UNIVERSITY of VIRGINIA LIBRARY

Development of Repository Architecture and Services at the University of Virginia Libraries

Leslie Johnston
Head of Digital Access Services
University of Virginia Library

DigCCurr 2007

UNIVERSITY of VIRGINIA LIBRARY

First, we had to Build Fedora

- The University of Virginia Library began working with Cornell University on Fedora™ in 1999.
- After UVa completed a reference implementation, 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 potential lies in that customization.
- Fedora was initially released in 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.

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

UNIVERSITY of VIRGINIA LIBRARY

We Defined Architecture Assumptions

- 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.

UNIVERSITY of VIRGINIA LIBRARY

We Defined Service Assumptions

- The Repository will be a curated repository.
- The UVa community is the primary user 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.

UNIVERSITY of VIRGINIA LIBRARY

UVA Repository Development Planning

- Legacy digital collections were surveyed to identify file formats and presence and format of metadata.
- Working groups were formed to identify functional requirements for the delivery interface.
- The output of both these processes were transformed into formal specifications that were translated into holistic local production standards (for migration and new production), Fedora *content models* (classes of content objects) and *disseminators* (mechanisms for delivering objects), and interface functionality.
- We developed the underlying object architecture using Fedora, designed the repository workflows, and wrote the necessary code.

UNIVERSITY of VIRGINIA LIBRARY

UVA Repository Development Process

- The local mantra "If digital collections cannot be used, then they have not been preserved" was foremost in our minds when we started to develop discovery services and end-user tools for the Repository.
- In summer 2003 a Phase 1 alpha prototype (not using Fedora) was developed to test functionality and look and feel.
- Feedback received informed a Phase 2 beta Fedora prototype which launched for limited testing in Fall 2004. Additional features and an expanded test group were brought in Fall 2005.
- Phase 2 development included service enhancements identified through testing as well as infrastructure enhancements.
- The Repository launched for the full UVa community in 2007.

http://lib.virginia.edu/digital/collections/

UNIVERSITY of VIRGINIA LIBRARY

UVA Digital Collection Repository Contents

- Much of the collections are the product of fifteen years of internal digital production, creating surrogates of the Library's physical collections. Some are licensed from vendors.
- The UVa Library Repository collections currently consist of digital images, electronic texts (transcriptions and/or page images), and EAD finding aids.
- New formats in process include printed music, video, datasets, audio, and GIS.
- Upcoming collections include born-digital scholarship created by faculty, often integrating Library materials.
- All the objects bring relationships with them, whether simple relationships between media files and metadata or that of page images to a text volume transcription, more complex relationships such as those between issues of a newspaper or volumes in a series, 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.

UNIVERSITY of VIRGINIA LIBRARY

What is "Digital Curation" at UVA?

- While the subject of digital preservation is central to the larger discussion of what is called digital curation, intellectual curation and digital preservation are both represented in our curatorial and operational assumptions.
  - UVA is creating a collection that supports our community's teaching and research, a collection that we manage and preserve not just for its current use, but for future scholarly uses and technologies that we have not yet even imagined. The overarching aim of our curation efforts is to build collections and preserve and enable the use of the objects. If an object cannot be used, it has not been preserved.
- No one person is or could be the "digital curator" – it's a team of staff with varied expertise and roles.

UNIVERSITY of VIRGINIA LIBRARY

What Did we Need to Know to Develop the Architecture?

- We had to be up-to-date on technologies and initiatives in the community that could be useful in helping us model our own projects.
- We had to understand what might constitute an object – number and formats of files, metadata, and behaviors.
- We had to be conversant with the general requirements of OAIS and with current strategies for digital file preservation to develop our storage and object architecture. We needed to map the underlying transactions of the repository's operations to OAIS and identify the required technical and administrative metadata.
- 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 Fedora disseminators required familiarity with the file formats, metadata standards, and the functional requirements of the discovery and delivery services as set by curators and users to assure that we were taking all needs into account at an architectural level.
- We needed to recognize that there is no such thing as a final version of an object or its metadata, given updating and migration of metadata and formats, and to support object versioning and citation of versions.
- We needed to create an architecture where we could ingest content to manage and preserve it, but also to get content OUT so our content is useful and usable in other systems and contexts.

UNIVERSITY of VIRGINIA LIBRARY

What Did we Need to Know to Develop the Standards and Workflows?

- We were required to be familiar with metadata and format standards used in the greater community to select appropriate standards for local use that can hopefully support future preservation of content. It's not just the formats, it's the metadata documenting the content and process of its creation.
- We had to be familiar with procedures for digital production and metadata creation to design workflows, and implement as many automated processes as possible.
- We had to recognize that there is no such thing as a final standard or workflow, and expect to review and revise our standards and workflows on a regular basis.
- Familiarity with processes and technologies to map and transform metadata and formats between standards for interoperability and migration.

UNIVERSITY of VIRGINIA LIBRARY

What Did we Need to Know to Develop the Services?

