archivesDIRECT

# The ArchivesDirect pilot: Road-testing Archivematica hosting in DuraCloud

Amy Rushing (UTSA), Julianna Barrera-Gomez (UTSA) and Courtney C. Mumma (Artefactual Systems)

TCDL, Austin, TX, April 27, 2015, Session 4A

# Overview

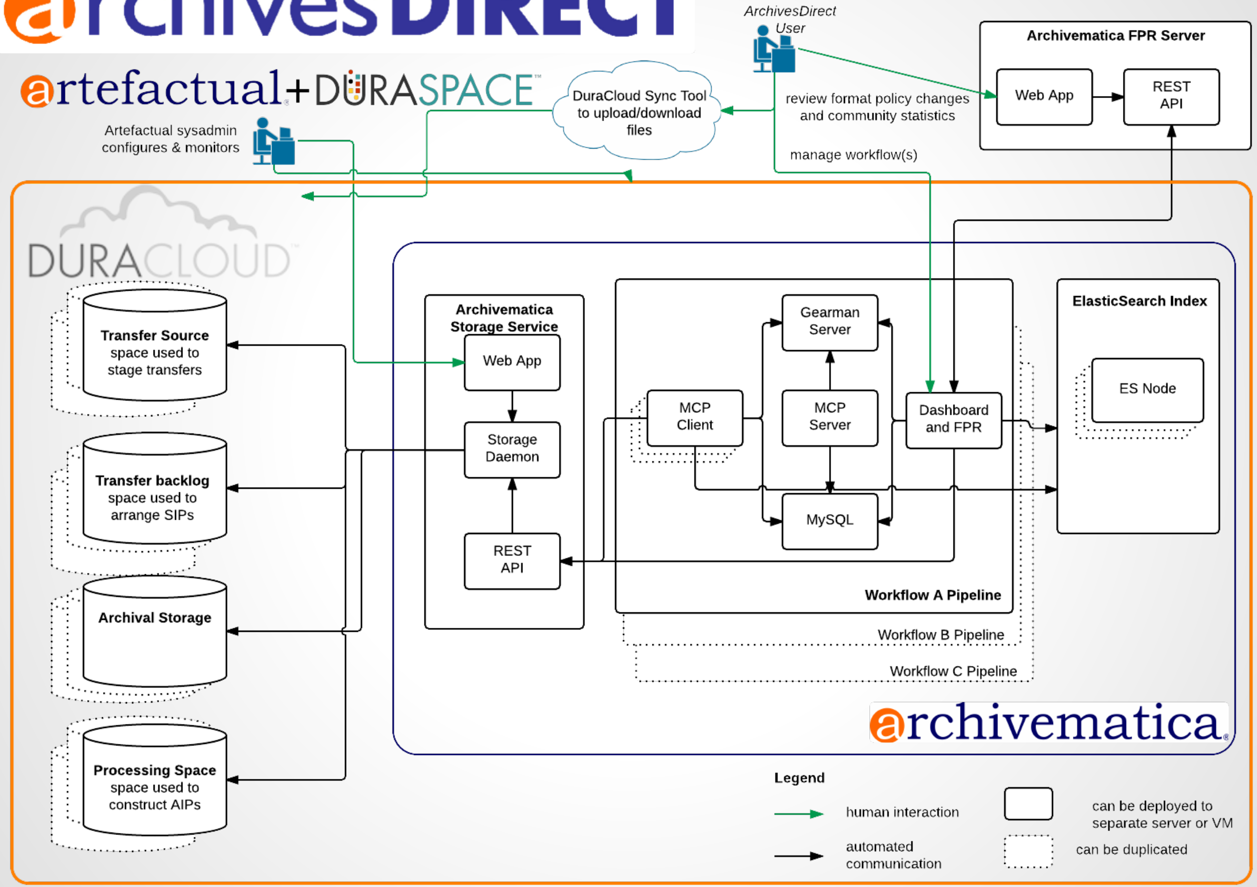
- The pilot
- Administrative experience - UTSA partner
- Operational experience
- Q&A and discussion

# What were the pilot’s goals?

- standards-based digital preservation packages in secure long-term storage
  - Archivematica + DuraCloud
  - hosted, web-based solution
  - customizable to match your needs/workflows/content
  - no vendor/content lock-in
- identify service needs and scope

# archivesDIRECT

artefactual+DURASPACE



ArchivesDirect User

Archivematica FPR Server

Web App

REST API

review format policy changes and community statistics

manage workflow(s)

DuraCloud Sync Tool to upload/download files

Artefactual sysadmin configures & monitors

DURACLOUD

**Transfer Source** space used to stage transfers

**Transfer backlog** space used to arrange SIPs

**Archival Storage**

**Processing Space** space used to construct AIPs

**Archivematica Storage Service**

Web App

Storage Daemon

REST API

Gearman Server

MCP Client

MCP Server

Dashboard and FPR

MySQL

**Workflow A Pipeline**

Workflow B Pipeline

Workflow C Pipeline

**ElasticSearch Index**

ES Node

archivematica

**Legend**

human interaction

automated communication

can be deployed to separate server or VM

can be duplicated

# Who were the pilot volunteers?

Berea College

Huntington Library

Illinois-Wesleyan University

Kansas State University

North Carolina Dept of Cultural Resources

Pepperdine University

Phillips Academy Andover

University of Texas at San Antonio

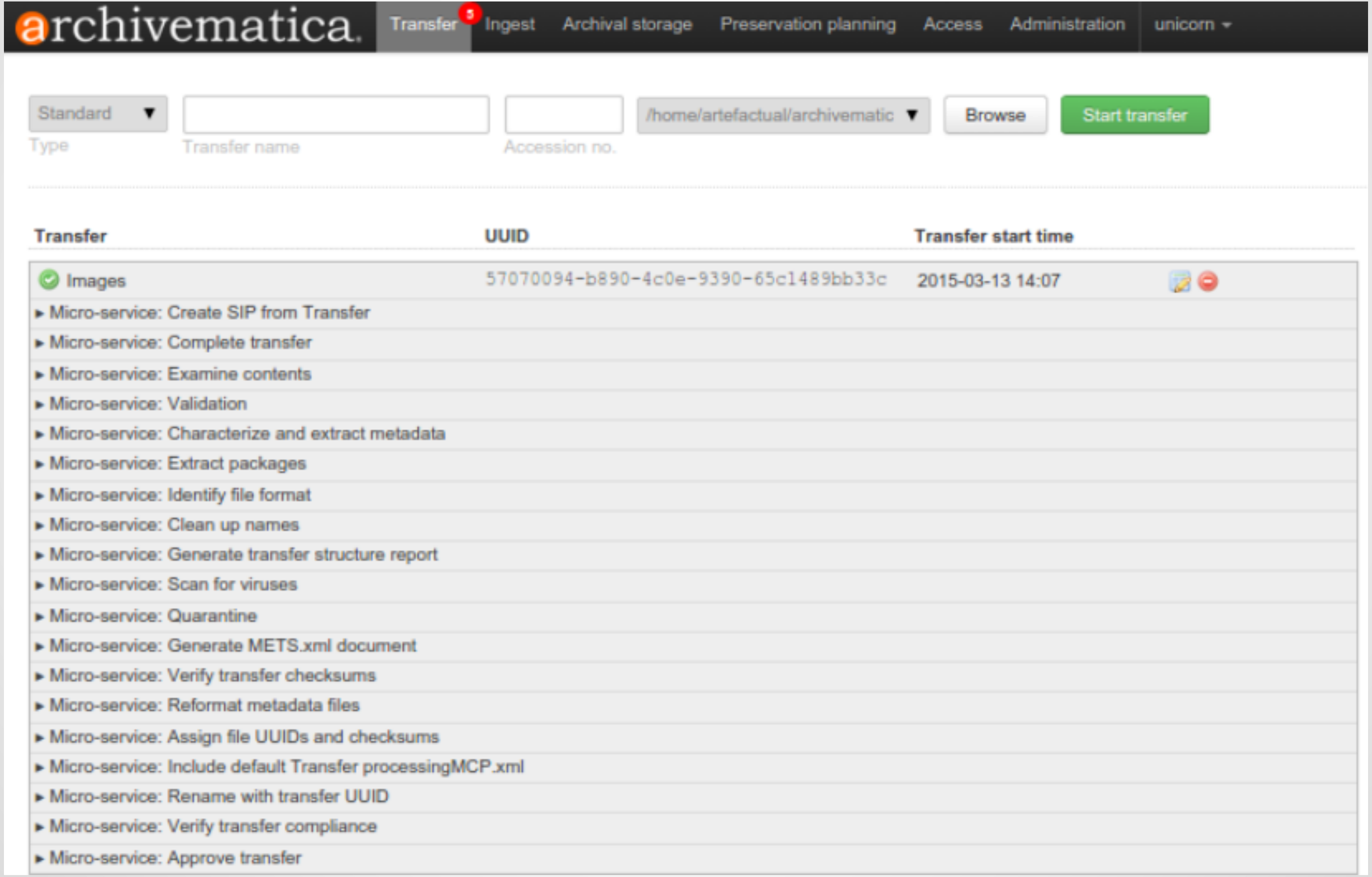
University of Washington

# What was the pilot structure?

- assess content
    - born-digital, digitized, MD and submission documentation
    - size and scale

- assess workflows for distinct content
    - preservation and/or access copies? service masters? manual normalization?
    - transcription, forensic analysis, directory structure for OO?
    - format identification tool, error handling, compression and extraction preferences?

