



Julianna Barrera-Gomez, Digital Archivist, UTSA Special Collections

Archivematica Camp, Houston, TX

November 15, 2018

# Who we are and what we do

## About UTSA Libraries Special Collections

**VISION** “to bring national recognition to the university for distinctive research collections documenting the diverse histories and development of San Antonio and South Texas.”

### **COLLECTING PRIORITIES**

- History of the African American, Mexican American and LGBTQ communities in our region
- History of women and gender in Texas
- Activism
- Urban planning
- UTSA's University Archives

# Our background

## **2014**

Began first assessment of digital holdings on servers & removable media/hard drives

## **2015**

Calculated nearly 25 TB of potential content

## **2016**

- Creation of Libraries Digital Curation Task Force
- Interim Preservation Plan & Workflow established
- ArchivesDirect pilot testing
- Implementation of ArchivesDirect with AIP storage in DuraCloud

**2017-2018** selected first collections to process

# Our Chosen Collections

## The Marquise Collection

- Local LGBTQ publication
- 948 digitized photographs (335 MB total)
- 184 digitized publications  
(7.13 GB total 4913 files, highly compressed jpgs)
- All items uploaded to CONTENTdm with descriptive metadata



*The Marquise*, June 1994

<http://digital.utsa.edu/cdm/ref/collection/p16018coll7/id/11681>



# Our Chosen Collections

## San Antonio Light Photographs Collection

- Local newspaper photographs, date range 1920s-1990s
- 47,359 negatives digitized by NEDCC
- Creation of descriptive metadata (.xlsx) an ongoing project
- Files are uploaded to CONTENTdm as they are described



MS 359: L-3514-A: Lydia Mendoza with guitar, 1948  
<http://digital.utsa.edu/cdm/ref/collection/p9020coll2/id/375>

# archivesDIRECT

artefactual + DURASPACE™

Artefactual sysadmin  
configures & monitors

DuraCloud Sync Tool  
to upload/download  
files

ArchivesDirect  
User

review format policy changes  
and community statistics  
manage workflow(s)

