IMPROVING DATA CURATION SERVICES WITHIN AN INSTITUTIONAL REPOSITORY

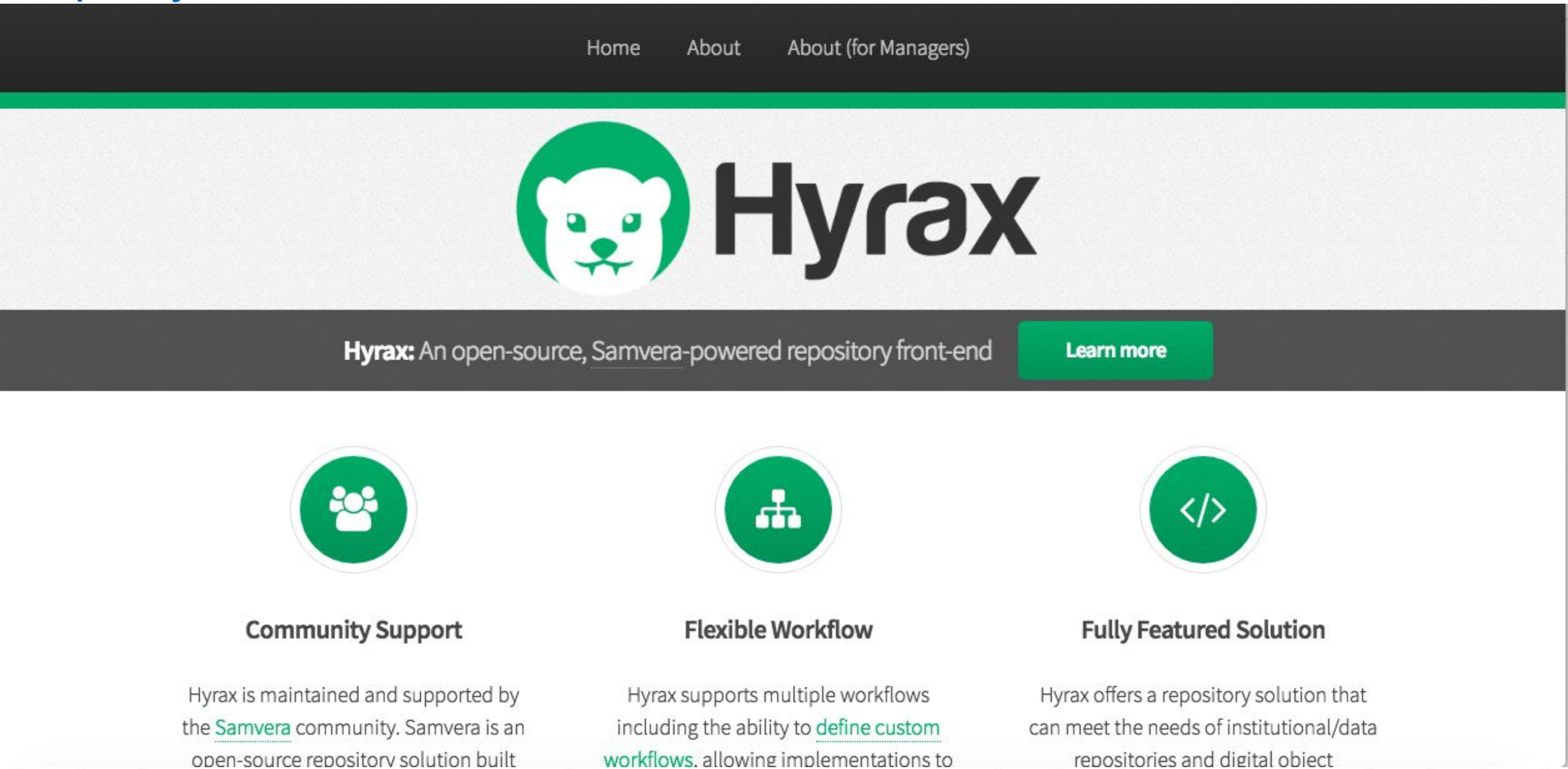
UNC Chapel Hill Libraries
Julie Rudder - Repository Program Librarian
Rebekah Kati - Institutional Repository Librarian

Overview

- Introduction
- Levels of Data Curation Activities Exercise
- Using the Levels of Data Curation at UNC
- Hyrax Audit for Data
- Next Steps


About
Contact
Login
UNC
UNIVERSITY
LIBRARIES
Carolina Digital Repository
Advanced Search
Search all collections
Search
Home
Collections
Departments
Deposit
Deposit, share, and promote your scholarly and creative work.
Deposit Your Work
Articles
Student Papers
Datasets
Other Deposits

http://hyr.ax/


Home
About
About (for Managers)
Hyrax
Hyrax: An open-source, Samvera-powered repository front-end
Learn more
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

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/

- 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.

