“Development of Repository Architecture and Services at the University of Virginia Library”

Leslie Johnston
Head, Digital Access Services
University of Virginia Library
Alderman Library, Box 400112
Charlottesville, VA 22904-4112
johnston@virginia.edu

Abstract:

This paper proved an overview of the development of a digital repository architecture and end-user repository services at the University of Virginia. The architecture was developed on top of Fedora, and is based on a set of local assumptions. Local standards were identified, and work flows were developed. Services were developed in conjunction with public services staff, and introduced over a two year period through an iterative beta release. An unplanned outcome of the project was the identification of a series of local principles of digital curation, which encapsulates the goals and activities of our repository project, as well as providing potential metrics for assessment.

In 1999 the University of Virginia Library began working with Cornell University on Fedora™, [1] after discovering an article (Payette and Lagoze, 1998) that described the Flexible and Extensible Digital Object Repository Architecture. After UVa completed a reference implementation (Staples and Wayland, 2000), UVa and Cornell secured grants from the Andrew W. Mellon Foundation in 2001 and again in 2004 to develop Fedora into an advanced open source digital library architecture.

The goal of the Fedora project is to develop a generalized digital asset management (DAM) architecture upon which many types of digital library systems might be built. Fedora is the underlying architecture for a digital repository, not a complete management, indexing, discovery, and delivery application. Fedora includes the software tools to ingest, manage, and provide basic delivery of objects with little or no customization, but Fedora's real flexibility and potential lies in that customization. Fedora was initially released in 2003 (Staples, Wayland, and Payette, 2003), and since that time a number of projects and institutions internationally have built local systems, open-source applications, and commercial systems on top of Fedora.

In 2002, the University of Virginia Library began working toward the development of a Digital Collections Repository on top of Fedora (Johnston, 2004; Johnston, 2005). The development process did not leap into software implementation, but began with a set of assumptions about

and specifications for the architecture and services to be built. There were six key assumptions about the architecture:

- The Repository will be a part of a global network that will be built by libraries, governments and corporations.
- All media and all content types will be integrated into one Repository collection.
- There will be simple objects and complex objects with many relationships, and we will need to manage both the objects and their relationships.
- We will be faced with born-digital scholarship incorporating both digital materials and context.
- Any given resource can be associated with and presented in any number of contexts.
- Searching and browsing are equally important.

There were six key assumptions about services:

- The Repository will be a curated repository.
- The UVa community is the primary users of the Repository.
- All of the UVa Library's digital collections will eventually be managed and delivered by the Repository.
- The Repository will be part of the solution to create a single point-of-access to the print and digital collections together.
- The Repository will have a public interface to support discovery and use of the collections by the UVa community.
- The Repository will provide tools for the use of the collections in instruction and research.

These assumptions framed all work on the Repository. Following the development of the assumptions, two working groups were formed to identify functional requirements for digital image and electronic texts, and two working groups reviewed existing media files to outline proposals for the Fedora architecture content models for images, texts, and finding aids.

The specifications for functionality and delivery were documented in different ways. The content models for the different format types include specifications for behaviors that objects should be able to present, such as delivering subsets of their content or metadata, delivery of static files or on-the-fly transformations (such as raw XML delivery versus styled HTML), or supporting the download of image files. The functional specifications for the searches and the results and presentation formatting were documented in detailed screen-by-screen descriptions. Great care was needed to determine that the desired functionality was matched to behaviors in the underlying content models, and that the correct number and types of media files and metadata were present to support the behaviors.

Specifications were set for the production of digital images, for the encoding of electronic texts in TEI (Text Encoding Initiative)[2], and for EAD (Encoded Archival Description) Finding Aids[3]. Because of the natural variation in production over time and with internal and external sources, varying content models were identified to handle the variability of objects because media files could not be migrated to meet a single standard. Three image content models and production variations were agreed upon, as were three variations in TEI files covering the presence or absence of page images and full transcriptions. One content model and production standard each were set for image metadata in the local GDMS (General Descriptive Modeling Scheme)[4] format, and for EAD finding aids.

Another part of the specification process was the documentation of metadata standards for all media and metadata formats. The UVa Library had already developed its GDMS, UVa DescMeta[5], and UVa AdminMeta[6] DTDs for the encoding of descriptive, administrative, and technical metadata. A Metadata Steering Group was formed to review all the applicable metadata formats, document use guidelines, and provide mappings to UVa DescMeta, which could serve as the crosswalk for use in ingesting and delivering objects through the Repository.

