IMPROVING DATA CURATION SERVICES WITHIN AN INSTITUTIONAL REPOSITORY

UNC Chapel Hill Libraries
Julie Rudder - Repository Program Librarian
Rebekah Kati - Institutional Repository Librarian
Overview

• Introduction
• Levels of Data Curation Activities Exercise
• Using the Levels of Data Curation at UNC
• Hyrax Audit for Data
• Next Steps
Hyrax: An open-source, Samvera-powered repository front-end

Community Support
Hyrax is maintained and supported by the Samvera community. Samvera is an open-source repository solution built

Flexible Workflow
Hyrax supports multiple workflows including the ability to define custom workflows, allowing implementations to

Fully Featured Solution
Hyrax offers a repository solution that can meet the needs of institutional/data repositories and digital object

http://hyr.ax/
Why research data curation, why now?

• Support FAIR data, Open Science
• Support faculty goals with grants and data sharing
• OA policy (resources and renewed interest in data)
• New staff
Sure, we do data curation! Specific & shared understanding
Triangle Research Libraries Network (TRLN) Institute

- [https://osf.io/preprints/lissa/zj5pq/](https://osf.io/preprints/lissa/zj5pq/)

Data Curation Team at Duke University

- **Digital Repository Content Analysts:** Moira Downey, Susan Ivey
- **Senior Research Data Management Consultants:** Sophia Lafferty-Hess, Jennifer Darragh
Goals for TRLN Institute

• provide more specificity around data curation within our individual contexts
• determine a method to discuss our service model
• identify gaps we would like to fill
• determine what is currently out of scope for our repositories.
“respondents conflated data curation activities with research data management services...this indicates that a common understanding of data curation is not widespread or ubiquitous”

(Hudson-Vitale et al., 2017)
A Definition for Research Data Curation

“the encompassing work and actions taken by curators of a data repository in order to provide meaningful and enduring access to data.”

Date Curation Network (Johnston et al., 2016).
For full definitions of data curation activities see the Data Curation Network: Data Curation Terms and Activities.

<table>
<thead>
<tr>
<th>Level 1 Curation</th>
<th>Level 2 Curation</th>
<th>Level 3 Curation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Ingest</strong></td>
<td><strong>Appraisal</strong></td>
<td><strong>Appraisal</strong></td>
</tr>
<tr>
<td>• Authentication</td>
<td>• Rights Management (DUAs)</td>
<td>• Risk Management (remediation)</td>
</tr>
<tr>
<td>• Chain of Custody</td>
<td>• Risk Management (file review)</td>
<td></td>
</tr>
<tr>
<td>• Deposit Agreement</td>
<td>• Selection</td>
<td></td>
</tr>
<tr>
<td>• Documentation</td>
<td><strong>Curate</strong></td>
<td><strong>Curate</strong></td>
</tr>
<tr>
<td>• File Validation</td>
<td>• Contextualize</td>
<td>• Code Review</td>
</tr>
<tr>
<td>• Metadata</td>
<td>• Curation Log</td>
<td>• Conversion (Analog)</td>
</tr>
<tr>
<td><strong>Appraisal</strong></td>
<td>• File Format Transformations</td>
<td>• Data Cleaning</td>
</tr>
<tr>
<td>• Rights Management (licenses)</td>
<td>• File Renaming</td>
<td>• De-identification</td>
</tr>
<tr>
<td><strong>Curate</strong></td>
<td>• Quality Assurance</td>
<td>• Interoperability</td>
</tr>
<tr>
<td>• Arrangement &amp; Description</td>
<td>• Restructure</td>
<td>• Peer Review</td>
</tr>
<tr>
<td>• File Inventory or Manifest</td>
<td><strong>Access</strong></td>
<td>• Quality Assurance</td>
</tr>
<tr>
<td>• Indexing</td>
<td>• Restricted Access (mediated requests)</td>
<td></td>
</tr>
<tr>
<td>• Persistent ID</td>
<td><strong>Preservation</strong></td>
<td><strong>Access</strong></td>
</tr>
<tr>
<td><strong>Access</strong></td>
<td>• Repository Certification</td>
<td>• Data Visualization</td>
</tr>
<tr>
<td>• Contact Information</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Data Citation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Discovery Services</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Embargo</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• File Download</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Full Text-Indexing</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Metadata Brokerage</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Terms of Use</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Use Analytics</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Restricted Access (system automated)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Preservation</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• File Audit</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Migration</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Secure Storage</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Versioning</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Succession Planning</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Tech/Monitoring Refresh</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Cease Data Curation</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Level 1 (UNC) - Systems and Policies