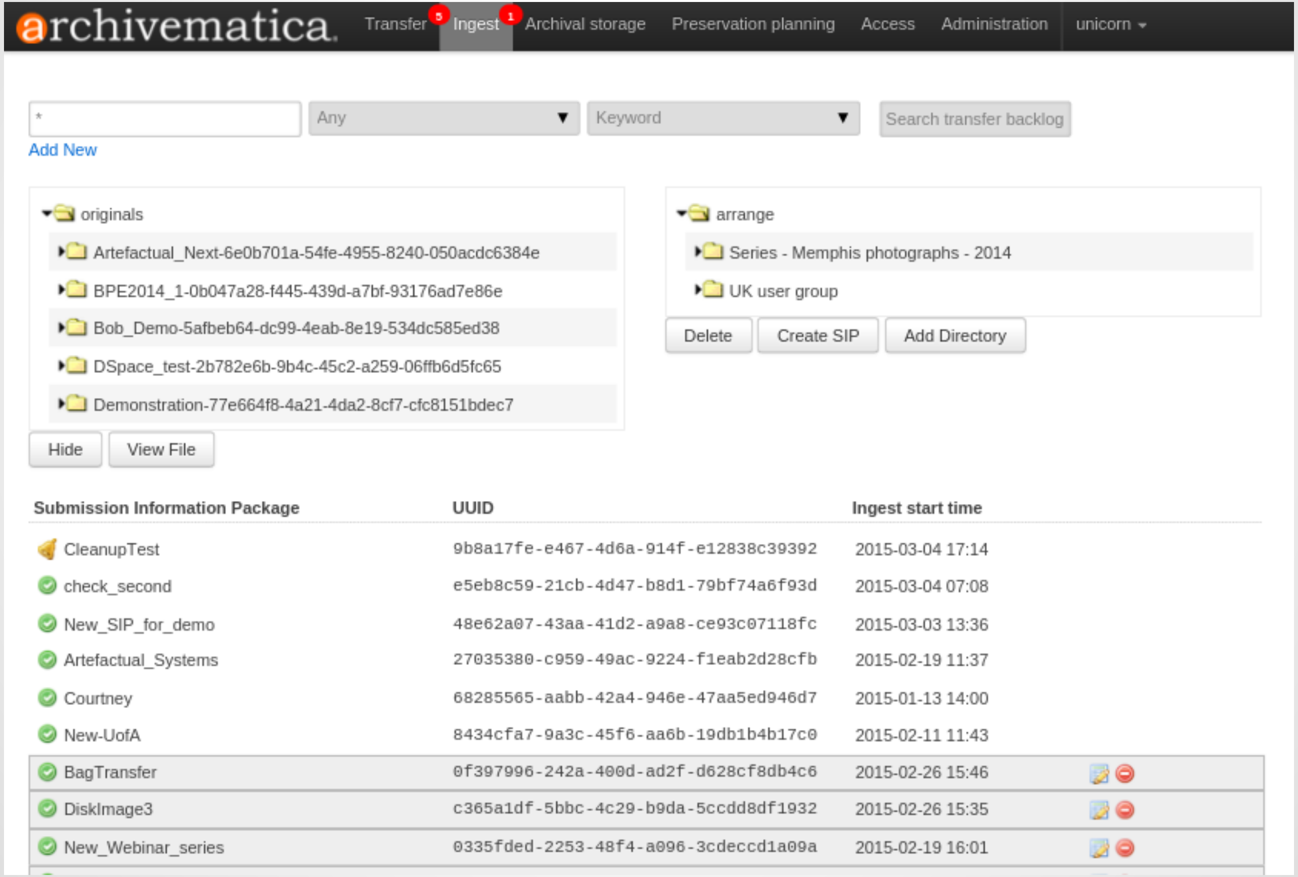
# Structure, cont'd



| Transfer | UUID | Transfer start time |
|---|---|---|
| Images | 57070094-b890-4c0e-9390-65c1489bb33c | 2015-03-13 14:07 |

- ▸ Micro-service: Create SIP from Transfer
- ▸ Micro-service: Complete transfer
- ▸ Micro-service: Examine contents
- ▸ Micro-service: Validation
- ▸ Micro-service: Characterize and extract metadata
- ▸ Micro-service: Extract packages
- ▸ Micro-service: Identify file format
- ▸ Micro-service: Clean up names
- ▸ Micro-service: Generate transfer structure report
- ▸ Micro-service: Scan for viruses
- ▸ Micro-service: Quarantine
- ▸ Micro-service: Generate METS.xml document
- ▸ Micro-service: Verify transfer checksums
- ▸ Micro-service: Reformat metadata files
- ▸ Micro-service: Assign file UUIDs and checksums
- ▸ Micro-service: Include default Transfer processingMCP.xml
- ▸ Micro-service: Rename with transfer UUID
- ▸ Micro-service: Verify transfer compliance
- ▸ Micro-service: Approve transfer

# Structure, cont’d



- originals
  - Artefactual_Next-6e0b701a-54fe-4955-8240-050acdc6384e
  - BPE2014_1-0b047a28-f445-439d-a7bf-93176ad7e86e
  - Bob_Demo-5afbeb64-dc99-4eab-8e19-534dc585ed38
  - DSpace_test-2b782e6b-9b4c-45c2-a259-06ffb6d5fc65
  - Demonstration-77e664f8-4a21-4da2-8cf7-cfc8151bdec7

- arrange
  - Series - Memphis photographs - 2014
  - UK user group

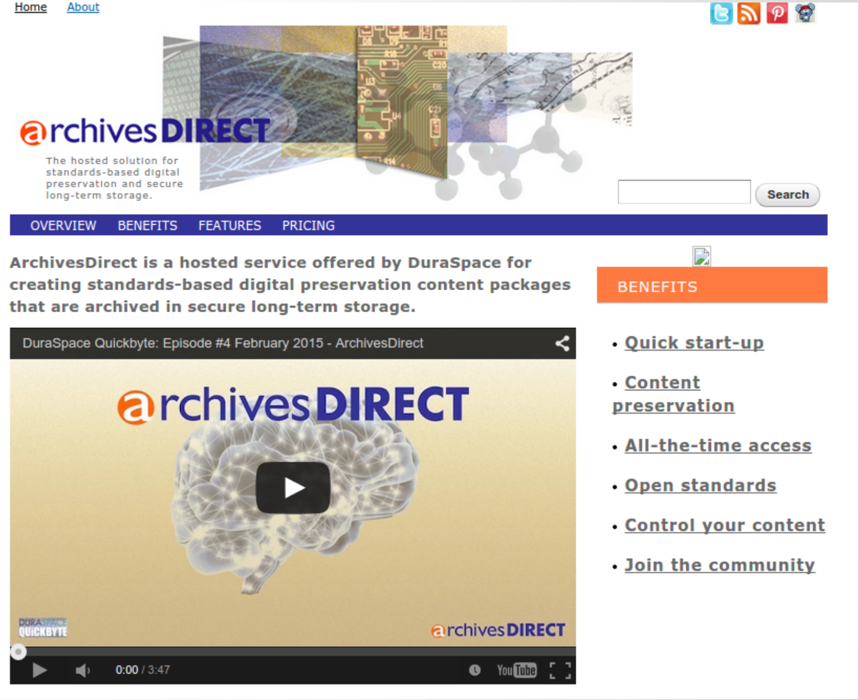
| Submission Information Package | UUID | Ingest start time |
|---|---|---|
| CleanupTest | 9b8a17fe-e467-4d6a-914f-e12838c39392 | 2015-03-04 17:14 |
| check_second | e5eb8c59-21cb-4d47-b8d1-79bf74a6f93d | 2015-03-04 07:08 |
| New_SIP_for_demo | 48e62a07-43aa-41d2-a9a8-ce93c07118fc | 2015-03-03 13:36 |
| Artefactual_Systems | 27035380-c959-49ac-9224-f1eab2d28cfb | 2015-02-19 11:37 |
| Courtney | 68285565-aabb-42a4-946e-47aa5ed946d7 | 2015-01-13 14:00 |
| New-UofA | 8434cfa7-9a3c-45f6-aa6b-19db1b4b17c0 | 2015-02-11 11:43 |
| BagTransfer | 0f397996-242a-400d-ad2f-d628cf8db4c6 | 2015-02-26 15:46 |
| DiskImage3 | c365a1df-5bbc-4c29-b9da-5ccdd8df1932 | 2015-02-26 15:35 |
| New_Webinar_series | 0335fded-2253-48f4-a096-3cdeccd1a09a | 2015-02-19 16:01 |

# Structure, cont’d

- intensive training and content analysis guidance
- pilot wiki for sharing workflows and results
- pilot-support and pilot-discussion lists
- agile development processes with rapid analysis/ upgrade/testing cycles

# What were the pilot outcomes?

- ArchivesDirect public offering and price structure
- toolkits and documentation
- DuraCloud integration and scalability enhancements to Archivematica
- training and support strategy



The hosted solution for standards-based digital preservation and secure long-term storage.

