The Hydraulics Project:
Empowering Communities to Build a Digital Library Utilizing Fedora and an Event-Driven Service-Oriented Messaging Framework

PAPER PROPOSAL for Open Repositories 2011

Recent collaborative efforts including the Hydra Project and development of Blacklight have made significant strides towards creating revolutionary tools for indexing and accessing digital content managed by a Fedora repository. Comparatively, less attention has been paid to creating applications and standardized workflows designed to easily ingest content into repositories. This issue demands attention because a key measure of a repository's usefulness is the quality and quantity of digital content made available to the public.

A central problem facing institutions eager to deploy a Fedora-based repository is the technological complexity of object creation and maintenance. Despite the Fedora development community's expansion and documentation of APIs for creating and maintaining content and the Hydra development community's recent upgrade and improvement of the ActiveFedora gem, the barrier to entry remains high. The skill-sets and experiences available amongst the heterogenous staff employed by most institutions are largely insufficient to utilize the software. It is the responsibility of technologists to develop tools and work with staff to create the workflows and policies that can more easily enable non-technical staff to claim a larger ownership stake in their institutional repository.

Digital Curation Services at the University of Virginia Library has made significant strides towards this goal by designing Hydraulics. For the last three years, a hybrid team of non-technical staff has collaborated with programmers to develop software and workflows that empower staff from multiple departments within the library to participate in the creation and management of digital objects throughout the production and maintenance lifecycle without the burden of learning the underlying technology. The result is a powerful web application that integrates a request module, a management system for digital production and automated workflows for archiving files and delivering content to both patrons and a Fedora-based digital library.

Development History

The Hydraulics project grew organically and metamorphosed over time to accommodate the needs of its users. Development on its predecessor, Tracksys, began in 2007. At the initial design meetings, the programming team selected Ruby on Rails as the web framework upon which to build the software; this decision was prescient and would enable developers to leverage the Hydra stack in years to come. The initial goal was to create an order management system for Digitization Services, the department charged with digitizing rare and unique materials housed in the University of Virginia's Albert H. Small Special Collections Library. The design specifications called for a web application that would:

- Allow University faculty, staff and students as well as patrons outside the University to place requests for digitization,
- Enable production staff to manage the digital assets created for these orders,
- Retain information about previously digitized material and serve as an intermediate repository for collections before ingestion into the library's digital repository.

The first production deployment went live in 2008 and during the subsequent three years programmers adhered to the principles of agile development by frequently meeting users, eliciting comments and honoring requests for additional functionality. Some notable features implemented at this time were:

- Integration of bibliographic records with the Solr index underlying the library's Blacklight instance. Rather than repeating bibliographic-level metadata, digitized items could now be linked to the MARC XML representation of the catalog record through a unique identifier.
- Population of descriptive and technical metadata for digital objects with XML. Longstanding use of Microsoft Expression Media (originally named Iview) by Digitization Services necessitated creating an import module and XSLT stylesheet to populate the database record for each object. By utilizing Expression Media, Digitization Services continued to use a low cost, easy to use product that produced information exportable as XML.
- Ability to capture, at the time of metadata entry, granular information about a manuscript collection and to associate images with individual components of an EAD guide.
- Authentication for both request form and administrative interface provided by querying centralized LDAP server.

Workflow Engine

The core of Hydraulics is an event-driven service-oriented workflow engine integrated with the request and data management modules of Tracksys. The system design comprises:

- Ruby on Rails,
- The ActiveMessaging gem to build processors that consume messages, perform some service based on the incoming data and then emitting one or more messages to the next processor,
- ActiveMQ to broker messaging queues and ensure message persistence,
- ActiveFedora to create and manage objects and datastreams in a Fedora repository,
- Fedora 3.4.2.

This workflow, designed to meet the current needs of users participating in the care and creation of the University's digital library, automates:

- Quality assuring digital images and metadata created by production staff against internal policies and standards,
- Populating the underlying database with records for newly created content,
- Archiving all materials to a redundant offsite hierarchical storage management (HSM) system,
- Creating lower resolution deliverable images for patrons (if requested),
- Delivering orders to patrons along with invoices and metadata manifests (if necessary),
- Creating and posting to Fedora objects and metadata datastreams recommended by the Hydra Project guidelines, and
- Creating and posting JPEG-2000 deliverables for all image objects.

By the time Open Repositories 2011 is held, Digital Curation Services will have ingested into the digital library approximately 200,000 digital objects (totaling approximately 1.8 million datastreams), and delivered over 500 orders to patrons utilizing the automated workflows of Hydraulics.

Enabling Communities

Understanding that no one person or department will have all the information or expertise required to select the most valuable scholarly content and create the rich metadata necessary to accurately describe that content, Hydraulics and its workflows were designed to enable collaboration. By yoking together all aspects of digital library creation into one application, Hydraulics enables the following parties to engage in these functions:

- **Selection** (Scholars, Patrons and Subject Librarians) - The request module allows those persons, who have an interest in making a particular collection available, to select materials for digitization and ingestion. By outsourcing the selection process to experts, the growth of the University's digital library will be demand-driven and primarily contain those materials whose scholarly worth is certain. Additionally, content requested by patrons for personal research or publication, while not expressly requested for inclusion in the digital library, may be deemed worthy for ingestion simply as a result of being requested.
- **Production** (Digitization Services) - Many of the tasks related to production are now greatly simplified by automated workflows tailored to current needs. The atomized design of the workflows is flexible and gives both staff and technologists the ability to quickly change the architecture to reflect current digitization procedures and policies.
- **Digital Object Creation and Maintenance** (Digitization Services, Metadata Services, Special Collections, University Counsel) - All objects in Hydraulics that can be ingested into a repository have some attributes that must and others that may be assigned by a knowledgable user before a service can process it. Some important attributes include access policy, transcription text, descriptive metadata (in the form of MODS XML), a Solr <add> document, Dublin Core (DC) and RDF for both RELS-EXT and RELS-INT. If these

attributes are not populated before ingest they will either inherit from their parent objects or the workflow can generate the information utilizing a template and information added during the metadata and production workflows. After ingest, Hydraulics retains a local copy of the datastreams for each object to enable a metadata specialist to review, enhance and update the object in the Fedora repository from within Hydraulics.

Towards Inter-Institutional Collaboration

Given the success of Hydraulics, the University of Virginia Library aims to unveil an open source and generalized version of Hydraulics at Open Repositories 2011. The Hydraulics project will hopefully influence conversations amongst interested institutions. Certainly production practices and workflows differ from institution to institution, but by bringing those differences together and trying to generalize many of those features into a ubiquitous workflow, the Hydraulics project stands to concretize a series of best practices for digitization into a powerful workflow and web interface.

A presentation of the University of Virginia's instance of Hydraulics at Open Repositories 2011 would highlight the automated workflows related to ingesting objects into a Fedora repository. Additionally, this presentation would provide links to the source code and documentation for the application and provide the presenter an opportunity to initiate discussions on operationalizing and formalizing collaborative efforts to enhance the capabilities of Hydraulics.

The Hydraulics Project

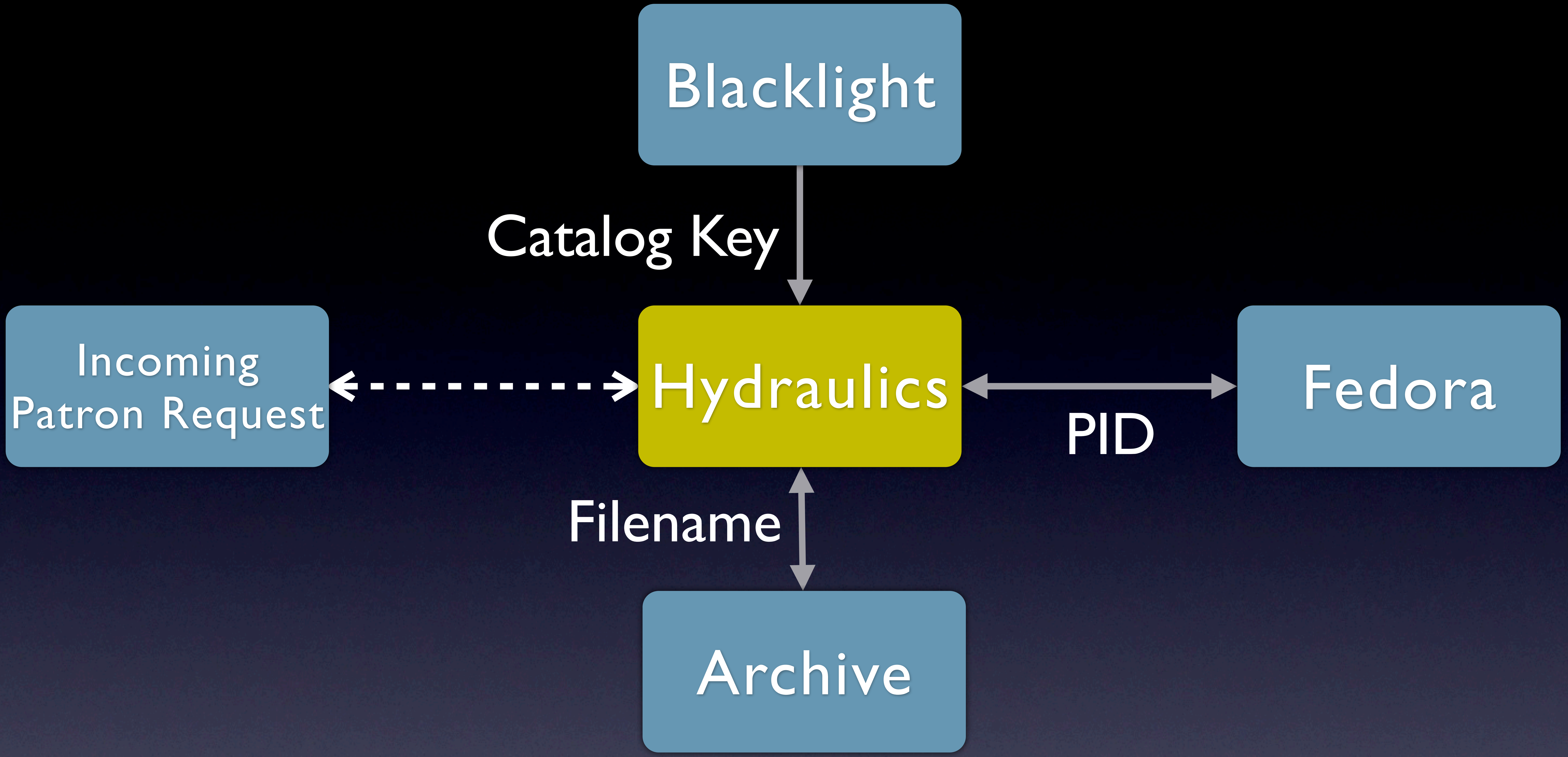
Empowering Communities to Build a Digital Library Utilizing Fedora and an Event-Driven Service-Oriented Messaging Framework

Andrew Curley
University of Virginia Library
Open Repositories 2011

What is Hydraulics?

End-to-end digitization workflow management tool that integrates:

- **Request module**
- **Management system for digital production**
- **Archive management workflows**
- **Patron delivery workflows**
- **Fedora ingestion workflows**



Hydraulics as Connection

Hydraulics serves as the canonical source by facilitating communication with Blacklight, Fedora and an archival filesystem.

