A Collaborative Approach to Digital Preservation for the Five Colleges

Aaron Rubinstein
University and Digital Archivist
Special Collections and University Archives
University of Massachusetts Amherst

Shaun Trujillo
Digital Collections and Metadata Lead
Digital Assets and Preservation Services
Mount Holyoke College

FIVE COLLEGES INCORPORATED

The Five Colleges

Amherst College

Hampshire College

Mount Holyoke College

Smith College

UMass Amherst

Founded in 1965

Strong collaborative infrastructure

Digital resource collaboration new and experimental

Digital Preservation at the Five Colleges

- Digital Preservation Task Force formed in 2011
- First phase: introspection, self assessment, and research

Lesson learned:
Unless all institutions commit to a similar level of readiness, collaboration is impossible.

Three-Pronged Plan

- **Education**
  Digital Preservation Management Workshop
  POWRR workshop
  Readiness guide *

- **Best Practices/Standardization**
  Stakeholders and decision making

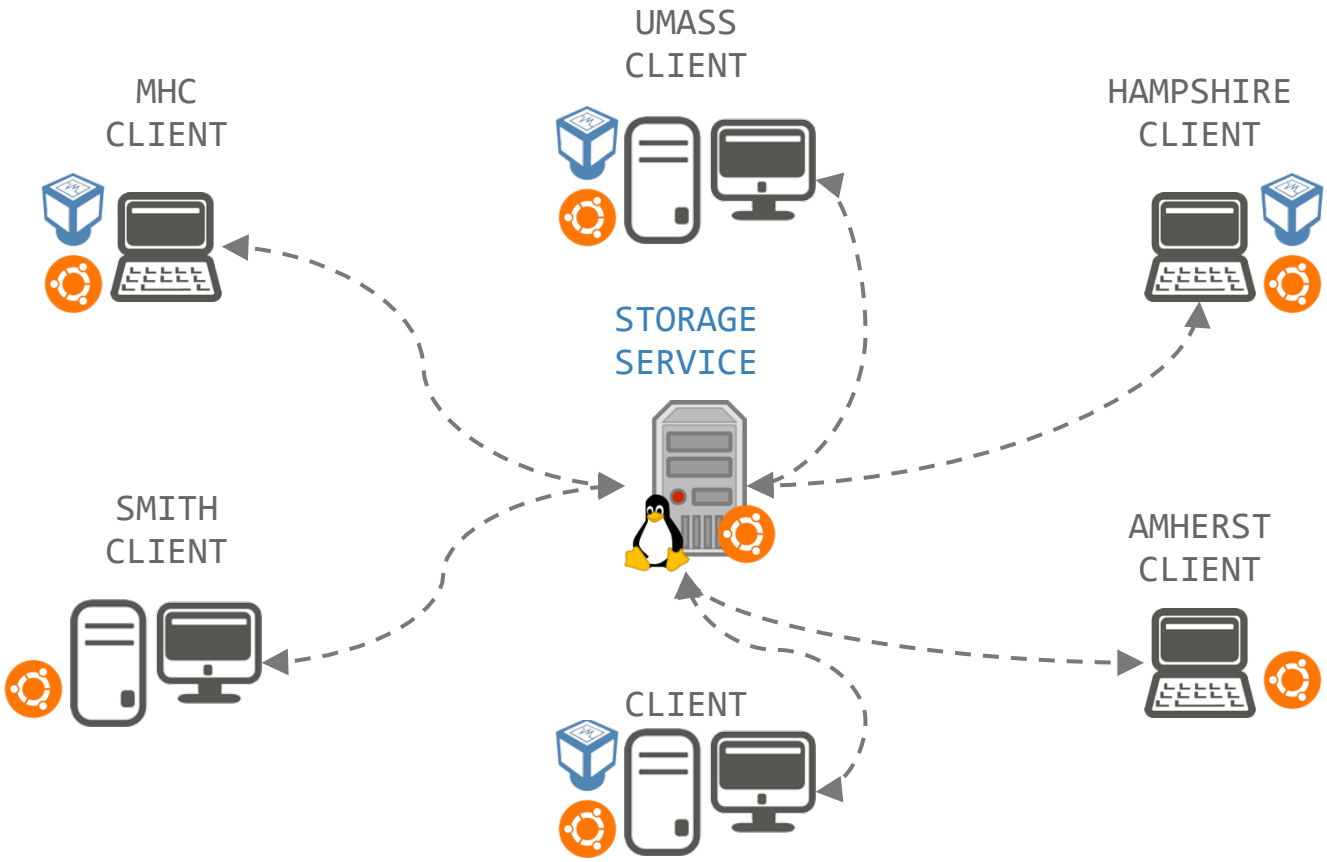
- **Experimentation**
  Archivematica pilot project