Archivematica FPR Server

Web App

REST API

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

MCP  
Client

Gearman  
Server

MCP  
Server

MySQL

Dashboard  
and FPR

ElasticSearch Index

ES Node

Workflow A Pipeline

Workflow B Pipeline

Workflow C Pipeline

archivematica

Legend

→ human interaction

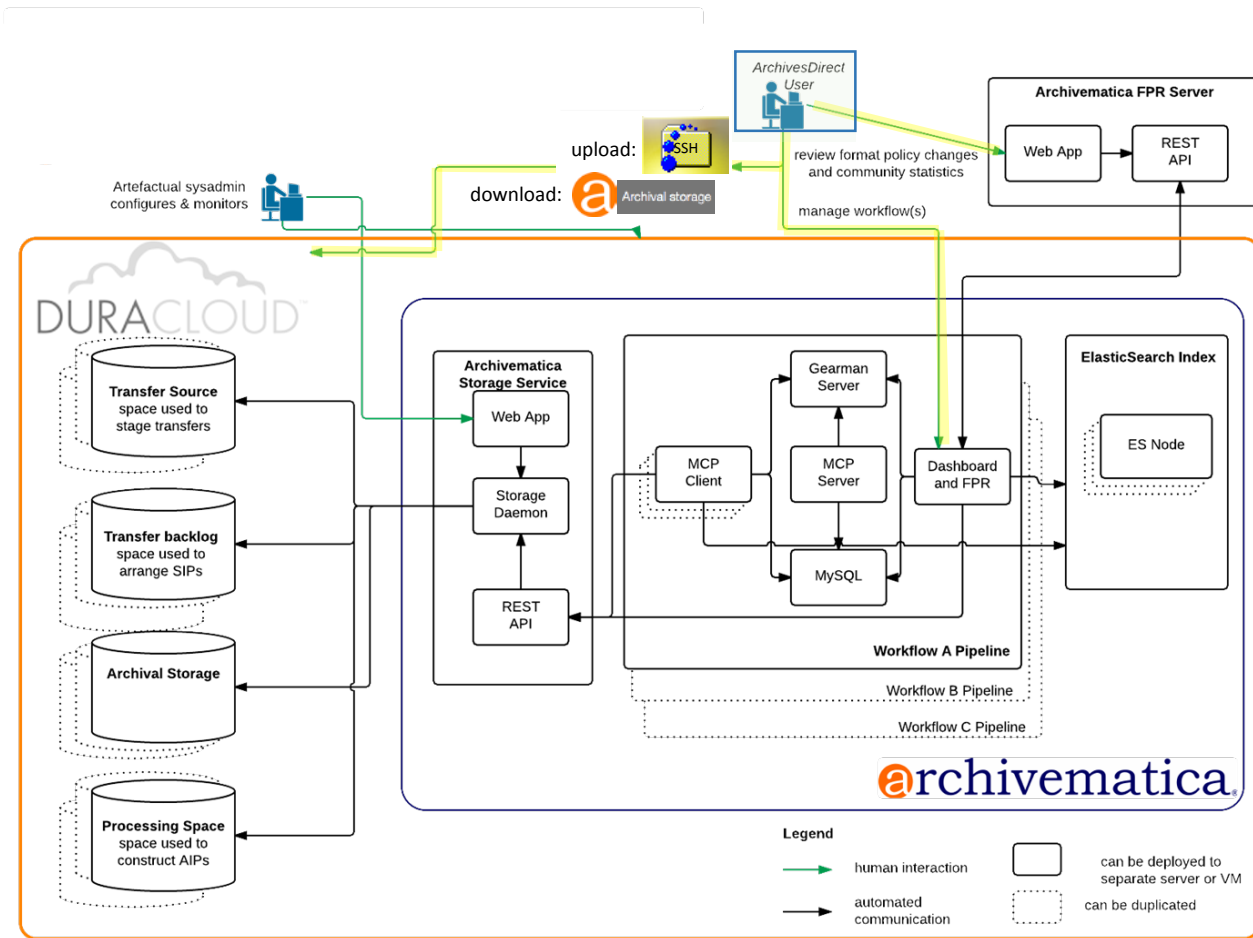
→ automated  
communication



can be deployed to  
separate server or VM



can be duplicated



# Workflow: Goals

- Testing processing possibilities with Archivemata & analyzing results
- Adopting an MPLP “good enough” approach
- Arranging transfers to maximize automated steps from SIP to AIP
- Creating processing tasks and outputs that other archivists can use & understand
- Maximizing available storage space in DuraCloud
- Staying mindful of future reprocessing opportunities

archivemata

Transfer

Ingest

Archival storage

Preservation planning

Access

Administration

julianna.barrera-gomez

Standard

Type

Transfer name

Accession no.

/home

Browse

Start transfer

Transfer



UUID

Transfer start time

metadataembedtest

a58c910b-75a1-4884-bd65-de2388c9e025


2016-10-21 11:31

- Micro-service: Characterize and extract metadata
- Micro-service: Update METS.xml document
- Micro-service: Extract packages
- Micro-service: Identify file format


Job: Identify file format

Completed successfully




Job: Determine which files to identify

Completed successfully




Job: Select file format identification command

Completed successfully



Job: Move to select file ID tool


Completed successfully



- Micro-service: Clean up names
- Micro-service: Generate transfer structure report


Job: Save directory tree

Completed successfully




Job: Generate transfer structure report

Completed successfully




Job: Move to processing directory

Completed successfully



Job: Move to generate transfer tree


Completed successfully



- Micro-service: Scan for viruses
- Micro-service: Quarantine


Job: Workflow decision - send transfer to quarantine [?]

Completed successfully




Job: Find type to process as

Completed successfully




Job: Move to workflowDecisions-quarantineSIP directory

Completed successfully



Job: Designate to process as a standard transfer

Completed successfully



- 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



# Workflow: Metadata options

- Adding Exif/XMP metadata to files
  - Embedding descriptive metadata into files prior to processing
  - Extracted into the METS record for the AIP
- Adding descriptive metadata to METS
  - Import DC metadata from CONTENTdm into the METS file for AIPs
  - .csv file uploaded with AIP is extracted into METS record

```
METS.cba8502e-30d9-46d4-a45f-7100981bbcb.xml (no sy...lected) #
<?xml version='1.0' encoding='ASCII'?>
<mets:mets xmlns:mets="http://www.loc.gov/METS/"
xmlns:xlink="http://www.w3.org/1999/xlink"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.loc.gov/METS/
http://www.loc.gov/standards/mets/version18/mets.xsd">
  <mets:metsHdr CREATEDATE="2016-10-21T17:22:09"/>
  <mets:dmdSec ID="dmdSec_1">
    <mets:mdWrap MDTYPE="DC">
      <mets:xmlData>
        <dcterms:dublincore xmlns:dc="http://purl.org/dc/elements/1.1/"
xmlns:dcterms="http://purl.org/dc/terms/"
xsi:schemaLocation="http://purl.org/dc/terms/
http://dublincore.org/schemas/xmls/qdc/2008/02/11/dcterms.xsd">
          <dc:title>San Antonio Pride parade and picnic 1998</dc:title>
          <dc:creator>Ted Switzer</dc:creator>
          <dc:subject>Marquise (San Antonio, Tex.)</dc:subject>
          <dc:subject>Gays--Texas--San Antonio--Periodicals.</dc:subject>
          <dc:subject>Lesbians--Texas--San Antonio--Periodicals.</dc:subject>
          <dc:description>San Antonio Pride parade and picnic 1998 held in
downtown San Antonio and Travis park. The picnic was in its 16th year,
whereas the parade was the first Pride parade in San Antonio since
1976. Marquise photographs were taken to document events and provide
images for The Marquise.</dc:description>
          <dc:publisher>University of Texas at San Antonio</dc:publisher>
          <dc:date>1998-06</dc:date>
          <dc:format>photograph</dc:format>
          <dc:identifier>txsau_ms00418_00690</dc:identifier>
          <dc:source>The Marquise</dc:source>
          <dc:language>eng</dc:language>
          <dc:relation>The Marquise Collection, MS 418</dc:relation>

          <dc:rights>http://lib.utsa.edu/specialcollections/reproductions/
copyright</dc:rights>
        </dcterms:dublincore>
      </mets:xmlData>
    </mets:mdWrap>
  </mets:dmdSec>
```