Hydraulics vs. Tracksys

Hydraulics

- Generalized, open-sourced, derived from Tracksys

Tracksys

- UVA implementation, tied to local practices and needs

Hydraulics : Tracksys ::
Blacklight : VIRGO

Hydraulics Stack

Ruby on Rails

- ActiveMessaging gem

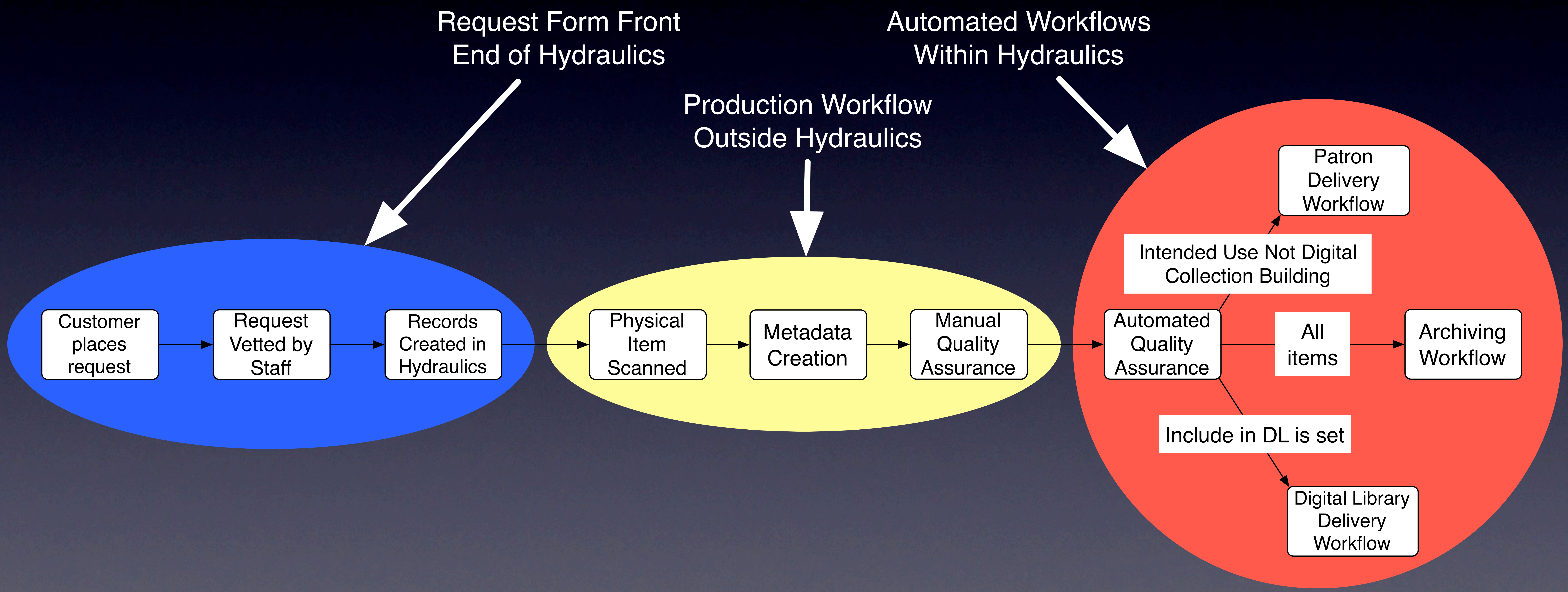
Fedora

- 3.4.2 REST API Compatible

Solr

ActiveMQ

The Hydraulics Model



Request Form Front End of Hydraulics

Production Workflow Outside Hydraulics

Automated Workflows Within Hydraulics

Customer places request

Request Vetted by Staff

Records Created in Hydraulics

Physical Item Scanned

Metadata Creation

Manual Quality Assurance

Automated Quality Assurance

Intended Use Not Digital Collection Building

Patron Delivery Workflow

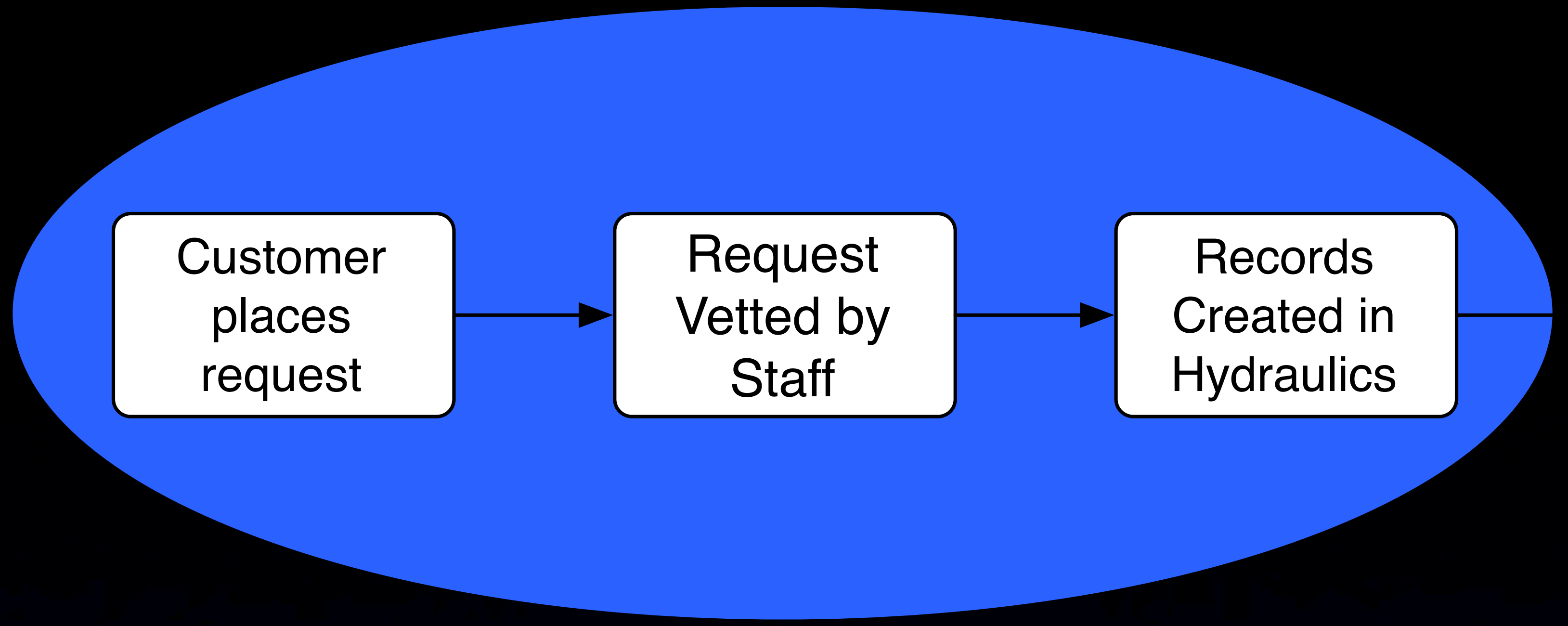
All items

Archiving Workflow

Include in DL is set

Digital Library Delivery Workflow

Request Module



Public facing request form integrates with underlying database

Distinguish internal and external patrons

Facilitates engagement between Special Collections and Preservation/Conservation staff and patron

Allows librarians to link digital objects with canonical metadata

Fee and billing management

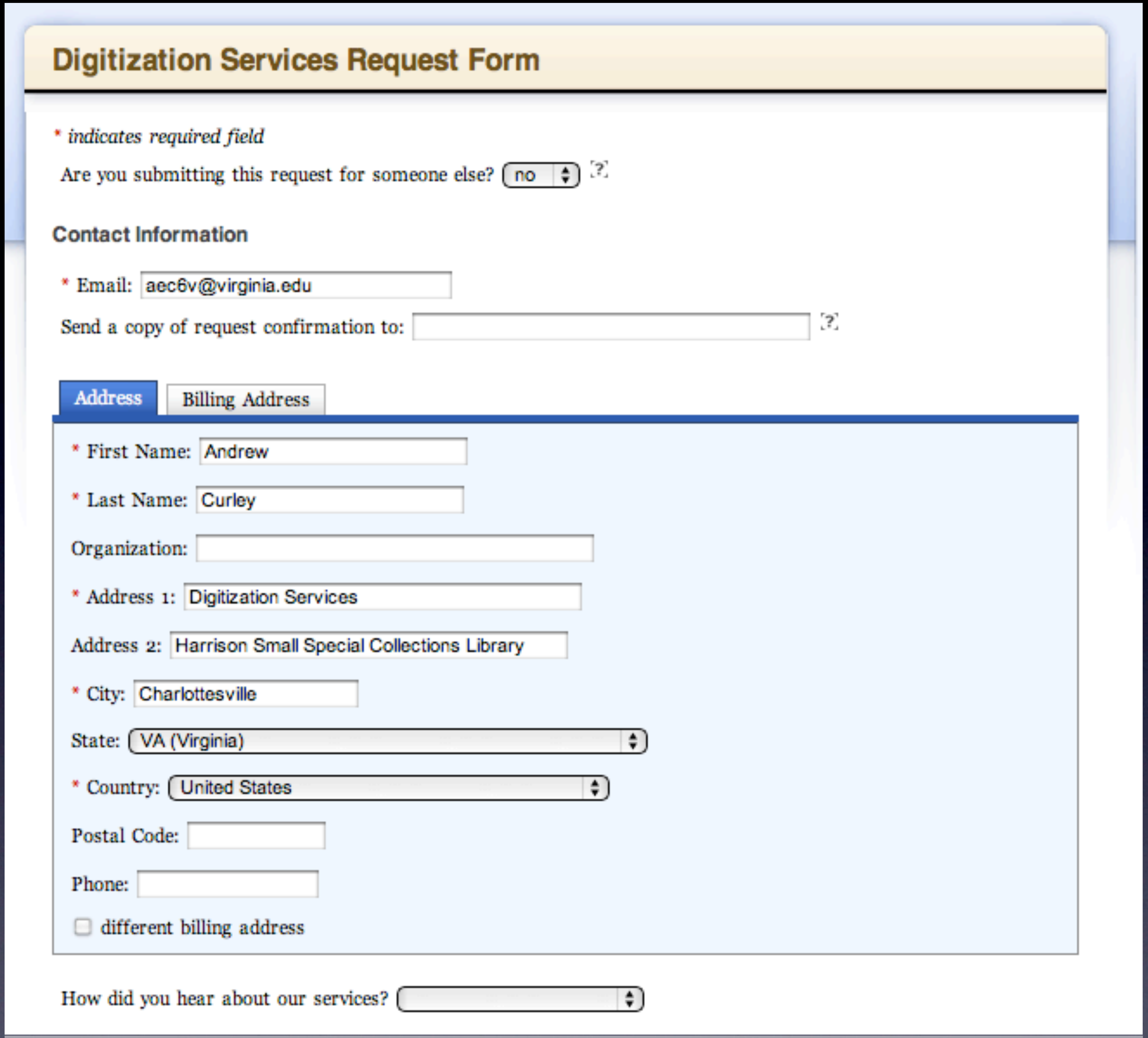
Request Form: User Information

Integration with local LDAP

Populate with existing information

Request can be made on behalf of another person

Billing address information



Digitization Services Request Form

* *indicates required field*

Are you submitting this request for someone else? no [?]

Contact Information

* Email: aec6v@virginia.edu

Send a copy of request confirmation to: [?]

* First Name:	Andrew
* Last Name:	Curley
Organization:	
* Address 1:	Digitization Services
Address 2:	Harrison Small Special Collections Library
* City:	Charlottesville
State:	VA (Virginia)
* Country:	United States
Postal Code:	
Phone:	

□ different billing address

How did you hear about our services?