The development of new standards required changes in daily activities for a number of units. In many cases the desired functionality for the Repository led to alterations of the DTDs and cataloging practices, such as strategies for recording seriality or identifying membership of an object in a defined set to support browsing via "aggregation objects," rules-based collection objects where the objects that belong to the set are identified and assembled at dissemination time rather than through explicit collection relationships, and the content data stream contains XQuery or XPath statements with the rules needed for the aggregation.

Even with the development of multiple content models and production standards, objects created as recently as a year earlier required migration. File names had to be changed to meet new naming standards. XML was updated to parse against the new DTDs and meet new encoding standards. A number of transformation scripts were required to update the legacy files as well as automate production as much as possible. Once the first identified set of files were migrated to meet the standards and tools for future production were in place, the transition of the UVa Library's Repository from an R&D project to a production operation could begin in earnest.

Digital curation is the creation of a viable social and technical infrastructure for managing and preserving valuable data without significant loss or degradation (Digital Curation Centre, 2005; Hank, 2006). The ultimate aim of our digital curation efforts is to both preserve and to enable the use of the objects in our collections. In identifying our metadata and format standards and starting the work to migrate content to meet those standards we are improving our ability to manage, preserve, and deliver the materials. With a controlled set of standards and object classes, we have fewer types of files to manage, deliver, and preserve, also limiting the scope of future format migrations. Variation is allowed for legacy collections, including low quality versus high quality images, electronic texts with or without transcriptions or pages images, video with or without transcriptions, etc. There is strong desire and need for an environment where data resources are interoperable, easily discovered, and with appropriate appraisal mechanisms in place for the selection of resources by searchers. The use of common standards and open standards is vital for this interoperability.

To coordinate this activity, we were required to be familiar with metadata and format standards used in the greater community to select standards for local use that were appropriate for production and interoperability. We worked with format and metadata staff from across the Library to identify those standards and map between local metadata standards and community standards. We knew that we needed to transition from project-based production models to more sustainable and higher volume production. We worked with format, production, and systems specialists from across the Library to identify necessary tasks, diagram workflows, and implement as many automated processes as possible.

We needed to be familiar with both our physical and digital collections and work closely with the subject librarians to understand the curricular and research needs of our faculty, something that changes with every term, so we could prioritize both new production and migration of legacy

collections. We developed “Guidelines for Digitization,” a “Production Prioritization Review of Collections” guide, and a “Technical Assessment” data entry form for subject librarians. We created an inventory of legacy collections and a list of current and upcoming special production projects for specific sets or collections. We combined the lists and, working with subject librarian and production staff input as to need, available resources, and ease of work, we prioritized all projects. The queue is reviewed quarterly for completions, additions, and re-prioritization, such as when the Library introduced the theme “The Experience of Race in America” for a special project to identify collections that could be used to develop workflow and delivery systems for new format types. Alongside the project-based production we also have ongoing queues of individual text titles and images to be used in research or instruction that are identified by subject librarians in consultation with faculty.

The current UVa Library Digital Collections Repository collections consist of digital images, electronic texts, and EAD finding aids. Digital video, audio, printed music, datasets, and GIS are part of the Library’s collections, and migration of those formats is in various stages of implementation. Many of the collections come from over a decade of internal digital production, the creation of surrogates of the Library’s physical collections. Some are licensed from vendors. Some are born-digital scholarship created by faculty, often integrating Library materials. Some comes from open access sources, such as Federal and state datasets. All the objects, when brought into the Repository, bring relationships with them, whether simple relationships between media files and metadata, more complex relationships, such as that of page images to a text volume transcription, the relationships between issues of a newspaper, or more complex relationships still, such as the organizational context that a scholar overlays onto a digital archive in a web site. The objects and their relationships are part of the Repository.

The development of our Repository’s architecture followed the guidelines of the OAIS reference model for trusted repositories (ISO, 2002). Repository architecture must validate objects, document objects, enforce rights through programmatic rights policies, and run in a managed server environment. It is expected that as the range of media formats that we manage increases, we will need to introduce representation format registries[7] into our operations. The UVA Digital Collections Repository manages the delivery versions of our digital resources, and all the metadata about them, including basic representation information, and all the computer programs needed for representation or rendering for the user. We use a system of persistent identifiers for all files in the Repository, which includes changing references to external files that are embedded in XML files or in databases.