OVERVIEW BENEFITS FEATURES PRICING

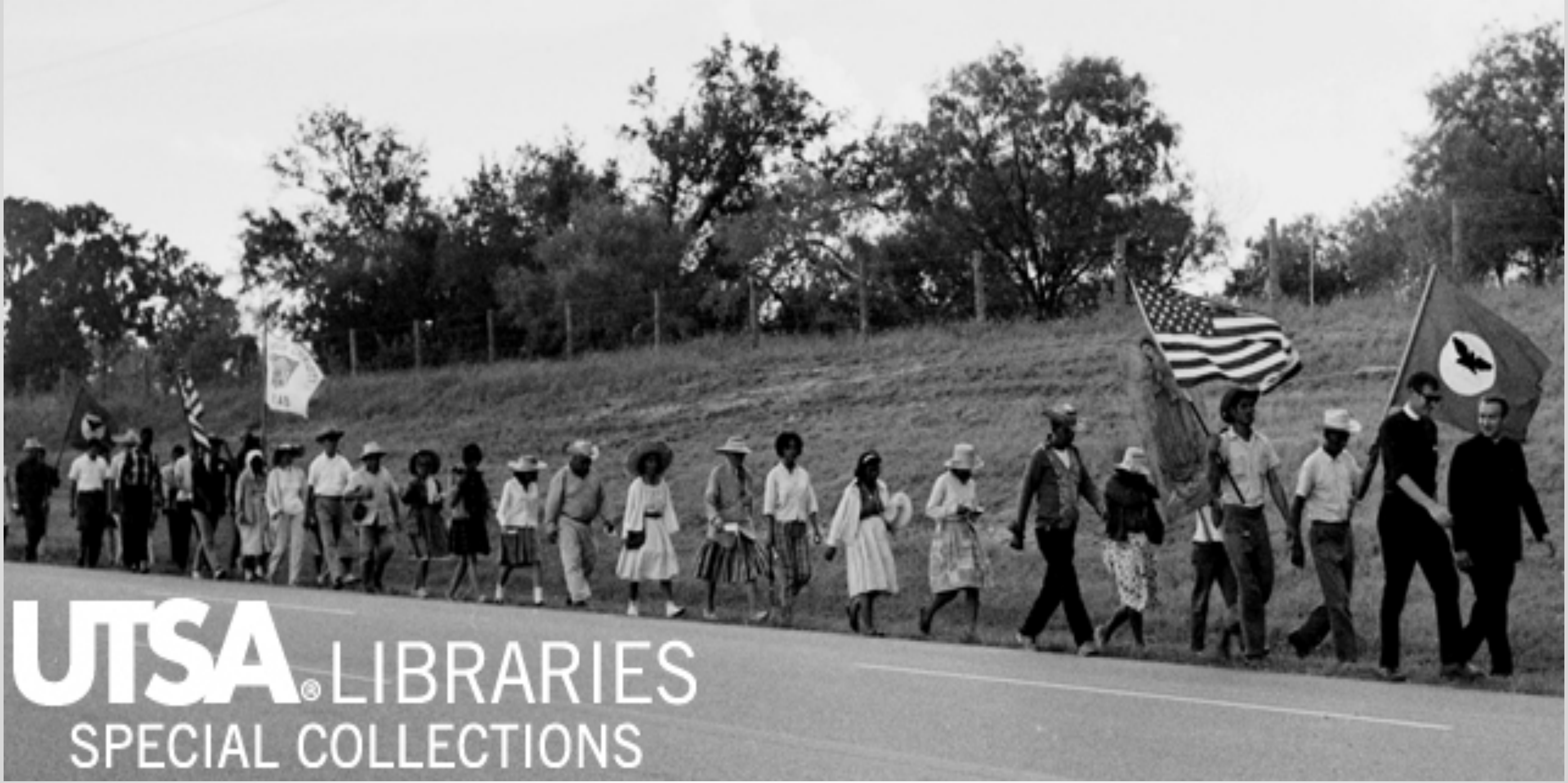
ArchivesDirect is a hosted service offered by DuraSpace for creating standards-based digital preservation content packages that are archived in secure long-term storage.

DuraSpace Quickbyte: Episode #4 February 2015 - ArchivesDirect

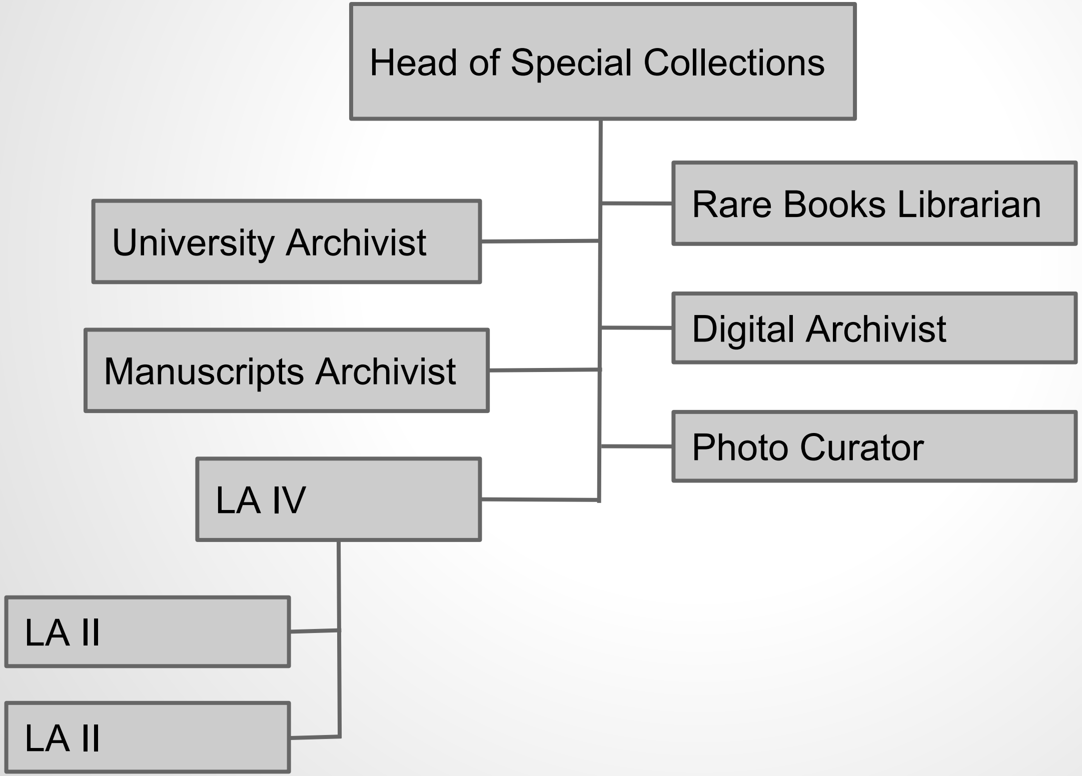
BENEFITS

- Quick start-up
- Content preservation
- All-the-time access
- Open standards
- Control your content
- Join the community