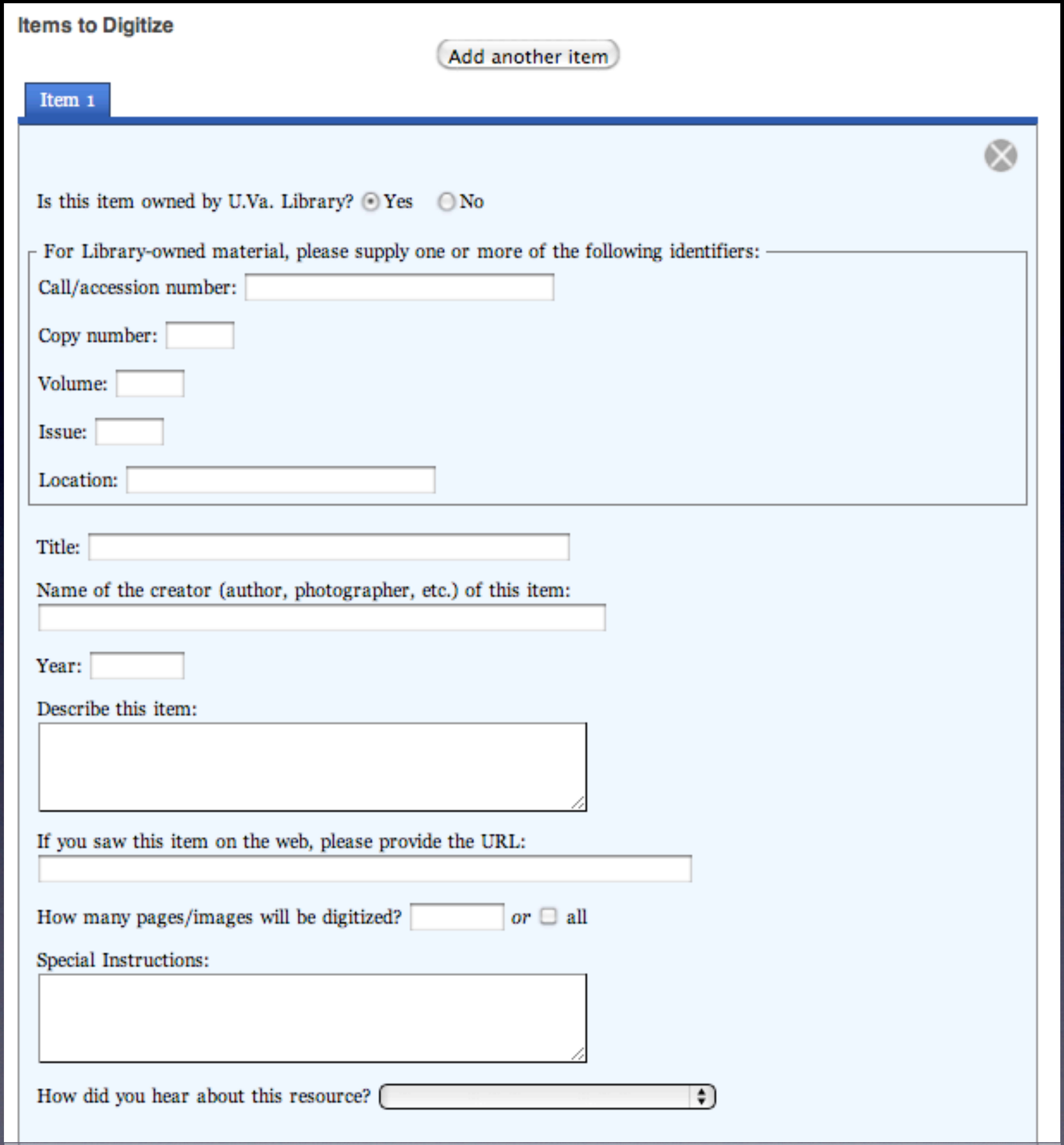
Request Form: Bibliographic Info

Multiple free text fields

Flexible, non-mapped fields

Intended to guide staff in locating material

Based on years of experience handling patron request of Special Collections materials



Items to Digitize

Is this item owned by U.Va. Library? ● Yes ○ No

For Library-owned material, please supply one or more of the following identifiers:

Call/accession number:

Copy number:

Volume:

Issue:

Location:

Title:

Name of the creator (author, photographer, etc.) of this item:

Year:

Describe this item:

If you saw this item on the web, please provide the URL:

How many pages/images will be digitized? *or* □ all

Special Instructions:

How did you hear about this resource?

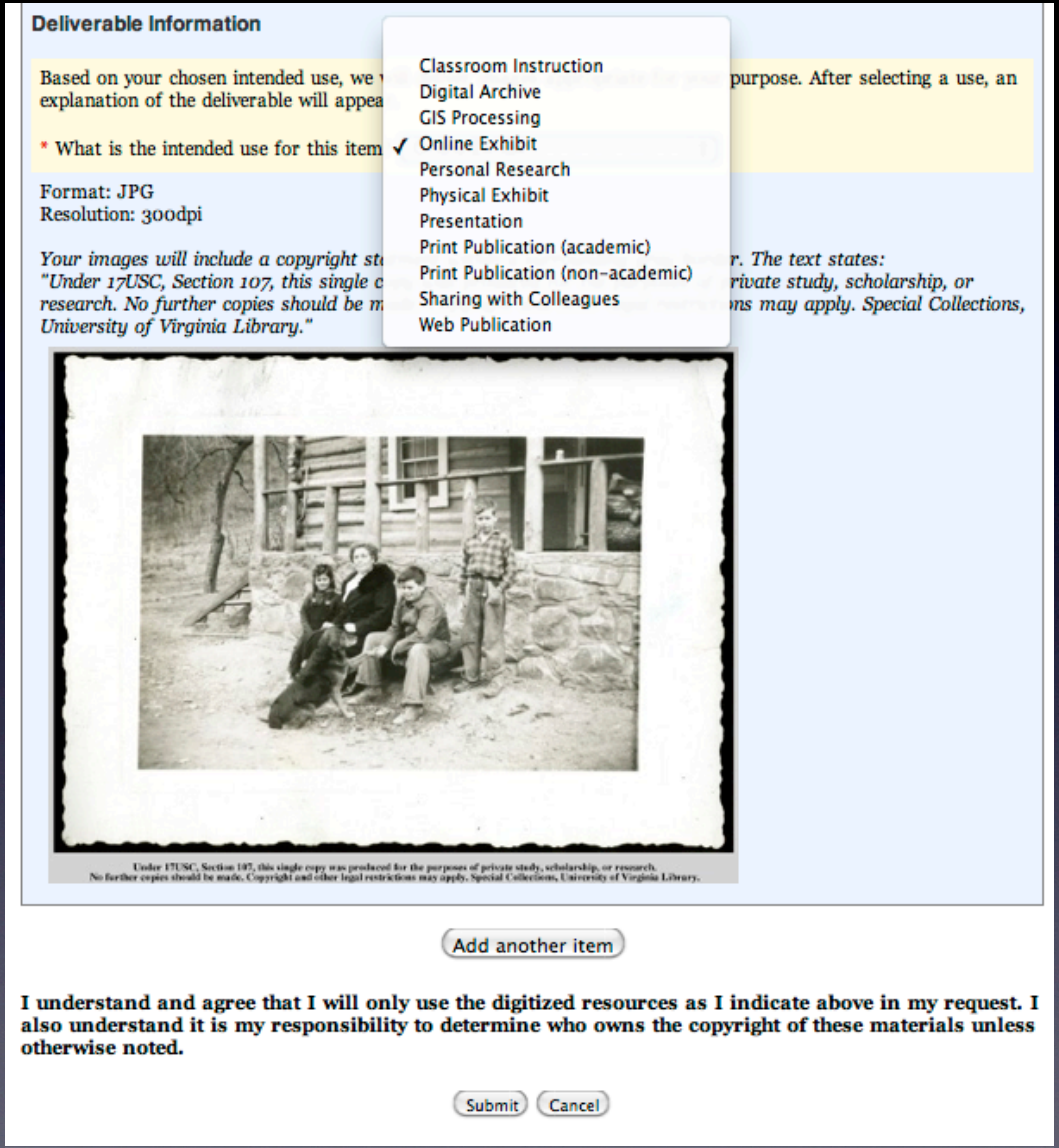
Request Form: Intended Use

Intended uses govern type of deliverable given to patron

Guide user in choosing their deliverable type rather than explicitly asking for technical specifications

Two kinds of deliverables:

- 300dpi watermarked JPEG
- Highest possible dpi TIF



Deliverable Information

Based on your chosen intended use, we w purpose. After selecting a use, an explanation of the deliverable will appea

* What is the intended use for this item

Format: JPG
Resolution: 300dpi

Your images will include a copyright st r. The text states: "Under 17USC, Section 107, this single c rivate study, scholarship, or research. No further copies should be m ns may apply. Special Collections, University of Virginia Library."

Under 17USC, Section 107, this single copy was produced for the purposes of private study, scholarship, or research.
No further copies should be made. Copyright and other legal restrictions may apply. Special Collections, University of Virginia Library.

I understand and agree that I will only use the digitized resources as I indicate above in my request. I also understand it is my responsibility to determine who owns the copyright of these materials unless otherwise noted.

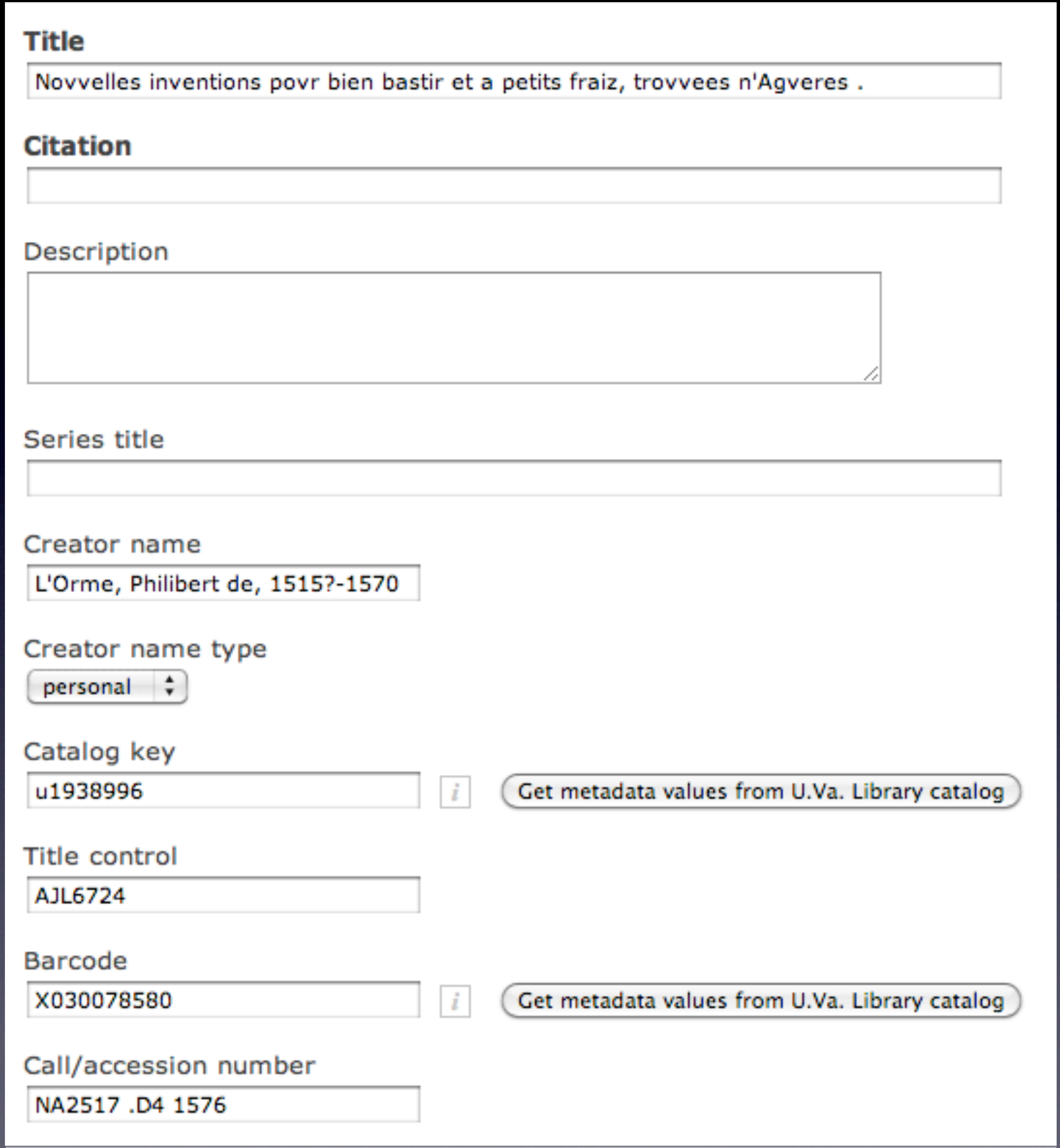
Request Approval: Bibliographic Records

Integration with Blacklight

Association accomplished by catalog key and/or barcode

Barcode is only unique value!

Bibliographic records must be “item” records; local practice is to make “manifestation” records, with “item” information in MARC 999 field



Title

Novvelles inventions povr bien bastir et a petits fraiz, trovvees n'Agveres .

