CUL Digital Preservation Service Ingest Workflows

John Gostick

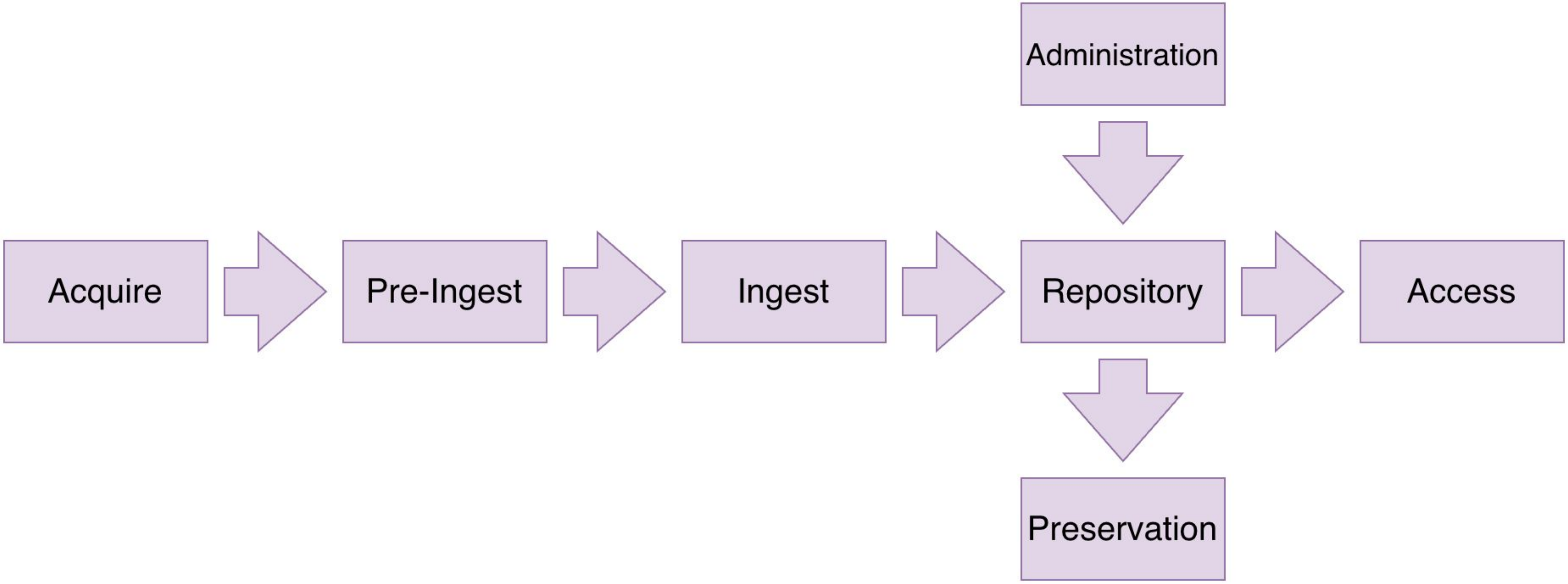
Technical Lead Digital Preservation

Cambridge University Library

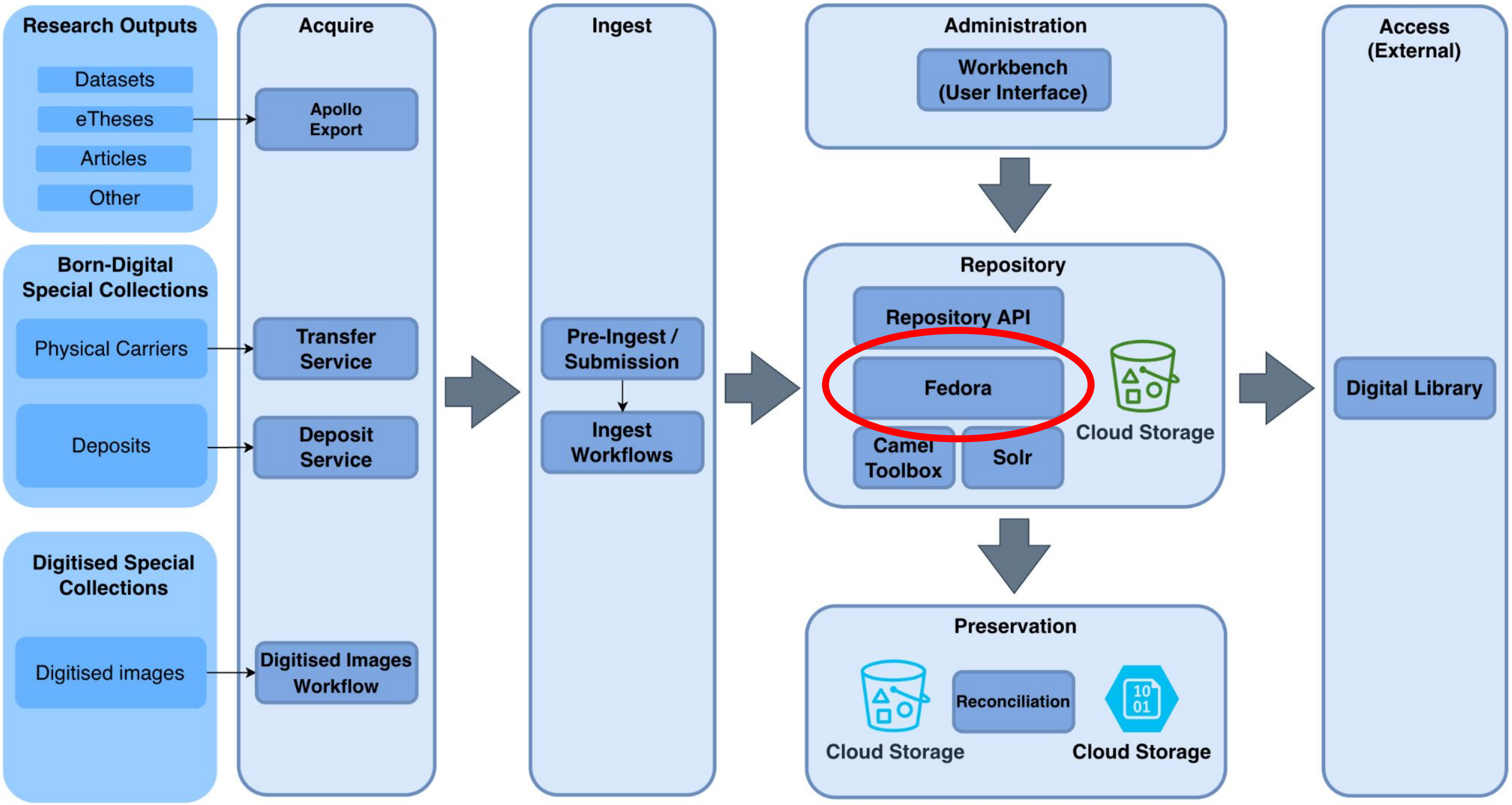
What is the DPS?