archivesdirect.org


UTSA® LIBRARIES
SPECIAL COLLECTIONS

# Special Collections



Head of Special Collections

University Archivist

Manuscripts Archivist

LA IV

LA II

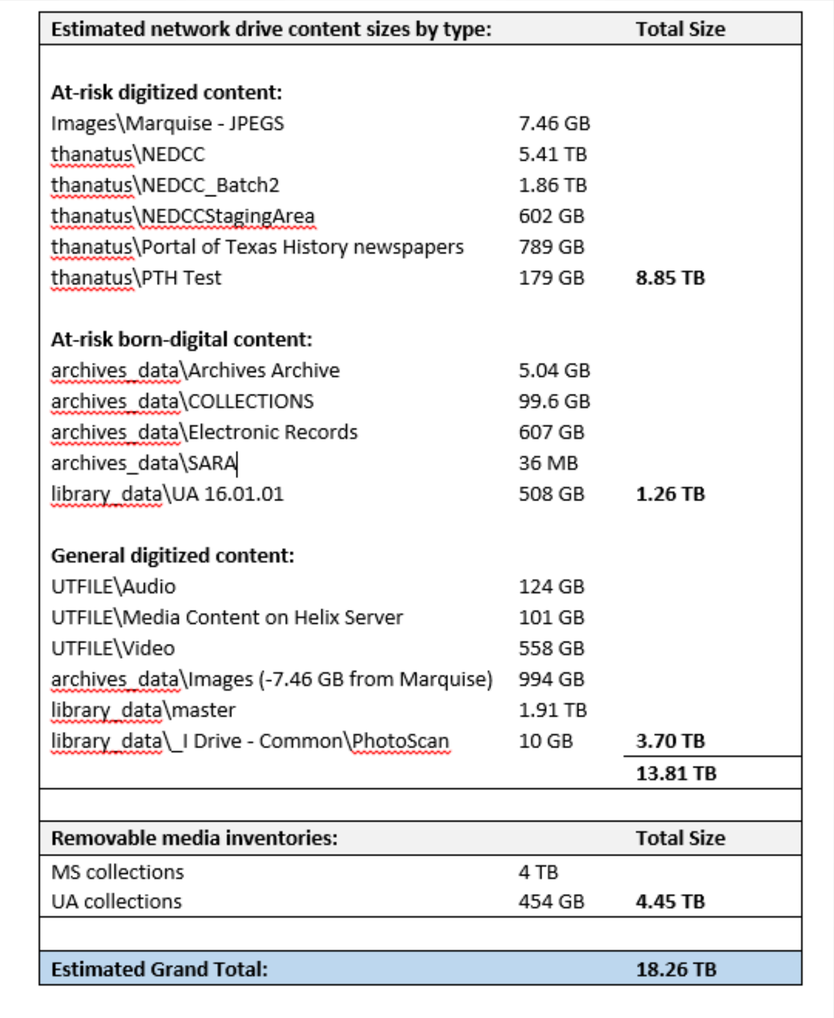
LA II

Rare Books Librarian

Digital Archivist

Photo Curator

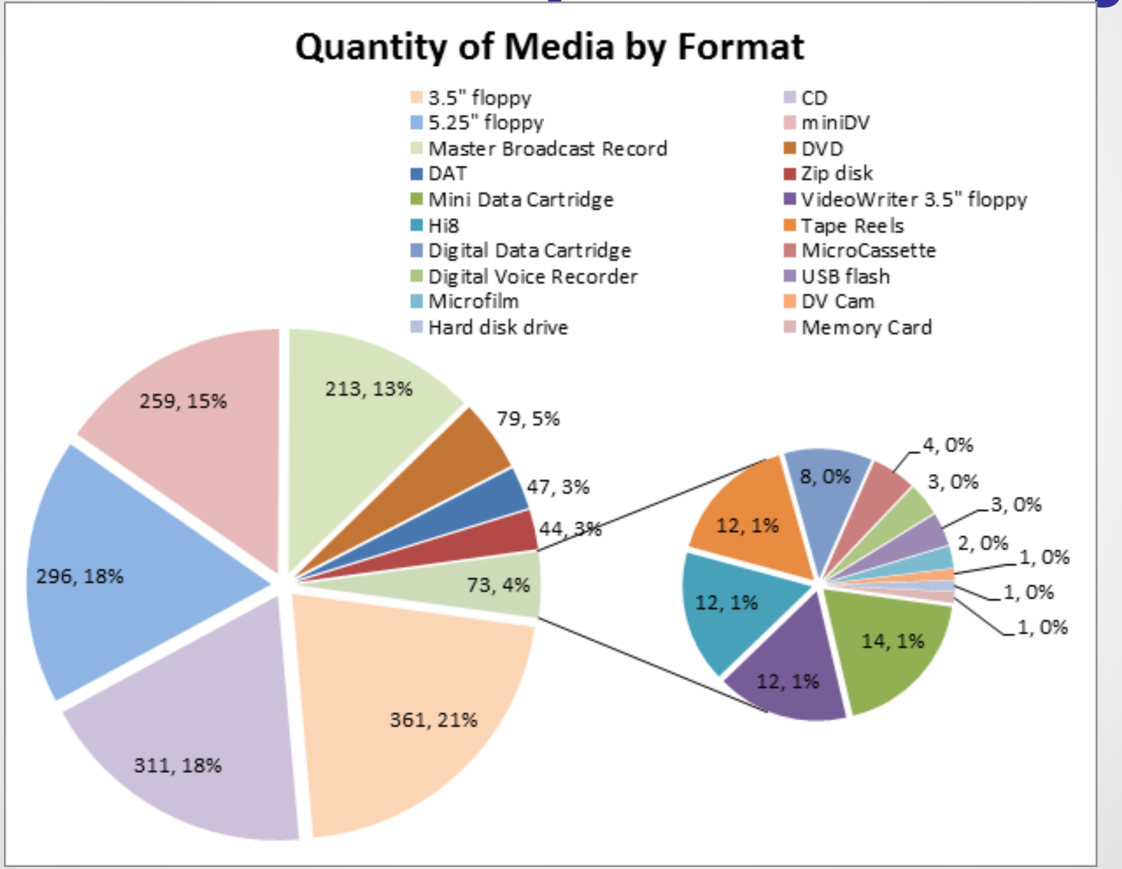
# IT @ UTSA



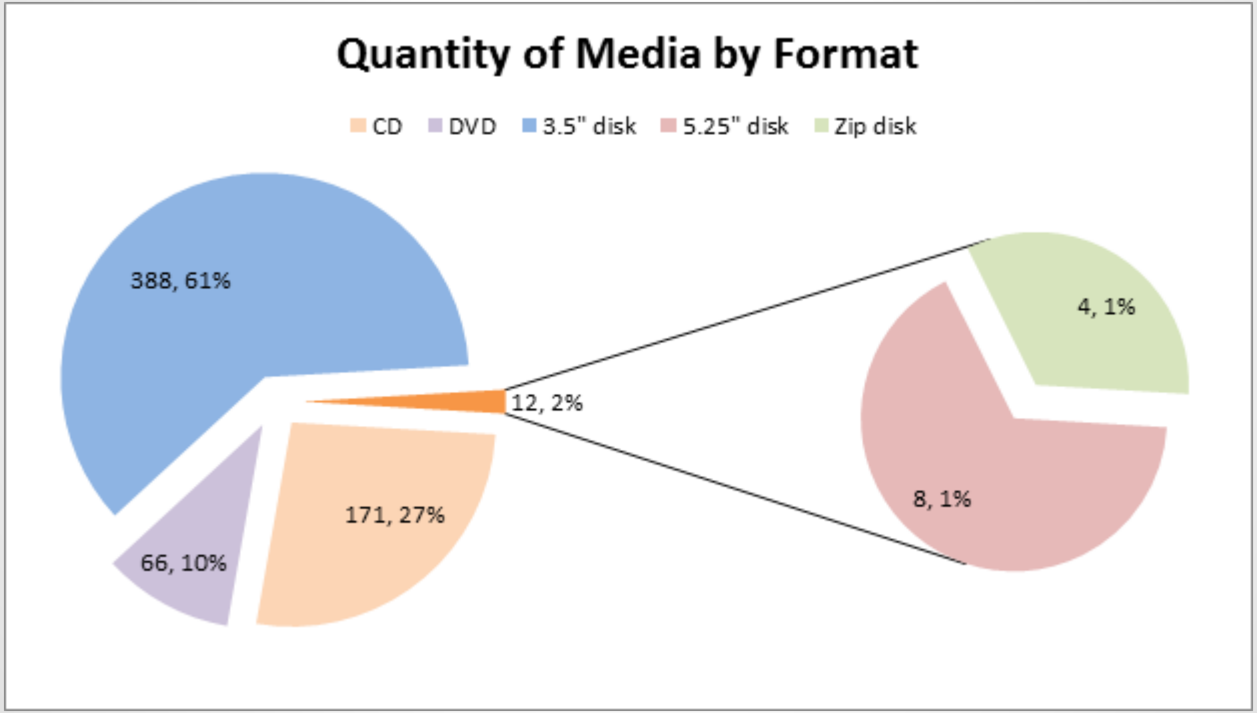
| Estimated network drive content sizes by type: | | Total Size |
|---|---|---|
| **At-risk digitized content:** | | |
| Images\Marquise - JPEGS | 7.46 GB | |
| thanatus\NEDCC | 5.41 TB | |
| thanatus\NEDCC_Batch2 | 1.86 TB | |
| thanatus\NEDCCStagingArea | 602 GB | |
| thanatus\Portal of Texas History newspapers | 789 GB | |
| thanatus\PTH Test | 179 GB | **8.85 TB** |
| **At-risk born-digital content:** | | |
| archives_data\Archives Archive | 5.04 GB | |
| archives_data\COLLECTIONS | 99.6 GB | |
| archives_data\Electronic Records | 607 GB | |
| archives_data\SARA | 36 MB | |
| library_data\UA 16.01.01 | 508 GB | **1.26 TB** |
| **General digitized content:** | | |
| UTFILE\Audio | 124 GB | |
| UTFILE\Media Content on Helix Server | 101 GB | |
| UTFILE\Video | 558 GB | |
| archives_data\Images (-7.46 GB from Marquise) | 994 GB | |
| library_data\master | 1.91 TB | |
| library_data_I Drive - Common\PhotoScan | 10 GB | **3.70 TB** |
| | | **13.81 TB** |
| **Removable media inventories:** | | **Total Size** |
| MS collections | 4 TB | |
| UA collections | 454 GB | **4.45 TB** |
| **Estimated Grand Total:** | | **18.26 TB** |

