Migrating the Carolina Digital Repository to Hyrax

Anna Goslen, Rebekah Kati & Jennifer Smith

From February 2019 through May 2019, UNC-Chapel Hill Libraries migrated our institutional repository content from custom Fedora to Hyrax. Here are some highlights:

Challenges

- Permissive MODS → RDF
- Merging Masters Papers from 10 worktypes to 1
- Ongoing submissions
- Collection-based model to worktypes

By the numbers

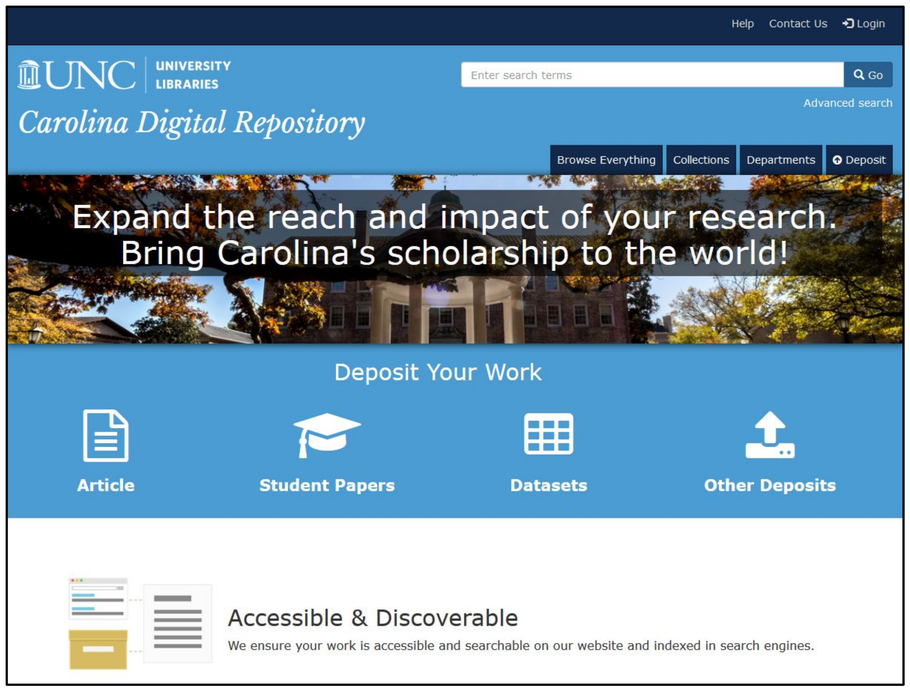
- 28K Works
- 132K Objects
- March 19 - May 30, 2019
- 18 Collections into 9 Worktypes
- 14 admin sets

Metadata

- 67 fields across 10 worktypes
- Assessment: OpenRefine, XPath
- Remediation: Regex, XSLT
- Overlay: Batch export and import in source system
- Create 9 worktype application profiles
- Migrate PREMIS and original MODS as private files
- Add Resource Types to existing objects
- Add RightsStatements.org

Order of migration

- Collection size
- Active deposits
- Stakeholder input needed



Fedora Collections	Hyrax Worktypes
3D Images, Images, Audio, Video	Multimedia
Articles	OA Articles
Data	Dataset
ETDs	Dissertation (hidden)
Honors Theses	Honors Theses
Journals, Book	Journal/Book
Masters Paper (10 worktypes)	Masters Paper (1 worktype)
Poster, Presentation, White Paper, Report	Poster, Presentation, or Paper
Everything else	General (hidden)

QA process

- Spot check files, based on size of collection
  - Metadata migrated as expected
  - Visibility settings
  - Admin set membership
  - Worktype assignment
- Check off on spreadsheet

Trouble spots

- Solr crashes due to full text extraction on 1GB+ files
- Thumbnail generation
- Content edge cases:
  - Hog Data (1 work w/2040 files)
  - Special characters in abstracts
  - Unique visibility settings
- Metadata
  - Multiple date fields/formats
  - Legacy data missing required fields
  - Keyword breaks

Lessons learned

- Data will not be uniform
- Build in a lot of QA time
- Anticipate multiple passes
- Manual remediation will be necessary