Extending Preservation Functionality in Hyrax 3, Fedora 4, and AWS

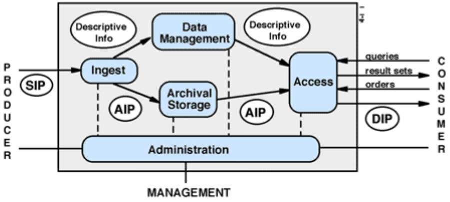
Emily Porter, Digital Repository Program Manager

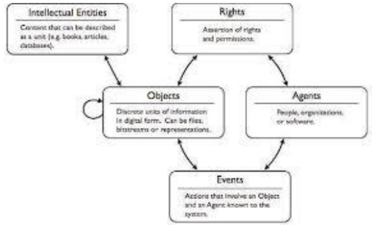
Devanshu Matlawala, Software Engineer

EMORY UNIVERSITY

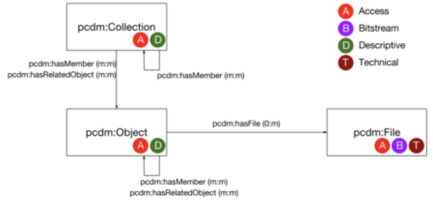
digital.library.emory.edu

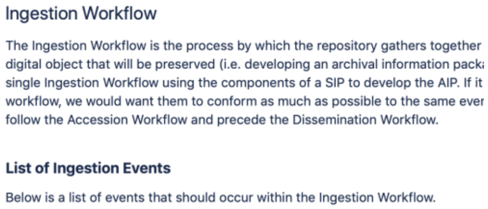
Complex Foundations





Intellectual Entities
Content that can be described as a unit (e.g. books, articles, databases).
Rights
Assertion of rights and permissions.
Objects
Discrete units of information in digital form. Can be files, bitstreams or representations.
Agents
People, organizations, or software.
Events
Actions that involve an Object and an Agent known to the system.



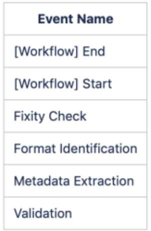


Ingestion Workflow

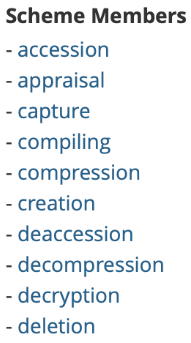
The Ingestion Workflow is the process by which the repository gathers together digital object that will be preserved (i.e. developing an archival information packa single Ingestion Workflow using the components of a SIP to develop the AIP. If it workflow, we would want them to conform as much as possible to the same ever follow the Accession Workflow and precede the Dissemination Workflow.

List of Ingestion Events

Below is a list of events that should occur within the Ingestion Workflow.

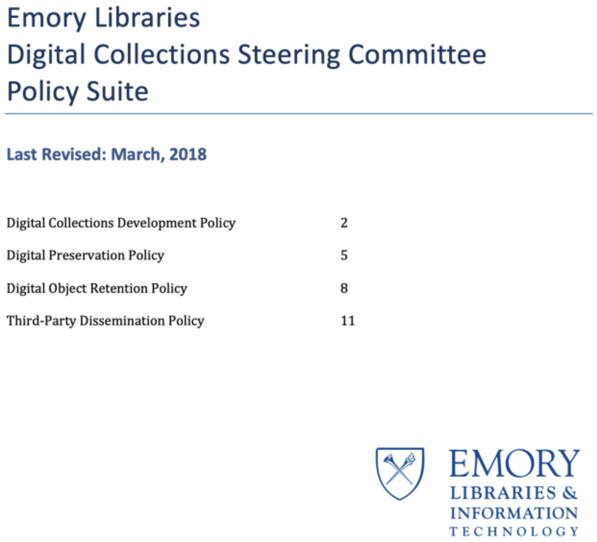


Event Name
[Workflow] End
[Workflow] Start
Fixity Check
Format Identification
Metadata Extraction
Validation



Scheme Members

- accession
- appraisal
- capture
- compiling
- compression
- creation
- deaccession
- decompression
- decryption
- deletion



Emory Libraries
Digital Collections Steering Committee
Policy Suite

Last Revised: March, 2018

EMORY
LIBRARIES &
INFORMATION
TECHNOLOGY

References:
OAIS, PREMIS, PCDM, LC Preservation Event Types, Local Emory Policy

Preservation v.2: Implementation Goals

1. Actualize our Preservation Policy!
2. Build our locally defined Archival Information Package structure
3. Leverage existing Fedora 4.x capabilities
4. Work within our institution’s AWS and S3 capabilities
5. Provide a human readable audit trail for significant activities occurring on assets
6. Manage as many preservation activities as possible in a Hyrax-based interface

Our local infrastructure

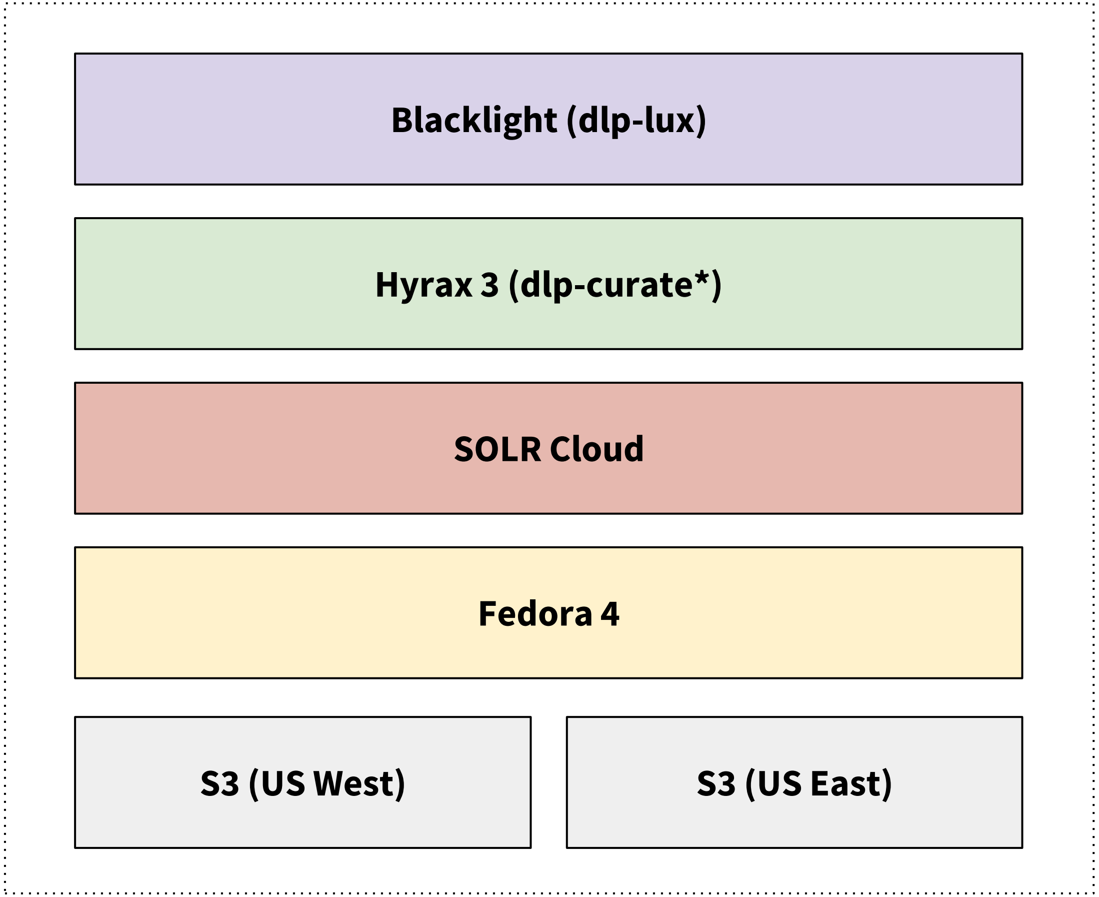
Hosted in an Emory University-managed AWS environment (some AWS Services not enabled)