Citation

Description

Series title

Creator name

L'Orme, Philibert de, 1515?-1570

Creator name type

personal

Catalog key

u1938996

Title control

AJL6724

Barcode

X030078580

Call/accession number

NA2517 .D4 1576

Get metadata values from U.Va. Library catalog
Get metadata values from U.Va. Library catalog

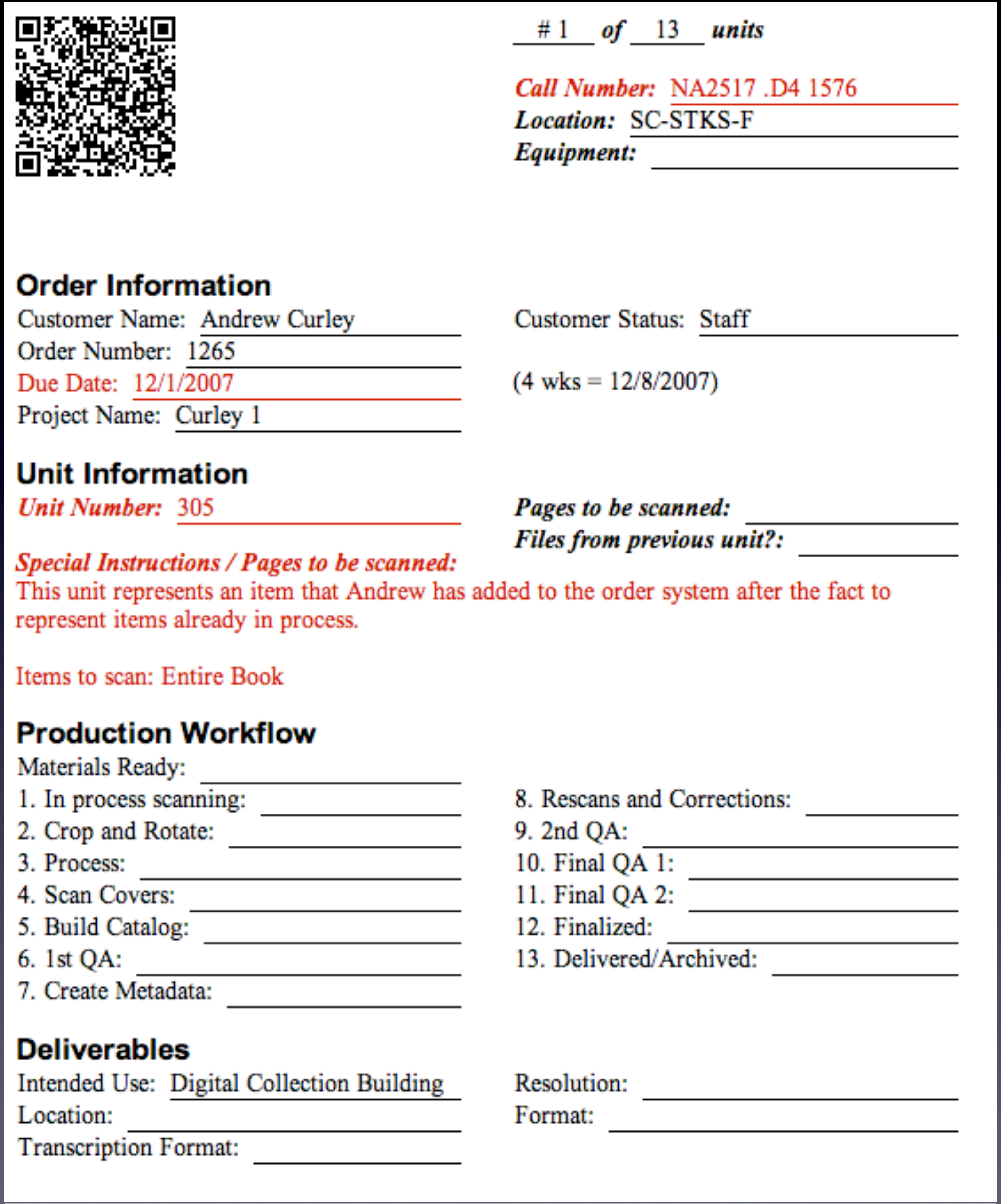
Request Approval: Routing Slip

Printed by staff after order vetting is complete

Provides workflow template and order metadata to production staff

Allows for notes or problems to be passed along throughout production

Allows production coordinators to manage priorities at a glance



# 1 of 13 units

Call Number: NA2517 .D4 1576
Location: SC-STKS-F
Equipment:

Order Information

Customer Name: Andrew Curley
Order Number: 1265
Due Date: 12/1/2007
Project Name: Curley 1

Customer Status: Staff

(4 wks = 12/8/2007)

Unit Information

Unit Number: 305

Pages to be scanned:
Files from previous unit?:

Special Instructions / Pages to be scanned:
This unit represents an item that Andrew has added to the order system after the fact to represent items already in process.

Items to scan: Entire Book

Production Workflow

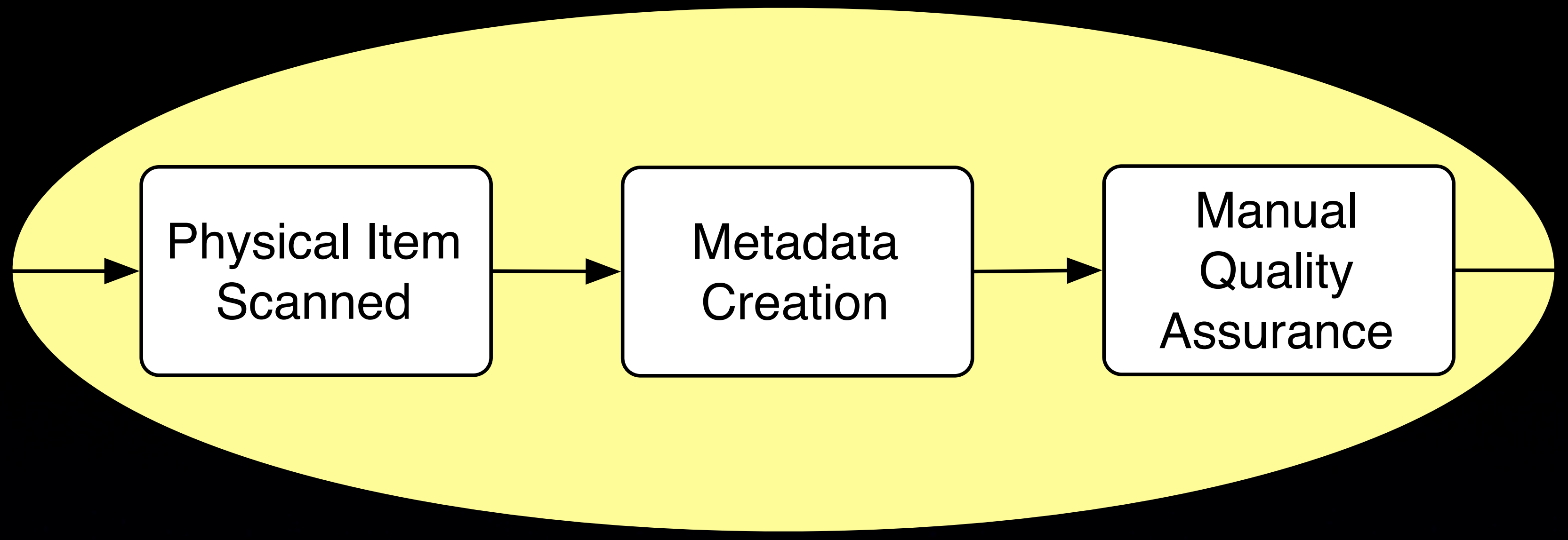
Materials Ready:
1. In process scanning:
2. Crop and Rotate:
3. Process:
4. Scan Covers:
5. Build Catalog:
6. 1st QA:
7. Create Metadata:
8. Rescans and Corrections:
9. 2nd QA:
10. Final QA 1:
11. Final QA 2:
12. Finalized:
13. Delivered/Archived:

Deliverables

Intended Use: Digital Collection Building
Location:
Transcription Format:

Resolution:
Format:

Production Workflow



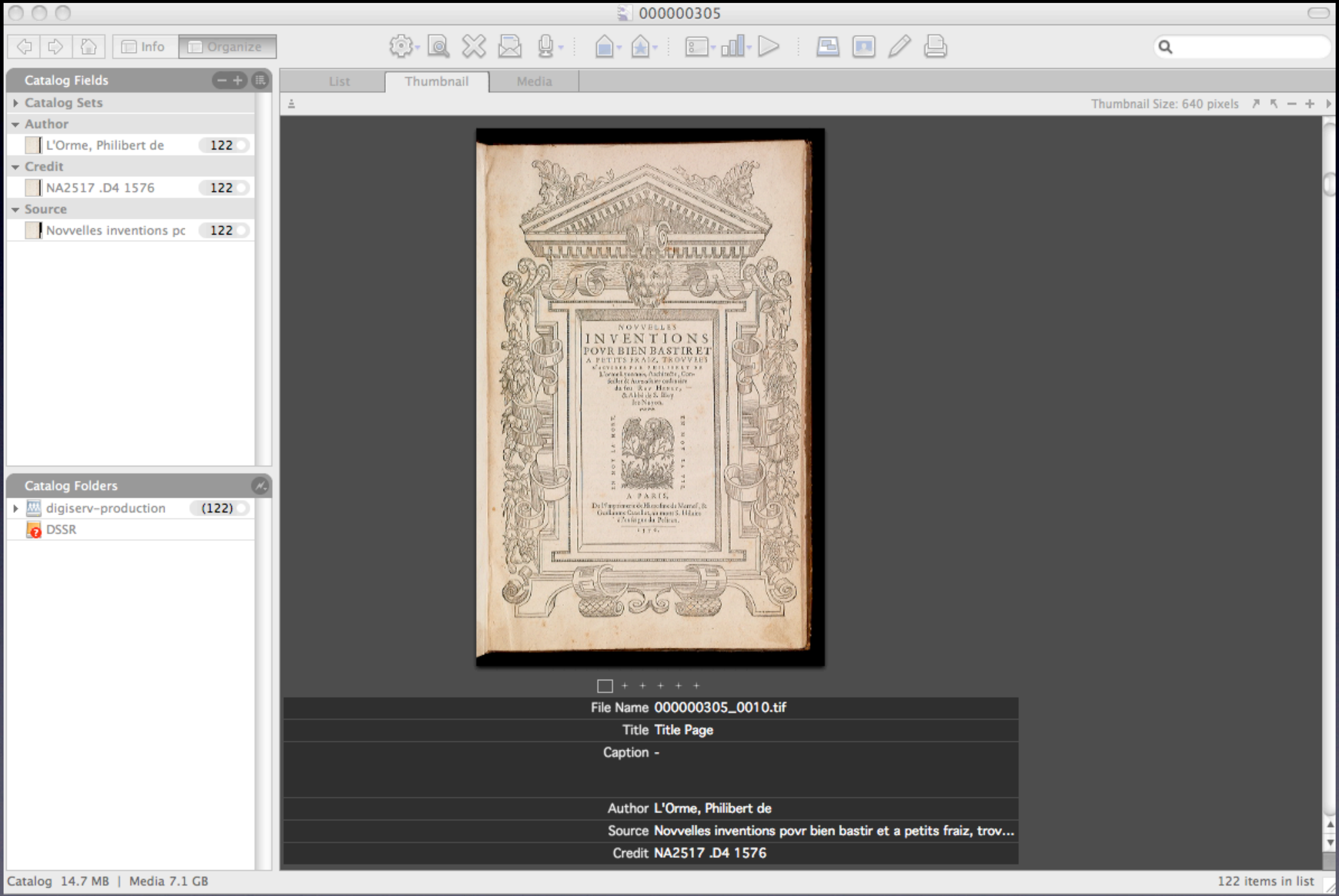
Workflows local to institutions, based on available equipment, computing and software

UVA Digitization Services uses:

- Phase One Capture One DB
- MS Expression Media (soon-to-be Phase One Media Pro, also formerly Iview)

