Developing an integrated technical infrastructure for archives at NC State

Trevor Thornton
Digital Library Initiatives,
NC State University Libraries

go.ncsu.edu/ltw_forum
Overview

1. Some background

2. A few key projects
   - Collection Guides – finding aids and discovery
   - Circa – request management and tracking
   - Wonda – digitization workflow orchestration

3. Lessons learned and future work
Background - NC State University Libraries

NC State University Libraries
- 2 main libraries & 3 branch libraries serving over 35,000 students
- Over 5.3 million volumes, approx. 123,000 serial subscriptions

Special Collections Research Center
- 27,000 linear feet of records in all formats
- Includes University archives
- Collecting priorities aligned with NC State academic strengths

Digital Library Initiatives
- 9-member team developing tools to facilitate research and support library services

NC State Libraries IT
- 22+ staff supporting core discovery systems, high-tech learning spaces, all patron and staff computing, and enterprise operations
Your friendly archives management tool

- We implemented in 2014, migrating from Archivists’ Toolkit
- Member institution since 2015
- Contributions to core code and documentation
- Currently on version 2.5, update to 2.7 in progress
- We made a Ruby gem to facilitate communication with backend API: github.com/NCSU-Libraries/archivesspace-api-utility

International Image Interoperability Framework™

- NC State University Libraries is a full founding member of the IIIF Consortium
- Image API implemented in 2015, later incorporated into our own image server (Eyebright)
- Presentation API manifests available via Libraries’ Digital Collections
- Content Search API implemented by Ocracoke (our OCR processing application)
- Currently developing tools to work with IIIF annotation data
Functional overview

Collections

Collection management and description (ArchivesSpace)

Digitization

Publication (Web)

Born-digital records processing

Digital Preservation

Access

Discovery

Researchers
Collection Guides

Finding aids and discovery

- Ruby on Rails, MySQL/MariaDB, Solr
- Data imported directly from ArchivesSpace via its API, with ongoing, automated updates
- Digital images presented in context via IIIF APIs

lib.ncsu.edu/findingaids
Automated updates from ArchivesSpace

1. Hourly query of ArchivesSpace index returns new or updated resources with finding aid status ‘completed’

2. Resource tree data for new/updated resources (including linked subjects, agents and digital objects) is retrieved, indexed, converted to a form optimized for display and stored in the database

3. New/updated finding aids are published
Adding images to finding aids via IIIF

1. Collection Guides finds digital objects with URLs pointing to NCSU Digital Collections

2. URL appended with '/manifest' and used to retrieve IIIF presentation manifest from digital collections

3. Data from manifest is used to retrieve thumbnail images, displayed in the finding aid

NCSU Digital Collections  →  Collection Guides  →  URL appended with '/manifest'  →  IIIF presentation manifest from digital collections  →  Image server  →  Data from manifest is used to retrieve thumbnail images, displayed in the finding aid
Circa

Request management and tracking

- All types of requests (independent research, course reserves, reproduction, processing, loans, etc.)
- Tracking transfers of materials between locations and use by researchers
- Detailed usage data and reporting

Open source: github.com/NCSU-Libraries/circa

Customization hooks allow implementers to extend functionality
Requesting materials from ArchivesSpace in Circa

1. New request (order) is created
2. User copies archival object URI from ArchivesSpace, pastes URI into Circa
3. Data is retrieved from ArchivesSpace API
4. Top container (e.g. box) is added to the request
5. Container and location information is imported for tracking
Requests for images published online

1. User creates new reproduction order and selects the option to add items from Digital Collections
2. User copies URL of the resources found in Digital Collections
3. User pastes URL into Circa and IIIF manifest is retrieved
4. User optionally selects specific images from a multi-page resource
5. Image identifiers are added to order, enabling staff to locate and retrieve masters for delivery
Wonda

Workflow orchestration for digitization

- Automates publication and asset management via API integration with other services
  - JP2 (processing JPG2000 derivatives)
  - Ocracoke (OCR)
  - SCPS (Preservation DAMS)
  - Digital Collections staff interface (Publication)
  - ArchivesSpace (again)

- Implementation agnostic – services can be added/replaced with minimal configuration
ArchivesSpace browser

Wonda provides an integrated way to select an archival object record from ArchivesSpace

1. Search to find the parent resource
2. Browse the resource hierarchy to find the record you need
3. Select the record

Available as a Rails Engine gem - configure it to do whatever you need with the data after you select the record

github.com/NCSU-Libraries/archivesspace_browser_rails

Developed in conjunction with 'The Animal Turn: Digitizing Animal Protection and Human-Animal Studies Collections', a three-year project in partnership with the ASPCA with funding from the Council on Library and Information Resources (CLIR)
Shared storage infrastructure

Desksops (Mac, Windows, Linux)

- Wonda
- JP2 Service
- Preservation DAMS
- Image server

VMware vSphere (virtual servers)

working storage
preservation staging
access images

Dell Compellent tiered storage array

virtual file servers
Digitization and publication workflow

User initiates new digitization project in Wonda

Working storage (TIFFs) -> Wonda

OCR

Image server

access storage (JPG2000s)

Digital Collections

Digital Collections (staff) -> Digital Collections (public)

Preservation DAMS

Collection Guides

ArchivesSpace
Digitization and publication workflow

User links new project to an archival object record in ArchivesSpace, descriptive metadata is retrieved.