- Digital Preservation Service (working title)
- Part of the 5-year Digital Preservation Programme
- Programme goals
    - A central place to store and manage digital assets
    - Improved end to end workflows reducing manual steps
    - Migration to the cloud, optimised where possible, reducing operational costs
    - Improvements to search, discovery and access across collections

DPS Conceptual Model

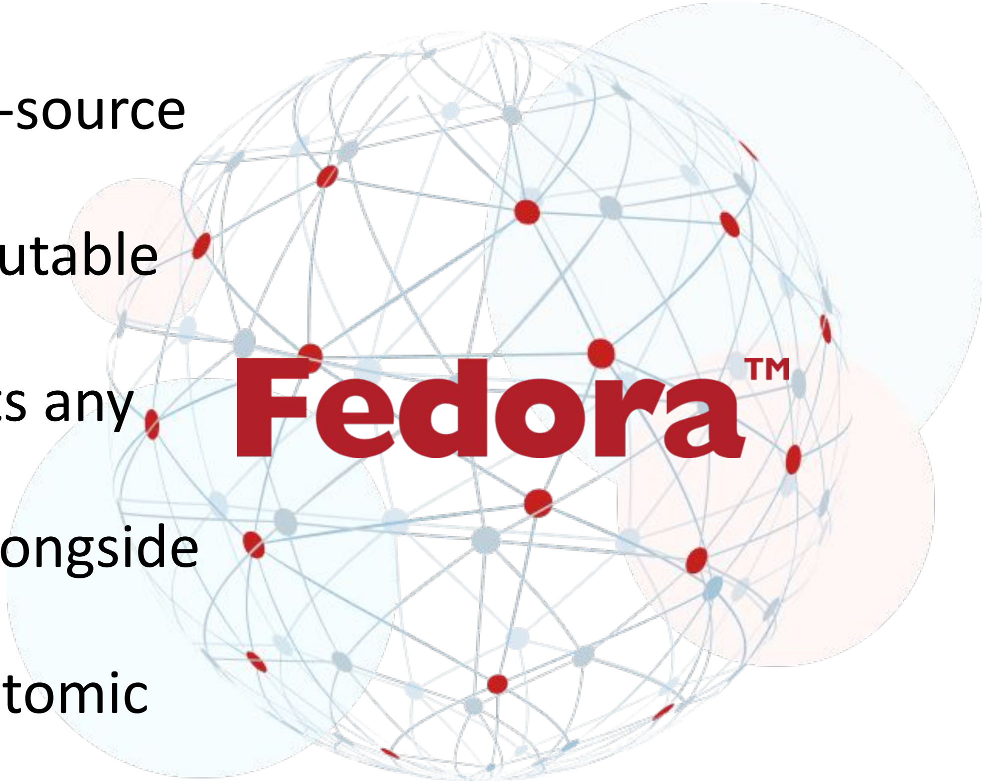


DPS Overview

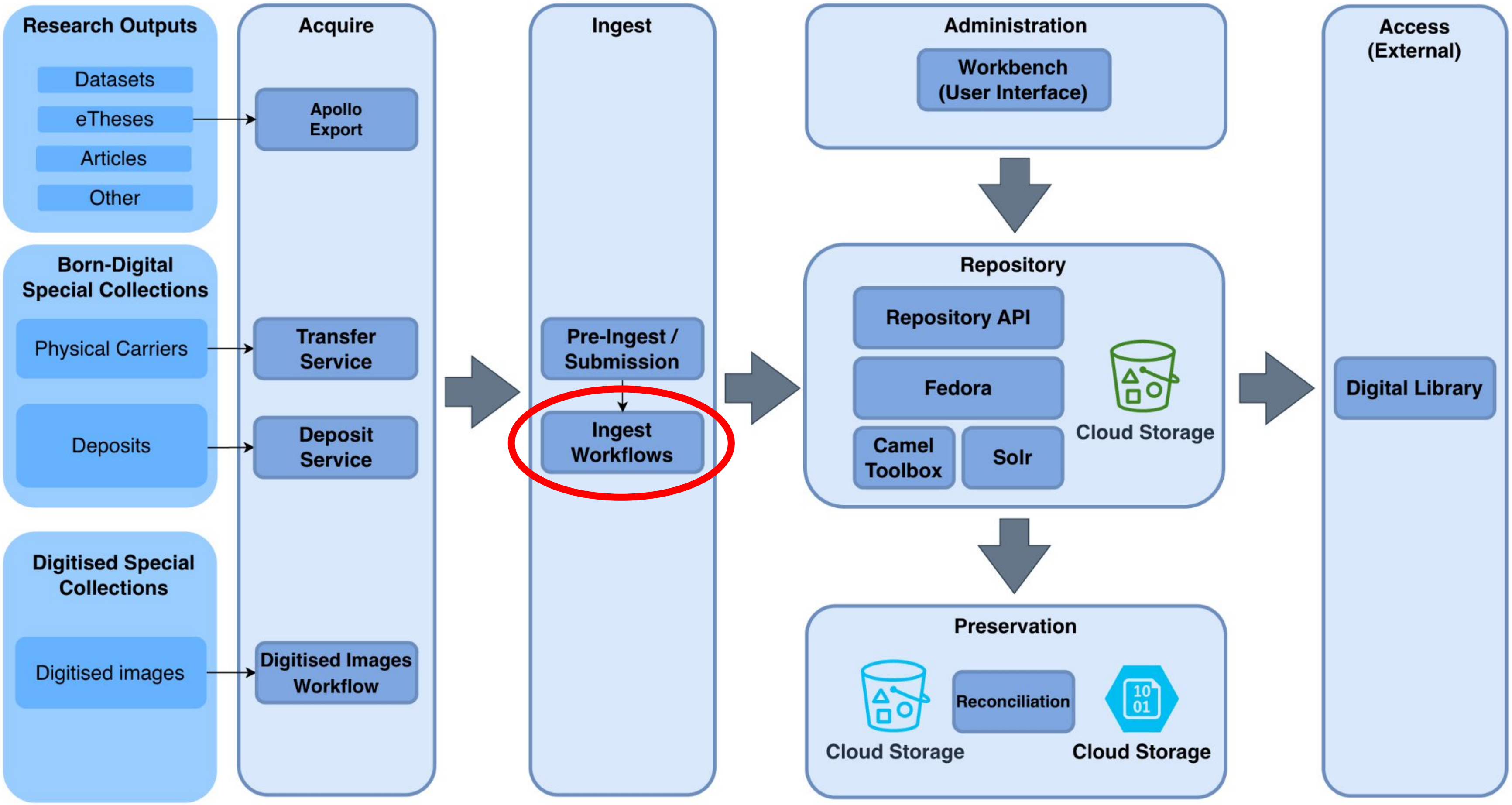


Repository: Fedora 6

- standards-based and open-source
- OCFL on s3
- supports versioning and mutable head extension
- extremely flexible, supports any metadata schema
- stores metadata in OCFL alongside binary data
- supports transactions for atomic operations
- verifies checksums on submission



