

Integrated Content Management

By Dmitri Tcherevik

As companies seek to generate new revenues, control costs, and increase market share, they're increasingly searching for solutions that could be used to integrate disparate content repositories and implement enterprisewide business processes. Existing solutions aimed at aggregating Web pages or integrating applications are ill-equipped for the purpose. Instead, a new type of middleware platform is required.





We broadly define content as that consisting of structured data normally saved in a database, and rich media assets such as documents, video clips, audio clips, images, and other types of content that cannot be straightforwardly mapped to rows and columns.

Content management is becoming increasingly important in the modern enterprise. This is reinforced by three trends that cut across many different industries:

- There's a strong trend toward adoption of paperless processes. Contracts are digitized, documents are posted on Websites, insurance quotes are issued online, high-resolution images are produced by digital cameras and printed on digital presses, and so on. Many companies are discovering that

memory, and hundreds of gigabytes of disk storage. A machine of this size can easily store the entire catalog of print ads and video commercials produced by an advertising agency, for example, and serve this content to tens or even hundreds of users simultaneously.

If history is telling, these trends will only accelerate. The power of computer chips has been doubling every 18 months. The capacity of computer networks has been doubling every six months, or three times faster. Storage capacity has been growing exponentially, too. So, in 18 months, the same amount of money will buy a machine at least twice as powerful as the one it can buy today. And, in 18 months, today's best machine will cost hundreds of dollars less.

"How you gather, manage, and use information will determine whether you win or lose."

Bill Gates, Chief Software Architect, Microsoft

their core assets exist only in digital form and are stored on disk farms and tape libraries instead of warehouses and folder cabinets.

- An Internet presence is now a requirement for every company. Many companies maintain distinct Websites for customers, partners, and employees. In the old days, it was sufficient to just throw a Website together from a collection of static Web pages. Then we saw the emergence of dynamic Websites that were built as spaghetti of Java or Visual Basic code. Today, everyone realizes that content posted on a Website must be managed separately from the infrastructure used to run the site. So we now observe a growing number of ongoing Web content management projects at many different companies.
- Powerful hardware and broadband network connectivity are becoming increasingly affordable. A moderately priced server now ships with a 2+ GHz CPU, a few gigabytes of main

Inevitably, as more people and companies cross the digital divide, the world will see an exponentially growing volume of digital content.

Challenges

Business and technology visionaries recognized these trends years ago and have been working hard to develop new processes and information systems. Their efforts are bearing fruit. Companies in many industries became hugely successful by reengineering themselves around digital content and processes. This success, while helping generate new revenues and control costs, has also produced new challenges.

In larger companies, visionaries and enthusiasts working in smaller groups, departments, and divisions have driven adoption of content management technologies. It's not unusual to see several different content management systems used in a single company. Now, as companies try to establish enterprisewide business processes, this situation pre-

sents several challenges:

- Different content repositories must be integrated and consolidated.
- These repositories must be integrated with traditional business applications such as Customer Relationship Management (CRM), Enterprise Resource Planning (ERP), help desk, and others.

An enterprise portal product can address the problem, and we've seen impressive growth in the number of portal installations. Simultaneously, portal products sometimes breed the same kind of problems they're intended to resolve.

Many vendors have recently entered the portal market. You can now acquire portal products from:

- Pure-play portal vendors such as Plumtree or Epicentric
- Database vendors such as Oracle or Sybase
- Application server vendors such as IBM or BEA
- Operating system vendors such as Microsoft
- Enterprise application vendors such as SAP or PeopleSoft
- Content management vendors such as Vignette or BroadVision.

Often, a portal server is bundled with a different software product such as an application server or an ERP system. So companies sometimes have multiple incompatible portal installations in a single data center.

A complicating factor is that the difference between an enterprise content management system and enterprise portal server is often blurred. Many portals offer extensive content management capabilities. A company attempting to consolidate content repositories by installing a portal server may face additional confusion:

- What does it mean to publish a document through a portal server?
- Is it copied to a portal's content repository or accessed directly?
- If it's copied, how can one ensure that multiple copies of the document are maintained in synchronization?
- Does one copy all documents at once or one by one?
- If a document isn't copied, how can one ensure that only authorized per-

sonnel can view documents in the original content repository?