Working storage (TIFFs) → JP2 → OCR → Image server → Access storage (JPG2000s) → Digital Collections Solr index

Wonda → Digital Collections (staff) → Digital Collections (public)

Preservation DAMS → ArchivesSpace → Collection Guides
Digitization and publication workflow
Digitization and publication workflow
Digitization and publication workflow

User notifies Wonda that scanning and QC are complete, initiates publication process

- Working storage (TIFFs)
- JP2
- OCR
- Image server
  - Access storage (JPG2000s)
  - Digital Collections Solr index
- Digital Collections (staff)
- Digital Collections (public)
- Preservation storage
- DAMS
- ArchivesSpace
- Collection Guides
Digitization and publication workflow

Wonda sends request to JP2 to initiate derivatives processing.
Digitization and publication workflow

JP2 generates JPEG2000 derivatives created from TIFF masters in working storage...

...and saves them to access storage

Image server

access storage (JPEG2000s)

Digital Collections Solr index

Digital Collections (public)

Digital Collections (staff)

preservation storage

Preservation DAMS

Wonda

Collection Guides

ArchivesSpace

ArchivesSpace
Digitization and publication workflow

JP2 notifies Wonda when derivatives processing is complete.
Digitization and publication workflow

Wonda sends request to Ocracoke to initiate OCR processing.
Digitization and publication workflow
Digitization and publication workflow
Digitization and publication workflow

- Wonda sends request to Digital Collections to publish the resource
- Image server
  - access storage (JPG2000s)
  - Digital Collections Solr index
- Digital Collections (staff)
- Digital Collections (public)
- Preservation DAMS
- Working storage (TIFFs)
- JP2
- OCR
- Preservation storage
- Collection Guides
Digitization and publication workflow
Digitization and publication workflow
Digitization and publication workflow

- Working storage (TIFFs)
- JP2
- OCR
- Image server
- Access storage (JPG2000s)
- Digital Collections Solr index
- Digital Collections (staff)
- Digital Collections (public)
- Preservation DAMS
- ArchivesSpace
- Wonda
- Collection Guides

Digital objects created in ArchivesSpace (one for the public resource, one for the preserved masters)
Digitization and publication workflow
Wonda impact so far

Since launching in mid-November 2019:

- 164 new resources published online (mostly as part of a CLIR digitization grant project)
- 123 of those included OCR processing, and of those:
  - Average pages/files per resource: 79
  - Average time to publication (after scanning and metadata complete): 25 minutes
Some things we’re working on (now or soon)

- Workflow orchestration for born-digital processing
- Integrating Wonda and Circa for reproduction requests
- New AV service for Wonda
- Universal Viewer embedded into Collection Guides
- Investigating patron-initiated Circa requests from Collection Guides
A few design principles/lessons learned

- Separation of concerns – smaller, specialized services focused on distinct areas of work are easier to build and maintain
- Make core functionality available via API requests to support integration
- Modularity – services can be replaced or updated provided API functionality remains consistent
- Document APIs and write automated tests to ensure changes don’t break integrations
- Shared storage makes things easier (but can be hard to implement)
Links to some things we’ve made

Projects on the web:
Collection Guides: lib.ncsu.edu/findingaids
Digital Collections: d.lib.ncsu.edu/collections
‘The Animal Turn’: Digitizing Animal Protection and Human-Animal Studies Collections
(CLIR-funded digitization project with the ASPCA): lib.ncsu.edu/animal-turn

Github:
Circa: github.com/NCSU-Libraries/circa
Ocracoke (OCR): github.com/NCSU-Libraries/ocracoke
Eyebright (image server): github.com/NCSU-Libraries/eyebright
ArchivesSpace API utility (Ruby gem): github.com/NCSU-Libraries/archivesspace-api-utility
ArchivesSpace browser (Rails engine): github.com/NCSU-Libraries/archivesspace_browser_rails
Annona (Javascript tools for IIIF annotations): github.com/ncsu-libraries/annona
Thank you and good luck!
Trevor Thornton
trthorn2@ncsu.edu

Questions about IIIF:
Jason Ronallo – jnronall@ncsu.edu
Niqui O’Neil (annotations) – doneill@ncsu.edu

Questions about storage:
Mike Kastellec – makastel@ncsu.edu

Slides: go.ncsu.edu/ltw_forum