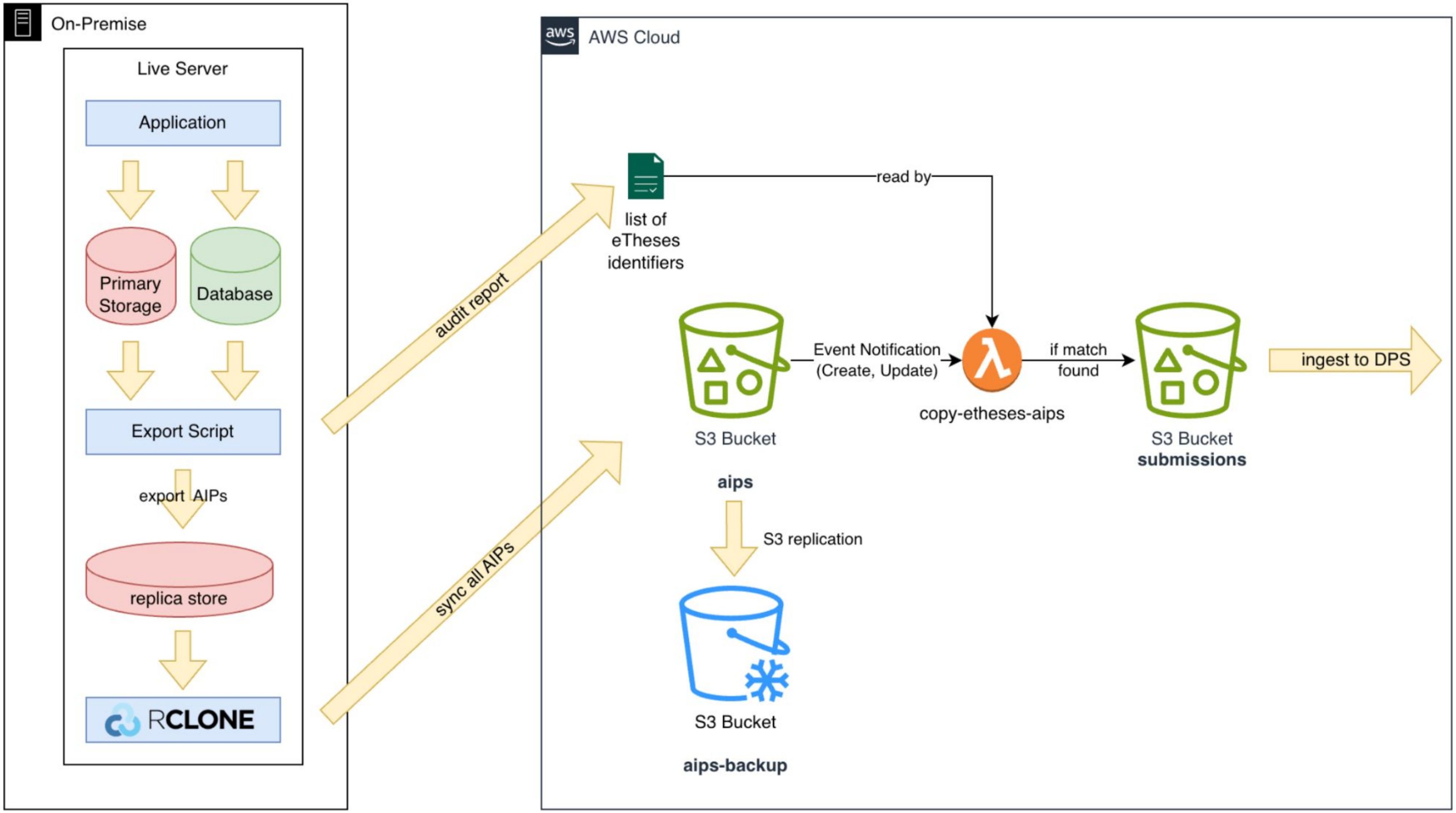
DPS Overview



Ingest Workflow (v1): eTheses

- First MVP ‘eTheses’ workflow built in 2023
- Ingests AIP packages from our institutional repository
- Unpacks then stores the content and metadata in our Fedora 6 repository
- Copies OCFL data into preservation storage
- Cloud native, serverless, event driven - automatically triggered, costs nothing when not running
- Initial backfill ingest of 11832 eTheses (1.2TB) completed in Feb 2024
- BAU ingest since Nov 2024
- v2 for additional streams coming soon!...