Many institutions are now thinking not only of the sustainability of media objects, but of the scholarly contexts created to organize, annotate, and deliver those objects. This can be accomplished with a flexible, granular approach to managing data as objects with multiple relationships. This must be enabled at a core object architecture level — objects are not monolithic, and their components can be part of multiple contexts and can be added into new contexts by the librarians and scholars who work with them. As an example, in the UVA object architecture a manuscript is an object (a work object), but every page image that makes up that manuscript is also an object (a media object) that can be part of the manuscript’s context and part of other contexts, such as a collection of architectural drawings (an aggregation object). In such an architecture, objects are essentially free agents, true to their original contexts but not solely bound to them. UVA has the beginnings of an authoring environment on top of the collections that is capable of taking advantage of not only the objects but the relationships between them, building a new network of contexts and relationships that we will want to collect and preserve on top of the original objects.

We had to be conversant with the general requirements of OAIS and with current strategies for digital file preservation to develop our architecture. We needed to map the underlying transactions of the repository's operations to OAIS as well as take into account the technical and administrative metadata that was required. We had to be exceptionally familiar with Fedora architecture to develop our granular object management, and to design content models and disseminators around those granular objects. The development of the disseminators required familiarity with the file formats, metadata standards, and the functional requirements of the discovery and delivery services to assure that we were taking all needs into account at an architectural level.

The local mantra "If digital collections cannot be used, then they have not been preserved" was foremost in our minds when we set out the assumption that the Library had to develop both discovery services and end-user tools for the Repository. In 2003 a "Phase 1" prototype discovery interface was released for review by Library staff. The prototype was not operating on top of Fedora, but was a proof-of-concept to guide the Fedora Repository development. Input was solicited from Library staff on the design, functionality, and usability. The project team collated over 130 comments into 23 recommendations in four broad categories: User Interface, General User Functionality, Image-specific Functionality, and Text-Specific Functionality. The identification of a minimum set of contents was the top priority; developing hierarchical, thesaurus-based searching ranked last. Other highly rated priorities included search limits by format, collection, or topical set; cross-format searching; support for collection browsing in addition to searching; and the availability of both keyword searching and advanced fielded searching with Boolean operators for all discovery options.

The "Phase 2" Repository built on top of Fedora was released to Library staff and selected faculty in 2004 as a beta. What followed was a two-year iterative process where feedback was collected via web forms and email, focus groups were held with faculty and library staff, and faculty taught classes using the service and tools. A key part of the Phase 2 Repository was a digital object collector tool, now named "Collectus," that allows users to create personal portfolios of objects. This is a Java application for the client machines (updated automatically via Java WebStart) that provides the ability to collect images and texts into personal portfolios and generate slide shows or electronic reserve websites that include pointers to the images and metadata in the Repository. The slide shows and electronic reserves deliver the images wrapped in an ImageViewer that allows zooming, rotation, and other on-the-fly image manipulation. As new formats are added to the collections the tool will be updated, functioning as a sort of combination shopping cart and low-level authoring tool for the Repository. Over the course of that two-year beta period feedback was constantly reviewed to prioritize redevelopment. The interface was updated, functionality was augmented, bugs were fixed, the server infrastructure was upgraded, collections were added, and workflows were stabilized. The production Digital Collections Repository became available to the full University community in 2007.

This process required close working relationships between the project leaders, the implementation team who were part of many Library units, and public services staff. The key skills were good communication – you can never communicate enough, even if it's to say that there hasn't been any progress – and a tangential communication skill: "translation" between the public services staff and the programmers. There were many requests for new services that could not be accomplished in the identified time frame due to technical challenges. Specifications were not always documented at the level that the programmers expected. We often served as translators, iterating through feature request discussion to develop

specifications for the programmers, and explaining technical challenges in an accessible way to the public services librarians to see if functional requirements could be changed.

Five years ago we started thinking about curation in the most traditional sense of the word – that digital collections would be evaluated and selected using the same subject-based criteria as the physical collections, augmented by technical assessment of the media and metadata. At the time, curation of the collection seemed a different effort than the stewardship of the digital objects. As work has progressed, the definition of local digital curation principles has expanded and evolved to encompass not only intellectual curation, but issues of standards and preservation that are enforced through best practices and systems architecture. Digital curation is the ongoing creation of a collection that supports our community's teaching and research, a collection that we add value to, manage, and preserve not just for its current use, but for future scholarly uses and technologies that we have not yet even imagined.

This overarching process led the UVa Library to an unplanned outcome —identification of a set of local principles of digital curation (Johnston, 2007):

- Principles for Selection of the Collections
    - Support teaching and research.
    - Promote and improve access to unique and rare items.
    - Look for valued-added possibilities when selecting material to be digitized.
    - Preservation of the physical is a selection criterion for the digital.
- Principles for the Use of Standards
    - Preservation of the digital is one of the ultimate goals, but underneath that goal is a standards issue.
    - Enforcement of standards and best practices creates a more controlled environment for preservation.
- Principles for Trustworthiness
    - The users must be able to trust the objects in the Repository.
    - Appropriate authentication, authorization, rights management and security are not just aspects of the architecture; they are part of the establishment of trust.
- Principles for Preservation and Sustainability
    - Enable use and sustainability of the Repository collections.
    - Build a trusted digital repository architecture.
    - Governance and operational policies are of equal importance to standards and architecture.