- How much does it cost to move all content from disparate content repositories to the content repository of the portal server?
- Can the portal server handle the terabytes of digitized documents now stored on my mainframe system?
- Can it handle terabytes of video stored in a tape library?

These questions have no easy answers. Consolidation of content can be an expensive, troublesome proposition. Often, leaving content at its present location is the only option available. Simultaneously, enabling unified access to distributed content repositories and implementing processes based on this content can be a challenging undertaking. Today's portal products, systems designed to handle content stored centrally and aggregate Web pages from different Websites, aren't well-equipped to address this problem.

A new type of middleware, a content integration platform, is required. Unlike middleware developed for enterprise application integration, this platform must:

- Be aware of different types of rich media — such as documents, images, video, audio, XML, and others
- Provide universal access to content saved in databases and content management systems from Documentum, Microsoft, Artesia, Oracle, IBM, Vignette, and others
- Be extensible enough to allow access to content saved in proprietary and mainframe systems.

Before we consider the features and capabilities of the content integration platform, let's consider a few scenarios where this type of middleware can solve real problems.

Advertising

XYZ is a hypothetical advertising company. It distributes a variety of products in digital form. Video commercials, for instance, can be digitized and streamed over the Internet.

XYZ intends to boost productivity by implementing a comprehensive content management solution. The solution will let account managers, brand planners, creative personnel, and production

personnel store, find, distribute, and publish digital assets such as images, documents, audio, and video clips.

XYZ faces a challenge. Many of its account teams already have content management systems from different vendors; the systems aren't integrated. Should XYZ standardize on a single content management system or integrate existing systems as part of a larger content management solution?

Standardizing has obvious benefits but presents challenges. One is the cost of moving content from the existing content management systems to the new enterprisewide system. Metadata formats and media formats used in different systems are dramatically different. Moving content will require many expensive and non-repeatable data and media format transformation processes.

So XYZ decides to use a combination of the two approaches. It will deploy an enterprisewide content management system. All account teams that don't yet have a content management solution will be invited to use this system. Content management systems used by other teams will be integrated with the enterprisewide system to form a comprehensive solution. The integration work will be performed at a fraction of the cost that would otherwise be required to consolidate content in one physical repository.

Music Publishing

ZYX is a hypothetical music label working on the release of a new music album. A music album is a complex product, and many different teams contribute to its creation. Some of the steps include negotiations with the artist, recording the songs, designing the graphics, developing and implementing a marketing plan, manufacturing the compact disc, and managing distribution.

Managing such a complex process is challenging. Album releases are often carefully timed to coincide with events such as holidays or a band's anniversary. Deadlines cannot be moved; millions of dollars may be at stake. Different teams contributing to the release of a music album use different information and content management systems. These systems aren't integrated. Often, they're also not Web-enabled.

ZYX decides to pursue a strategic content and data integration project to deliver accurate, timely information

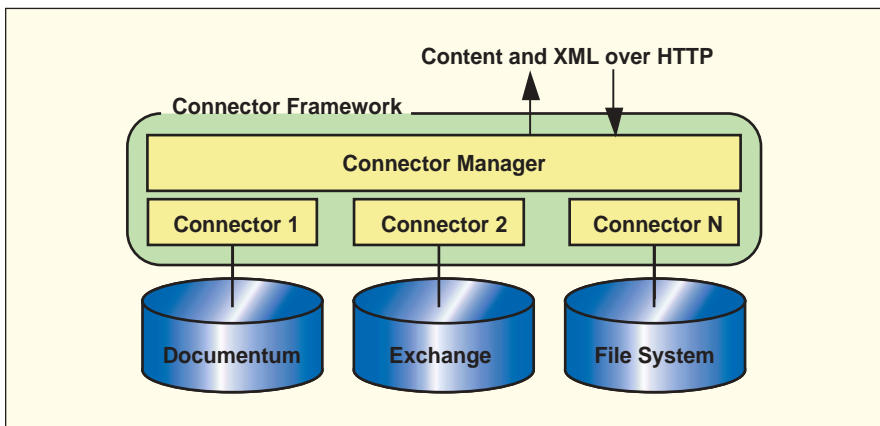


Figure 1 — Content Connector Framework

describing its products at every stage in the release cycle. The resulting solution will help ZYX executives make timely, informed decisions.