Expression Media used for quality assurance and metadata entry

Production Workflow: Metadata Entry



Author
L'Orme, Philibert de 122

Credit
NA2517 .D4 1576 122

Source
Novvelles inventions pc 122

digiserv-production (122)
DSSR

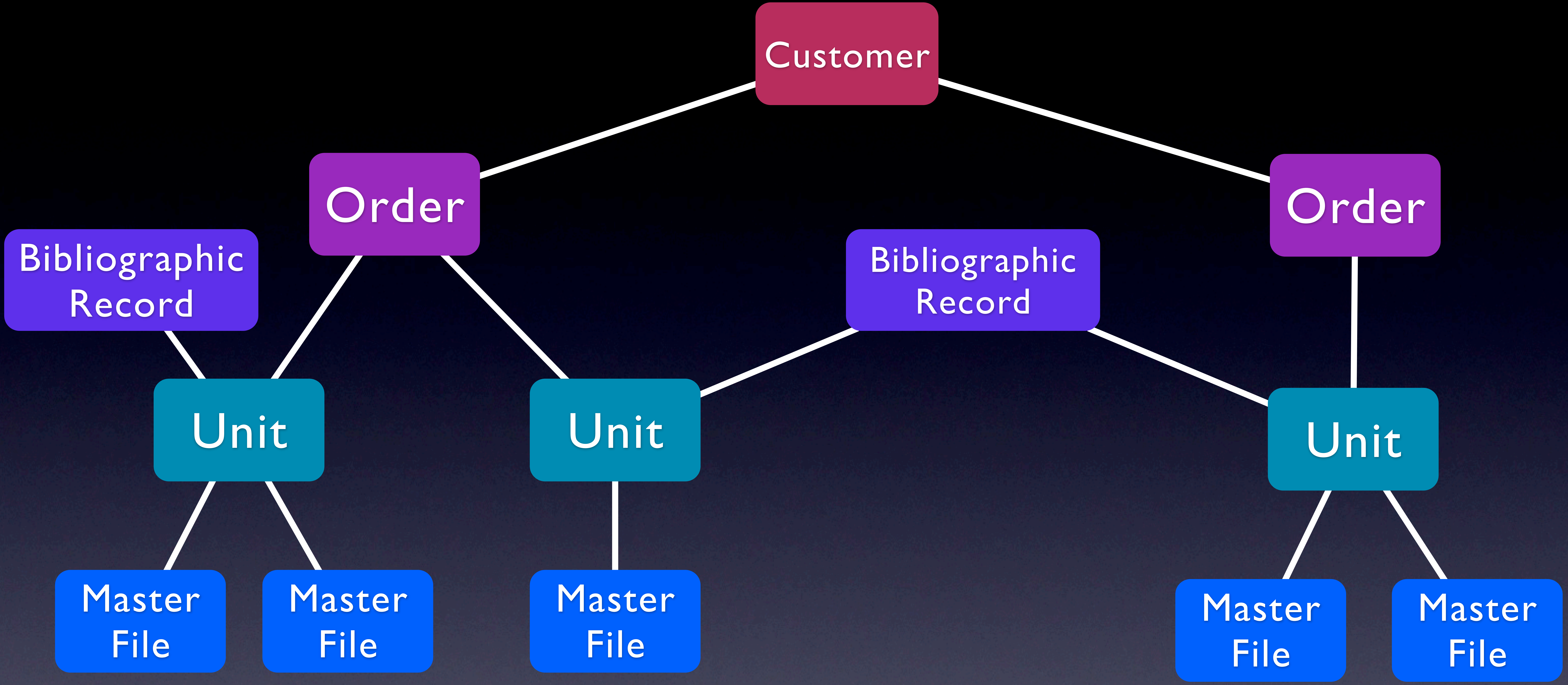
File Name 000000305_0010.tif
Title Title Page
Caption -
Author L'Orme, Philibert de
Source Novvelles inventions povr bien bastir et a petits fraiz, trov...
Credit NA2517 .D4 1576

Catalog 14.7 MB | Media 7.1 GB

122 items in list

Production Workflow: Metadata Export

```
<MediaItem>
        <AssetProperties>
                <Filename>000001483_0134.tif</Filename>
                <Filepath>000001483:000001483_0134.tif</Filepath>
                <UniqueID>427</UniqueID>
                <Label>0</Label>
                <Rating>0</Rating>
                <MediaType>TIFF</MediaType>
                <FileSize unit="Bytes">48510315</FileSize>
                <Created>2008:08:27 15:08:47</Created>
                <Modified>2008:11:13 15:29:34</Modified>
                <Added>2008:08:27 15:18:45</Added>
                <Annotated>2008:11:14 15:55:28</Annotated>
        <ThumbnailSource>000001483_0134.jpg</ThumbnailSource>
        </AssetProperties>
        <MediaProperties>
                <Width unit="Pixels">3320</Width>
                <Height unit="Pixels">4862</Height>
                <Resolution unit="DPI">600</Resolution>
                <Depth unit="Bits">24</Depth>
                <ViewRotation>1</ViewRotation>
                <SampleColor>R:D0 G:D0 B:C0</SampleColor>
                <Pages>1</Pages>
                <ColorSpace>RGB </ColorSpace>
                <Compression>65537</Compression>
                <PrimaryEncoding>TIFF (Uncompressed)</PrimaryEncoding>
                <ColorProfile>Adobe RGB (1998)</ColorProfile>
        </MediaProperties>
        <AnnotationFields>
                <Headline>119</Headline>
                <Author>Phillips, David Graham</Author>
                <Credit>Taylor 1917 .P55 S8</Credit>
                <Source>Susan Lenox, her fall and rise, with a portrait of the
author</Source>
                <Location>SC-STKS</Location>
        </AnnotationFields>
        <MetaDataFields>
                <Maker>Phase One</Maker>
                <Model>P 45+</Model>
                <Software>Adobe Photoshop CS2 Macintosh</Software>
                <SourceURL></SourceURL>
                <ExifVersion>2.2</ExifVersion>
                <CaptureDate>2008:08:22 13:29:23</CaptureDate>
                <ISOSpeedRating>50</ISOSpeedRating>
                <ExposureBias>+0.0</ExposureBias>
                <ExposureTime>1/15</ExposureTime>
                <Aperture>f11.0</Aperture>
                <FocalLength>120.0</FocalLength>
        </MetaDataFields>
</MediaItem>
```



Hydraulics Core Database Layout

http://www.flickr.com/photos/niallkennedy/1780421312/

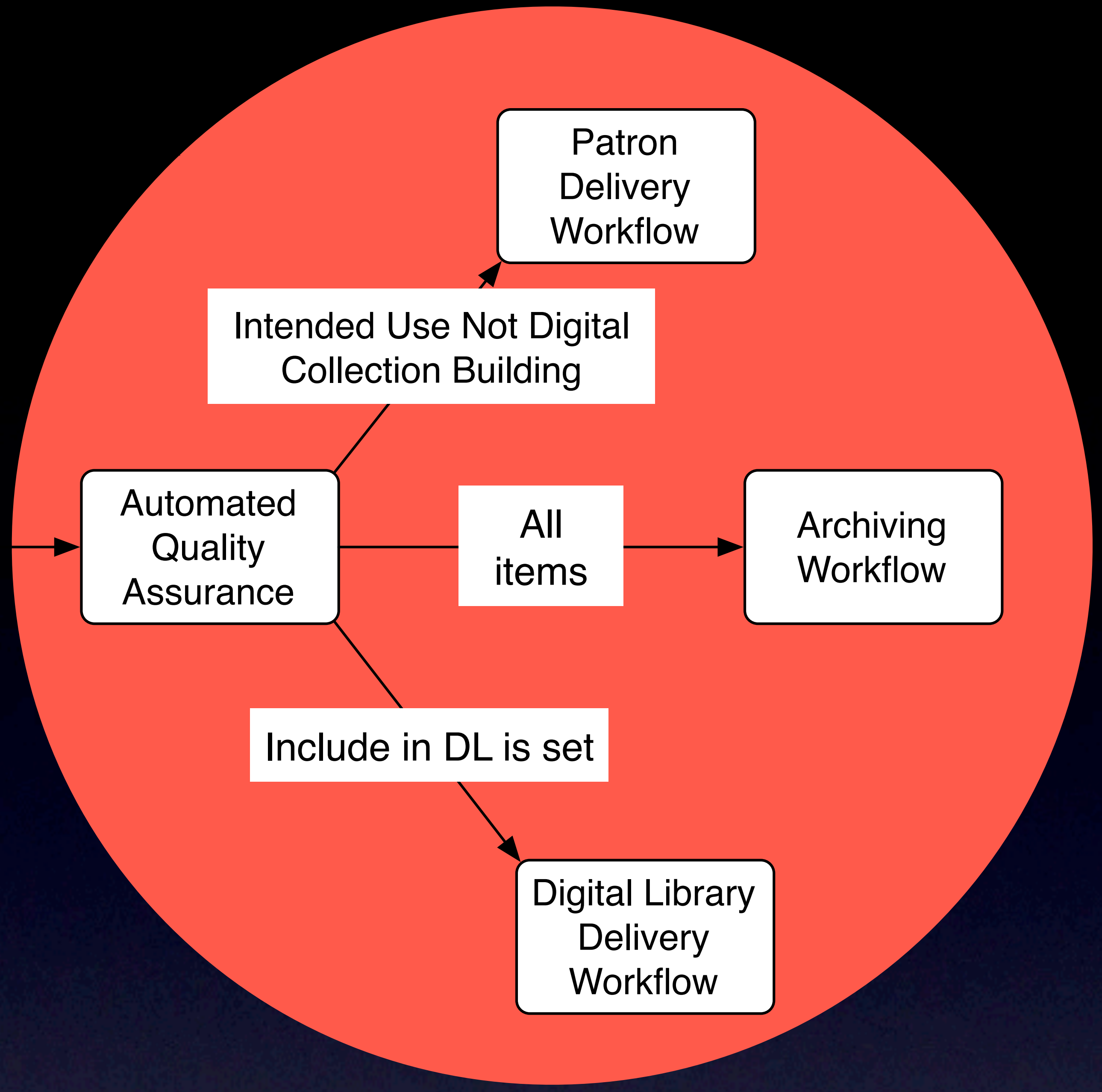
Units As Coffee Beans

Finalization Workflows

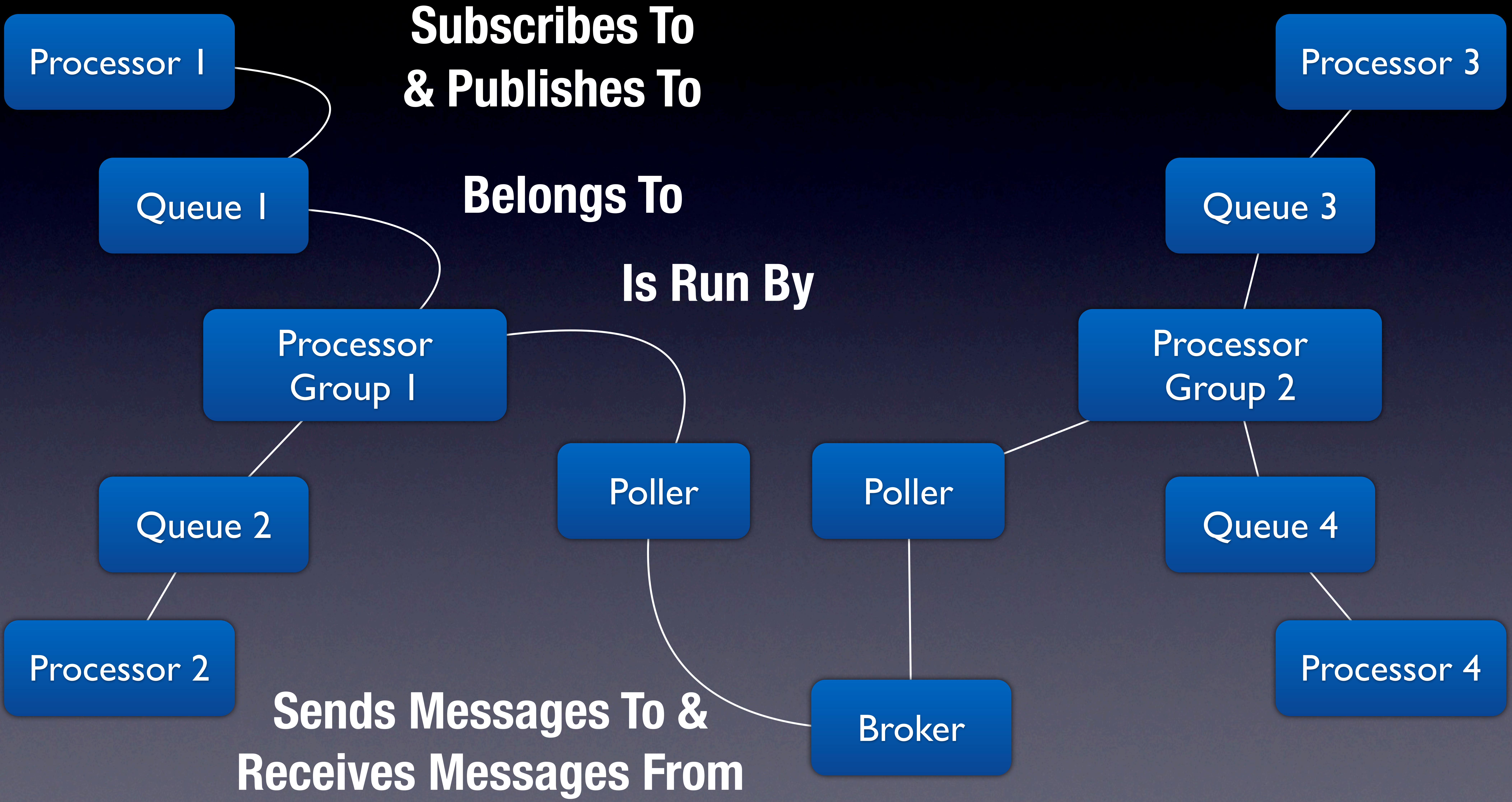
Three distinct workflows:

1. Archiving
2. Patron Delivery
3. Digital Library Delivery

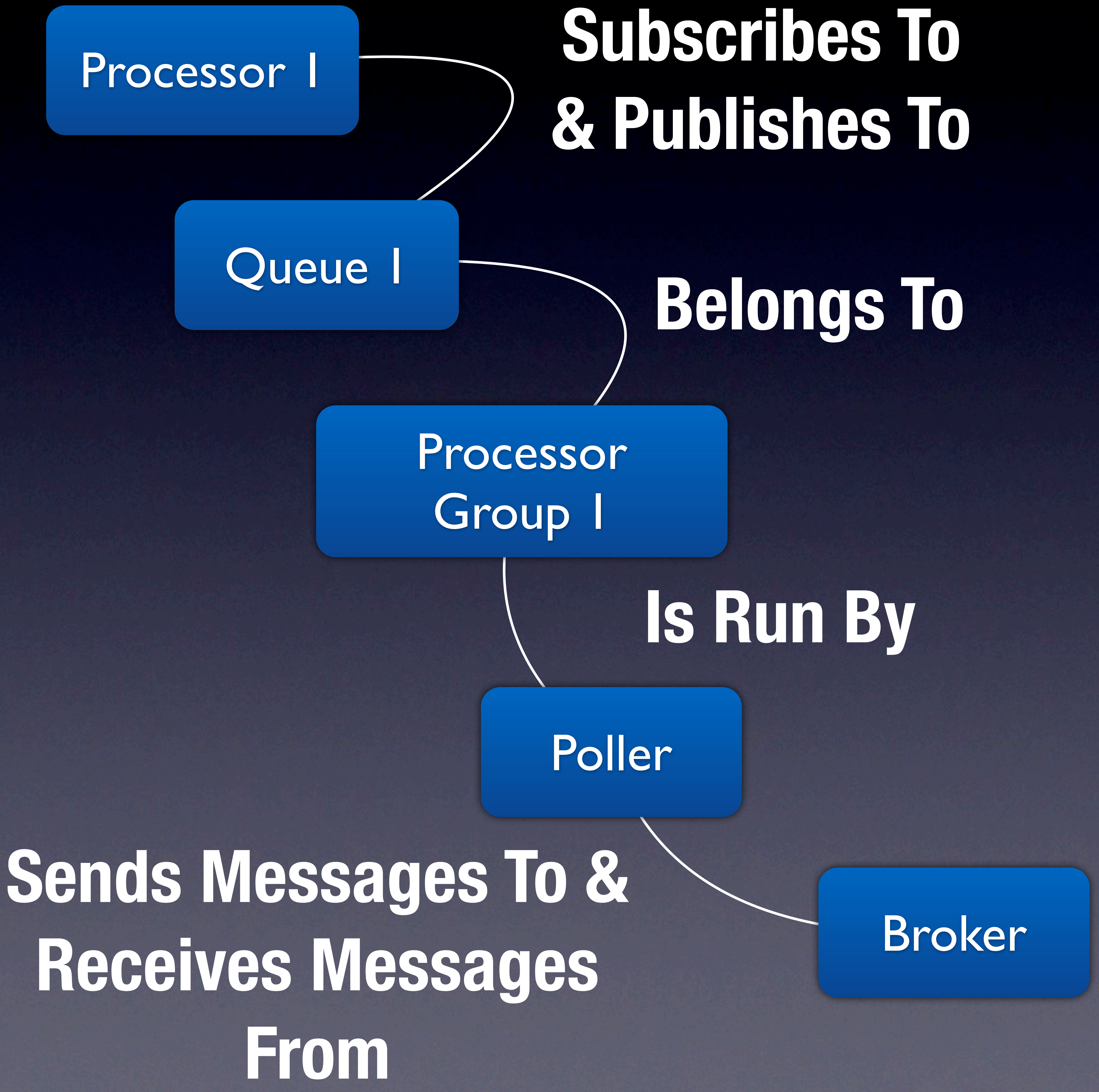
Decisions that govern what workflows each Unit undergoes are data-driven


Patron Delivery Workflow
Intended Use Not Digital Collection Building
Automated Quality Assurance
All items
Archiving Workflow
Include in DL is set
Digital Library Delivery Workflow

Messaging in Ruby on Rails



Messaging in Ruby on Rails



Processor 1

Subscribes To & Publishes To

Queue 1

Belongs To

Processor Group 1

Is Run By

Poller

Sends Messages To & Receives Messages From

Broker

- ActiveMessaging gem
  - Creates Processor class
  - Defines queues in config/messaging.rb
  - Controls processor message handling
- Daemons gem
  - Runs Ruby processes in background
  - Similar to Linux ‘service’ command
- ActiveMQ
  - Messaging broker implementing JMS
  - Message Persistence

Logging Messages in Hydraulics: AutomationMessageClass

Write ‘success’, ‘failure’ and ‘error’ messages to database

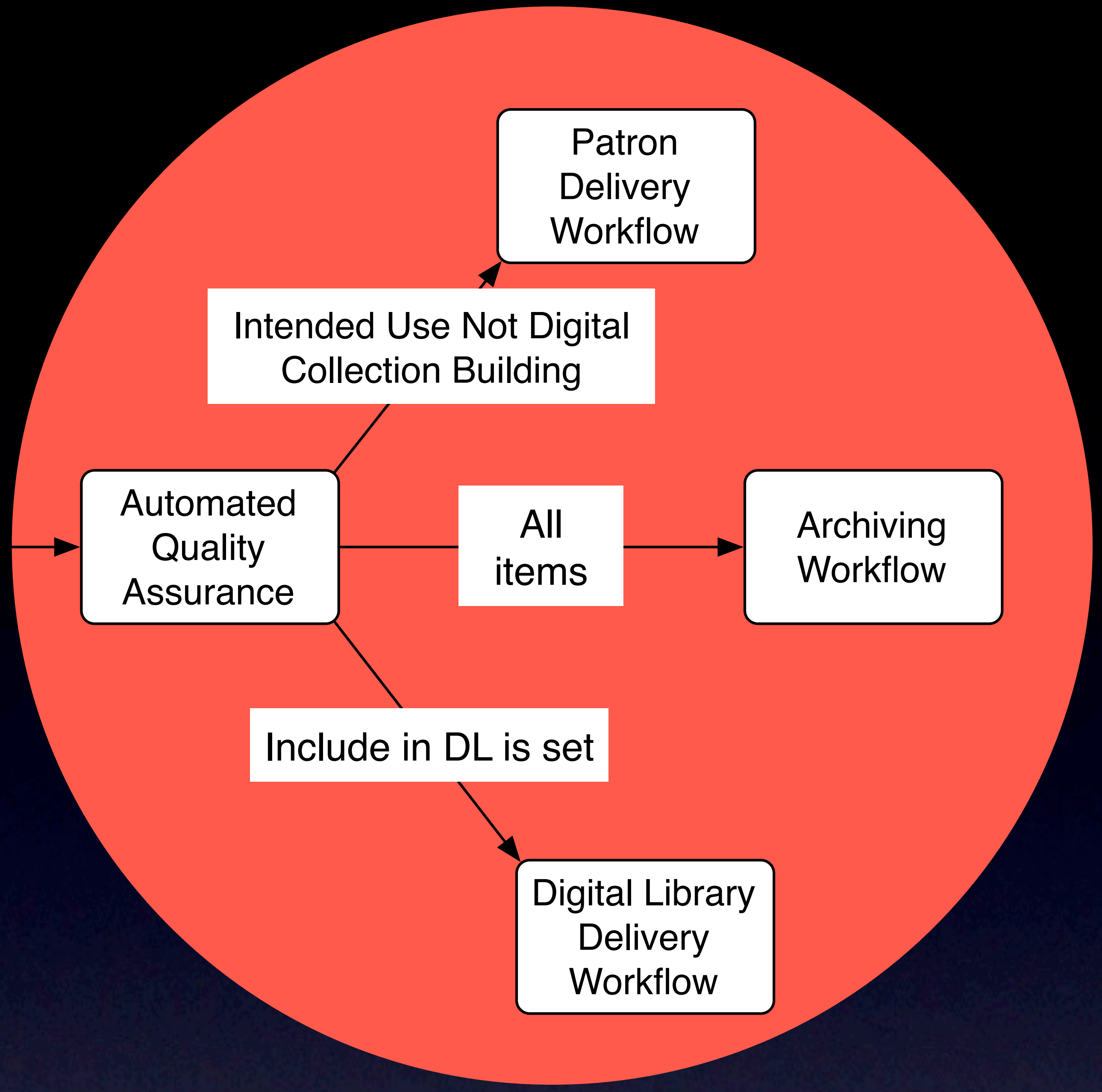
Associated to an Order, Unit, Master File or Bibliographic record

Error messages are flagged; display on administrative home page; contain both stack trace and diagnostic message

Provide persistent audit trail of completed work

Tracksys has processed 2.2 million messages as of 6/10/2011

Finalization Workflows: Archiving


Patron Delivery Workflow
Intended Use Not Digital Collection Building
Automated Quality Assurance
All items
Archiving Workflow
Include in DL is set
Digital Library Delivery Workflow

All Units archived to HSM for long-term preservation

- **Quantum Stornext**

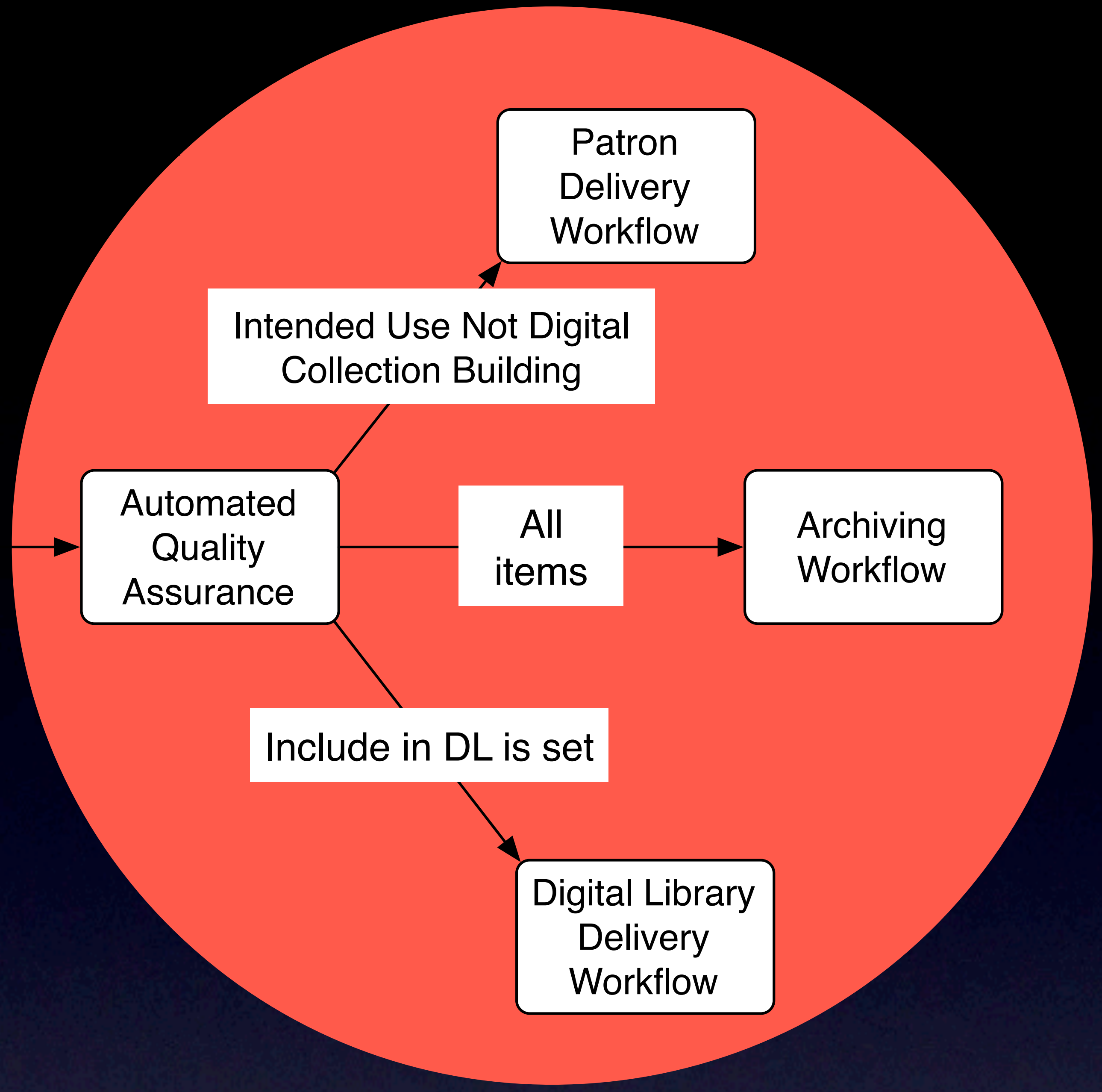
Checksums created

File paths systemized

One-click download of files for future redelivery

Able to ingest content into Fedora repository directly from archive

Finalization Workflows: Patron Delivery


Patron Delivery Workflow
Intended Use Not Digital Collection Building
Automated Quality Assurance
All items
Archiving Workflow
Include in DL is set
Digital Library Delivery Workflow