Ingest Workflow (v1): Content Submission

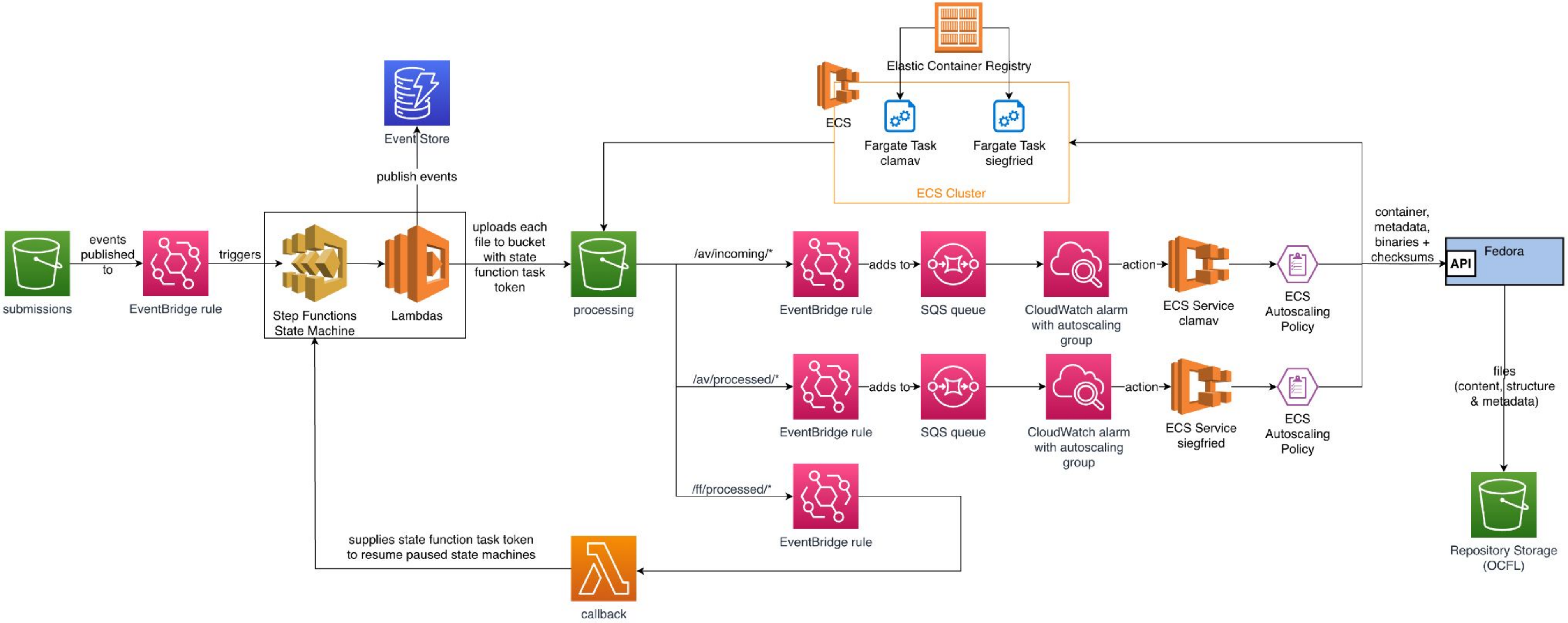


Ingest Workflow (v1): What does it do?

- Check for duplicate submissions
- Unpack submission package and embedded containers
- Parse any supplied metadata
- Scan for viruses
- Identify file formats