Solution Architecture

Now let's consider components of a platform that facilitate integrated content management:

- Connector framework
- Content directory
- Indexing and search engine
- Workflow management system
- Security and single sign-on
- Portal integration toolkit.

A content integration platform may also include a full-fledged content repository with extensive media and library services. Here, however, our focus is on the integration aspects of content management.

Connector Framework

Content can be stored in a:

- Document management system
- Digital asset management system
- Collaboration and messaging system
- Database management system
- Plain old file system.

Metadata describing this content can be embedded in digital assets, stored in a database, or saved in a file system directory as a collection of XML documents or Microsoft Excel spreadsheets. The purpose of the connector framework is to erase differences that exist between different methods of content and metadata management. This is achieved by developing a set of connectors.

Every connector performs three functions:

- Data mapping
- Object model mapping
- Application Program Interface (API) mapping.

Data mapping is needed to convert primitive data types, such as dates and currencies, to the format the integration platform uses. Object-model mapping is required to map between elements of the metadata schema. A video clip, for instance, can be called a "video clip" in one system and an "advertising spot" in another system. Connectors used to access these systems must ensure that a consistent set of terms is used across the entire integrated solution. Finally, the API mapping performed by connectors ensures that all systems can be accessed in a uniform manner from user inter-

faces, portals, and programming tools.

Besides the three types of mapping, the connectors are also responsible for propagating events related to metadata and content. For instance, when a piece of content is updated in a content management system, its connector must capture the corresponding event and hand it over to other elements of the integration platform for processing.

The connector framework assigns a universally unique identifier to each piece of content that can be accessed through its connectors. This identifier is identical to a Universal Resource Locator (URL) and can be used to access content from anywhere on the network. When presented with a URL, the connector framework automatically loads an appropriate connector and retrieves content from one of the repositories under its control. This is shown in Figure 1.

The content integration platform includes connectors for several popular content management systems. It also offers a toolkit that can be used to develop new connectors for proprietary or legacy systems.

Content Directory

A content directory is used to categorize, organize, and describe the vast volume of content that can be accessed through connectors of the integration platform. It works as shown in Figure 2.