ARL SPEC KIT 354

"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](#).

Level 1 Curation	Level 2 Curation	Level 3 Curation
Ingest • Authentication • Chain of Custody • Deposit Agreement • Documentation • File Validation • Metadata Appraisal • *Rights Management (licenses)* Curate • Arrangement & Description • File Inventory or Manifest • Indexing • Persistent ID 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 • Versioning • Succession Planning • Tech/Monitoring Refresh • Cease Data Curation	Appraisal • *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)* Preservation • Repository Certification	Appraisal • *Risk Management (remediation)* Curate • Code Review • Conversion (Analog) • Data Cleaning • De-Identification • Interoperability • Peer Review • *Quality Assurance* Access • Data Visualization

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

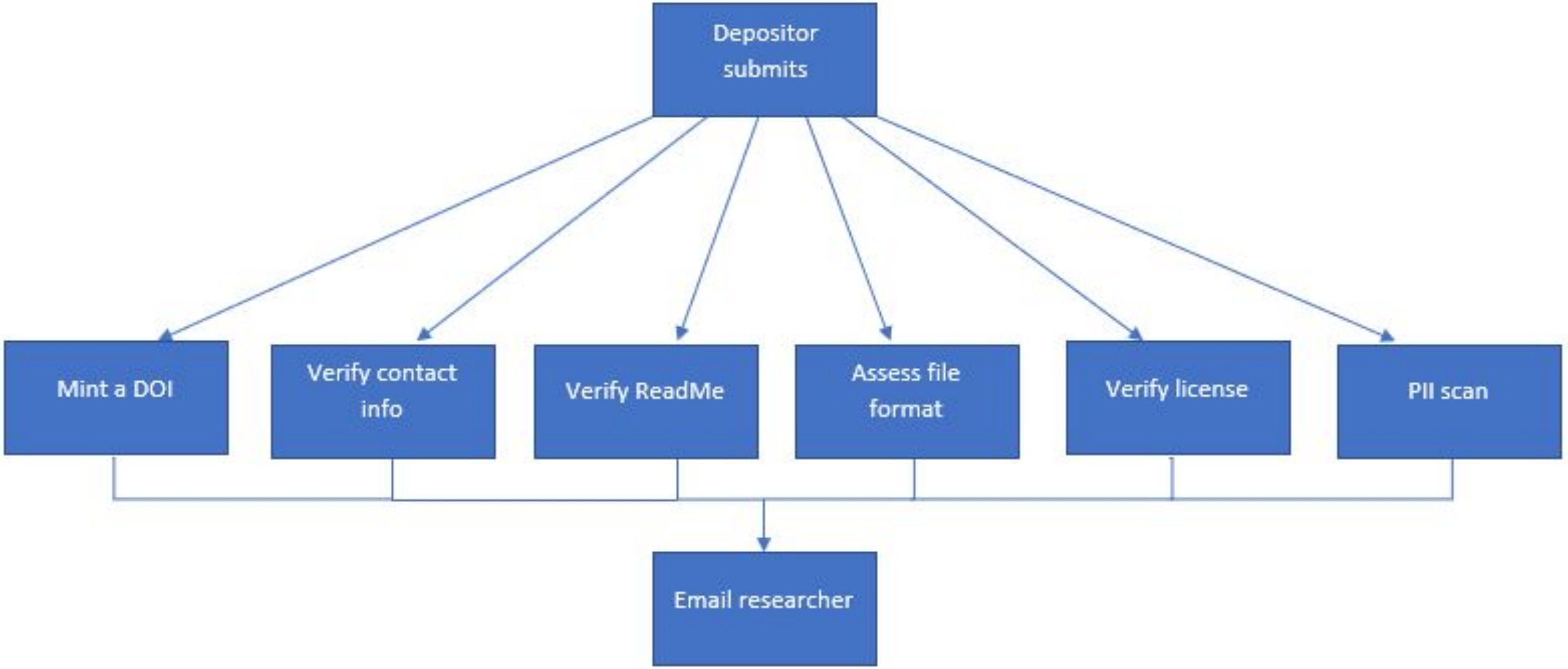
What is in the new data policy?