All three interrelate

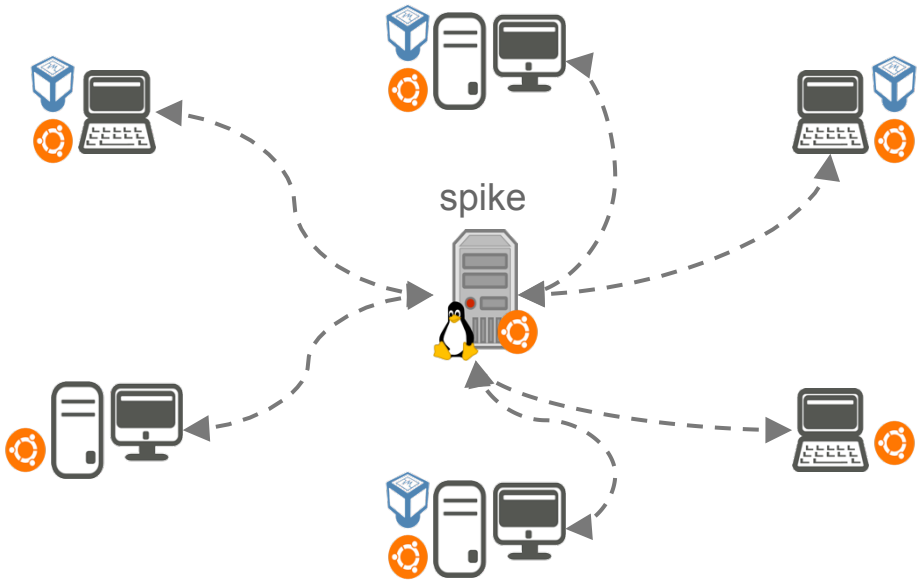
*https://www.fivecolleges.edu/libraries/digital-preservation/digital-preservation-a-guide-for-the-five-colleges

Enter Archivematica

- Micro-Service model of DP
- Excels at born-digital accessioning
- Customizable workflow
- Runs on Ubuntu Linux OS
- Two-part architecture:
  - Client (Pipeline)
  - Storage Service


UMASS
CLIENT
MHC
CLIENT
HAMPSHIRE
CLIENT
STORAGE
SERVICE
SMITH
CLIENT
AMHERST
CLIENT
CLIENT

Consortial Model



- Centralized Storage Service
  - Server hosted at MHC (spike)
- Pipelines - Local Clients running on VirtualBox virtual machine emulation (or not, physical Ubuntu machine)
- Clients connect to spike via VPN
  - reduces complication of two-way SSH traffic and VM network configuration
  - use NAT connection and sign in over VPN (no bridging, no port forwarding)
- Project Leads administer the Storage Service
  - gain experience assigning and administering transfer and storage of AIPs & DIPs, i.e. spaces and locations
- Working Group collaborates on policies and use case workflows for their respective institutions. Configures local client to reflect those decisions.

Benefits of an Archivematica Pilot

- Applied Five College collaboration
- Cross Committee Working Group
- Jumpstart digital preservation conversations and decision making by focusing on something tangible
- Uncover and learn about implicit practices at the Five Colleges
  - Articulate practices in place
  - Align practices with policy/requirements
  - Define policy where there is none
  - Define content streams
- Create a ‘baseline’ for digital preservation in the Five Colleges

archivematica Transfer Ingest 2 Archival storage Preservation planning Access Administration admin

- Processing configuration
- General
- Failures
- Transfer source locations
- AIP storage locations
- AtoM DIP upload
- CONTENTdm DIP upload
- Archivists Toolkit DIP upload
- PREMIS agent
- REST API
- Users

Processing configuration

☑ Send transfer to quarantine	No
☐ Approve normalization	Yes
☐ Store AIP	Yes
☐ Remove from quarantine after	days
☐ Create SIP(s)	--Actions--
☑ Extract packages	Extract
☐ Normalize	--Actions--
☑ Reminder: add metadata if desired	Continue
☑ Select file format identification command (Transfer)	Fido version 1 PUID runs Identify
☑ Select file format identification command (Ingest)	Use existing data
☑ Delete packages after extraction	Delete
☑ Select compression algorithm	7z using bzip2
☑ Select compression level	5 - normal compression mode
☐ Store AIP location	--Actions--

Save

spike.mtholyoke.edu/fpr/fprule/normalization/

Apps | 10 PRINT CHR$ (205 | acoburn's Gists | Activate X-Ray Gogg | http://asteria.fiveco | Alakazam | Amazon S3 | Atlas

archivematica | Transfer | Ingest | Archival storage | Preservation planning | Access | Administration | strujill