- Verify then store files and metadata in repository (Fedora)
- Uses Fedora transactions to ensure completeness
- Create a version in Fedora/OCFL
- Update workflow database table

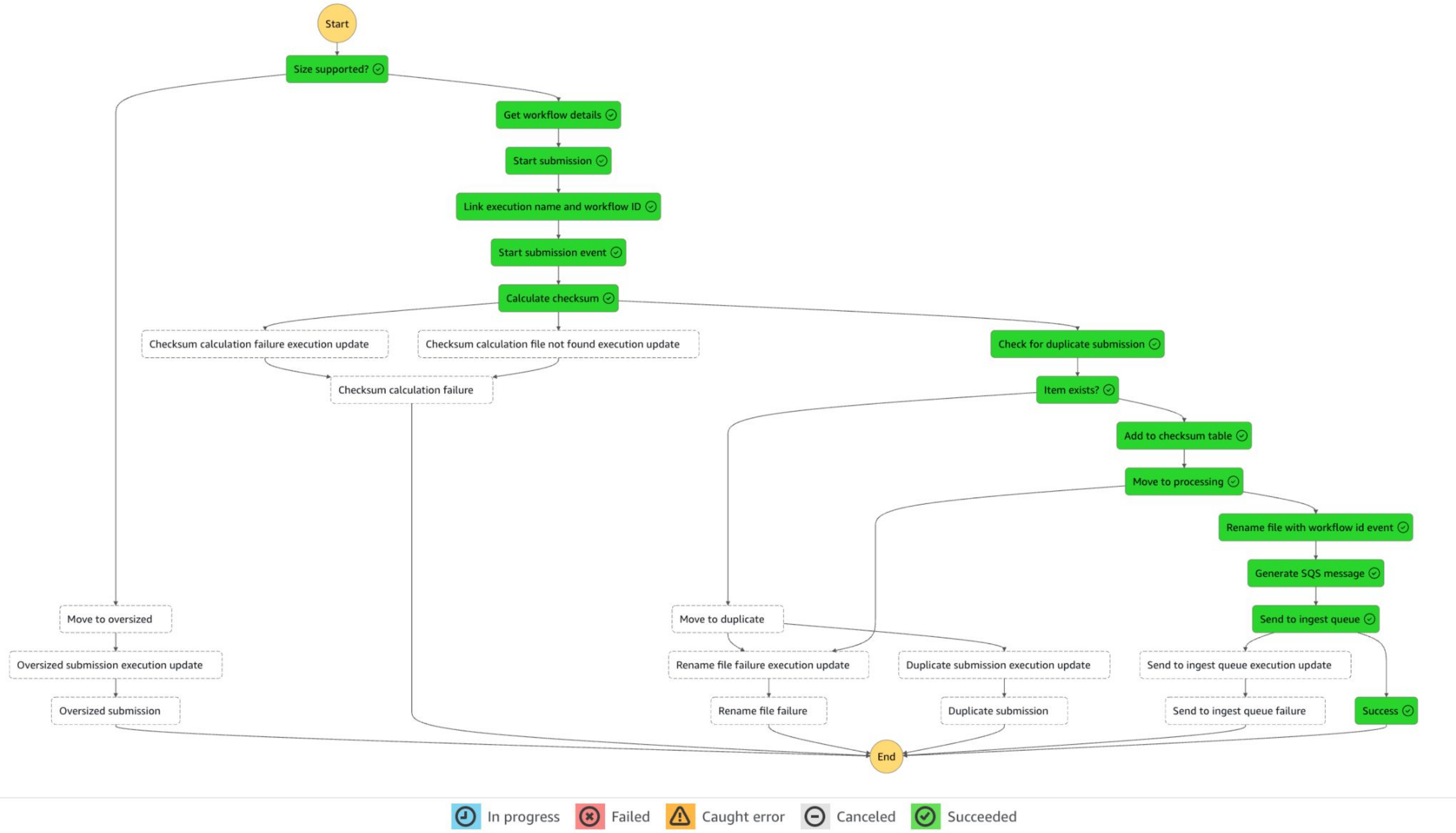
Find out more on our blog:

Ingest Workflow (v1): AWS Architecture



Find out more on our blog:

Ingest Workflow (v1): Submission Example

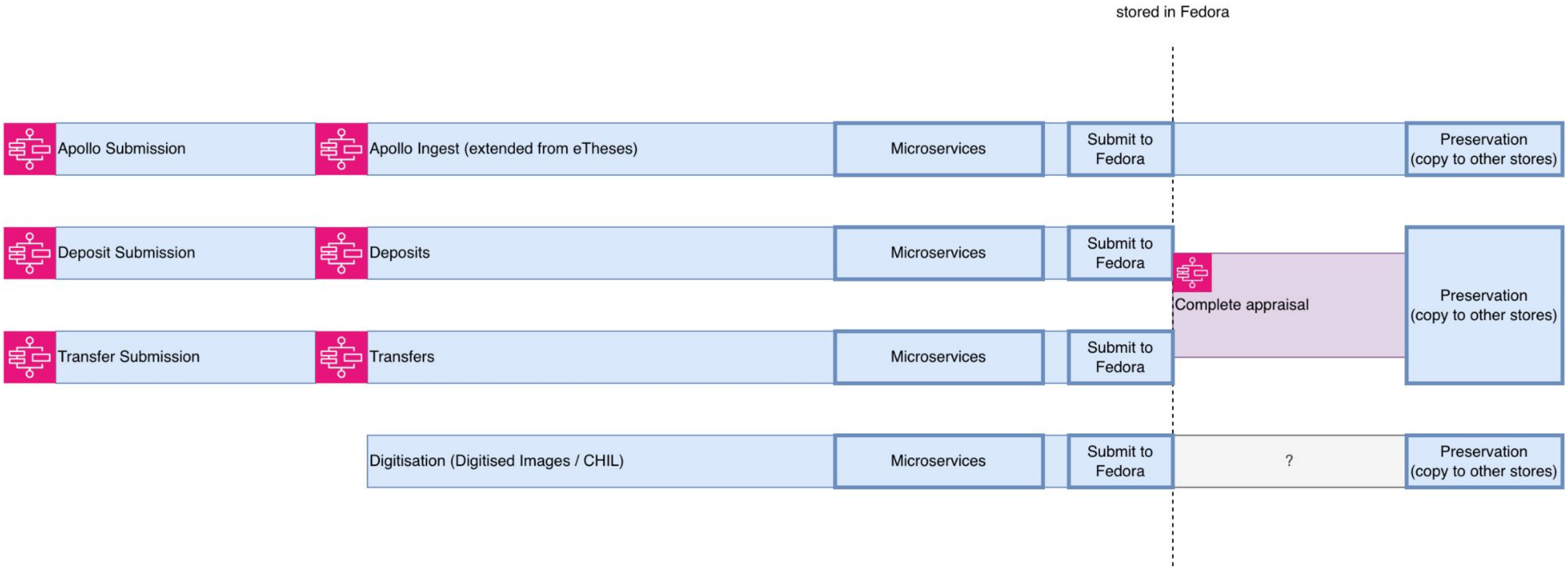

Start
Size supported?
Get workflow details
Start submission
Link execution name and workflow ID
Start submission event
Calculate checksum
Checksum calculation failure execution update
Checksum calculation file not found execution update
Check for duplicate submission
Checksum calculation failure
Item exists?
Add to checksum table
Move to processing
Rename file with workflow id event
Generate SQS message
Move to oversized
Move to duplicate
Send to ingest queue
Oversized submission execution update
Rename file failure execution update
Duplicate submission execution update
Send to ingest queue execution update
Oversized submission
Rename file failure
Duplicate submission
Send to ingest queue failure
Success
End
In progress
Failed
Caught error
Canceled
Succeeded

Find out more on our blog:

Ingest Workflow (v2): What’s changed?

- Re-architected first version to support multiple content sources
- Modularised into shared, reusable components:
    - Lambdas
    - Copy to Preservation and Extraction Step Functions (State Machines)
    - Executions table
    - Configuration table
- Addition of Apache Tika ‘microservice’ for embedded metadata and language extraction
- Support ingest for appraisal and accessioning before preservation
- Removal of Event Store
- Further workflows can now be added with a simple ‘wrapper’

Ingest Workflow (v1)