# Jump in Initiative

# Jump In Manuscripts Inventory



# Jump In University Archives Inventory



# Pilot Content overview

*San Antonio Light* digitized negatives

- Tiffs and jpegs, plus metadata in .csv

Born Digital material from Dr. Ellen Riojas Clark

- Facebook export, Outlook email

Born Digital material from Dean George Perry

- .msg email files with attachments, Word files

Born Digital material from Dean Dan Gelo

- Word and .pdf

*The Marquise* digitized newspapers

- jpegs

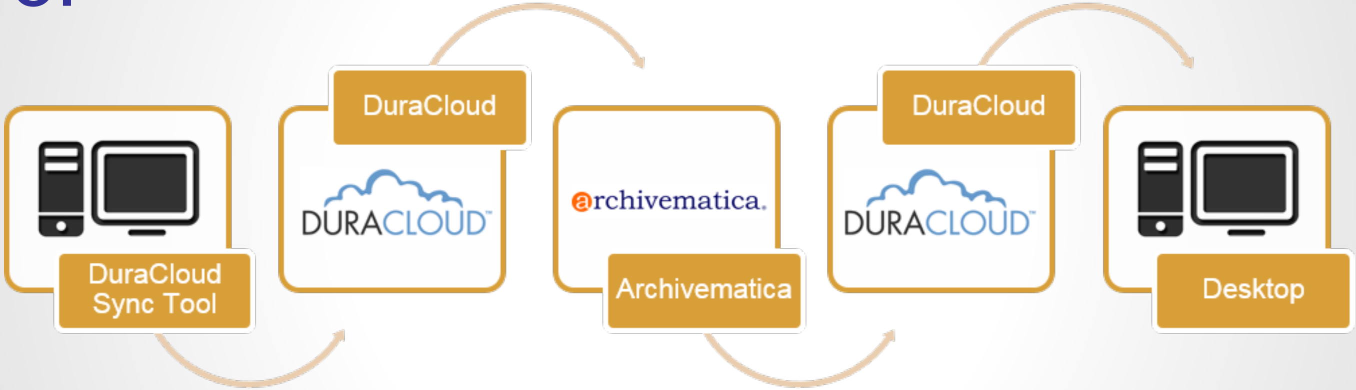
Born Digital architecture files from Killis Almond & Associates

- variety of CAD file types

# Processing POV

# Workflow

Macro:

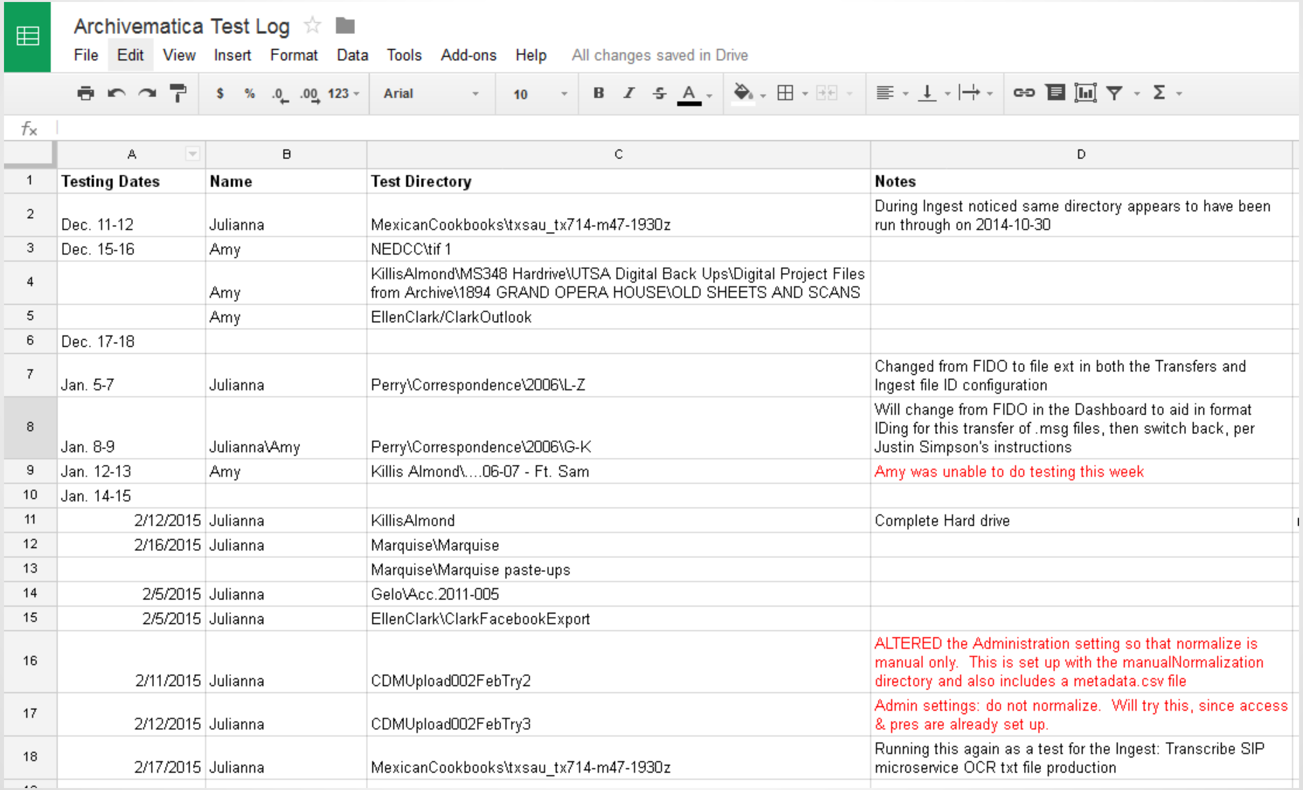


# Workflow con't.

Micro:

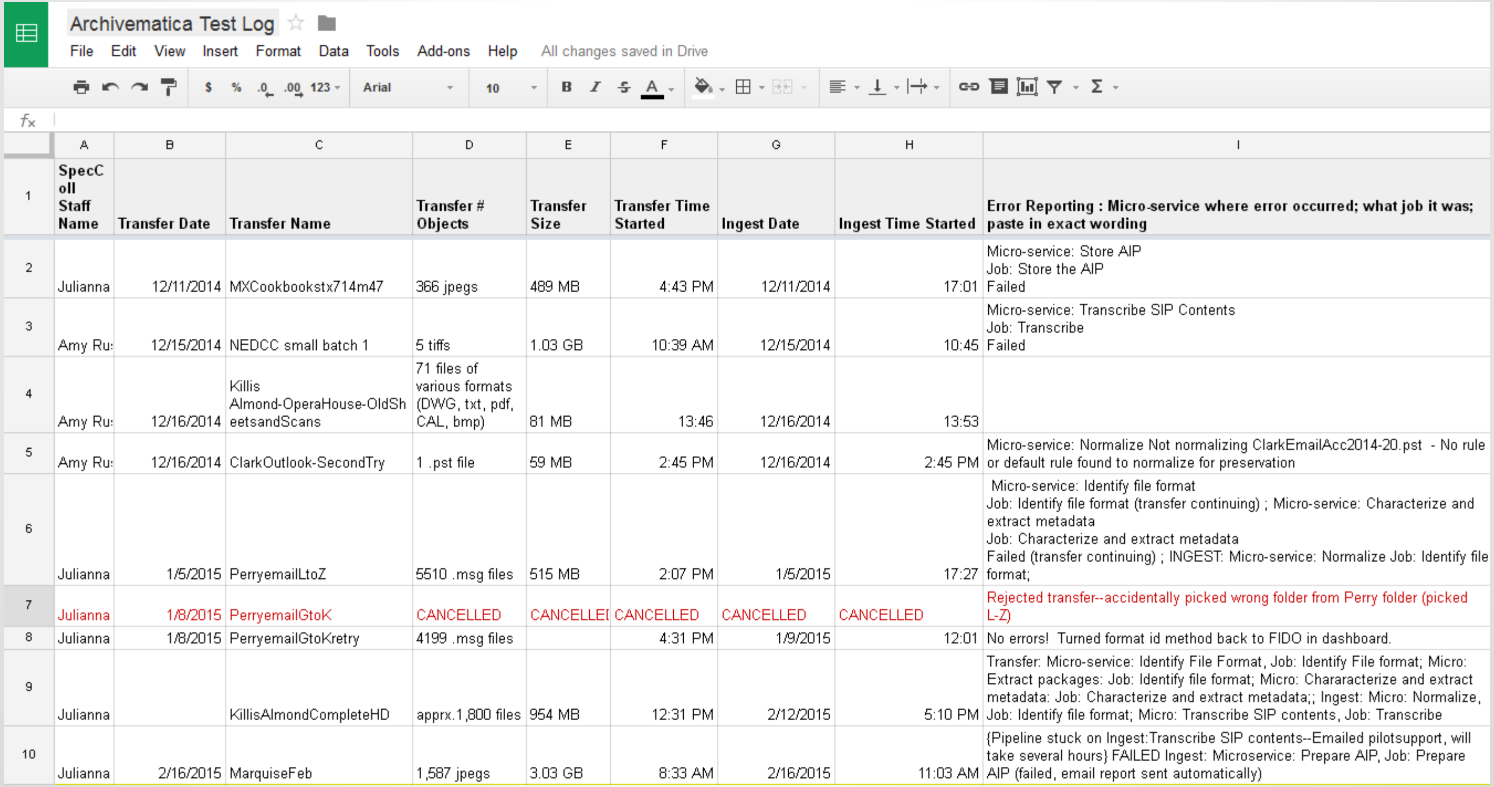
- Set up
- Assemble testers
- Coordinate tests
- Generate transfers, SIPs and AIPs
- Track results
- Track issues