Hyrax v3.0.0-rc1

Binaries not stored in Fedora

Additional mediated deposit application (future)

** Our apologies/compliments to Notre Dame's Curate!*



Blacklight (dlp-lux)

Hyrax 3 (dlp-curate*)

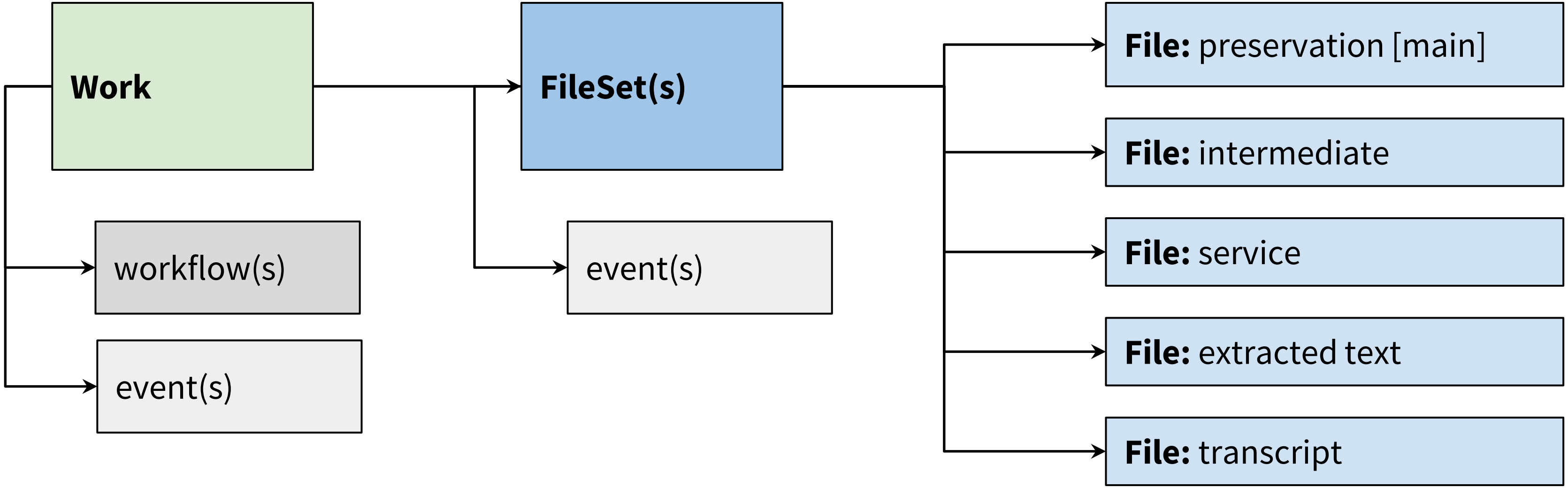
SOLR Cloud

Fedora 4

S3 (US West)

S3 (US East)

Generic Preservation Data Model



Works hold Descriptive, Rights, Administrative metadata plus preservation workflow and event objects

FileSets hold minimal Descriptive metadata and a “use” context for the Set itself, extended technical metadata, and event objects

Files receive a use/role within the Set and minimal technical metadata

FileSet Extensions

Harmful Language Note

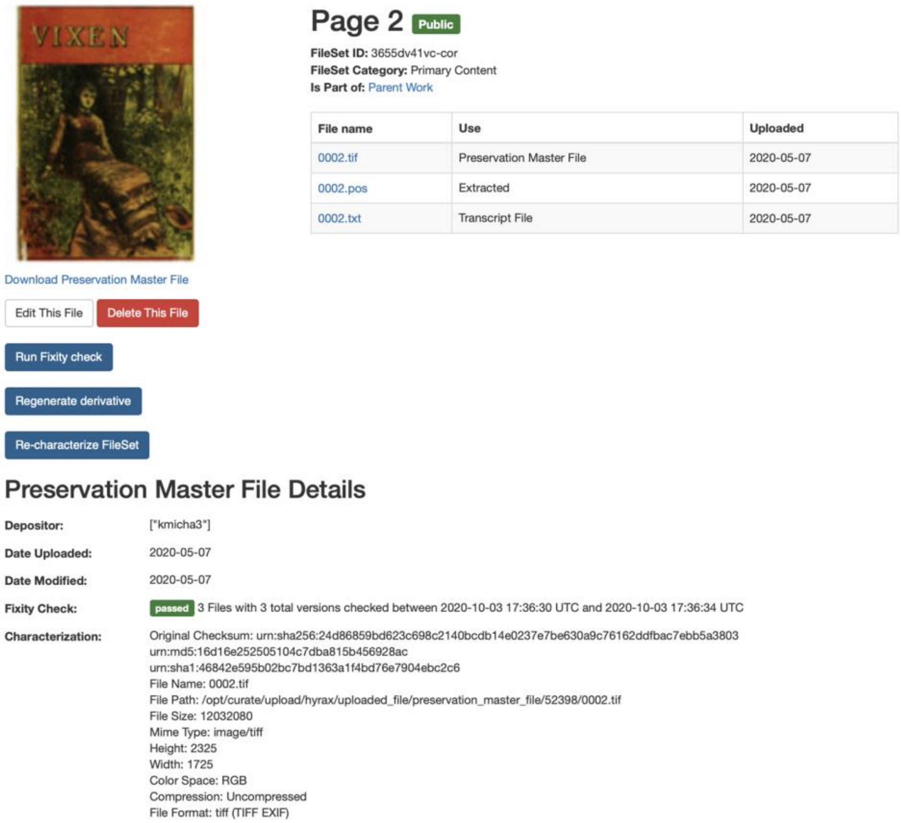
“Preservation Master File”

As a community, can we find a better term?