stored in Fedora

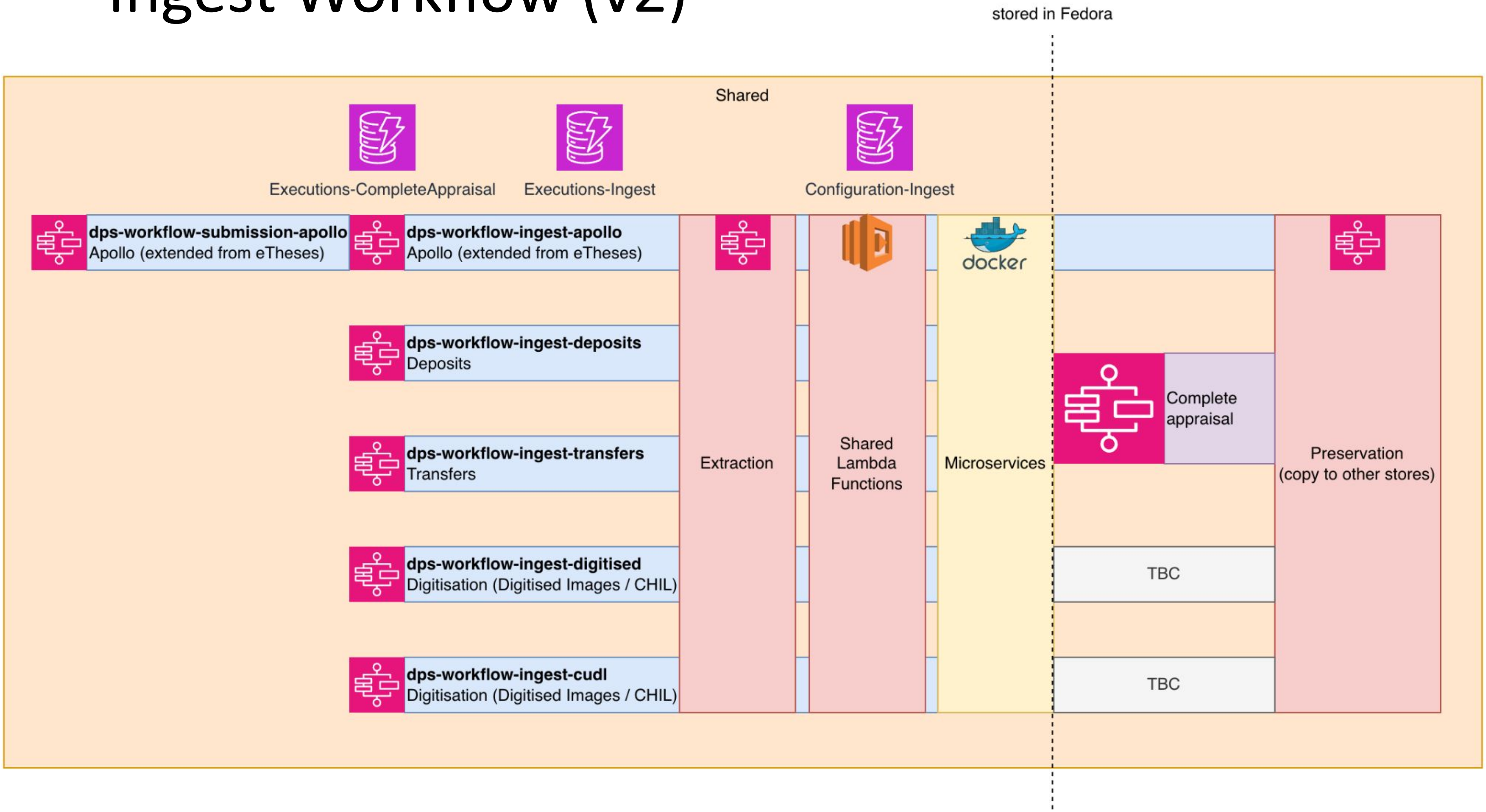
Apollo Submission | Apollo Ingest (extended from eTheses) | Microservices | Submit to Fedora | Preservation (copy to other stores)

Deposit Submission | Deposits | Microservices | Submit to Fedora | Complete appraisal | Preservation (copy to other stores)

Transfer Submission | Transfers | Microservices | Submit to Fedora

Digitisation (Digitised Images / CHIL) | Microservices | Submit to Fedora | ? | Preservation (copy to other stores)

Ingest Workflow (v2)



stored in Fedora

Shared

Executions-CompleteAppraisal

Executions-Ingest

Configuration-Ingest

dps-workflow-submission-apollo
Apollo (extended from eTheses)

dps-workflow-ingest-apollo
Apollo (extended from eTheses)

dps-workflow-ingest-deposits
Deposits

dps-workflow-ingest-transfers
Transfers

dps-workflow-ingest-digitised
Digitisation (Digitised Images / CHIL)

dps-workflow-ingest-cudl
Digitisation (Digitised Images / CHIL)

Extraction

Shared Lambda Functions

Microservices

Complete appraisal

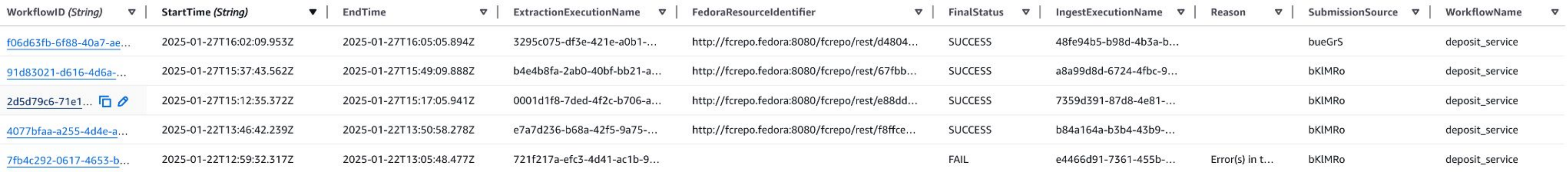
TBC

TBC

Preservation (copy to other stores)