Deliverable images created on a Unit-by-Unit basis

- Type determined by intended use

Manifest of digital objects and invoice (PDF-format)

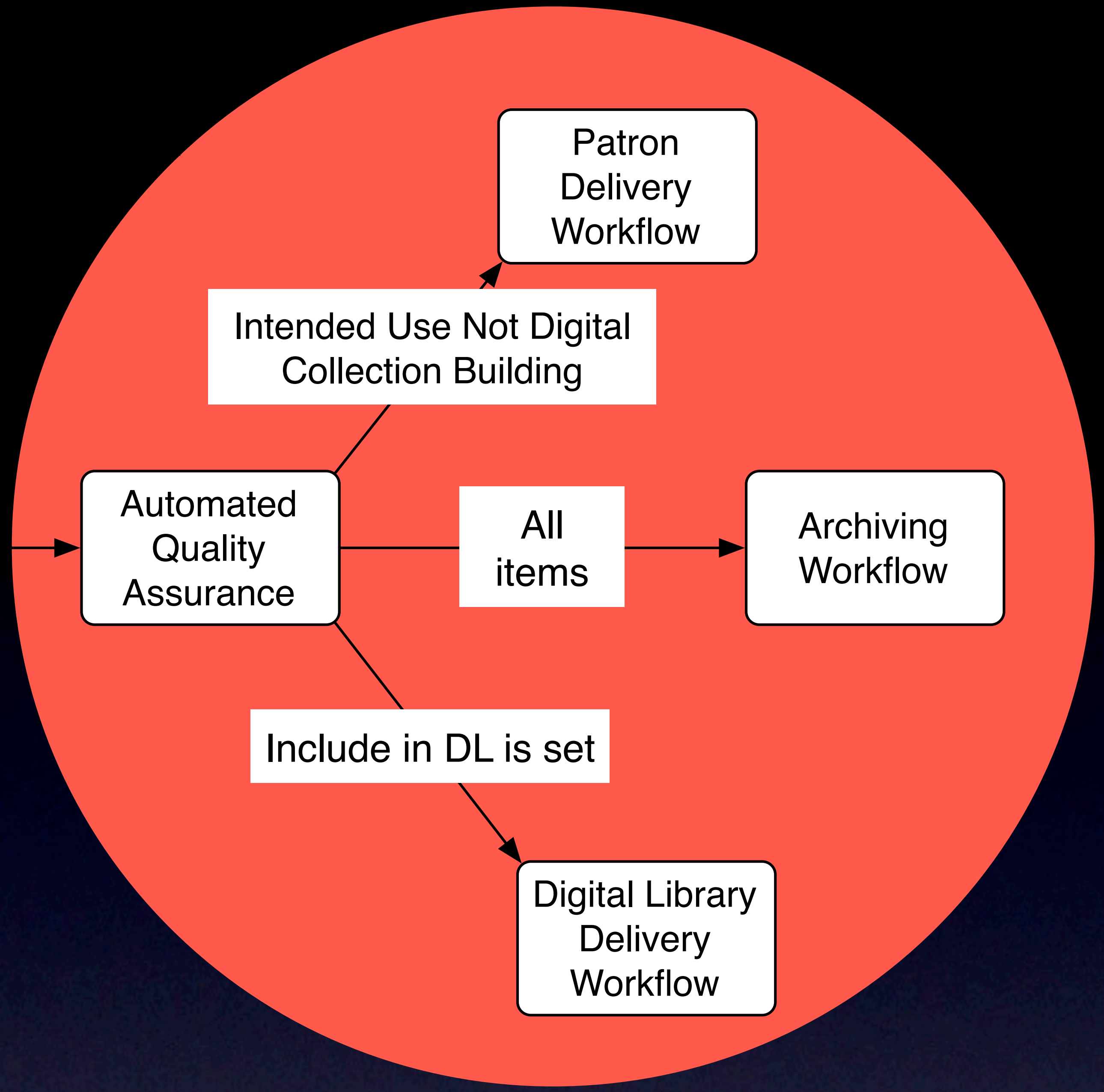
Zip archive of deliverables made web accessible

Email sent to customer providing URL for image pickup

Optional DVD pickup; requires staff override

Automated server cleanup

Finalization Workflows: Digital Library Delivery


Patron Delivery Workflow
Intended Use Not Digital Collection Building
Automated Quality Assurance
All items
Archiving Workflow
Include in DL is set
Digital Library Delivery Workflow

Atomistic model, datastreams match Hydra guidelines

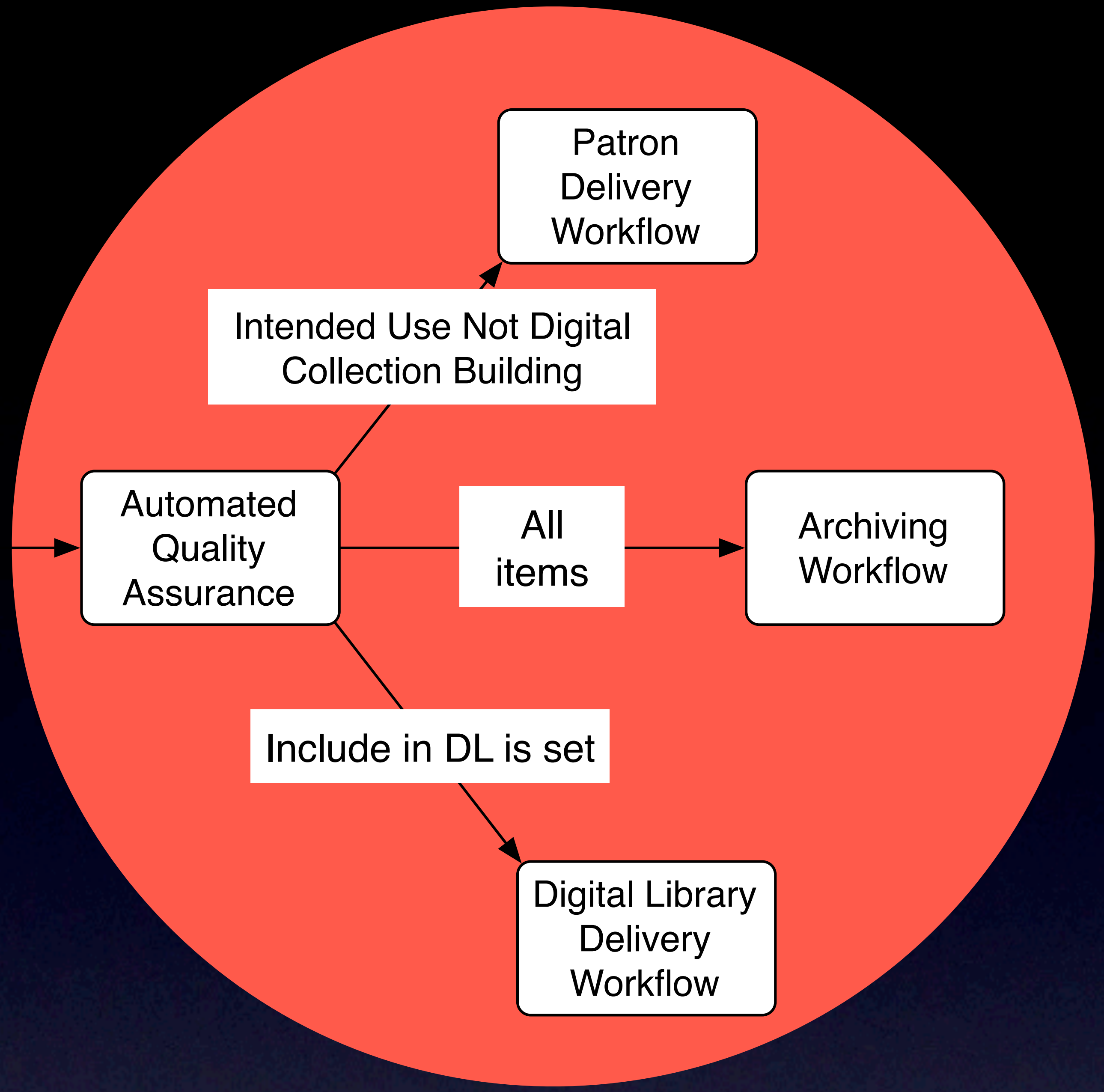
Fedora Objects will be created for:

- **Bibliographic Records**
- **Master File**

Object creation through API calls; No FOXML

Flexible access rights and 'discoverability' set at Unit, inherited by Master File and Bibliographic objects.

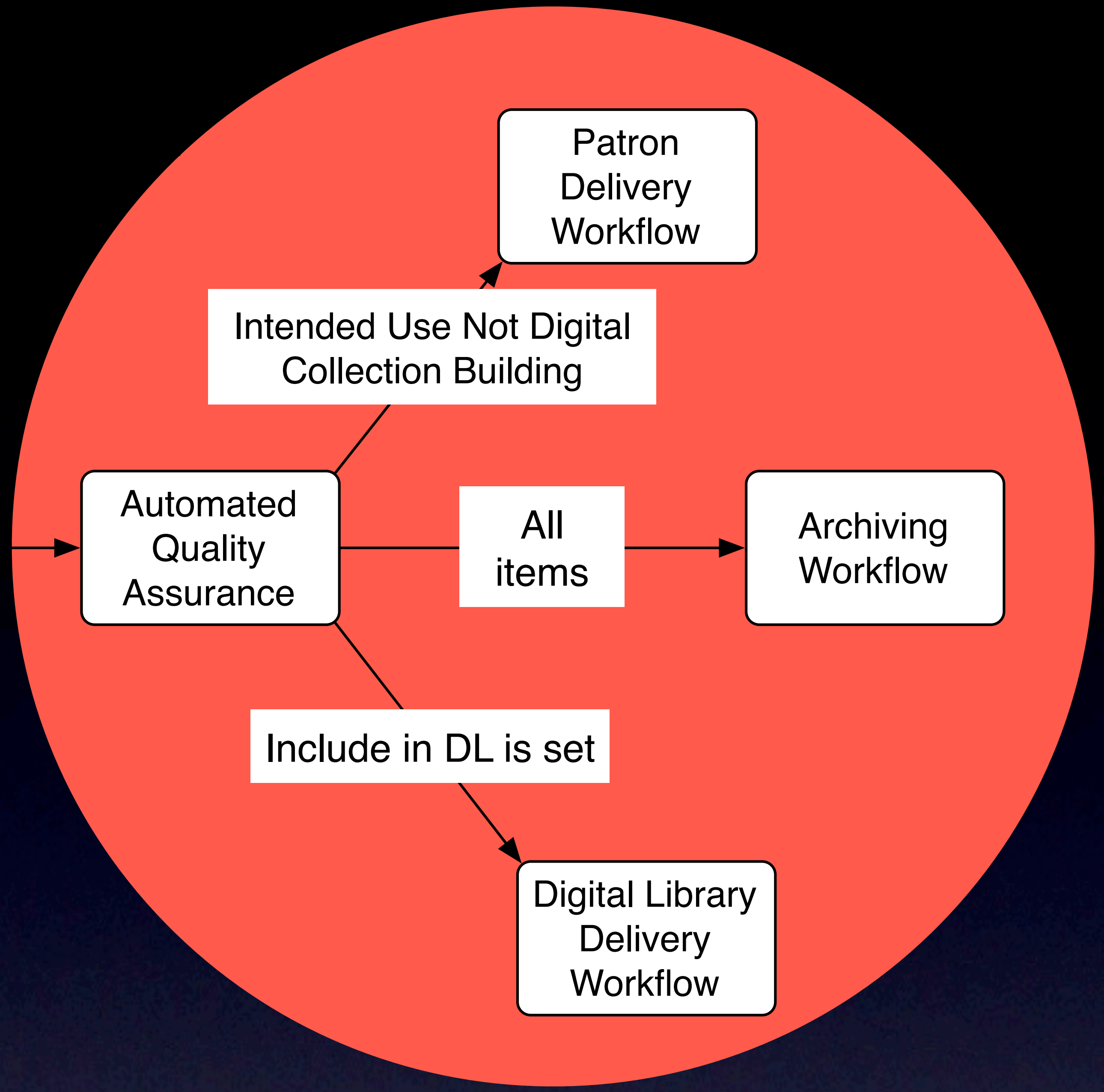
Finalization Workflows: Digital Library Delivery