Support for primary file and multiple derivatives, based on PCDM use extension

Support for Primary/Supplemental Content

Additional technical metadata indexed from FITS



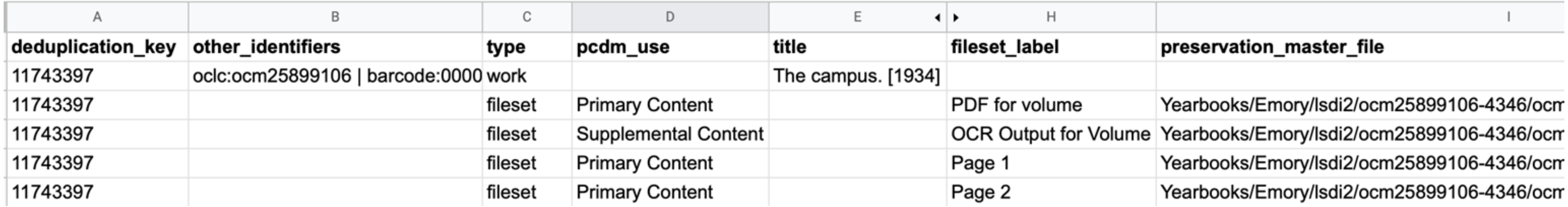
Page 2 Public

FileSet ID: 3655dv41vc-cor
FileSet Category: Primary Content
Is Part of: Parent Work

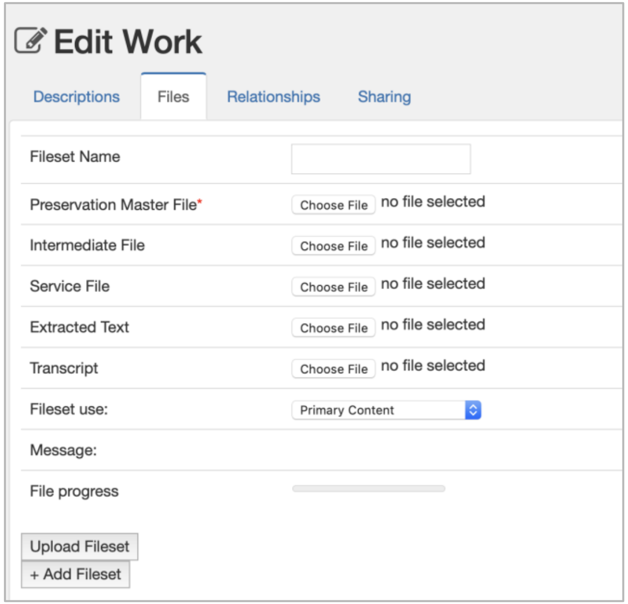
File name	Use	Uploaded
0002.tif	Preservation Master File	2020-05-07
0002.pos	Extracted	2020-05-07
0002.txt	Transcript File	2020-05-07

Preservation Master File Details

Depositor:	["kmicha3"]
Date Uploaded:	2020-05-07
Date Modified:	2020-05-07
Fixity Check:	passed 3 Files with 3 total versions checked between 2020-10-03 17:36:30 UTC and 2020-10-03 17:36:34 UTC
Characterization:	Original Checksum: urn:sha256:24d86859bd623c698c2140bcdb14e0237e7be630a9c76162ddfbac7ebb5a3803 urn:md5:16d16e252505104c7dba815b456928ac urn:sha1:46842e595b02bc7bd1363a1f4bd76e7904ebc2c6 File Name: 0002.tif File Path: /opt/curate/upload/hyrax/uploaded_file/preservation_master_file/52398/0002.tif File Size: 12032080 Mime Type: image/tiff Height: 2325 Width: 1725 Color Space: RGB Compression: Uncompressed File Format: tiff (TIFF EXIF)



A	B	C	D	E	H	I
deduplication_key	**other_identifiers**	**type**	**pcdm_use**	**title**	**fileset_label**	**preservation_master_file**
11743397	oclc:ocm25899106 | barcode:0000	work		The campus. [1934]		
11743397		fileset	Primary Content		PDF for volume	Yearbooks/Emory/lsdi2/ocm25899106-4346/ocm
11743397		fileset	Supplemental Content		OCR Output for Volume	Yearbooks/Emory/lsdi2/ocm25899106-4346/ocm
11743397		fileset	Primary Content		Page 1	Yearbooks/Emory/lsdi2/ocm25899106-4346/ocm
11743397		fileset	Primary Content		Page 2	Yearbooks/Emory/lsdi2/ocm25899106-4346/ocm



Edit Work

Fileset Name	
Preservation Master File*	Choose File no file selected
Intermediate File	Choose File no file selected
Service File	Choose File no file selected
Extracted Text	Choose File no file selected
Transcript	Choose File no file selected
Fileset use:	Primary Content
Message:	
File progress	

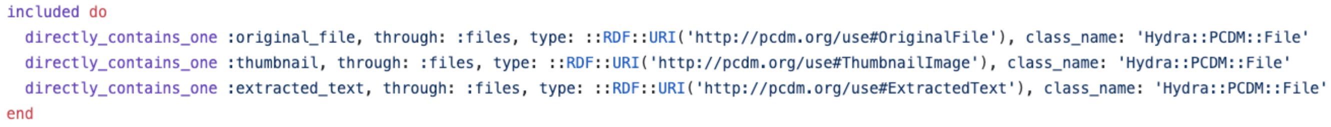
Fileset Creation:

1. CSV for bulk-import (developed by DCE) contains distinct rows for works and filesets; provides file-level use information

1. Edit Work > Files tab allows users to attach one or more filesets and assign the use for each file as well as overall categorization for the FileSet itself (Primary/Supplemental)

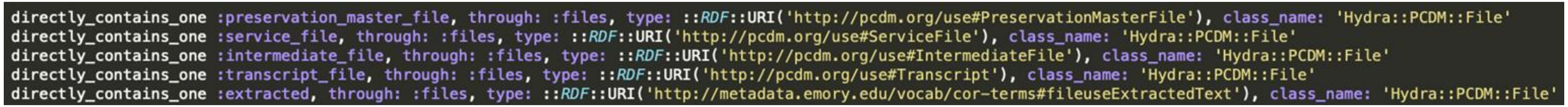
Implementation Notes: Code Changes

Hyrax filesets contain three files per fileset (original file, thumbnail, extracted text):