**Ingest:** Authenticity, Chain of Custody, Deposit Agreement, Documentation, File Validation, Metadata

**Appraisal:** *Rights Management (licenses only)

**Curate:** Arrangement & Description, File Inventory or Manifest, Indexing, Persistent ID, Transcoding

**Access:** Contact information, Data Citation, Discovery Services, Embargo, File Download, Full Text-Indexing, Metadata Brokerage, Terms of Use, Use Analytics, *Restricted Access (system automated)

**Preservation:** File Audit, Migration, Secure Storage, Succession Planning, Tech/Monitoring Refresh, Versioning, Cease Data Curation
Level 2 (Duke) - Human intervention, data knowledge

**Appraise/Accept:** *Rights Management (DUAs), *Risk Management (file review), Selection

**Curate:** Contextualize, Curation Log, File Format Transformations, File Renaming, *Quality Assurance, Restructure

**Access:** *Restricted Access (mediated requests)

**Preserve:** Repository Certification
Level 3: Human intervention, domain-specific, data knowledge

**Appraise/Accept:** *Risk Management (remediation)*

**Curate:** *Code Review Conversion (Analog), Data Cleaning, De-Identification, Interoperability, Peer Review, Quality Assurance, Software Registry*

**Access:** Data Visualization

**Preserve:** Emulation
Using the Levels in the CDR

- Feasibility of FAIR with current staffing levels and generalist subject knowledge?
- What is our sweet spot? Where can we add value?
- Is our new system Hyrax good for data? Capabilities?
- How do we get institutional support?
- What training do we need?
What did we do?

- Training
- Created policies
- Created documentation
- Created new workflow
- Assessed existing deposits
- Evaluated Hyrax for data
Training

- Data management online course
- Literature review
- Review of existing services
- Conference attendance
- Workshops
Why a new data policy?

- Formalized and documented service activities
  - Defines activities in Levels
  - Transparency in services
- Large deposit approval took a long time
- Data definition was vague
- Administrative and library leadership buy-in for service and staff time
What is in the new data policy?

- Data definition
- Review of large deposit requests is now reviewed by data librarians
- 10 year retention review
- Tombstone record
- Submission review by CDR staff
- ReadMe or other documentation is now required
Documentation

• CDR staff documentation
  • Mediated deposit questions
  • Workflow for self and mediated submission
  • Comparison document contrasting old practice with new policy

• Depositor documentation
  • Data deposit FAQs
  • ReadMe template
Workflow

1. Depositor submits
2. Mint a DOI
3. Verify contact info
4. Verify ReadMe
5. Assess file format
6. Verify license
7. PII scan
8. Email researcher
Existing Data Deposit Assessment Project Goals

• Determine compliance with new policy
• Test scope of service
• Test new workflow
• Target areas for improvement
Results

• Compliance with new policy
  • ReadMe needed
  • Open formats needed
  • DOIs and licenses are good!

• Scope of the service
  • Covers self and mediated deposit, not supplemental data

• Test of new workflow

• Further questions
Where Do We Go Next?

• Locally
  • Launch Hyrax
  • Evaluate level of data support as submissions increase
  • Investigate subject area data curation support
  • Assess Level Two activities to plan future service expansion

• Community
  • Audit
  • Hyrax work to support data
Hyrax Audit for FAIR/Level One
Repository Data Audit Community Template

https://docs.google.com/spreadsheets/d/1NXyELgm1YVnE4mLtNTyuhPWP3-aSRCz5SyWnicX8i2M/edit#gid=0

- Hyrax Capability
- Local Solution Needed
- Policy Needed
### Instructions for conducting an Audit:

1. **Using the Level 1 criteria for Data Curation**
   Activities and FAIR guiding principles, specify how your system, institution, policies meet these criteria.

2. **Duplicate the Repository Data Audit Template**
   Create a tab for your repository and fill in your criteria.