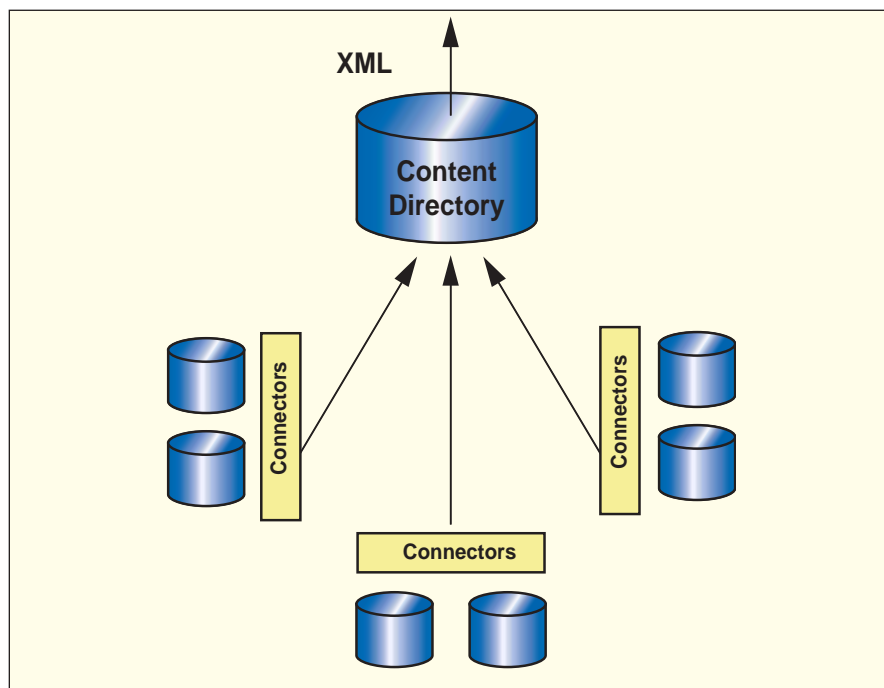


Figure 2 — Distributed Content Directory

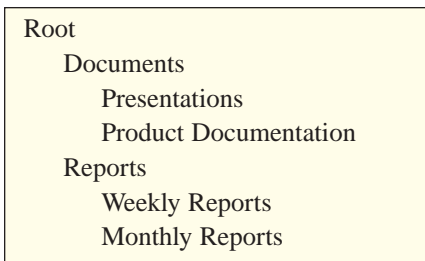


Figure 3 — Sample Content Directory

The content directory is a hierarchical structure that consists of links and folders. Links point to digital assets or collections of assets stored in various content repositories. Folders are used to organize links and folders. Figure 3 shows a sample content directory.

A link can appear in multiple folders. This allows creation of several orthogonal category trees mapping to a single body of content. Documents, for example, can be categorized by type, date, or products they describe. Different types of queries are best answered with different types of categories.

The content directory can manage metadata describing content. Every link or folder saved in the directory can be associated with a collection of properties. Users are free to define new properties and templates describing collections of properties.

Imagine, for instance, that you'd like to assign a product code to digital assets related to a set of products. You could modify the metadata schema in every content repository that's used to manage the assets. Or, you could define a property called "Product Code" in one place, the content directory, and then associate this property with links that point to assets associated with the products. Once a property is defined in the directory, it can be used to define queries that span multiple content repositories. For example, one could ask the directory to retrieve links to all digital assets related to a particular product.

Besides managing data describing content, the directory can maintain relationships among different digital assets. A word document saved in one system, for example, can contain a script for a video file saved in a different system. The central content directory can be used to establish a link between these two pieces of content. A directory user can jump directly from a video clip to the document containing its script.

It's important that directory contents

be synchronized with content of the content repositories. When a piece of content is removed from a repository, we'd like a link to this piece of content to be automatically removed from the directory. The content integration platform achieves this with help from the distributed event propagation mechanism. When the content directory receives an event indicating a piece of content was created, modified, or destroyed, it adjusts the related links and folders accordingly.

The content discovery mechanism of the integration platform can be used to populate the content directory in batch mode. When a new content repository is connected to the platform, it's scanned and links to its content are added to the directory. It's possible to tune the granularity of content discovery. The directory may point to individual digital assets or to collections containing hundreds and even thousands of assets.

Distributed Search and Indexing

The content directory allows content browsing based on metadata saved centrally in the directory. Often, however, content must be retrieved based on information saved in individual content repositories. It may be required, for example, to do a keyword-based search for all images that contain a certain object. The list of keywords associated with an image is typically stored with that image in its home repository. We can only leverage this information by

sending a search request directly to that repository.

Different content repositories have different search capabilities. On one end of the spectrum, we have a full-fledged enterprise content management system with extensive content and metadata-based search capabilities. On the other end, there's a file system directory that can be used only to look up an asset by the name of its file.

The connector framework of the integration platform does a good job at masking these differences. For every repository connected to the framework, it exposes a full set of search methods. Some of these methods are forwarded to the repository. Some of them are implemented in the connector that's used to access the repository. The file system connector, for example, can filter assets based on underlying file system metadata. It also works in tandem with the integration platform search engine to implement content-based search over documents saved in file system directories.