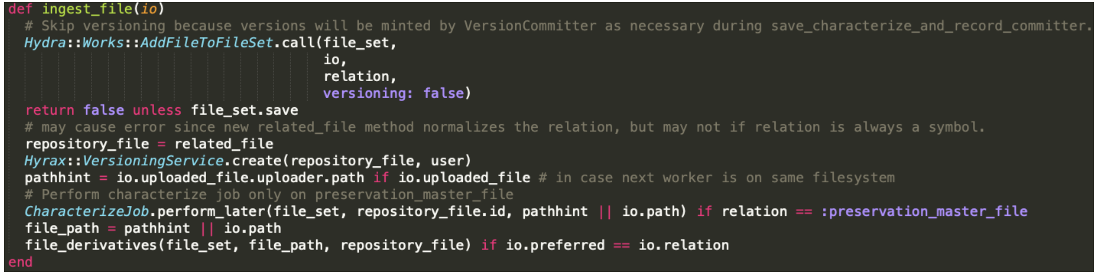
```ruby
included do
  directly_contains_one :original_file, through: :files, type: ::RDF::URI('http://pcdm.org/use#OriginalFile'), class_name: 'Hydra::PCDM::File'
  directly_contains_one :thumbnail, through: :files, type: ::RDF::URI('http://pcdm.org/use#ThumbnailImage'), class_name: 'Hydra::PCDM::File'
  directly_contains_one :extracted_text, through: :files, type: ::RDF::URI('http://pcdm.org/use#ExtractedText'), class_name: 'Hydra::PCDM::File'
end
```

Curate filesets contain the three above plus five new files per fileset:



```ruby
directly_contains_one :preservation_master_file, through: :files, type: ::RDF::URI('http://pcdm.org/use#PreservationMasterFile'), class_name: 'Hydra::PCDM::File'
directly_contains_one :service_file, through: :files, type: ::RDF::URI('http://pcdm.org/use#ServiceFile'), class_name: 'Hydra::PCDM::File'
directly_contains_one :intermediate_file, through: :files, type: ::RDF::URI('http://pcdm.org/use#IntermediateFile'), class_name: 'Hydra::PCDM::File'
directly_contains_one :transcript_file, through: :files, type: ::RDF::URI('http://pcdm.org/use#Transcript'), class_name: 'Hydra::PCDM::File'
directly_contains_one :extracted, through: :files, type: ::RDF::URI('http://metadata.emory.edu/vocab/cor-terms#fileuseExtractedText'), class_name: 'Hydra::PCDM::File'
```

Characterization only for preservation [main] file in FileActor



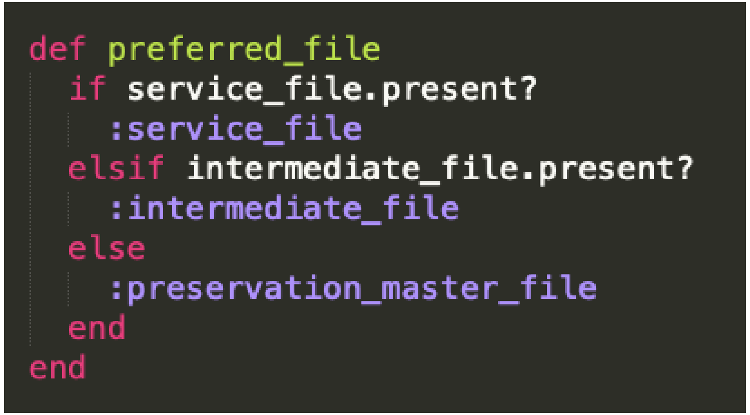
```ruby
def ingest_file(io)
  # Skip versioning because versions will be minted by VersionCommitter as necessary during save_characterize_and_record_committer.
  Hydra::Works::AddFileToFileSet.call(file_set,
                                      io,
                                      relation,
                                      versioning: false)
  return false unless file_set.save
  # may cause error since new related_file method normalizes the relation, but may not if relation is always a symbol.
  repository_file = related_file
  Hyrax::VersioningService.create(repository_file, user)
  pathhint = io.uploaded_file.uploader.path if io.uploaded_file # in case next worker is on same filesystem
  # Perform characterize job only on preservation_master_file
  CharacterizeJob.perform_later(file_set, repository_file.id, pathhint || io.path) if relation == :preservation_master_file
  file_path = pathhint || io.path
  file_derivatives(file_set, file_path, repository_file) if io.preferred == io.relation
end
```

Implementation Notes, Continued

Code changes:

- Logic for detecting preferred file for thumbnail generation and Universal Viewer display

- Curate ingest process documentation/readme



```
def preferred_file
  if service_file.present?
    :service_file
  elsif intermediate_file.present?
    :intermediate_file
  else
    :preservation_master_file
  end
end
```

Preservation Event Entities

(Mostly) automated system actions that occur relative to a larger workflow; actions are captured in a human-readable event entry