# Test log



| Testing Dates | Name | Test Directory | Notes |
|---|---|---|---|
| Dec. 11-12 | Julianna | MexicanCookbooks\txsau_tx714-m47-1930z | During Ingest noticed same directory appears to have been run through on 2014-10-30 |
| Dec. 15-16 | Amy | NEDCC\tif 1 | |
| | Amy | KillisAlmond\MS348 Hardrive\UTSA Digital Back Ups\Digital Project Files from Archive\1894 GRAND OPERA HOUSE\OLD SHEETS AND SCANS | |
| | Amy | EllenClark/ClarkOutlook | |
| Dec. 17-18 | | | |
| Jan. 5-7 | Julianna | Perry\Correspondence\2006\L-Z | Changed from FIDO to file ext in both the Transfers and Ingest file ID configuration |
| Jan. 8-9 | Julianna\Amy | Perry\Correspondence\2006\G-K | Will change from FIDO in the Dashboard to aid in format IDing for this transfer of .msg files, then switch back, per Justin Simpson's instructions |
| Jan. 12-13 | Amy | Killis Almond\....06-07 - Ft. Sam | Amy was unable to do testing this week |
| Jan. 14-15 | | | |
| 2/12/2015 | Julianna | KillisAlmond | Complete Hard drive |
| 2/16/2015 | Julianna | Marquise\Marquise | |
| | | Marquise\Marquise paste-ups | |
| 2/5/2015 | Julianna | Gelo\Acc.2011-005 | |
| 2/5/2015 | Julianna | EllenClark\ClarkFacebookExport | |
| 2/11/2015 | Julianna | CDMUpload002FebTry2 | ALTERED the Administration setting so that normalize is manual only.  This is set up with the manualNormalization directory and also includes a metadata.csv file |
| 2/12/2015 | Julianna | CDMUpload002FebTry3 | Admin settings: do not normalize.  Will try this, since access & pres are already set up. |
| 2/17/2015 | Julianna | MexicanCookbooks\txsau_tx714-m47-1930z | Running this again as a test for the Ingest: Transcribe SIP microservice OCR txt file production |

# Test log: issues


Archivematica Test Log
File Edit View Insert Format Data Tools Add-ons Help
All changes saved in Drive
$ % .0 .00 123
Arial
10
B I S A
fx

| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | SpecColl Staff Name | Transfer Date | Transfer Name | Transfer # Objects | Transfer Size | Transfer Time Started | Ingest Date | Ingest Time Started | Error Reporting : Micro-service where error occurred; what job it was; paste in exact wording |
| 2 | Julianna | 12/11/2014 | MXCookbookstx714m47 | 366 jpegs | 489 MB | 4:43 PM | 12/11/2014 | 17:01 | Micro-service: Store AIP<br>Job: Store the AIP<br>Failed |
| 3 | Amy Ru: | 12/15/2014 | NEDCC small batch 1 | 5 tiffs | 1.03 GB | 10:39 AM | 12/15/2014 | 10:45 | Micro-service: Transcribe SIP Contents<br>Job: Transcribe<br>Failed |
| 4 | Amy Ru: | 12/16/2014 | Killis Almond-OperaHouse-OldSheetsandScans | 71 files of various formats (DWG, txt, pdf, CAL, bmp) | 81 MB | 13:46 | 12/16/2014 | 13:53 | |
| 5 | Amy Ru: | 12/16/2014 | ClarkOutlook-SecondTry | 1 .pst file | 59 MB | 2:45 PM | 12/16/2014 | 2:45 PM | Micro-service: Normalize Not normalizing ClarkEmailAcc2014-20.pst - No rule or default rule found to normalize for preservation |
| 6 | Julianna | 1/5/2015 | PerryemailLtoZ | 5510 .msg files | 515 MB | 2:07 PM | 1/5/2015 | 17:27 | Micro-service: Identify file format<br>Job: Identify file format (transfer continuing) ; Micro-service: Characterize and extract metadata<br>Job: Characterize and extract metadata<br>Failed (transfer continuing) ; INGEST: Micro-service: Normalize Job: Identify file format; |
| 7 | Julianna | 1/8/2015 | PerryemailGtoK | CANCELLED | CANCELLEI | CANCELLED | CANCELLED | CANCELLED | Rejected transfer--accidentally picked wrong folder from Perry folder (picked L-Z) |
| 8 | Julianna | 1/8/2015 | PerryemailGtoKretry | 4199 .msg files | | 4:31 PM | 1/9/2015 | 12:01 | No errors! Turned format id method back to FIDO in dashboard. |
| 9 | Julianna | | KillisAlmondCompleteHD | apprx.1,800 files | 954 MB | 12:31 PM | 2/12/2015 | 5:10 PM | Transfer: Micro-service: Identify File Format, Job: Identify File format; Micro: Extract packages: Job: Identify file format; Micro: Chararacterize and extract metadata: Job: Characterize and extract metadata;; Ingest: Micro: Normalize, Job: Identify file format; Micro: Transcribe SIP contents, Job: Transcribe |
| 10 | Julianna | 2/16/2015 | MarquiseFeb | 1,587 jpegs | 3.03 GB | 8:33 AM | 2/16/2015 | 11:03 AM | {Pipeline stuck on Ingest:Transcribe SIP contents--Emailed pilotsupport, will take several hours} FAILED Ingest: Microservice: Prepare AIP, Job: Prepare AIP (failed, email report sent automatically) |

# Support methods

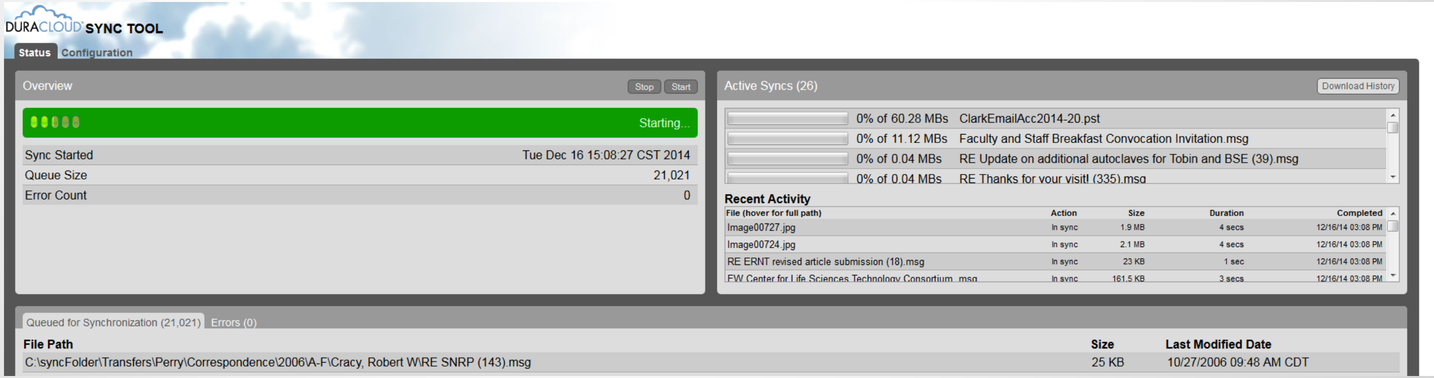
- ArchivesDirect:
  - DuraSpace group wiki
  - Pilot support email list
  - Consultations

- Archivematica:
  - User manual 1.2 wiki
  - Archivematica Google group

# Testing experience

## Assembling transfers

- Metadata set-up
- Directory structure
- DuraCloud Sync tool



Overview

Starting...

|  |  |
|---|---|
| Sync Started | Tue Dec 16 15:08:27 CST 2014 |
| Queue Size | 21,021 |
| Error Count | 0 |

Active Syncs (26)

| | |
|---|---|
| 0% of 60.28 MBs | ClarkEmailAcc2014-20.pst |
| 0% of 11.12 MBs | Faculty and Staff Breakfast Convocation Invitation.msg |
| 0% of 0.04 MBs | RE Update on additional autoclaves for Tobin and BSE (39).msg |
| 0% of 0.04 MBs | RE Thanks for your visit! (335).msg |

