Fedora Open Source Repository Software:

White Paper

Fedora Development Team

October 28, 2005

The Problem: Managing Rich Content

Institutions and organizations face increasing demands to deliver rich digital content. A scan of the web reveals complex multi-media content that combines text, images, audio, and video. Much of this content is produced dynamically through the use of servlet technology and distributed web services.

Delivery of rich content is possible through a variety of technologies. But, delivery is only one aspect of a suite of content management tasks. Content needs to be created, ingested, and stored. It needs to be aggregated and organized in collections. It must be described with metadata. It must be available for reuse and refactoring. And, finally, it must be preserved.

Without some form of standardization, the costs of such management tasks become prohibitive. Content managers find themselves jury-rigging tasks onto each new form of digital content. In the end, they are faced with a maze of specialized tools, repositories, formats, and services that must be upgraded and integrated over time.

Content managers need a flexible content repository system that allows them to uniformly store, manage, and deliver all their existing content and that will accommodate new forms that will inevitably arise in the future.

Flexible Extensible Digital Object Repository Architecture

Fedora is an open source digital repository system that meets these challenges. It does this by combining a number of key features including:

- ***Powerful digital object model***: The digital objects, or units of information, in Fedora may combine any number and any variety of data streams. These data streams can be local to the repository or may reference content anywhere on the web. For example, one digital object may aggregate a scholarly document in multiple text formats, and another may combine the text, images, and video that are the basis of a rich web page.
- ***Extensible metadata management****:* Because metadata and data are treated uniformly in the digital object model, any number and variety of metadata formats may be stored as data streams, alongside content, in a digital object.
- ***Expressive inter-object relationships****:* Digital objects contain metadata that can express any type of relationships such as membership in collections, structural associations like articles in journals or pictures in albums, or taxonomic

relationships. Relationship metadata is indexed and can be searched using semantic web query languages.

- ***Web service integration****:* Fedora fits in with n-tier applications because of two types of web service integration:
    - *Dynamic content delivery:* Web services can be associated with any of the data streams in a digital object. As a result, a digital object can deliver dynamic content: the output of a web service processing data in the digital object. For example, an image processing service can be associated with a digital object that contains an archival tiff image, making it possible to convert the image to other formats (jpeg, gif) or resize and scale the image. Or, a metadata crosswalk service can be associated with a digital object that contains MODS metadata, making it possible to deliver other metadata formats such as Dublin Core.
    - *Management and Access APIs:* A Fedora repository runs as a service within a web server. All of its functionality and all features of its digital object model are accessible through well-defined REST and SOAP interfaces. Thus, a Fedora repository can be easily integrated into a variety of application environments with different user interfaces.
- ***Version management:*** Fedora stores a history of all modifications to digital objects. The full history is accessible through the Fedora access API.
- ***Configurable security architecture:*** Access to all aspects of the Fedora management and access API can be controlled by fine-grained XML-based access-control policies. These policies define sets of rules to permit or deny access by users and groups to repository operations.
- ***OAI-PMH conformance:*** Fedora repositories are fully conformant with the interoperability framework defined by the Open Archives Initiative Protocol for Metadata Harvesting. The Fedora OAI-PMH service exploits Fedora's extensible metadata management, supporting harvest of any form of metadata delivered by digital objects.
- ***Preservation worthy*****:** Fedora repositories incorporate a number of features that facilitate the complex tasks associated with digital preservation. Internally all Fedora digital objects are represented in the file system as files in an open XML format. These XML files include data and metadata for the objects plus relationships to services and other objects. The entire structure of a Fedora repository can be rebuilt from the information in these files. In addition, Fedora repositories are compliant with the Reference Model for an Open Archival Information System (OAIS) due to their ability to ingest and disseminate Submission Information Packages (SIPS) and Dissemination Information Packages (DIPS) in standard container formats such as METS and MPEG-DIDL.

Open source for multiple applications

Fedora is open-source software covered under the Educational Community License 1.0 (ECL). This license makes the software openly usable by anyone for any purpose. The only requirement imposed is that anyone using the source code, making changes and distributing the changed source, is required to make it clear that the new distribution is different and to describe the changes made.