- Understanding of our diverse users and their demanding needs. We often make the mistake of referring to "our users," when we have users from and across myriad disciplines with many requirements depending upon pedagogy, research practices, and formats and tools used.
- We worked with subject librarians – aka content curators – to identify discovery requirements, as well as identifying and communicating with the user community to identify their content needs and use needs.
- To the curators and the users the discussion of service functionality and collection contents went hand-in-hand – they identified collections to be included in the Repository and how they expected to be able to use them.
- Their input led us to develop tools to GATHER and USE the collections once discovered. We built the Collectus digital object collector tool and an ImageViewer to develop and organize personal sets, and create slideshows and web pages. A tool to integrated images from non-Repository collections is also under development. Once the gathering tools are in place, we will begin work on more elaborate creation/authoring tools.

UNIVERSITY of VIRGINIA LIBRARY

What Did we Need to Know to Develop the Collection and Policies?

- We had to be familiar with general principles of collection development and content assessment, but adding in technical assessment.
- We had to develop expertise in licensing and intellectual property issues, especially fair use.
- We had to learn what our curators had and what our users wanted.
- We developed a Collection Development Policy and "Guidelines for Digitization" to identify the scope of the collection building.
- We created a "Production Prioritization Review of Collections" guide, and Technical Assessment criteria and a data entry form for subject librarians to aid in the assessment of collections to identify production needs.
- 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 (from the technical assessment), we prioritized all projects.
  - The queue is reviewed quarterly for completions, additions, and re-prioritization.
  - 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.

UNIVERSITY of VIRGINIA LIBRARY

Our Process led us to identify and document our Principles of Digital Repository Curation.

UNIVERSITY of VIRGINIA LIBRARY

Principles for Selection of the Collections

- Support teaching and research.
  - Our primary principle of digital curation, or else we are creating a Repository of limited utility for our community. Subject librarians identify content that supports the curricular and research needs.
- Promote and improve access to unique and rare items.
  - In prioritizing content, one of the most important criteria, in addition to direct requests from faculty through subject librarians, is that the content that is rare or unique to UVa. This content can be existing digital surrogates of physical materials or physical materials to be digitized.
- Look for valued-added possibilities when selecting material to be digitized.
  - Identify high-use physical or digital materials, provide rich discovery metadata, provide the most granular markup possible, and opportunities to provide tools for the use of the objects.
- Preservation of the physical is a selection criterion for the digital.
  - The incorporation of preservation reformatting projects into Repository production ensures the continued use of our brittle collections.

UNIVERSITY of VIRGINIA LIBRARY

Principles for the Use of Standards

- Preservation of the digital is one of the ultimate goals, but underneath that goal is a standards issue.
  - The intellectual selection of materials for the Repository has been balanced with a technical assessment, where the materials are compared to the library's standards, assessed for migration to those standards, and appraised for viability and preservation over time.
- Enforcement of standards and best practices creates a more controlled environment for preservation.
  - With a controlled set of standards and object classes, the Library has fewer types of files to manage, deliver, and preserve, including limiting the scope of future format migrations.
  - 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.

UNIVERSITY of VIRGINIA LIBRARY

Principles for Trustworthiness

- Users must be able to trust the objects in the Repository.
  - The Library's role in the selection, production, documentation, and management of the digital objects in its Repository provides a perception of trustworthiness. Consistent format and metadata standards, consistent object behaviors, versioning, and the ability to cite versions of objects contribute to the real trustworthiness as seen by users.
- Appropriate authentication, authorization, rights management and security are not just aspects of the architecture; they are part of the establishment of trust.
  - Validation of datastreams during production and ingestion; assumption of authentic datastreams in delivery.
  - Documentation of rights and policies, authorization of users, and support of policy-based the use of objects based on their rights documentation.
  - Security of the infrastructure.

UNIVERSITY of VIRGINIA LIBRARY

Principles for Preservation and Sustainability

- Build a trusted digital repository architecture.
  - The Repository manages the delivery versions of our digital resources, and all the metadata about them, including basic representation information, plus all the scripts and programs needed for transformation, representation, or rendering for the user.
  - Our architecture validates objects, enforces rights through programmatic rights policies, and runs in a managed server environment.
- These core trusted Repository architecture attributes are key components in assuring our community that they can trust our Repository and that digital scholarship that we collect from them will be properly managed and preserved.
- Enable use and sustainability of the Repository collections.
  - Work is ongoing to identify levels of sustainability that can be promised for various types of objects, including the functionality that accompanies or is expressed by those objects. This effort goes hand-in-hand with the identification of the controlled set of formats that we will manage, and the ability to migrate objects to need those format and metadata standards.
- Governance and operational policies are of equal importance to standards and architecture.
  - Develop documented mission statement, policies, and workflows for the operation of the Repository.
  - Ensure that there is organizational support for the operations, adequate, appropriately skilled staffing, and adequate technical infrastructure.
  - Operational activities must include periodic review of the operational status, and an audit of the Repository.
  - Policies must include those that ensure the continued review and updating of the workflow and the policies themselves.

UNIVERSITY of VIRGINIA LIBRARY

What are the Outcomes?

- Created a collection development policy and digitization guidelines to build collections that increases access and use of our unique materials and provides faculty with what they want and need.
- Identified a set of circumscribed formats and minimum metadata standards to which our objects must adhere.
- Built a controlled environment that, in theory, simplifies our preservation tasks by minimizing the classes of objects that we must sustain.
- Created an architecture with which to manage objects and the relationships among them.
- Created a consistent, managed environment that makes the task easier to build discovery and delivery services, and tools for the use of the objects.
- 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.