- 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


Depositor submits
Mint a DOI
Verify contact info
Verify ReadMe
Assess file format
Verify license
PII scan
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

Repository Data Audit Template

File Edit View Insert Format Data Tools Add-ons Help All changes saved in Drive

SHARE

Instructions for conducting an Audit:

	A	B	C	D	E	F
1	**Instructions for conducting an Audit:**					
2	Using the Level 1 criteria for Data Curation Activities and FAIR guiding principles, specify how your system, institution, policies meet these criteria.					
3	Duplicate the Repositor Data Audit TemplateTab to create a tab for your repository and fill in your criteria					
4						
5						
6	**To learn more about criteria in the template:**					
7	Data Curation Network's Activities	https://conservancy.umn.edu/handle/11299/188638				
8	Level 1 Criteria for Data Curation Network Activities:	http://doi.org/10.17605/OSF.IO/ZJ5PQ				
9	FAIR Guiding Principles:	https://www.force11.org/group/fairgroup/fairprinciples				
10						
11						
12	**Level 1 Activities Defined**					
13	Authentication	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.				
14	Chain of Custody	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.				
15	Deposit Agreement	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 are fit for deposit into the repository. A deposit agreement may also include rights transfer to the repository for ongoing stewardship.				
16	Documentation	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).				
17	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				

README | Repository Data Audit Template | Hyrax-Audit

Repository Data Audit Template

File Edit View Insert Format Data Tools Add-ons Help All changes saved in Drive

SHARE

	A	B	C	D	E	F	G	H	I	J	K
1	**Data Curation Network Activity - Level One**	**FAIR Principle**	**Hyrax 2.1.x**	**Local Solution Needed**	**Policy Needed**	**Notes**					
2	Authentication			x		Shibboleth config docu:					
3	Chain of Custody					This could use a good look.					
4	Deposit Agreement		Deposit Aggrement Included		x						
5	Documentation		Supports Supplemental Files		x						
6	File Validation		Hyrax uses FITS for file characterization								
7	Metadata		Custom WorkTypes; RDF; GeoNames	Data work type creation							
8	Rights Management (licenses)		CC Licenses available	Additional License can be added	x						
9	Arrangement & Description		File ordering, Work Nesting, Collection Building								
10	File Inventory or Manifest		Supports Supplemental Files								
11	Indexing		SOLR; formatted date fields, ??								
12	Persistent Identifier		ActiveFedora::Noid gem to mint Noid-style identifiers	DOIs need to be locally implemented	x						
13	Transcoding		Supported via FFmpeg			May need to be tweaked from default settings					
14	Contact Information		Depositor and Creator information	Customize Worktype to collect additional information	x						
15	Data Citation		MLA, APA, Chicago suggested formatting								
16	Discovery Services		Blacklight, SOLR, SEO								
17	Embargo		Embargo supported		x						
18	File Download		File Download Supported		x						
19	Full Text-Indexing		Runs agains PDFs and office docs uses SOLR (needs to be configured)	could extend service to more types		https://github.com/samvera/hydra-derivatives/blob/3434afecb7261f63415be78d8b371cd20b514a7b/lib/hydra/derivatives/processors/full_text.rb	https://github.com/samvera/hyrax/blob/21117ad865f62a15668ad59f474764856a56dbdf/app/services/hyrax/file_set_derivatives_service.rb				
20	Metadata Brokerage		OAI-PMH; ResourceSync	Would need to register for any specific registries		OAI-PMH via Blacklight OAI Provider					
21	Restricted Access (system automated)		Robust Access Controls	Any HIPPA or PII restrictions	x						
22	Terms of Use		Licenses are displayed on the view pages		x						
			Hyrax provides support for capturing usage								

README | Repository Data Audit Template | Hyrax-Audit

Works Cited

FORCE11. *The FAIR Data Principles.* Retrieved June 1, 2018 from: https://www.force11.org/group/fairgroup/fairprinciples

Hudson-Vitale, Cynthia; Imker, Heidi; Johnson, Lisa R.; Carlson, Jake; Kozlowski, Wendy; Olendorf, Robert; and Stewart, Claire. (2017). *SPEC Kit 354 Data Curation*. Washington DC: Association of Research Libraries. https://doi.org/10.29242/spec.354

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

Johnson, Lisa R.; Carlson, Jacob; Hudson-Vitale, Cynthia; Imker, Heidi; Kozlowski, Wendy; Olendorf, Robert; and Stewart, Claire. (2018). How Important Are Data Curation Activities to Researchers? Gaps and Opportunities for Academic Libraries. *Journal of Librarianship and Scholarly Communication*, 6(General Issue), eP2198. https://doi.org/10.7710/2162-3309.2198

Lafferty-Hess, Sophia; Rudder, Julie; Downey, Moira; Ivey, Susan; and Darragh, Jen. (2018). *Conceptualizing Data Curation Activities Within Two Academic Libraries*. Retrieved from: https://share.osf.io/preprint/460F1-F92-9F9