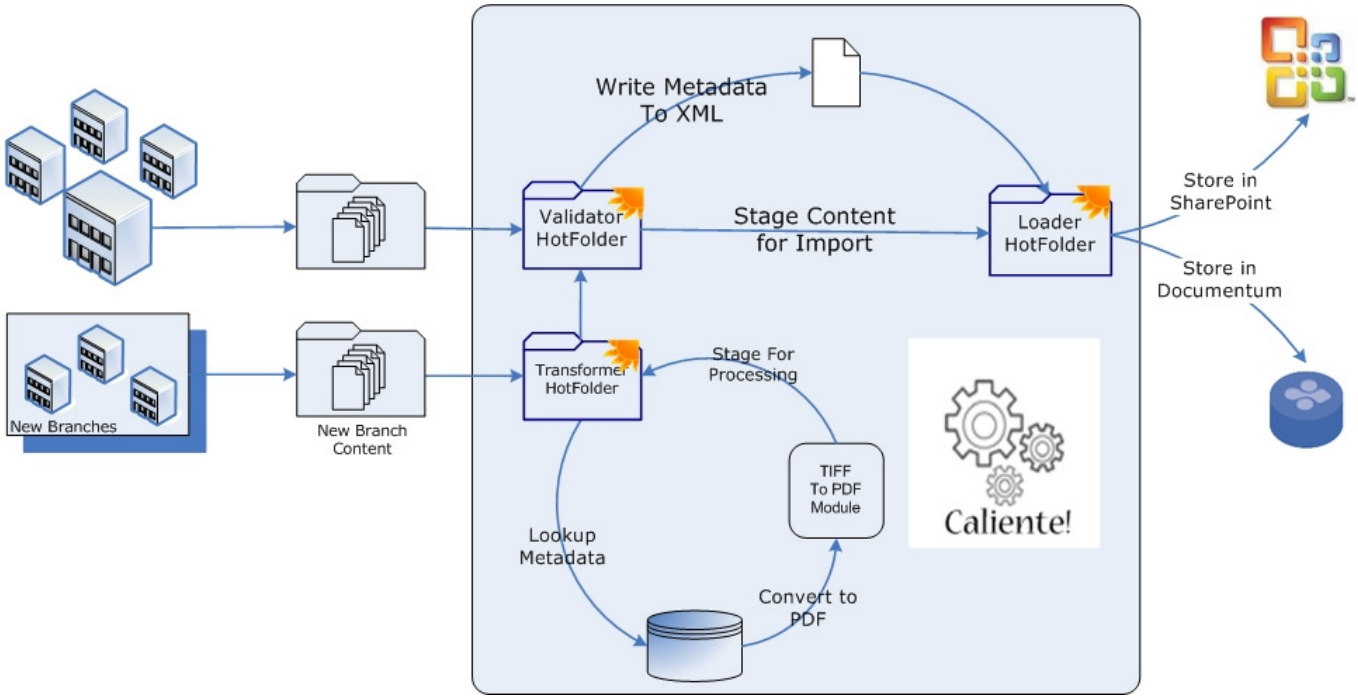
The business value of a platform approach to content management import can be demonstrated by measuring the level of effort required to adjust the implementation to accommodate new needs as they arise.

Changing Business Problems

A very common challenge for businesses today is dealing with mergers and acquisitions. The costs associated with modifying systems to accommodate this type of growth often delays or prevents the company from benefiting from the elimination of duplicate processes.

© 2008, Armedia LLC

For example, suppose that the loading of content into the repository does not need to change, but



the new branches do not support converting the content to PDF before sending it to the home office. It also may be necessary to retrieve metadata for content from a database instead of deriving it from the file path, as is the case in the existing system.

In the new scenario, both the loader and validator steps can remain unchanged. The difference would be that a transformer step would be inserted ahead of the validator to pre-process the new branch data before feeding it into the existing process. Naturally, more complex requirements could suggest changes to the core, but the framework modularizes the tasks so that only those components that need to change are affected.

Conclusion

Integration of multiple content sources with content management repositories is a growing business problem, and it cannot be addressed efficiently by collecting one-off application development projects or depending on user-driven migration tools. Repeatable and configurable middleware that simplifies the exchange of content between systems will reduce total cost of ownership and protect organizations from costly remediation when either the source or target systems have to be upgraded or replaced. Armedia Caliente! offers a low-cost option that balances the supportability of packaged integration with the flexibility of custom development. Because of its flexibility and open architecture, Caliente! can support the most common import scenarios facing companies today.

© 2008, Armedia LLC