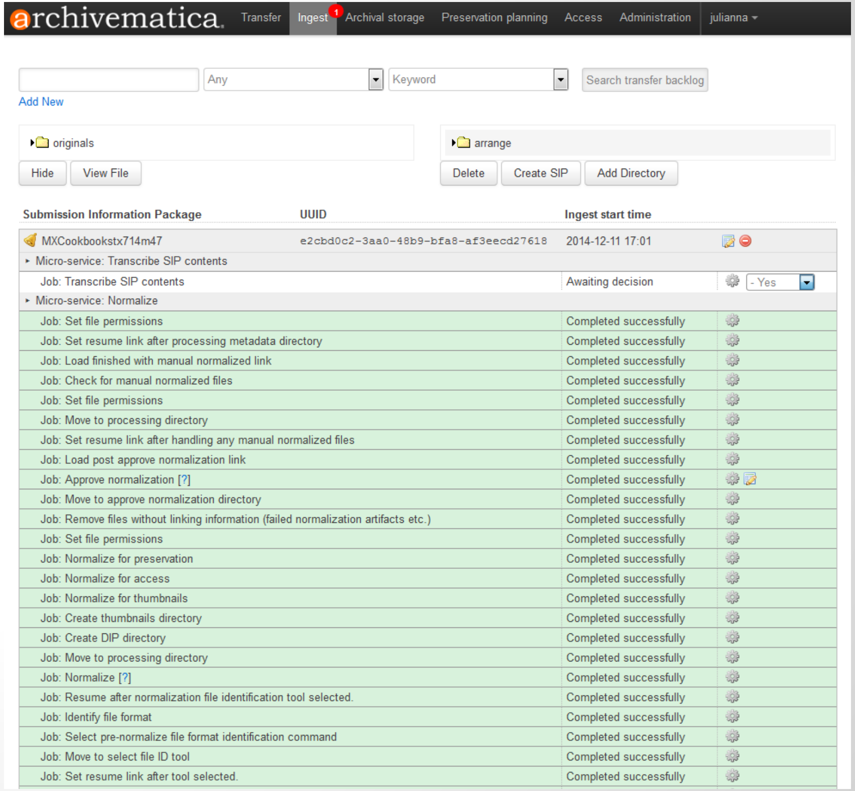
**Recent Activity**

| File (hover for full path) | Action | Size | Duration | Completed |
|---|---|---|---|---|
| Image00727.jpg | In sync | 1.9 MB | 4 secs | 12/16/14 03:08 PM |
| Image00724.jpg | In sync | 2.1 MB | 4 secs | 12/16/14 03:08 PM |
| RE ERNT revised article submission (18).msg | In sync | 23 KB | 1 sec | 12/16/14 03:08 PM |
| FW Center for Life Sciences Technology Consortium .msg | In sync | 161.5 KB | 3 secs | 12/16/14 03:08 PM |

| File Path | Size | Last Modified Date |
|---|---|---|
| C:\syncFolder\Transfers\Perry\Correspondence\2006\A-F\Cracy, Robert W\RE SNRP (143).msg | 25 KB | 10/27/2006 09:48 AM CDT |

# Archivematica testing

- Transfers
- Micro-services
- Specific tools
- Processing methods



| Submission Information Package | UUID | Ingest start time | |
|---|---|---|---|
| MXCookbookstx714m47 | e2cbd0c2-3aa0-48b9-bfa8-af3eecd27618 | 2014-12-11 17:01 | |
| Micro-service: Transcribe SIP contents | | | |
| Job: Transcribe SIP contents | | Awaiting decision | - Yes |
| Micro-service: Normalize | | | |
| Job: Set file permissions | | Completed successfully | |
| Job: Set resume link after processing metadata directory | | Completed successfully | |
| Job: Load finished with manual normalized link | | Completed successfully | |
| Job: Check for manual normalized files | | Completed successfully | |
| Job: Set file permissions | | Completed successfully | |
| Job: Move to processing directory | | Completed successfully | |
| Job: Set resume link after handling any manual normalized files | | Completed successfully | |
| Job: Load post approve normalization link | | Completed successfully | |
| Job: Approve normalization [?] | | Completed successfully | |
| Job: Move to approve normalization directory | | Completed successfully | |
| Job: Remove files without linking information (failed normalization artifacts etc.) | | Completed successfully | |
| Job: Set file permissions | | Completed successfully | |
| Job: Normalize for preservation | | Completed successfully | |
| Job: Normalize for access | | Completed successfully | |
| Job: Normalize for thumbnails | | Completed successfully | |
| Job: Create thumbnails directory | | Completed successfully | |
| Job: Create DIP directory | | Completed successfully | |
| Job: Move to processing directory | | Completed successfully | |
| Job: Normalize [?] | | Completed successfully | |
| Job: Resume after normalization file identification tool selected. | | Completed successfully | |
| Job: Identify file format | | Completed successfully | |
| Job: Select pre-normalize file format identification command | | Completed successfully | |
| Job: Move to select file ID tool | | Completed successfully | |
| Job: Set resume link after tool selected. | | Completed successfully | |

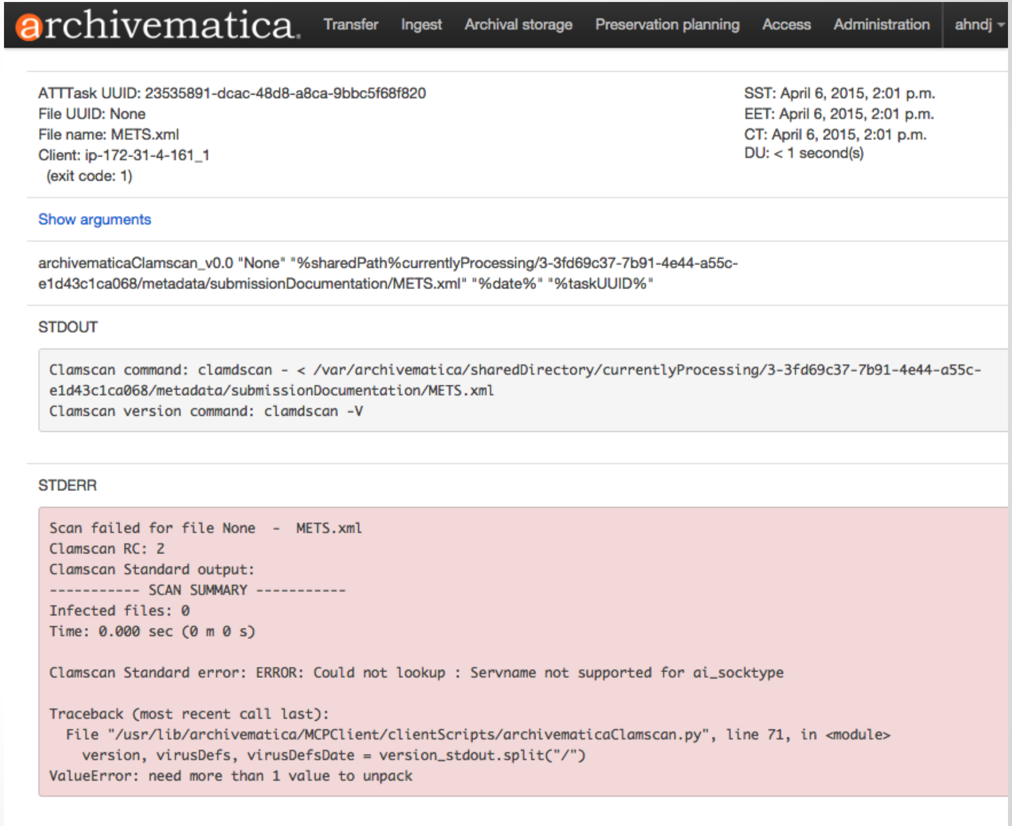
# Outcome

Errors & issues

# Outcome

## Errors & issues

- File-level errors



ATTTask UUID: 23535891-dcac-48d8-a8ca-9bbc5f68f820
File UUID: None
File name: METS.xml
Client: ip-172-31-4-161_1
(exit code: 1)

SST: April 6, 2015, 2:01 p.m.
EET: April 6, 2015, 2:01 p.m.
CT: April 6, 2015, 2:01 p.m.
DU: < 1 second(s)

Show arguments

```
archivematicaClamscan_v0.0 "None" "%sharedPath%currentlyProcessing/3-3fd69c37-7b91-4e44-a55c-e1d43c1ca068/metadata/submissionDocumentation/METS.xml" "%date%" "%taskUUID%"
```

STDOUT

```
Clamscan command: clamdscan - < /var/archivematica/sharedDirectory/currentlyProcessing/3-3fd69c37-7b91-4e44-a55c-
e1d43c1ca068/metadata/submissionDocumentation/METS.xml
Clamscan version command: clamdscan -V
```