This licensing arrangement and the flexibility of the software and its interfaces have led to the adoption of Fedora by a wide variety of institutions and organizations. Some examples of the use of Fedora are:

- ***Digital preservation:*** Rutgers University (New Jersey, USA) is using Fedora as the foundation for their digital preservation platform.
- ***Institutional repositories:*** The ARROW Project (Australia) and DEFF Project are two of a number of projects building large scale deployments of Fedora for storage and dissemination of scholarly literature.
- ***Commercial content systems:*** VTLS Inc. (Virginia, USA) has built two products. VITAL is a repository solution for universities, libraries, museums, archives and information centers for handling large text and rich content collections. VALET is a customizable, web-based interface that allows remote users to submit Electronic Theses & Dissertations into a Fedora digital object repository.
- ***Electronic records***: Yale University (USA) is developing a system using Fedora for storage and preservation of electronic university records.
- ***Multimedia web sites***: The Electronic Encyclopedia of Chicago (http://www.encyclopedia.chicagohistory.org/), developed by the Chicago Historical Society and Northwestern University (USA), uses Fedora to deliver a hyperlinked resource of maps, images, and text.
- ***Digital library collections:*** A number of university research libraries including the University of Virginia (USA), Tufts University (USA), the National Library of Wales (UK), and the University of Athens (Greece) are using Fedora to host and deliver digital collections.
- ***Distributed digital libraries*****:** The National Science Digital Library (USA) is a National Science Foundation project that is using Fedora to store an information network overlay that represents content, metadata, agents, services, and annotations for its multi-portal digital library.

A look under the hood

From the perspective of client access, digital objects in a Fedora repository have the same characteristics as any web-based content. They are accessible via standard URLs that return mime-typed streams. These URLs can be issued from any standard web browser and the results are viewable in the same browser (or, alternatively from any client that can issue REST URLs or SOAP calls).

This is illustrated in Figure 1, with a digital object identified as demo:10. This digital object is accessible in four representations: Dublin Core metadata, a thumbnail image, a color image, and a grayscale image. Each representation is available via a URL that can be issued from a browser.

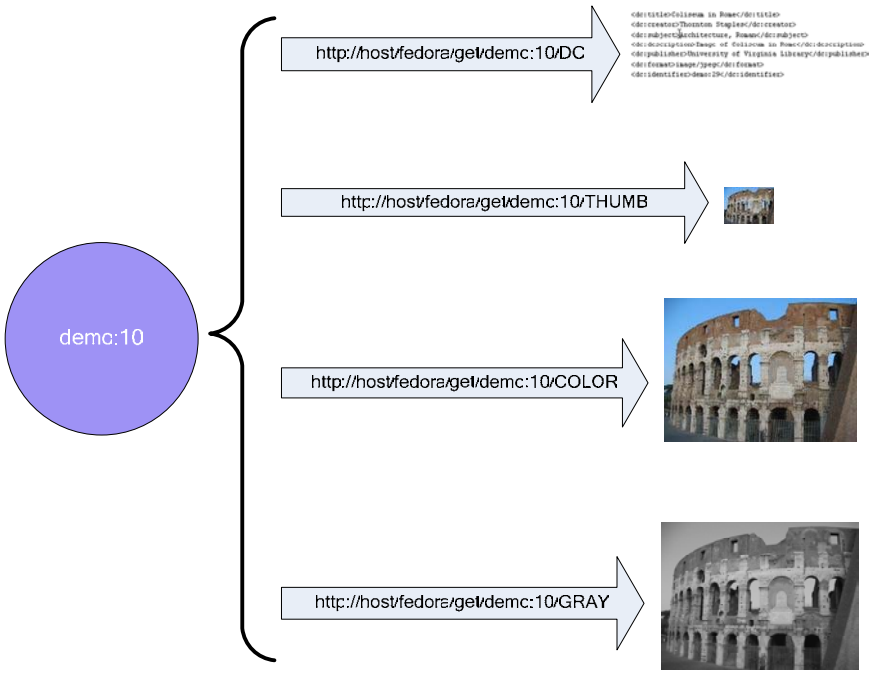

Figure 1 - Accessing digital object representations via URLs

Digital object internals

Uniform access via URLs to digital object representations hides the underlying structure of digital objects. In its simplest form a digital object is an aggregation of content items, where each content item maps to a representation. The Fedora object model defines a component known as a datastream to represent a content item. A datastream

component either encapsulates bytestream content internally or references it externally. In either case that content may be in any media type.