How do these principles reflect our activities? We outlined policies to build collections that increase access and use of our unique materials and provide faculty with what they want and need. We identified a set of circumscribed formats and minimum metadata standards to which all objects must adhere. We have a controlled environment that, in theory, simplifies our preservation tasks by minimizing the classes of objects that we must sustain. There is a scaleable architecture with which to manage objects and the relationships among them, operating in a consistent, managed environment that makes the task easier to build discovery and delivery services, and tools for the use of the objects. The principles encapsulate the goals and activities of our repository project, as well as providing potential metrics for assessment.

Success of a repository can only be assessed against the purpose that the repository serves in its operating environment; no repository can rated as successful unless it fulfills its purpose. The collections, services, and tools have been tested by our faculty and we have heard that we are giving them what they want – persistent, trusted collections that contain content that they

find useful in their teaching and research, and the tools that they need to use them. These are the foundations for a sustainable repository and collection.

References

Digital Curation Centre. "Digital Curation and Preservation: Defining the research agenda for the next decade, Warwick Workshop 7/8 November 2005." *Curation Services and Technologies Session Report* (November 2005). <http://www.dcc.ac.uk/training/warwick_2005/Warwick_Workshop_report.pdf>

Hank, Carolyn. "Digital Curation and Institutional Repositories: Seeking Success: JCDL 2006 Workshop Report." *D-Lib Magazine*, vol. 12, no. 7/8 (July/August 2006). <http://www.dlib.org/dlib/july06/hank/07hank.html>

International Standards Organization. "Reference Model for an Open Archival Information System (OAIS)." January 2002. <http://public.ccsds.org/publications/archive/650x0b1.pdf>

Johnston, Leslie. "An Overview of Digital Library Repository Development at the University of Virginia Library." *OCLC Systems & Services: International Digital Library Perspectives* 20, Number 4 (2004): pp. 170-173.

Johnston, Leslie. "Development and Assessment of a Public Discovery and Delivery Interface for a Fedora Repository." *D-Lib Magazine*, vol. 11, no. 10 (October 2005). <http://www.dlib.org/dlib/october05/johnston/10johnston.html>

Johnson, Leslie. "Principles and Activities of Digital Curation for Developing Successful and Sustainable Repositories." In *Strategies for Sustaining Digital Libraries* , edited by Martin Halbert and Katherine Skinner. Atlanta, Ga: Emory University, *forthcoming*.

Payette, Sandy and Carl Lagoze. "Flexible and Extensible Digital Object and Repository Architecture (FEDORA)," presented at Second European Conference on Research and Advanced Technology for Digital Libraries, Heraklion, Crete, 1998. <http://www2.cs.cornell.edu/payette/papers/ECDL98/FEDORA.html>

Staples, Thornton, and Ross Wayland. "Virginia Dons Fedora: A Prototype for a Digital Object Repository." *D-Lib Magazine* (July/August 2000). <http://www.dlib.org/dlib/july00/staples/07staples.html>

Staples, Thornton, Ross Wayland, and Sandra Payette. "The Fedora Project: An Open-source Digital Object Repository Management System." *D-Lib Magazine* (April 2003). <http://www.dlib.org/dlib/april03/staples/04staples.html>

[1] Fedora software releases and documentation are available at: <http://www.fedora.info/>

[2] Information about the Text Encoding Initiative (TEI) is available at <http://www.tei-c.org/>.

[3] Information about the Encoded Archival Description (EAD) format is available at: <http://www.loc.gov/ead/>.

[4] Information about GDMS is available at: <http://www.lib.virginia.edu/digital/metadata/gdms.html>

[5] Information about UVa DescMeta is available at: <http://www.lib.virginia.edu/digital/metadata/descriptive.html>

[6] Information about UVa AdminMeta is available at: <http://www.lib.virginia.edu/digital/metadata/administrative.html>

[7] Current representation information and file format registry projects include PRONOM <http://www.records.pro.gov.uk/pronom/>, the Global Digital Format Registry (GDFR) <http://hul.harvard.edu/gdfr/>, and the Presidential Electronic Records Project Operational System (PERPOS) <http://perpos.gtri.gatech.edu/>.