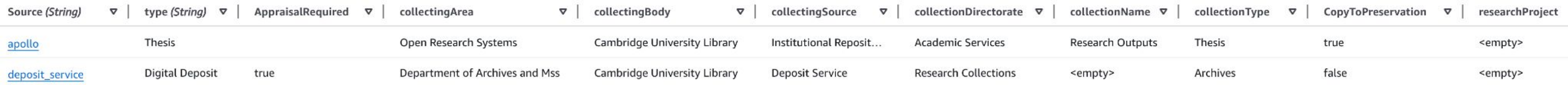
Shared Tables

Ingest Executions



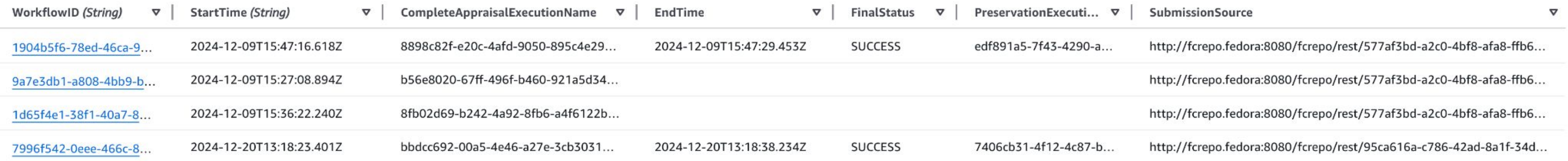
WorkflowID *(String)*	StartTime *(String)*	EndTime	ExtractionExecutionName	FedoraResourceIdentifier	FinalStatus	IngestExecutionName	Reason	SubmissionSource	WorkflowName
f06d63fb-6f88-40a7-ae…	2025-01-27T16:02:09.953Z	2025-01-27T16:05:05.894Z	3295c075-df3e-421e-a0b1-…	http://fcrepo.fedora:8080/fcrepo/rest/d4804…	SUCCESS	48fe94b5-b98d-4b3a-b…		bueGrS	deposit_service
91d83021-d616-4d6a-…	2025-01-27T15:37:43.562Z	2025-01-27T15:49:09.888Z	b4e4b8fa-2ab0-40bf-bb21-a…	http://fcrepo.fedora:8080/fcrepo/rest/67fbb…	SUCCESS	a8a99d8d-6724-4fbc-9…		bKlMRo	deposit_service
2d5d79c6-71e1…	2025-01-27T15:12:35.372Z	2025-01-27T15:17:05.941Z	0001d1f8-7ded-4f2c-b706-a…	http://fcrepo.fedora:8080/fcrepo/rest/e88dd…	SUCCESS	7359d391-87d8-4e81-…		bKlMRo	deposit_service
4077bfaa-a255-4d4e-a…	2025-01-22T13:46:42.239Z	2025-01-22T13:50:58.278Z	e7a7d236-b68a-42f5-9a75-…	http://fcrepo.fedora:8080/fcrepo/rest/f8ffce…	SUCCESS	b84a164a-b3b4-43b9-…		bKlMRo	deposit_service
7fb4c292-0617-4653-b…	2025-01-22T12:59:32.317Z	2025-01-22T13:05:48.477Z	721f217a-efc3-4d41-ac1b-9…		FAIL	e4466d91-7361-455b-…	Error(s) in t…	bKlMRo	deposit_service

Ingest Configuration



Source *(String)*	type *(String)*	AppraisalRequired	collectingArea	collectingBody	collectingSource	collectionDirectorate	collectionName	collectionType	CopyToPreservation	researchProject
apollo	Thesis		Open Research Systems	Cambridge University Library	Institutional Reposit…	Academic Services	Research Outputs	Thesis	true	<empty>
deposit_service	Digital Deposit	true	Department of Archives and Mss	Cambridge University Library	Deposit Service	Research Collections	<empty>	Archives	false	<empty>

Complete Appraisal



WorkflowID *(String)*	StartTime *(String)*	CompleteAppraisalExecutionName	EndTime	FinalStatus	PreservationExecuti…	SubmissionSource
1904b5f6-78ed-46ca-9…	2024-12-09T15:47:16.618Z	8898c82f-e20c-4afd-9050-895c4e29…	2024-12-09T15:47:29.453Z	SUCCESS	edf891a5-7f43-4290-a…	http://fcrepo.fedora:8080/fcrepo/rest/577af3bd-a2c0-4bf8-afa8-ffb6…
9a7e3db1-a808-4bb9-b…	2024-12-09T15:27:08.894Z	b56e8020-67ff-496f-b460-921a5d34…				http://fcrepo.fedora:8080/fcrepo/rest/577af3bd-a2c0-4bf8-afa8-ffb6…
1d65f4e1-38f1-40a7-8…	2024-12-09T15:36:22.240Z	8fb02d69-b242-4a92-8fb6-a4f6122b…				http://fcrepo.fedora:8080/fcrepo/rest/577af3bd-a2c0-4bf8-afa8-ffb6…
7996f542-0eee-466c-8…	2024-12-20T13:18:23.401Z	bbdcc692-00a5-4e46-a27e-3cb3031…	2024-12-20T13:18:38.234Z	SUCCESS	7406cb31-4f12-4c87-b…	http://fcrepo.fedora:8080/fcrepo/rest/95ca616a-c786-42ad-8a1f-34d…

Thank you

Any questions?

jag245@cam.ac.uk

digitalpreservation@lib.cam.ac.uk

https://digitalpreservation-blog.lib.cam.ac.uk/