Patron Delivery Workflow
Intended Use Not Digital Collection Building
Automated Quality Assurance
All items
Archiving Workflow
Include in DL is set
Digital Library Delivery Workflow

Datastreams of bibliographic record object:

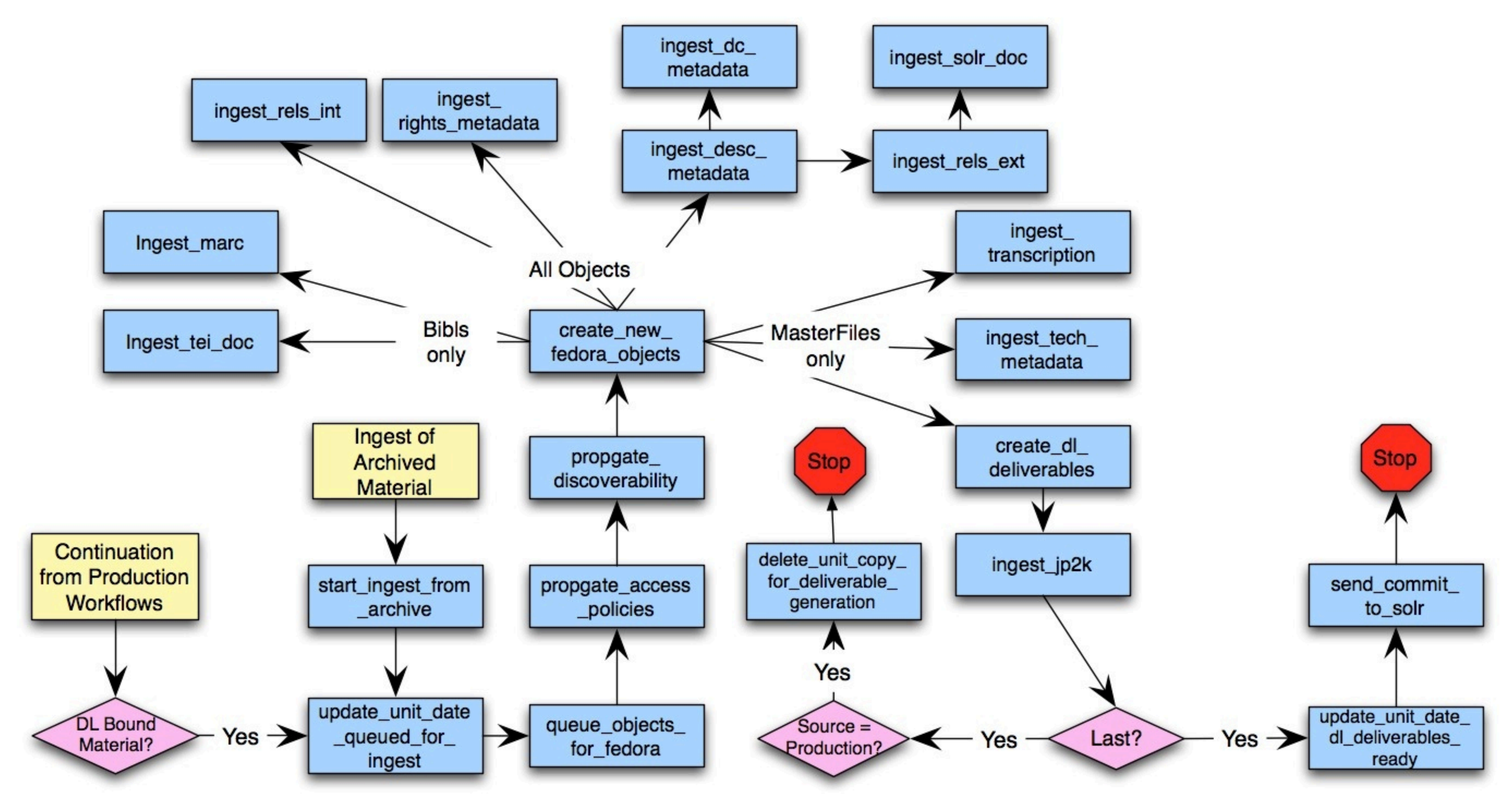
- MARC - External reference to Blacklight MARC XML view
- descMetadata - Transformation of MARC XML to MODS. LC stylesheet used, with local modification.
- DC - Dublin Core generated by MODS-to-DC LC stylesheet
- solrArchive - Record of solr <add> doc. Generated by custom and parameterized MODS-to-Solr XSLT
- POLICY and rightsMetadata - XACML policy
- RELS-EXT and RELS-INT

Finalization Workflows: Digital Library Delivery


Patron Delivery Workflow
Intended Use Not Digital Collection Building
Automated Quality Assurance
All items
Archiving Workflow
Include in DL is set
Digital Library Delivery Workflow

Datastreams of Master File objects

- descMetadata - MODS created from data in Hydraulics
- DC - Dublin Core generated by MODS-to-DC LC stylesheet
- solrArchive - Record of solr <add> doc. Generated by custom and parameterized MODS-to-Solr XSLT
- POLICY and rightsMetadata - XACML policy
- RELS-EXT and RELS-INT
- content - Binary JPEG2000
- techincalMetadata - MIX generated from Hydraulics.



Fedora Ingestion Workflow

Demonstration of Ingestion

Empowering Communities: Production

Automated delivery and quality assurance saves staff time; more focus on digital production

More efficient access to preservation archive

Easier redelivery of already digitized content; less harm to material

Empowering Communities: Selection

Subject specialist, librarians and scholars do the selection of content by placing requests

Demand-driven repository development

Newly accessed collections can be added to digitization queue shortly after cataloging

Flexibility to select content not originally requested for the digital library

Empowering Communities: Object Creation & Maintenance

By brokering all ingestion through a messaging service:

- **Metadata changes are now effectively done with a click of a button**
- **Policy/Access rights can be changed swiftly in response to IP violations or concerns**
- **Poor quality images can be replaced as soon as they are rescanned and re-archived**
- **New architectures can be instantiated quickly**

Additional Functionality

The following topics were not covered in this presentation, but are part of Hydraulics:

1. EAD metadata: Associating Master Files with Component
2. Non-EAD MSS metadata: Associating Master Files with Box/Folder-level information provided by archivists at the time of digitization
3. Associating transcription of images with Master File records; methods of ingesting and indexing this content
4. User access and privileges
5. Extensibility of Master File class beyond TIF images
6. Statistical analysis of production metrics (for you managers out there....)
7. Hand-crafted MODS, RDF, SOLR can be provided to ingestion workflow for all object types

Next Steps

1. Community outreach and engagement

2. Possible integration with the Hydra initiative

3. Redesign code for Rails 3

4. Expand Master Files to handle A/V material

5. Develop and release app as Rails plugin

Andrew Curley

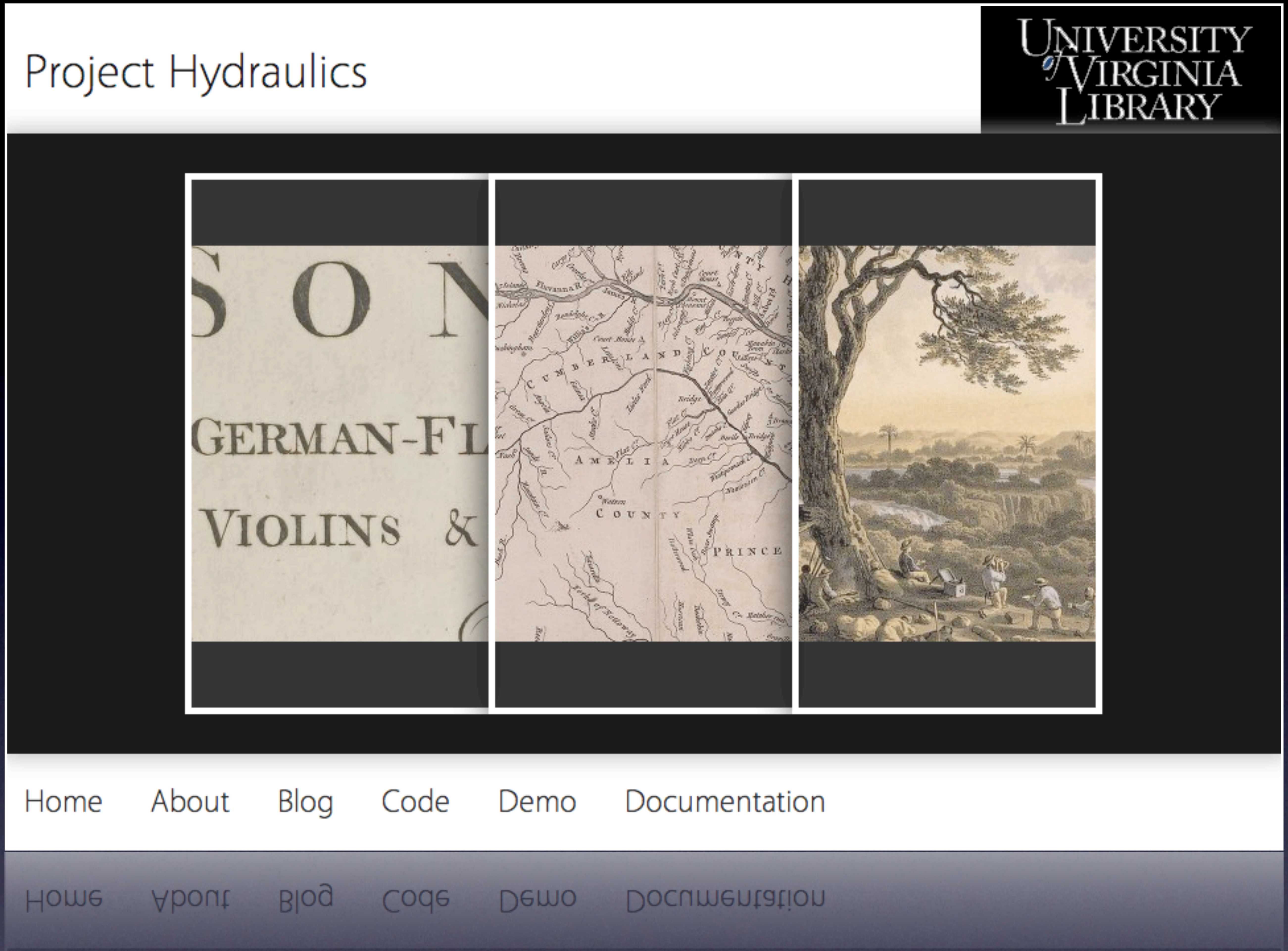
University of Virginia Library

Digital Curation Services

andrew.curley@gmail.com

@andrewcurley

#projecthydraulics



http://projecthydraulics.org/
http://demo.projecthydraulics.org/request
http://demo.projecthydraulics.org/admin

https://github.com/uvalib/hydraulics