Figure 2 expands on the previous picture, showing that each representation available via a URL is a direct transcription of a datastream stored within the digital object. As noted above, each of these datastreams could also be an external reference to web content. In that case the digital object serves as a local *surrogate* for remote content.

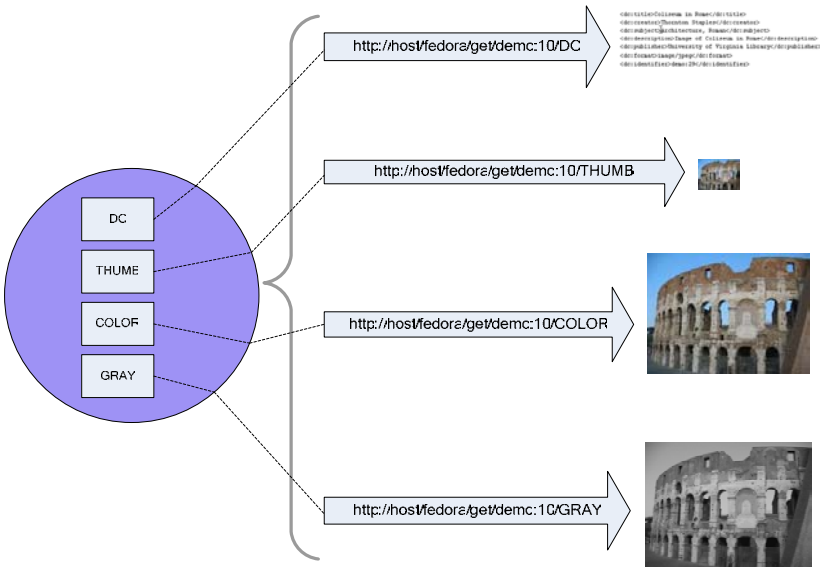

Figure 2 – Representations mapped to datastreams

In addition to theses representations, which are direct transcriptions of datastreams, the Fedora object model enables the definition of *virtual representations* of a digital object. A virtual representation, also known as a *dissemination*, is a view of an object that is produced by a service operation (i.e., a method invocation) that can take as input one or more of the datastreams of the respective digital object. As such, it is a means to deliver dynamic or computed content from a Fedora object.

Figure 3 shows a digital object with the same client-visible representations as the previous figures. But in this case, only the Dublin Core representation is stored as a datastream. The THUMB, JPEG, and GIF representations are produced via a *disseminator* that takes the stored TIFF as input and sends it to an external web service. As before, the TIFF datastream could be an external reference, in which case the digital object provides a surrogate for web service orchestration for external content.

Note that from a client perspective the structure of an object is invisible. Only the access URLs are exposed. This information hiding permits allows changes to underlying

content implementation and design, without affecting client and user access to the content.