Work-level

- Policy assignment
- Validation
- Modification

More events are targeted in future development cycles

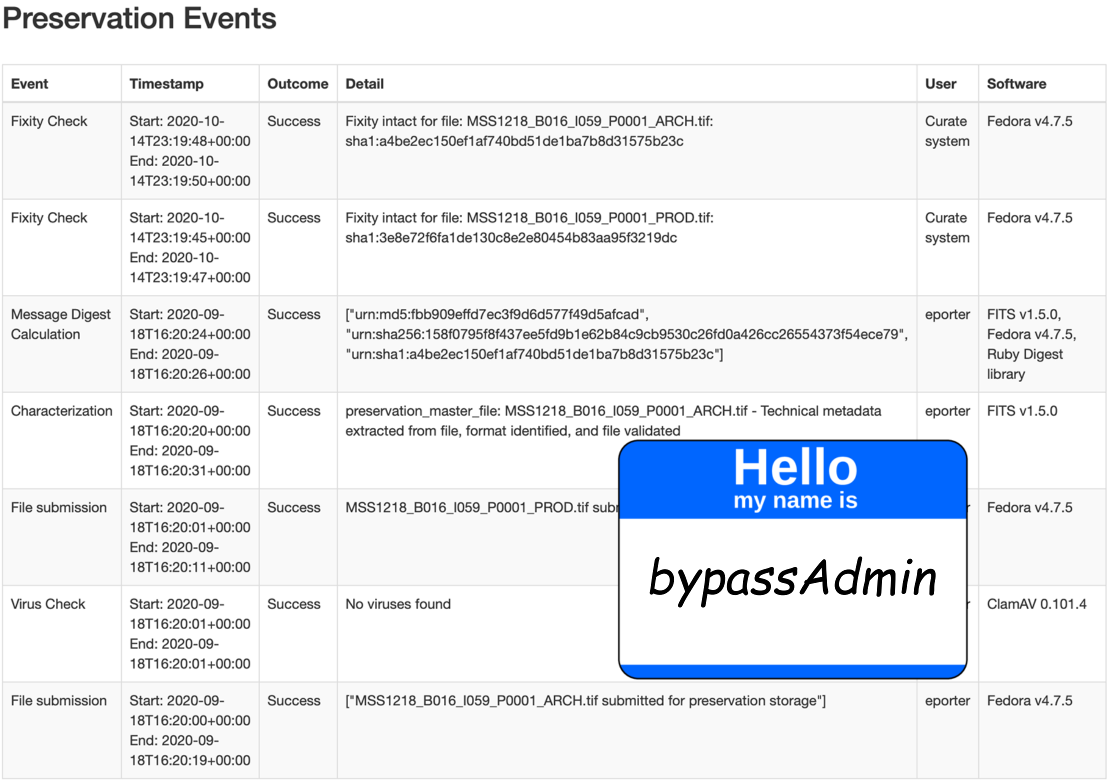
FileSet-level

- Virus scan
- Characterization
- Message digest calculation
- File submission
- Fixity check

Event Metadata

Local namespace defines:

1. eventType
2. eventUser
3. eventStart
4. eventEnd
5. eventOutcome
6. softwareVersion
7. eventDetails
8. eventIdentifier

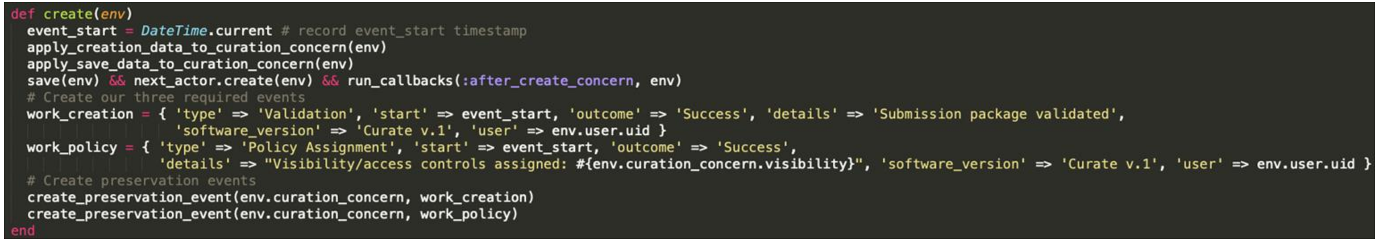


Preservation Events

Event	Timestamp	Outcome	Detail	User	Software
Fixity Check	Start: 2020-10-14T23:19:48+00:00 End: 2020-10-14T23:19:50+00:00	Success	Fixity intact for file: MSS1218_B016_I059_P0001_ARCH.tif: sha1:a4be2ec150ef1af740bd51de1ba7b8d31575b23c	Curate system	Fedora v4.7.5
Fixity Check	Start: 2020-10-14T23:19:45+00:00 End: 2020-10-14T23:19:47+00:00	Success	Fixity intact for file: MSS1218_B016_I059_P0001_PROD.tif: sha1:3e8e72f6fa1de130c8e2e80454b83aa95f3219dc	Curate system	Fedora v4.7.5
Message Digest Calculation	Start: 2020-09-18T16:20:24+00:00 End: 2020-09-18T16:20:26+00:00	Success	["urn:md5:fbb909effd7ec3f9d6d577f49d5afcad", "urn:sha256:158f0795f8f437ee5fd9b1e62b84c9cb9530c26fd0a426cc26554373f54ece79", "urn:sha1:a4be2ec150ef1af740bd51de1ba7b8d31575b23c"]	eporter	FITS v1.5.0, Fedora v4.7.5, Ruby Digest library
Characterization	Start: 2020-09-18T16:20:20+00:00 End: 2020-09-18T16:20:31+00:00	Success	preservation_master_file: MSS1218_B016_I059_P0001_ARCH.tif - Technical metadata extracted from file, format identified, and file validated	eporter	FITS v1.5.0
File submission	Start: 2020-09-18T16:20:01+00:00 End: 2020-09-18T16:20:11+00:00	Success	MSS1218_B016_I059_P0001_PROD.tif subr	r	Fedora v4.7.5
Virus Check	Start: 2020-09-18T16:20:01+00:00 End: 2020-09-18T16:20:01+00:00	Success	No viruses found	r	ClamAV 0.101.4
File submission	Start: 2020-09-18T16:20:00+00:00 End: 2020-09-18T16:20:19+00:00	Success	["MSS1218_B016_I059_P0001_ARCH.tif submitted for preservation storage"]	eporter	Fedora v4.7.5

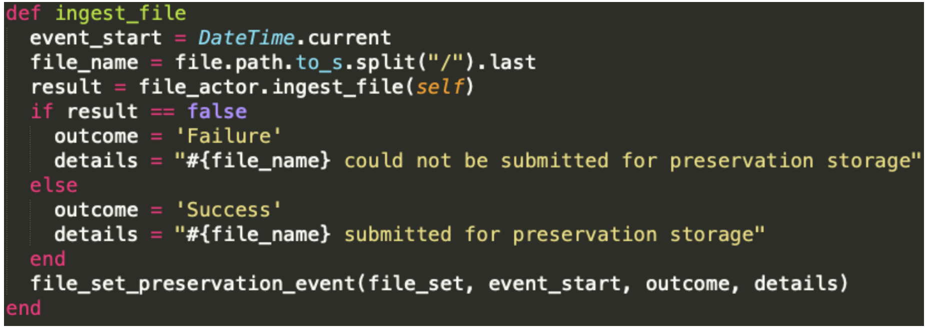
Implementation: Events

Work events (Validation and Policy Assignment during work creation; Modification post-ingest)



```ruby
def create(env)
  event_start = DateTime.current # record event_start timestamp
  apply_creation_data_to_curation_concern(env)
  apply_save_data_to_curation_concern(env)
  save(env) && next_actor.create(env) && run_callbacks(:after_create_concern, env)
  # Create our three required events
  work_creation = { 'type' => 'Validation', 'start' => event_start, 'outcome' => 'Success', 'details' => 'Submission package validated',
                    'software_version' => 'Curate v.1', 'user' => env.user.uid }
  work_policy = { 'type' => 'Policy Assignment', 'start' => event_start, 'outcome' => 'Success',
                  'details' => "Visibility/access controls assigned: #{env.curation_concern.visibility}", 'software_version' => 'Curate v.1', 'user' => env.user.uid }
  # Create preservation events
  create_preservation_event(env.curation_concern, work_creation)
  create_preservation_event(env.curation_concern, work_policy)
end
```

FileSet events (File submission in JobIoWrapper)