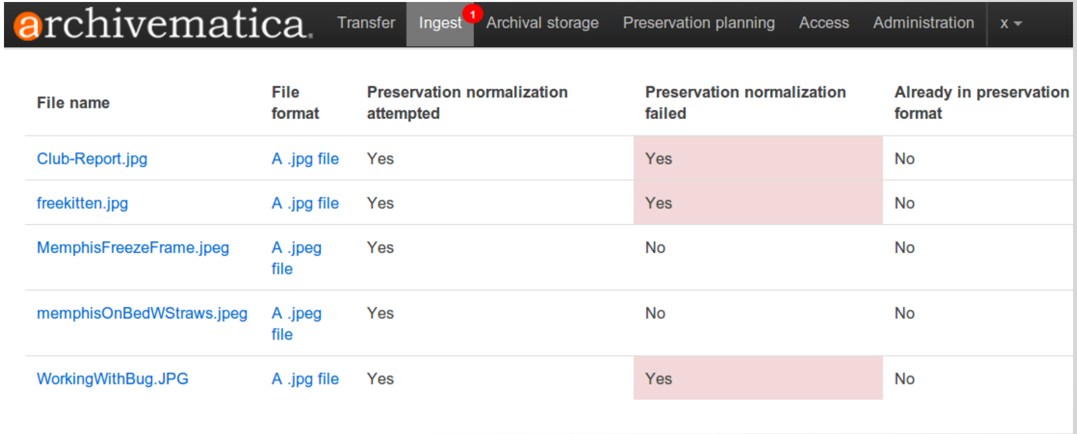
STDERR

```
Scan failed for file None  -  METS.xml
Clamscan RC: 2
Clamscan Standard output:
----------- SCAN SUMMARY -----------
Infected files: 0
Time: 0.000 sec (0 m 0 s)

Clamscan Standard error: ERROR: Could not lookup : Servname not supported for ai_socktype

Traceback (most recent call last):
  File "/usr/lib/archivematica/MCPClient/clientScripts/archivematicaClamscan.py", line 71, in <module>
    version, virusDefs, virusDefsDate = version_stdout.split("/")
ValueError: need more than 1 value to unpack
```

# Outcome

## Errors & issues

- File-level errors
- Normalization errors

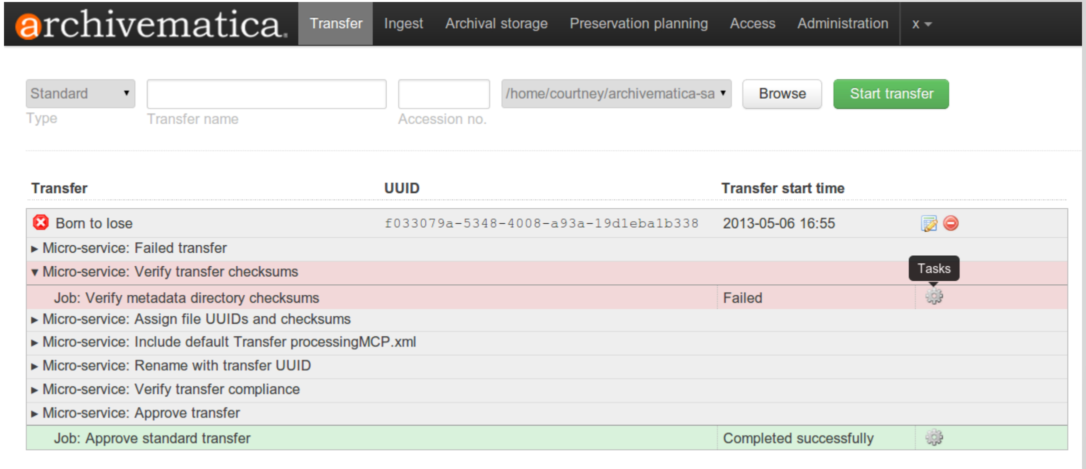


| File name | File format | Preservation normalization attempted | Preservation normalization failed | Already in preservation format |
|---|---|---|---|---|
| Club-Report.jpg | A .jpg file | Yes | Yes | No |
| freekitten.jpg | A .jpg file | Yes | Yes | No |
| MemphisFreezeFrame.jpeg | A .jpeg file | Yes | No | No |
| memphisOnBedWStraws.jpeg | A .jpeg file | Yes | No | No |
| WorkingWithBug.JPG | A .jpg file | Yes | Yes | No |

# Outcome

## Errors & issues

- File-level errors
- Normalization errors
- Metadata errors



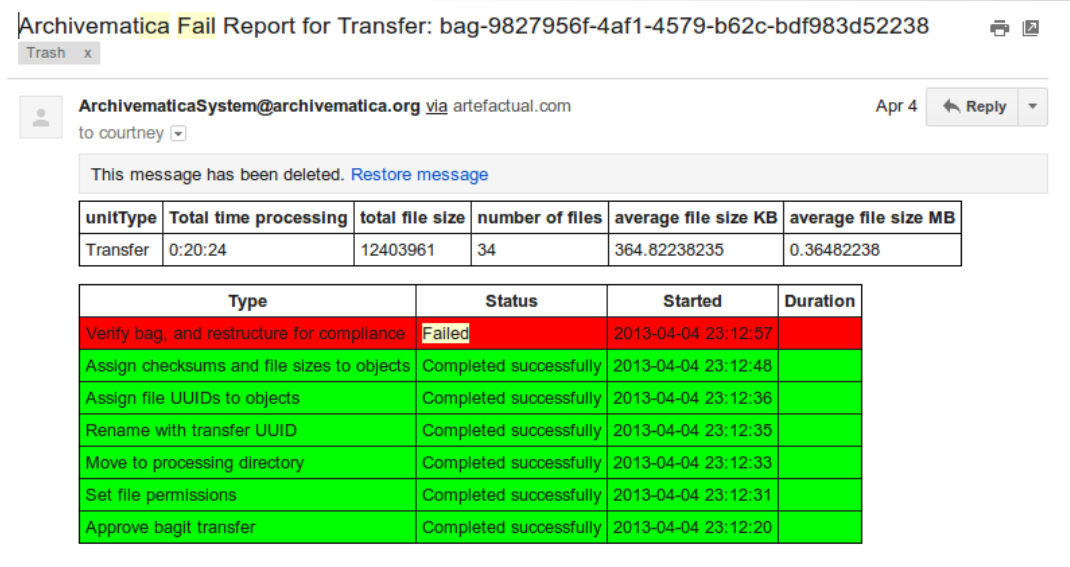
Type: Standard
Transfer name:
Accession no.:
/home/courtney/archivematica-sa

| Transfer | UUID | Transfer start time |
|---|---|---|
| Born to lose | f033079a-5348-4008-a93a-19d1eba1b338 | 2013-05-06 16:55 |
| ▸ Micro-service: Failed transfer | | |
| ▾ Micro-service: Verify transfer checksums | | |
| Job: Verify metadata directory checksums | | Failed |
| ▸ Micro-service: Assign file UUIDs and checksums | | |
| ▸ Micro-service: Include default Transfer processingMCP.xml | | |
| ▸ Micro-service: Rename with transfer UUID | | |
| ▸ Micro-service: Verify transfer compliance | | |
| ▸ Micro-service: Approve transfer | | |
| Job: Approve standard transfer | | Completed successfully |

# Outcome

## Errors & issues

- File-level errors
- Normalization errors
- Metadata errors
- Fatal errors



Archivematica Fail Report for Transfer: bag-9827956f-4af1-4579-b62c-bdf983d52238

ArchivematicaSystem@archivematica.org via artefactual.com

to courtney

This message has been deleted. Restore message

| unitType | Total time processing | total file size | number of files | average file size KB | average file size MB |
|---|---|---|---|---|---|
| Transfer | 0:20:24 | 12403961 | 34 | 364.82238235 | 0.36482238 |

| Type | Status | Started | Duration |
|---|---|---|---|
| Verify bag, and restructure for compliance | Failed | 2013-04-04 23:12:57 | |
| Assign checksums and file sizes to objects | Completed successfully | 2013-04-04 23:12:48 | |
| Assign file UUIDs to objects | Completed successfully | 2013-04-04 23:12:36 | |
| Rename with transfer UUID | Completed successfully | 2013-04-04 23:12:35 | |
| Move to processing directory | Completed successfully | 2013-04-04 23:12:33 | |
| Set file permissions | Completed successfully | 2013-04-04 23:12:31 | |
| Approve bagit transfer | Completed successfully | 2013-04-04 23:12:20 | |

# Outcome

## Errors & issues

- File-level errors
- Normalization errors
- Metadata errors
- Fatal errors
- General pain points

Connected ●

# Final thoughts/Next steps

- Piloting alternative ArchivesDirect workflow (not yet in production)
- Positive pilot outcome will result in alternative hosting option for ArchivesDirect users

# helpful links

archivesdirect.org

lib.utsa.edu/special-collections

Archivematica Birds-of-a-Feather Session (with boxed lunch) - PCL Map Room, Tues 12-130

archivematica.org - duracloud.org - artefactual.com - duraspace.org