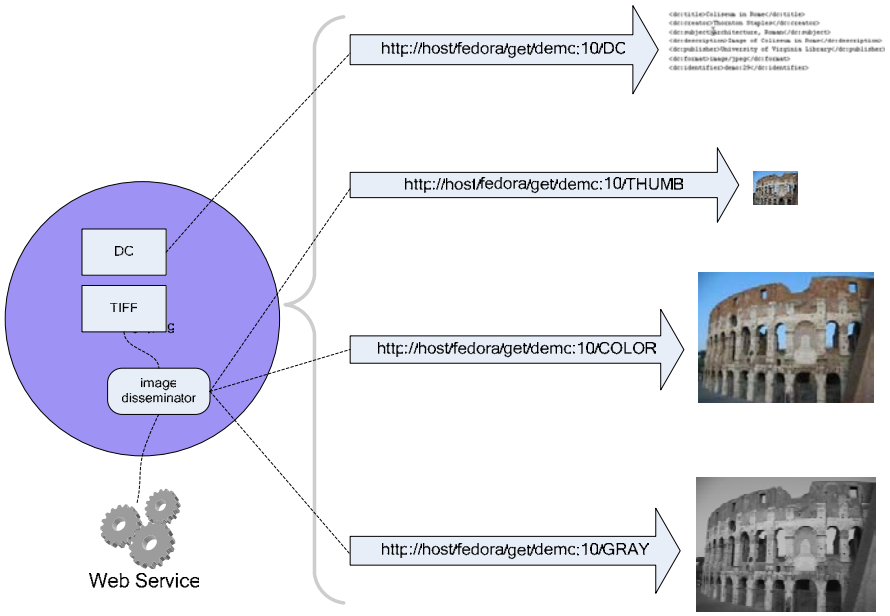

Figure 3 – Disseminations via service operations

Relationships among objects

As stated earlier, each Fedora digital object has a slot for metadata that expressing the relationship of the object to other digital objects, and in fact to information entities outside of Fedora (for example, external web pages). This feature makes use of state-of-the-art semantic web technologies. Relationships are stored within a special datastream in a digital object as statements encoded in the Resource Description Format (RDF) XML syntax. These relationships may be derived from any ontology, including a basic relationship ontology supplied with Fedora. The Fedora system automatically indexes the relationship metadata from all digital objects in a special database. This database can be queried using a query language specialized for extracting information from a relationship graph. This query interface is exposed as a service in the Fedora API, and in fact can be used for dynamic disseminations like any web service.

Figure 4 shows an example of the use of relationships. The figure shows four digital objects, two of which have relationship metadata. demo:10 is a collection with two members, demo:11 and demo:12. demo:10 ; and demo:13 is an annotation for demo:11.

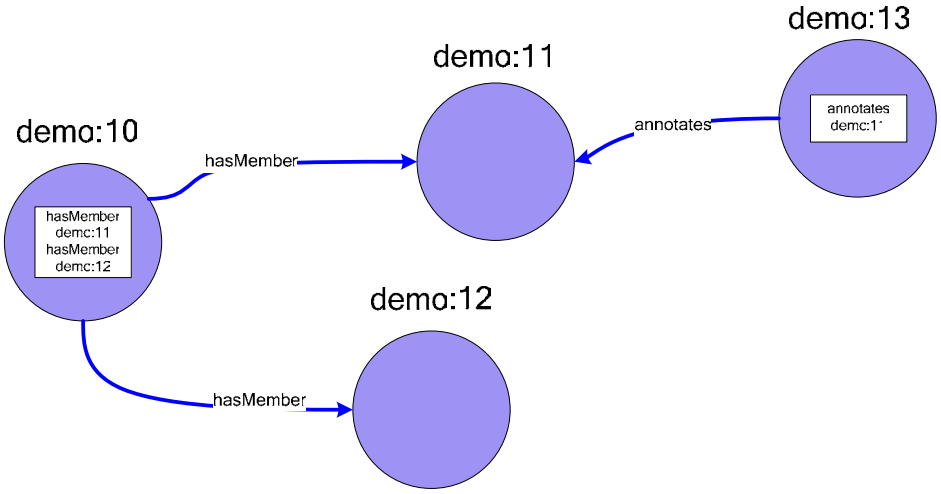

Figure 4 - Relationships among objects

This relationship metadata is automatically indexed into the relationship store. As a result, as shown in Figure 5, demo:10 can have a URL-accessible dissemination that queries the relationship store to return its list of members. Note that these queries can traverse the entire relationship graph, not just information local to the relationship metadata in an individual digital object. For example, a dissemination from demo:10 could return the list of all members that have annotations.