The integration platform search engine maintains a content-based search index over assets saved in repositories that don't offer content-based search functionality. This index is updated automatically with help from the event propagation mechanism. When an asset is created, modified, or destroyed, the index is updated accordingly. The index can also be populated in batch mode as part of the content discovery process the system performs.

To perform a content- or metadata-based search over information saved in

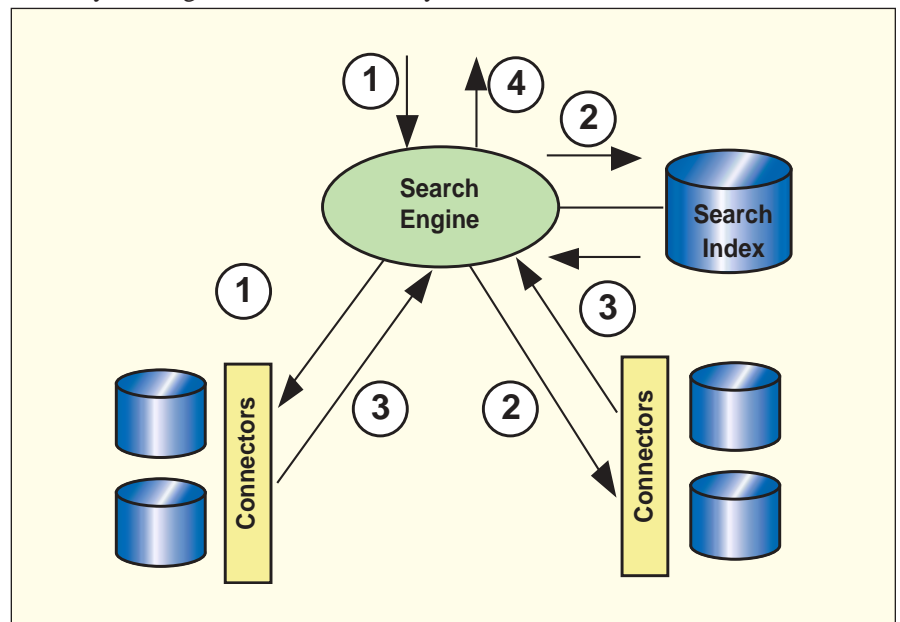


Figure 4 — Distributed Content Search

distributed content repositories, the user sends a request to the integration platform search engine. The engine maps this request to multiple search requests forwarded to individual content repositories. If necessary, it also consults its local search index. Information retrieved from the index is collated with information retrieved from various content repositories. This is shown in Figure 4.

Multiple instances of steps 2 and 3 in Figure 4 occur in parallel. The time required for a distributed search across multiple repositories is comparable to the time required for the same search in a single repository.

Workflow

Management of processes used to create, capture, transform, review, and publish digital assets is central to content management. So virtually every content management system includes a workflow engine. Unfortunately, workflow engines shipped with different content management systems are tightly coupled with these systems and cannot be used to orchestrate processes involving content from different content repositories.

We address this problem by providing a generic workflow engine as part of the content integration platform. The engine leverages the platform's event propagation mechanism. When a digital asset is created, modified, or deleted in any content repositories connected to the platform, a workflow process can start automatically.

This functionality is useful to auto-

mate non-trivial content management processes. For example, a workflow process could be used to ensure that when a video clip is submitted to one content repository, it's automatically converted to a lower resolution format and copied to a different repository or posted on a Website (see Figure 5).

The content integration platform workflow engine is integrated with e-mail and instant messaging systems. When a work item becomes available, workflow participants responsible for this type of work can be notified through multiple communication channels.

The workflow engine offers extensive reporting capabilities managers can use to review the list of active workflow processes, the state of each process, the list of work items assigned to participants, the list of unassigned work, and more. Based on this information, a manager can make timely decisions.

Finally, with help from the connector framework and some scripting, the generic workflow engine can be integrated with workflow engines of the various content repositories connected to the platform. Such integration may be

required, for instance, when an enterprise-wide workflow process is built as a combination of workflow processes defined in individual departments.