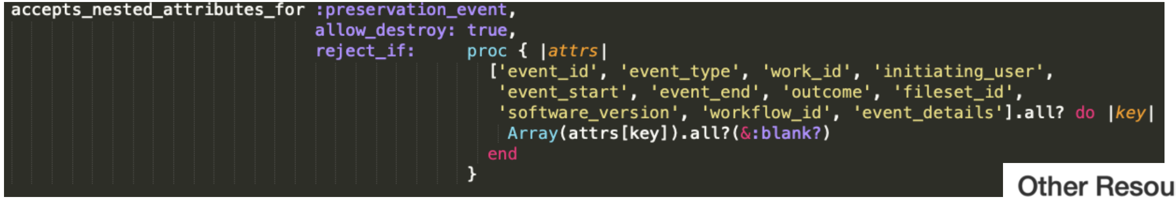


```ruby
def ingest_file
  event_start = DateTime.current
  file_name = file.path.to_s.split("/").last
  result = file_actor.ingest_file(self)
  if result == false
    outcome = 'Failure'
    details = "#{file_name} could not be submitted for preservation storage"
  else
    outcome = 'Success'
    details = "#{file_name} submitted for preservation storage"
  end
  file_set_preservation_event(file_set, event_start, outcome, details)
end
```

Some FileSet events only occur on the primary (main) file

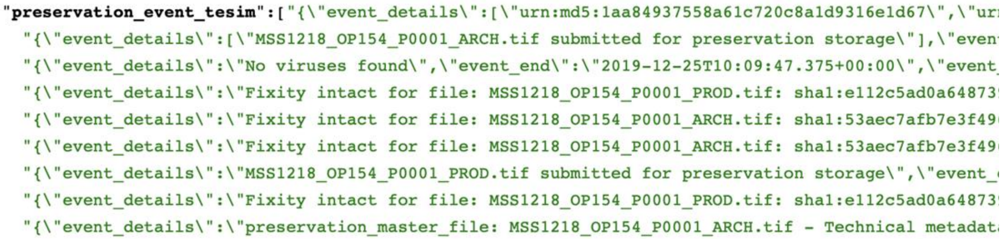
Implementation: Events

Implemented as nested Fedora objects (PreservationEvent has its own model class)



```
accepts_nested_attributes_for :preservation_event,
                              allow_destroy: true,
                              reject_if:     proc { |attrs|
                                               ['event_id', 'event_type', 'work_id', 'initiating_user',
                                                'event_start', 'event_end', 'outcome', 'fileset_id',
                                                'software_version', 'workflow_id', 'event_details'].all? do |key|
                                                 Array(attrs[key]).all?(&:blank?)
                                               end
                                             }
```

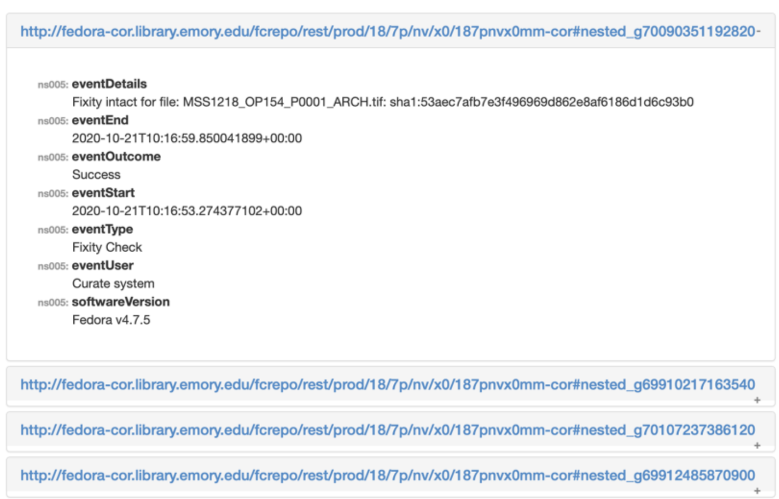
Events are indexed in Solr for display purposes



```
"preservation_event_tesim":["{\"event_details\":[\"urn:md5:1aa84937558a61c720c8a1d9316e1d67\",\"ur
  "{\"event_details\":[\"MSS1218_OP154_P0001_ARCH.tif submitted for preservation storage\"],\"even
  "{\"event_details\":\"No viruses found\",\"event_end\":\"2019-12-25T10:09:47.375+00:00\",\"event_
  "{\"event_details\":\"Fixity intact for file: MSS1218_OP154_P0001_PROD.tif: sha1:e112c5ad0a64873
  "{\"event_details\":\"Fixity intact for file: MSS1218_OP154_P0001_ARCH.tif: sha1:53aec7afb7e3f49
  "{\"event_details\":\"Fixity intact for file: MSS1218_OP154_P0001_ARCH.tif: sha1:53aec7afb7e3f49
  "{\"event_details\":\"MSS1218_OP154_P0001_PROD.tif submitted for preservation storage\",\"event_
  "{\"event_details\":\"Fixity intact for file: MSS1218_OP154_P0001_PROD.tif: sha1:e112c5ad0a64873
  "{\"event_details\":\"preservation_master_file: MSS1218_OP154_P0001_ARCH.tif - Technical metadat
```

Preservation Event readme documentation

Other Resources



http://fedora-cor.library.emory.edu/fcrepo/rest/prod/18/7p/nv/x0/187pnvx0mm-cor#nested_g70090351192820-

ns005: **eventDetails**
Fixity intact for file: MSS1218_OP154_P0001_ARCH.tif: sha1:53aec7afb7e3f496969d862e8af6186d1d6c93b0

ns005: **eventEnd**
2020-10-21T10:16:59.850041899+00:00

ns005: **eventOutcome**
Success

ns005: **eventStart**
2020-10-21T10:16:53.274377102+00:00

ns005: **eventType**
Fixity Check

ns005: **eventUser**
Curate system

ns005: **softwareVersion**
Fedora v4.7.5

http://fedora-cor.library.emory.edu/fcrepo/rest/prod/18/7p/nv/x0/187pnvx0mm-cor#nested_g69910217163540 +

http://fedora-cor.library.emory.edu/fcrepo/rest/prod/18/7p/nv/x0/187pnvx0mm-cor#nested_g70107237386120 +

http://fedora-cor.library.emory.edu/fcrepo/rest/prod/18/7p/nv/x0/187pnvx0mm-cor#nested_g69912485870900 +

Preservation Workflow Entities

Human-managed, human-readable context for major digital object lifecycle phases, including any associated rights determinations:

Accession: why did we decide to digitize/preserve this item?

Ingest: how and when was the object ingested?

Versioning*: who changed this object and when?