### To learn more about criteria in the template:

- **Data Curation Network’s Activities**
  - [https://conservancy.umn.edu/handle/11299/111820](https://conservancy.umn.edu/handle/11299/111820)

- **Level 1 Criteria for Data Curation Network Activities**
  - [http://dx.doi.org/10.17909/T75J3QZD](http://dx.doi.org/10.17909/T75J3QZD)

- **FAIR Guiding Principles**
  - [https://www.ifi11.org/group/ifi11/principles](https://www.ifi11.org/group/ifi11/principles)

### Level 1 Activities Defined

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td><strong>Authentication</strong></td>
<td>The process of confirming the identity of a person, generally the depositor, who is contributing data to the data repository. (e.g., password authentication or authorization via digital signature). Used for tracking provenance of the data files.</td>
</tr>
<tr>
<td>14</td>
<td><strong>Chain of Custody</strong></td>
<td>Intentional recording of provenance metadata of the files (e.g., metadata about who created the file, when it was last edited, etc.) in order to preserve file authenticity when data are transferred to third parties.</td>
</tr>
<tr>
<td>15</td>
<td><strong>Deposit Agreement</strong></td>
<td>The certification by the data author (or depositor) that the data conform to all policies and conditions (e.g., do not violate any legal restrictions placed on the data) and arefit for deposit into the repository. A deposit agreement may also include rights transfer to the repository for ongoing stewardship.</td>
</tr>
<tr>
<td>16</td>
<td><strong>Documentation</strong></td>
<td>Information describing any necessary information to use and understand the data. Documentation may be structured (e.g., a code book) or unstructured (e.g., a plain text “Readme” file).</td>
</tr>
</tbody>
</table>

### File Validation

- A computational process to ensure that the intended data transfer to a repository was perfect and complete, using means such as generating and validating file checksums (e.g., test if a digital file has changed at the bit level) and format validation to ensure that file types match their extensions.
<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Data Curation Network Activity - Level One</td>
<td>FAIR Principle</td>
<td>Hyrax 2.1.x</td>
<td>Local Solution Needed</td>
<td>Policy Needed</td>
</tr>
<tr>
<td>2</td>
<td>Authentication</td>
<td></td>
<td></td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Chair of Custody</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Deposit Agreement</td>
<td>Deposit Agreement Included</td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>5</td>
<td>Documentation</td>
<td>Supports Supplemental Files</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>File Validation</td>
<td>Hyrax uses FITS for file characterization</td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>7</td>
<td>Metadata</td>
<td>Custom WorkTypes; RDF; daeNames</td>
<td>Data work type creation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Rights Management (licensing)</td>
<td>CC Licenses available</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Arrangement &amp; Description</td>
<td>File ordering, Work Nailing, Collection Building</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>File Inventory or Manif</td>
<td>Supports Supplemental Files</td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>11</td>
<td>Indexing</td>
<td>SOLR, formatted data fields, ??</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Persistent Identifier</td>
<td>ActiveDirectory: Need to store Mds style identifiers</td>
<td>DOIs need to be locally implemented</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Transcoding</td>
<td>Supported via FRPEG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Contact Information</td>
<td>Depositor and Creator Information</td>
<td>Customized Worktype to collect additional information</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Data Citation</td>
<td>MLA, APA, Chicago suggested formatting</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Discovery Services</td>
<td>Blacklight, SOLR, SEO</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>Embargo</td>
<td>Embargo supported</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>File Download</td>
<td>File Download Supported</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>Full Text-indexing</td>
<td>Runs against PDFs and office docs uses SOLR (needs to be configured)</td>
<td>Could extend service to more types</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>Metadata Brokerage</td>
<td>OA-PMH, ResourceSync</td>
<td>Would need to register for any specific registries</td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>Restricted Access (system automated)</td>
<td>Rapid Access Controls</td>
<td>Any HIPAA or PII restrictions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>Terms of Use</td>
<td>Licenses are displayed on the view pages</td>
<td>Hyrax provides support for more information</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Works Cited


Johnson, Lisa R.; Carlson, Jake; Hudson-Vitale, Cynthia; Imker, Heidi; Kozlowski, Wendy; Olendorf, Robert; and Stewart, Claire. (2016, October 23). Data Curation Network: Data Curation Terms and Activities. Retrieved from: http://hdl.handle.net/11299/188638