Security and Single Sign-On

In today's world of Internet trading, the importance of protecting digital assets from theft and unauthorized access cannot be overestimated. Today, a music band whose latest album is stolen and posted on the Internet may lose millions in uncollected CD sales. Movie studios could face similar risks.

Content management systems from different vendors differ dramatically in their security capabilities. There are systems that store unencrypted digital assets in file system directories and let anyone with access to these directories download and copy the assets. There are also systems that use sophisticated digital rights management mechanisms to prevent unauthorized access to digital assets.

The content integration platform is designed to preserve security in systems where it's sufficiently strong and to enhance security in systems where it's below the desired level.

To see how the integration platform can enhance security, consider a content repository that's implemented as a file system directory. Typically, assets posted in this directory are shared by turning the directory into a shared disk drive. Anyone who can access this drive can download and copy the assets.

This presents a serious security risk. Once the content integration platform is deployed, however, all the assets saved in the file system directory will be accessed only through the platform's connector framework. The connector framework will act as a "firewall" that prevents people without sufficient privileges from accessing the assets.

Single sign-on is also addressed by the content integration platform. Consider a case when the platform is used to access content saved in three different repositories. Every repository may implement a

Content management systems differ dramatically in their security capabilities.

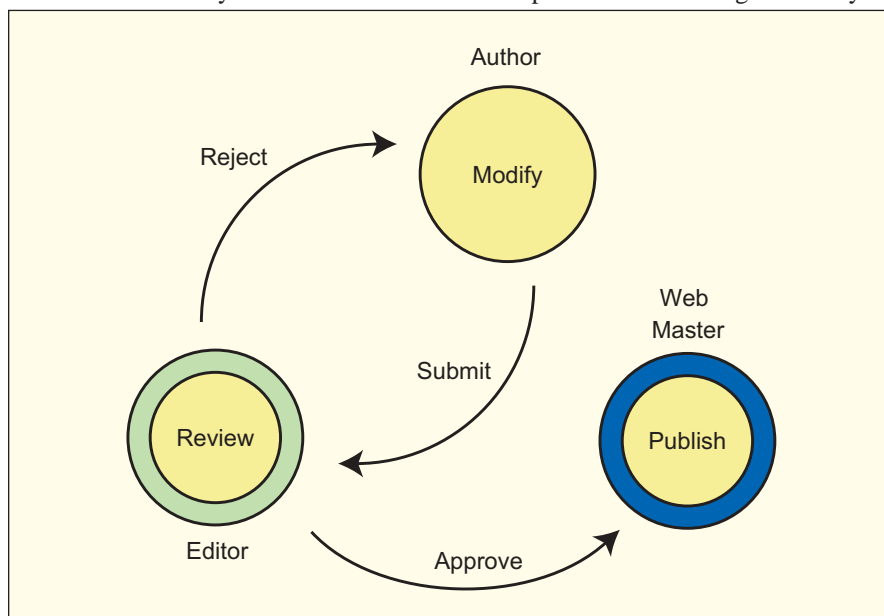


Figure 5 — Sample Workflow Process

different way of authenticating users and protecting assets. Without single sign-on, a user who wishes to view assets in all these repositories will be burdened with using three different user names and passwords.

The content integration platform single sign-on mechanism works as shown in Figure 6. The six steps involved are:

- The user provides a set of credentials to the integration platform security service.
- The credentials are verified against information saved in a user directory; the user receives a security token.
- This token is used to access assets in a content repository via an instance of the connector framework.
- The connector framework asks the security service to match the user's security token to credentials that can be used to access the desired content repository.
- Credentials provided by the security service are used to automatically log the user into the content repository.
- Assets retrieved from the content repository are sent to the user.

This process occurs automatically whenever a user tries to access a content repository. So a single set of credentials can be used to gain access to all content repositories connected to the integration platform.

Portal Integration Toolkit

The content integration platform offers tools and a Web-based user interface designed for:

- Solution developers
- Content contributors
- System administrators.