Preservation	Nikon Digital SLR Camera Raw Image File	Transcoding to tif with convert	0 out of 0	View | Replace
Preservation	Kodak Digital Camera Raw Image File	Transcoding to tif with convert	0 out of 0	View | Replace
Preservation	Sony RAW Image File	Transcoding to tif with convert	0 out of 0	View | Replace
Preservation	Canon CR2 Raw	Transcoding to tif with convert	0 out of 0	View | Replace
Preservation	Raw Bitmap	Transcoding to tif with convert	0 out of 0	View | Replace
Preservation	Generic TGA	Transcoding to tif with convert	0 out of 0	View | Replace
Preservation	Truevision TGA Bitmap 2.0	Transcoding to tif with convert	0 out of 0	View | Replace
Preservation	Truevision TGA Bitmap 1.0	Transcoding to tif with convert	0 out of 0	View | Replace
Preservation	Generic PICT	Transcoding to tif with convert	0 out of 0	View | Replace
Preservation	PICT 1.0	Transcoding to tif with convert	0 out of 0	View | Replace
Preservation	Adobe Photoshop	Transcoding to tif with convert	0 out of 0	View | Replace
Preservation	Exchangeable Image File Format (Uncompressed) 2.2	Transcoding to tif with convert	0 out of 0	View | Replace
Preservation	Exchangeable Image File Format (Uncompressed) 2.1	Transcoding to tif with convert	0 out of 0	View | Replace
Preservation	Exchangeable Image File	Transcoding to tif	0 out of 0	View |

Archivematica FPR Server | Archivematica Dashboard

localhost/transfer/#

Archivematica | ICA-AtoM | Elasticsearch B... | AM FAQ - Archi... | Archivematica ... | Other Bookmarks

archivematica | Transfer 3 | Ingest | Archival storage | Preservation Planning | Access | Administration | x | Connected

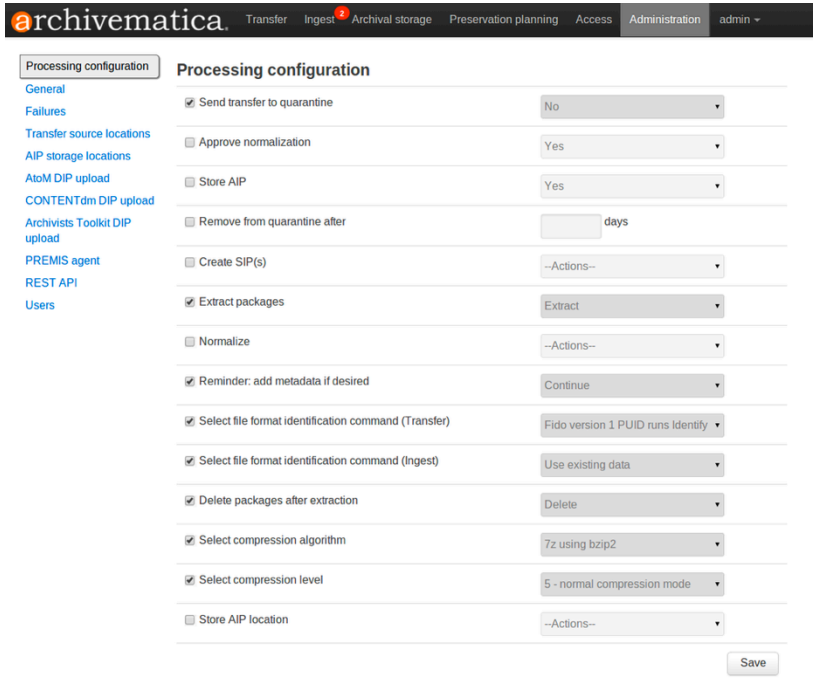
Standard (Type) | Transfer name | Accession no. | /home | Browse | Start transfer

Transfer	UUID	Transfer start time	
Sample series	89a46845-0bcd-4917-a482-ea004a798b9a	2013-10-10 13:06	
▸ Micro-service: Create SIP from Transfer			
Job: Create SIP(s) [?]		Awaiting decision	Actions
Job: Load options to create SIPs		Com	
Job: Check transfer directory for objects		Com	
▸ Micro-service: Complete transfer			
▸ Micro-service: Characterize and extract metadata			
Job: Load labels from metadata/file_labels.csv		Completed successfully	
Job: Characterize and extract metadata		Completed successfully	
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: Scan for viruses			

Actions
- Create SIP(s) manually
- Send to backlog
- Reject transfer
- Create single SIP and continue processing

localhost

Micro-Services Inform Decision Making



- **Characterization**: managing a panoply of file extensions
  - Which formats are common? Which are edge cases?
- **Normalization**: Master file format / access file format
  - Generalized file management / discreet file management
  - Legacy formats >> Data Loss via normalization
    - Acceptable data loss vs. critical characteristics
- **Versioning** - Master, Access 1, Access 2, etc. LOCKSS
- **Metadata** compliance - at the object level, folder level, item level?
- **Custom Actions**: plugin scripts for specific use cases
  - e.g. Exif metadata extraction with ExifTool
  - e.g. provide OCR for PDFs with Tesseract

Questions?