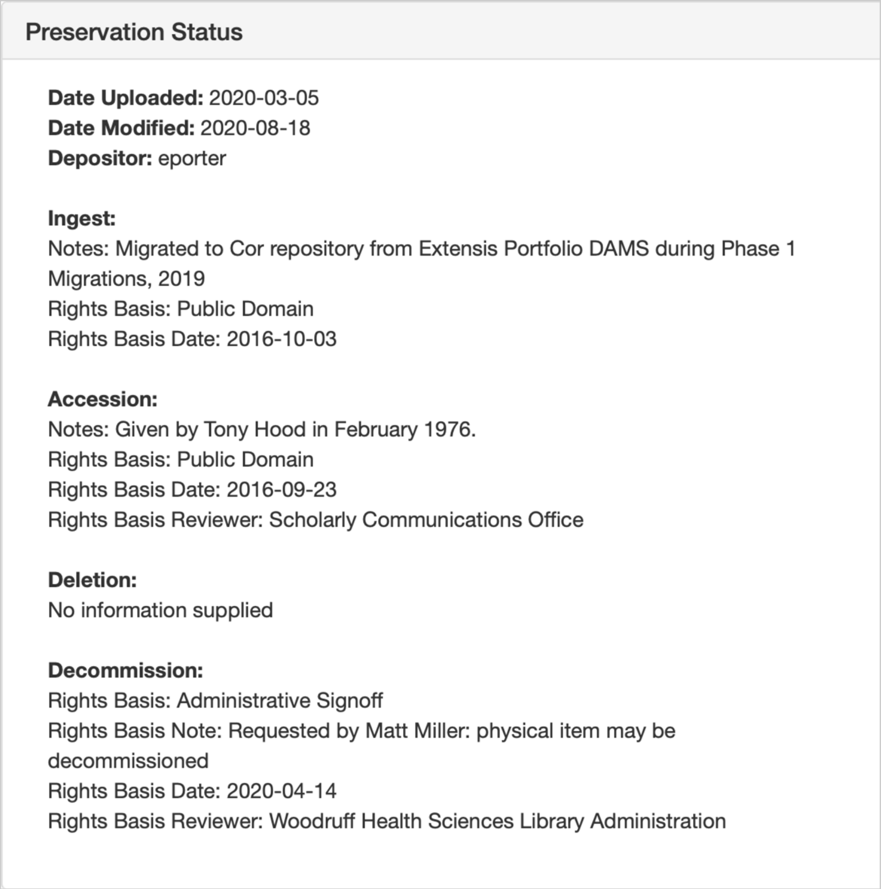
Decommission: why and when did we decide to remove access?

Deletion: why did we delete this content from the repository?

Workflow Metadata

Local namespace defines:

1. Workflow Type
2. Notes
3. Rights Basis
4. Rights Basis Note
5. Rights Basis Review Date
6. Rights Basis Reviewer
7. Rights Basis URI



Preservation Status

Date Uploaded: 2020-03-05
Date Modified: 2020-08-18
Depositor: eporter

Ingest:
Notes: Migrated to Cor repository from Extensis Portfolio DAMS during Phase 1 Migrations, 2019
Rights Basis: Public Domain
Rights Basis Date: 2016-10-03

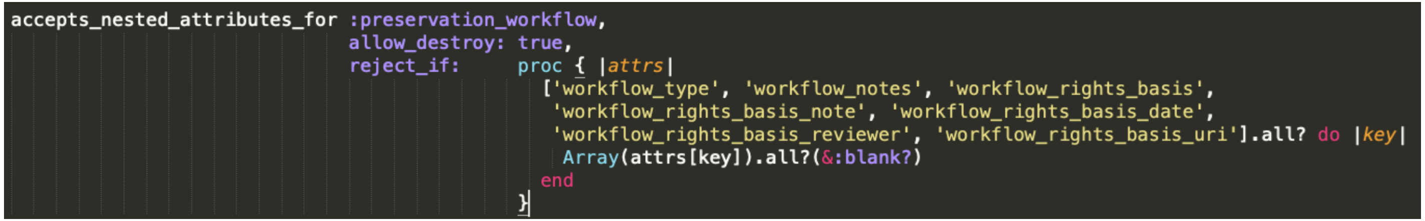
Accession:
Notes: Given by Tony Hood in February 1976.
Rights Basis: Public Domain
Rights Basis Date: 2016-09-23
Rights Basis Reviewer: Scholarly Communications Office

Deletion:
No information supplied

Decommission:
Rights Basis: Administrative Signoff
Rights Basis Note: Requested by Matt Miller: physical item may be decommissioned
Rights Basis Date: 2020-04-14
Rights Basis Reviewer: Woodruff Health Sciences Library Administration

Implementation: Workflows

Similar to Preservation Events, Workflows are also nested objects, however in the work model only. Like events, these are also indexed in Solr for display purposes

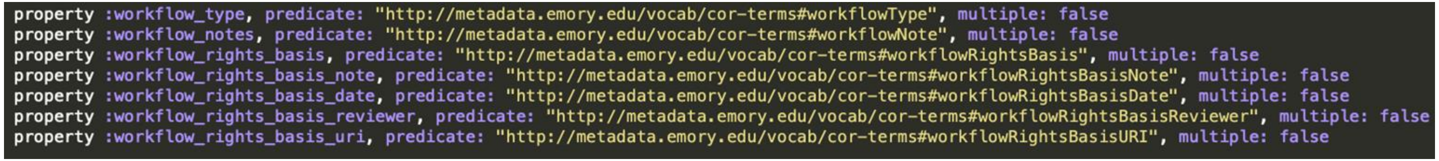


```
accepts_nested_attributes_for :preservation_workflow,
                              allow_destroy: true,
                              reject_if:     proc { |attrs|
                                ['workflow_type', 'workflow_notes', 'workflow_rights_basis',
                                 'workflow_rights_basis_note', 'workflow_rights_basis_date',
                                 'workflow_rights_basis_reviewer', 'workflow_rights_basis_uri'].all? do |key|
                                  Array(attrs[key]).all?(&:blank?)
                                end
                              }
```

Metadata is loaded *after* works are created for Accession, Ingest information (rake task)

Should only be populated once (not editable unless to correct a mistake).