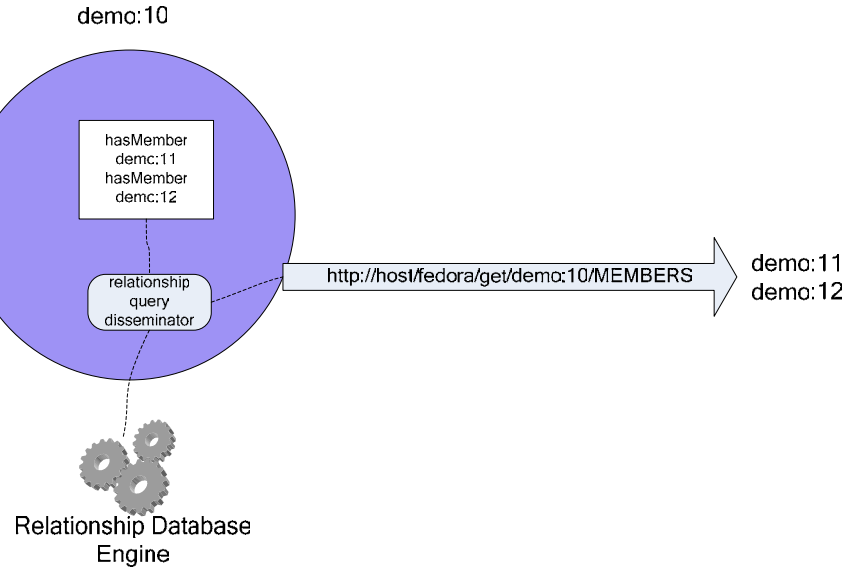

Figure 5 - Access to relationships

Fedora service framework

As described earlier, a Fedora repository runs as a service within a web server. All the functionality of Fedora is exposed as a set of web service interfaces. While Fedora provides the set of core repository services listed earlier in this document, there are many other services that are beneficial companions to a repository. These include specialized ingest services, workflow services, and preservation services.

The Fedora Service Framework facilitates the integration of new services with the Fedora repository. It takes a service-oriented architecture approach to adding new functionality around a Fedora repository, allowing new services to be built around the core repository as stand-alone web applications that run independently of the Fedora repository.

The Fedora development team has developed an initial set of services – a directory ingest and OAI-PMH service – and will continue to develop new services in the future, especially services for workflow, preservation, and search. New services will be part of the main Fedora distribution and will be kept up to date with new versions of the core Fedora repository distribution. Members of the Fedora community are also developing new services that will be shared through the Fedora web site.

Figure 6 illustrates the Fedora Repository Service in the context of the Fedora Service Framework with current and projected services and applications.

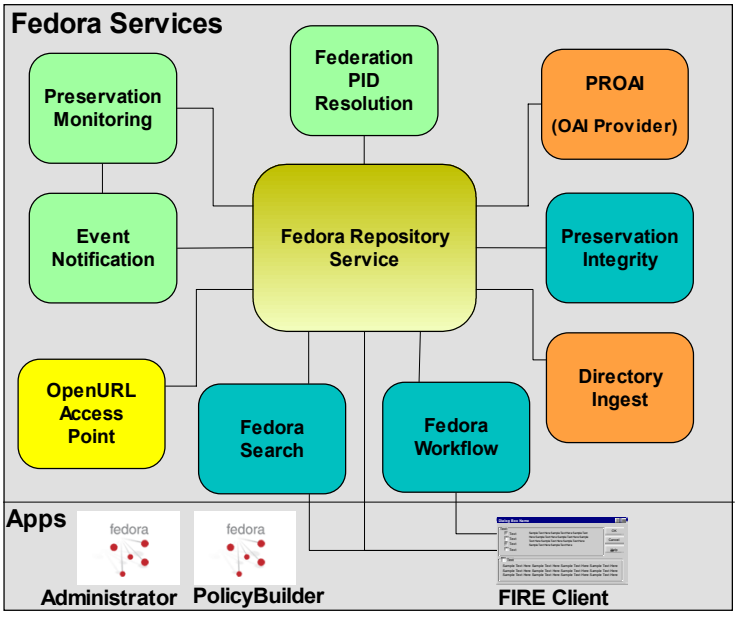

Figure 6 - Fedora service framework

The Fedora Project and Community

Fedora is developed and supported by the Fedora Project, which is located at Cornell University and University of Virginia. The Project developed from initial research work at Cornell funded by the The Defense Advanced Research Projects Agency (DARPA) and the National Science Foundation (NSF). Development for the open source project has been supported by the Andrew W. Mellon Foundation since 2002. The current funding extends through 2007.

The Fedora Community consists of a set of institutions and organizations that actively use Fedora and contribute to its development. The community includes a number of advisory groups that recommend future strategies to the Fedora development team.

Getting more information

The Fedora web site at http://www.fedora.info is the best source of information on the Fedora software, project, and community.