Solution developers implement a content management solution by integrating existing databases, applications, and content repositories. Content contributors create, review, annotate, transform, and share content. System administrators maintain the content management solution by managing users, backing up data, tuning configuration parameters, and performing other tasks.

The system's Web-based interface can be installed as a self-sufficient, stand-alone application or as a collection of semidependent portlets in a portal server. In the latter case, the portal

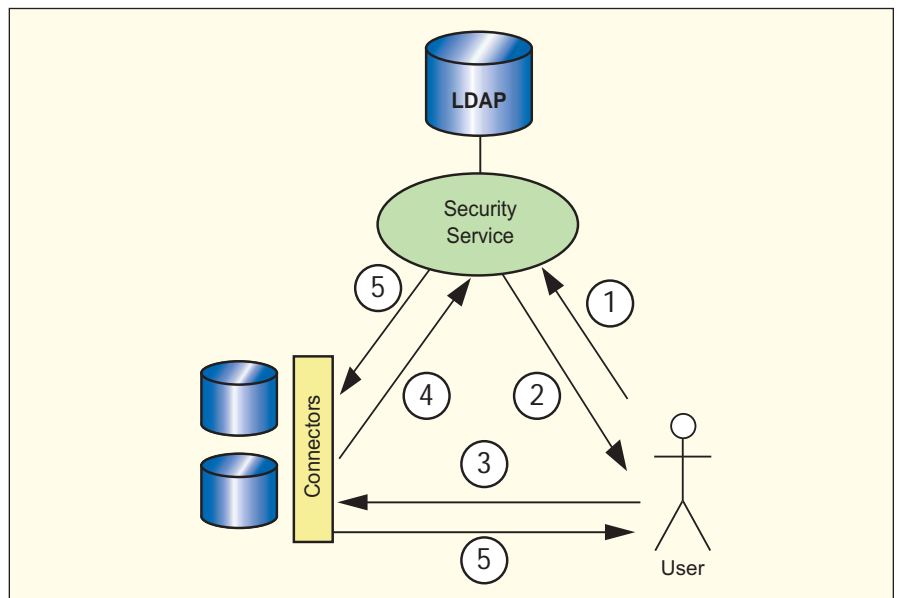


Figure 6 — Single Sign-On Mechanism

server personalization engine can be used to arrange portlets on the screen in a way that's convenient for each user.

The content integration platform also offers programmatic interfaces that can be used to access services of the platform from third-party applications. These interfaces support standard Web services protocols and can be accessed locally or remotely from applications written in Java, C#, Visual Basic, and other languages.

Summary


Companies in different industries are now completing multiple content management, digital asset management, and portal projects. Often, these are departmental projects. Now, as companies seek to control costs and increase revenues, they're increasingly searching for a solution that could be used to integrate disparate departmental systems and implement enterprisewide processes.

We've presented a content integration platform that can be used to build an integrated enterprisewide content management solution based on content saved in multiple databases, portals, applications, and content management systems. The platform consists of a:

- Connector framework to connect to proprietary systems from various vendors
- Content directory to organize, categorize, and describe content saved in multiple content repositories
- Search and indexing engine for sup-

porting content- and metadata-based searches across multiple content repositories

- Workflow management system to formalize, enforce, and automate both business and content publishing processes
- Portal integration toolkit that can be used as a stand-alone application or deployed in a portal server.

Besides integrating disparate content management systems, the platform can be used to fill gaps in their functionality such as content categorization, content-based search, metadata-based search, or workflow. The platform can also be used to Web-enable an existing content management solution and integrate it with a portal server. 

About the Author



Dmitri Tcherevik is a divisional vice president of research and development and technology strategist in the office of the chief technology officer at

Computer Associates. After extensive development experience in Ingres and the Jasmine object database, he led the development of CleverPath Enterprise Content Manager and Advantage Integration Server. e-Mail: Dmitri.Tcherevik@ca.com; Web-site: www.computerassociates.com.