Attributes are single-valued entries.



```
property :workflow_type, predicate: "http://metadata.emory.edu/vocab/cor-terms#workflowType", multiple: false
property :workflow_notes, predicate: "http://metadata.emory.edu/vocab/cor-terms#workflowNote", multiple: false
property :workflow_rights_basis, predicate: "http://metadata.emory.edu/vocab/cor-terms#workflowRightsBasis", multiple: false
property :workflow_rights_basis_note, predicate: "http://metadata.emory.edu/vocab/cor-terms#workflowRightsBasisNote", multiple: false
property :workflow_rights_basis_date, predicate: "http://metadata.emory.edu/vocab/cor-terms#workflowRightsBasisDate", multiple: false
property :workflow_rights_basis_reviewer, predicate: "http://metadata.emory.edu/vocab/cor-terms#workflowRightsBasisReviewer", multiple: false
property :workflow_rights_basis_uri, predicate: "http://metadata.emory.edu/vocab/cor-terms#workflowRightsBasisURI", multiple: false
```

Storage, Backup and Restoration

Content storage

- Pre-ingest shared drive (EFS); fixity checks performed in file transfer (rclone + md5)
- Ingested content files: first copy stored in S3 (US East)
- Second copy: Ansible role compares S3 inventory reports, generates a custom inventory report with new files, copies to second S3 bucket (US West) every 48 hours

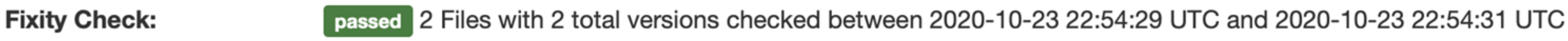
Lessons learned: don't enable S3 versioning; multiple tries needed on S3 batch operations

Backup and restoration testing:

- Application databases
- S3
- Fedora
- SOLR
- Redis (User Activity on FileSets)

Lessons learned: FileSet User Activity trail is stored in Redis; that needs to be backed up too

Fixity Checking



Fixity Check: passed 2 Files with 2 total versions checked between 2020-10-23 22:54:29 UTC and 2020-10-23 22:54:31 UTC

Implementation Notes:

- Using Fedora 4's fixity service: all files in a FileSet are checked using their sha1 checksum
- On-demand checking offered in the View FileSet page UI
- Rake task runs bi-monthly, checks files that haven't been checked in the last 90 days
- Fixity check outcomes logged as Preservation Events
- Investigating AWS serverless fixity service

Lessons learned:

- Hyrax notifications only sent to depositor
- When running in batch, throttle the number of requests to S3!
- Dedicating a fixity queue in sidekiq with a single thread utilized (one file at a time)
- Still figuring out false failures
- Fixity checking has benefited QA process for ingested files (identifying missing files, etc.)

Preservation Reporting

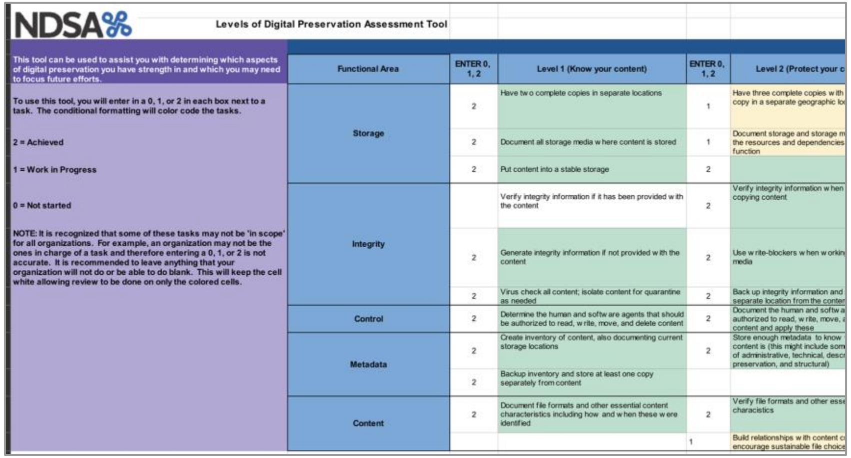
Manually aggregated results reported quarterly to stakeholders and collection stewards:

- Hyrax-supplied information: top mimetypes; total works; total FileSets
- New rake task to count works, filesets, and files per Collection
- AWS: Cloudwatch dashboard provides storage usage and binary file counts
- Fixity checking results (failures only: compiled from Hyrax/Sidekiq/SOLR)

Self-assessing

Gaps and planned features:

- Replication/Dissemination
  - 3+ copies: sending preservation copies to APTrust and other services
  - Fixity checking on additional copies
- Versioning (Works and FileSets; OCFL)



NDSA

Levels of Digital Preservation Assessment Tool

This tool can be used to assist you with determining which aspects of digital preservation you have strength in and which you may need to focus future efforts.

To use this tool, you will enter in a 0, 1, or 2 in each box next to a task. The conditional formatting will color code the tasks.

2 = Achieved

1 = Work in Progress

0 = Not started

NOTE: It is recognized that some of these tasks may not be 'in scope' for all organizations. For example, an organization may not be the ones in charge of a task and therefore entering a 0, 1, or 2 is not accurate. It is recommended to leave anything that your organization will not do or be able to do blank. This will keep the cell white allowing review to be done on only the colored cells.

Functional Area	ENTER 0, 1, 2	Level 1 (Know your content)	ENTER 0, 1, 2	Level 2 (Protect your c
Storage	2	Have two complete copies in separate locations	1	Have three complete copies with copy in a separate geographic lo
	2	Document all storage media where content is stored	1	Document storage and storage m the resources and dependencies function
	2	Put content into a stable storage	2	
Integrity		Verify integrity information if it has been provided with the content	2	Verify integrity information when copying content
	2	Generate integrity information if not provided with the content	2	Use write-blockers when workin media
	2	Virus check all content; isolate content for quarantine as needed	2	Back up integrity information and separate location from the conten
Control	2	Determine the human and software agents that should be authorized to read, write, move, and delete content	2	Document the human and softwa authorized to read, write, move, content and apply these
Metadata	2	Create inventory of content, also documenting current storage locations	2	Store enough metadata to know content is (this might include som of administrative, technical, desc preservation, and structural)
	2	Backup inventory and store at least one copy separately from content		
Content	2	Document file formats and other essential content characteristics including how and when these were identified	2	Verify file formats and other esse characistics
			1	Build relationships with content c encourage sustainable file choice

NDSA Levels of Digital Preservation Assessment Tool

Thank you!

Emory Libraries Digital Preservation Functional Requirements Group

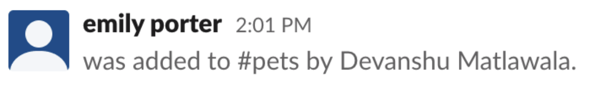
Emory Libraries Software Engineering and Middleware teams

DCE (current & alumni)

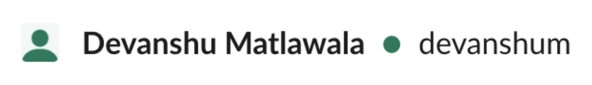
Samvera community consultations & code

emory-libraries/dlp-curate

https://wiki.service.emory.edu/display/DLPP



emily porter 2:01 PM
was added to #pets by Devanshu Matlawala.



Devanshu Matlawala ● devanshum