

















An iterative approach to building sustainable repository services on Fedora

Claire Stewart, Karen Miller, and Bill Parod
Northwestern University Library

February 2009

NORTHWESTERN UNIVERSITY LIBRARY

The Humphrey Winterton Collection of East African Photographs:1860-1960

Home	About this Site	Winterton in the Classroom	Browse
Collection Browser Decades Subjects Places Persons Keywords 'Albrecht 'Kilindini 'Nyanza' The 'Tings' 'gig' --Comment -Governor 103] 106] 109] 1890s-1930s 1920s-1930s 1950s Africa 1960s Arusha Avenue Africans Brackenhurst British Camels Camping Church Circa Coast Coffee Colony Drying East Egypt Elgon Falls Fort Fully Goods Great Hotel Jesus Kenya Kenya Kikuyu Kisumu Limuru Lion Market Mombasa Mount Mozambique Nairobi Namanga Nautical Nile Nyasaland Rift River Road Safari Salim Sixth Sudan Tanganyika The Thika Torch Turkana Uganda Valley Various back background base channel coast coffee exterior fermenting fording forest game gathering Search	 <p>View of the monastery...</p>  <p>Transport in old Cairo...</p>  <p>Pyramid Road...</p>  <p>Cart...</p>	 <p>The usual 'load'...</p>  <p>Street scene with car...</p>  <p>The Nile at Gezira...</p>  <p>View from halfway up the Cheo...</p>	 <p>The Sphinx...</p>  <p>The 'Noted' House...</p>  <p>Heliopolis Road...</p>  <p>At the barrage, 1936...</p>
	 <p>Bismillah, Piazza de del Dilday</p>  <p>The Nile, Gezira</p>  <p>City of Alexandria</p>  <p>Temp of the Calipha</p>		

Digitized collections context

In 2007, a significant reorganization strengthened the Northwestern University Library's commitment to building a robust and sustainable digital repository for unique library collections and other valuable institutional materials. The Library Information Technology division dedicated staff positions to repository development, a Scholarly Communications Librarian position was created in the new Academic Liaison Services department, and the Catalog Department was transformed into the Bibliographic Services Department, with a charge to provide metadata services for digital projects and, eventually, to members of the broader University community. A large new department, Digital Collections, was also formed and charged with building a Digital Image Library for the entire campus, managing the library's special digitization projects, and creating services to support faculty use of digital content in teaching and research.

Both in order to quickly address a significant accumulation of digitized content, and to work with a flexible internal audience, digitized library collections have been targeted for the first round of repository development. Digitized books, finding aids, photographic materials, and audio are among the top priorities. Fedora was selected as the Library's repository platform.

Repository Development Strategy

In the spring of 2008, two full time staff were assigned to begin building comprehensive repository services around the Fedora repository platform. Building a system that accommodates richly heterogeneous collections, empowers professional staff engaged in a variety of production workflows, and provides access commensurate with the richness materials afford is not achieved all at once. In order to support specific near-term project commitments while building broadly applicable content models and services for the long term, we devised a cyclic three-phase strategy. It is cyclic in that feedback and demand from users will result in revision and expansion of core models, services and associated end-user tools.

1. Implement models and services for ingest, preservation, and access of core content.

Develop content and service models for the core content types represented by our collections. Prioritize development by pending project commitments. Use these specific project materials and goals to provide detailed requirements for the ingest, preservation, discovery, access, and management of those materials.

2. Provide tools for staff to ingest and manage repository content.

Phase one provides models and developer mechanisms to ingest core content types and enable their discovery and presentation. Phase two incorporates these ingest mechanisms into end-user tools and workflows that staff can use to ingest and manage repository content.

3. Facilitate integration of repository materials with end-user tools and services.

The use of collections increases when users have good tools to collect, organize, annotate, embed, and otherwise incorporate materials in their scholarly practice. Sometimes users already have favored tools and practice that they use for these purposes. In some cases good tools are lacking. In Phase Three we will evaluate some favored approaches and consider integration/development prospects.