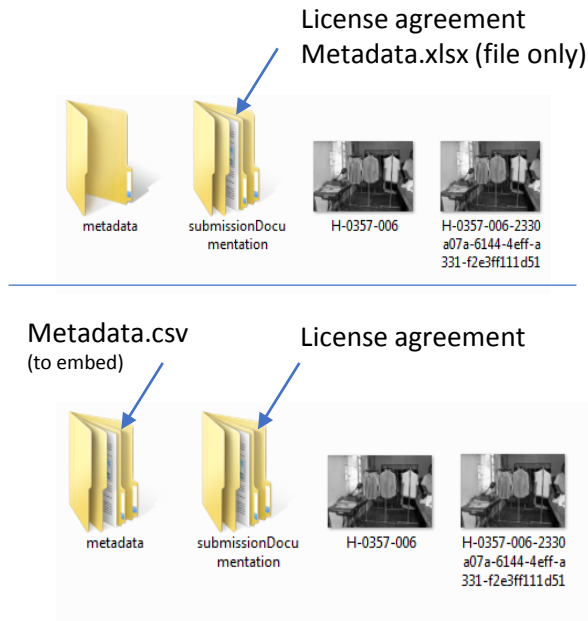
# Workflow: Metadata decisions

- Marquise Collection:

- Upload metadata as .csv (not embedded)
- For now: process the AIP into DuraCloud, reprocess later with re-ingest feature

- San Antonio Light Photos:

- Embedding select descriptive metadata into images prior to upload
- Batching files into AIPs + 1 full metadata “key” AIP
- Unify using AIC





# Workflow: Normalization options

- Normalize files using Archivematica
  - Original kept, preservation copy made, PREMIS record for each event
  - Additional copy uses 2x (up to 14x) space in DuraCloud
- Do not normalize
  - Keep original format (.tif) or normalize prior to transfer
  - Saves space



# Workflow: Normalization decisions

- Marquise Collection:

- Normalize .jpgs to .tif prior to transfer, skip normalization microservice
- Avoid unknown processing space issues and save 9 GB in DuraCloud

- San Antonio Light Photos:

- Keep original .tifs, skip normalization microservice
- Save processing time and 4.8 TB in DuraCloud

Name	Type	Size
metadata	File folder	
submissionDocumentation	File folder	
Image00002	JPEG image	1,812 KB
Image00002-969259bb-6ff5-43b9...	TIF File	25,438 KB

14x increase object size



# Workflow: SIP-AIP size options

- Testing:
  - Transfer sizes: balance efficient uploading
  - SIP sizes: arranging SIPs to maximize automated microservices
  - AIP sizes: creating AIPs that balance compression processing time with storage footprint in DuraCloud

## Browse archival storage

Total size: 22177.72 MB Total files: 156 indexed

AIP	Size	UUID
<a href="#">ComprLevTest1</a>	22.07 MB	473a3470-1f8e-4d
<a href="#">ComprLevTest2</a>	9.74 MB	f877fd09-f2cb-4e

2x decrease AIP size



# Workflow: SIP-AIP size decisions

- Marquise Collection:

- Uploading: keep original order
- SIP & AIP: process multiple SIPs into one single AIP
- Default AIP compression

- San Antonio Light Photos:

- Uploading: keep original order, batch into ~30 GB SIPs for processing
- SIP & AIP: process multiple AIPs, unify with AIC
- Compress AIP to highest level, saving 2x space in DuraCloud

# Next steps

- Processing all of *The Marquise* and *The San Antonio Light* Photograph Collections by December ~~2016~~ 2018...
- Standardizing metadata for AIPs & files to include in Submission Documentation
- Funding Sustainability Plan (in tandem with our Preservation Plan)
- Processing digitized A/V collections
- Shared our processing configurations with A-TEX!



# We're getting there!

## THANK YOU

Many slides borrowed from our PASIG 2016 presentation:

[https://figshare.com/articles/Jumping in and Staying Afloat/4141713](https://figshare.com/articles/Jumping_in_and_Staying_Afloat/4141713)

Julianna Barrera-Gomez

Digital Archivist

UTSA Libraries Special Collections

University of Texas at San Antonio

julianna.barrera-gomez@utsa.edu