Current Status

At this time we are mostly finished with phase one and well into phase two. One project that has been particularly influential in defining our models and services is the Institute for Museum and Library Services (IMLS)-funded project to digitize the Humphrey Winterton photograph collection.

The Humphrey Winterton Collection

The Humphrey Winterton Collection of East African photographs at the Herskovits Library of African Studies incorporates over 7,600 photographs grouped into 76 intellectual entities, each representing original provenance. The 76 entities are different physical groups of photographs, slides, or postcards. Each entity reflects the scope and interest of the collector, so we wanted to maintain this structure in the display. The entities are all different in terms of type and extent of containers; they consist of single photographs, single albums, boxes, or folders of photos, framed collections of photos, and groups of albums. EAD provides an ideal way to construct a finding aid for this hierarchically arranged collection.

We chose EAD because it is hierarchical in nature and allows for the description of a collection, subsets of the collection, and individual items in the collection, while maintaining the relationship that describes how they fit together. This allows us to maintain the provenance by keeping together the original holdings of individual collectors as single entities, and also to describe the groupings and individual items in each collection.

This project illuminated difficulties of providing item level access within a very large finding aid. Subject heading assignment impacted search results at the item level, and the size of the document itself made navigation difficult for end users. This implementation also raised issues regarding the update of controlled vocabularies and use of single source cataloging to generate derivatives in other metadata formats.

Technical Architecture

Our strategy for repository development has been to use the requirements of specific projects to inform content and service model design, do our best to abstract those designs to maximize reuse, and then refine those abstractions as needed to accommodate new projects. Winterton yielded content models and service stacks for large scale EAD encoded finding aids, hi-resolution images, virtual image crops, and component level discovery. These service methods expose specific content through Fedora disseminators.

EAD Object Model

With Winterton EAD containing around 10,000 components, content emphasis at the lowest levels, and descriptive emphasis at higher levels, the goal of responsive delivery for any component suggested the use of indexing retrieval software. Our implementation achieves quick random access to any component, with lean reference to its ancestors and descendants. The service model also includes metadata transformation for any component as full-page HTML, embedded-HTML, MODS, Dublin Core, and EAD fragment.

Hi-resolution Image Model

Winterton includes over 7,600 digital images accessible at any resolution. Their dissemination takes many forms whether in thumbnail grids of complete albums, image previews in search results, screen size format in web pages, or dynamic zooming viewers. The content model includes source (tiff) and delivery (jp2) formats, MIX technical metadata on each, and Dublin Core and MODS transformation of their associated EAD components for description.

Hi-resolution Image Crop Model

The Winterton collection includes many photographs that share an album page with other photographs. Each individual page is photographed and represented in the repository by a single image object, having its own technical metadata, descriptive metadata, and boundary information for cropping out color bars and barcodes.

When pages contain multiple photographs, each photograph is additionally represented in the repository as a lightweight ‘Crop’ object. Crop objects internally reference their associated image object for their image content. Crops expose all the same imaging methods and so behave the same as images, but they provide their own crop boundaries and descriptive metadata. And like all objects, they have their own identifiers and so are individually citable.

Searchable Service Model

Searchable is not an object class, but a service that any object can use to make itself searchable. The current implementation uses Solr and supports full-text, fielded, and faceted searching. The default implementation indexes MODS Collection streams. Any object that can provide a MODS Collection stream by itself or by proxy can expose this search functionality.

Supporting Services

The formal access models expressed with Fedora disseminators sit at the top of service stacks that encapsulate the technical implementations of content requests. We will describe these disseminators, the web services they invoke, and the underlying implementations for text and image extraction and processing.

General Applicability

We have used the same EAD service components on multiple collections as well as, with different query templates, for TEI dissemination. The Image model is in use with multiple EAD collections, scanned books, and various ad-hoc scanning requests.