Trials And Tribulations Of Moving Forward With Digital Preservation Workflows And Strategies

University of CINCINNATI

NDSA

GEORGETOWN UNIVERSITY LIBRARY

Thursday, October 26 • 8:30am - 10:00am
NDSA Digital Preservation 2017

- Discuss implementation of digital preservation plans and strategies at two research libraries
- Describe workflows and tools to ingest content into a shared digital preservation environment using open source tools and community solutions
- Focus on next steps after digital preservation planning and system selection

Presenters

Georgetown University Library and soon to be **MIT Libraries**

Joe Carrano

Georgetown University Library

Suzanne Chase || Salwa Ismail

University of Cincinnati Libraries

Linda Newman

Penn State University Libraries

Nathan Tallman

Agenda

1. The Platform - APTrust
2. Georgetown University
    a. Digital Preservation Strategy
    b. Landscape
    c. Workflows
    d. Resources, Training, & Roles
    e. Lessons Learned
3. University of Cincinnati
    a. Choosing
    b. Content Landscape
    c. Implementing Workflows
    d. Roles and Responsibilities
    e. Lessons Learned
4. Questions/Open Forum

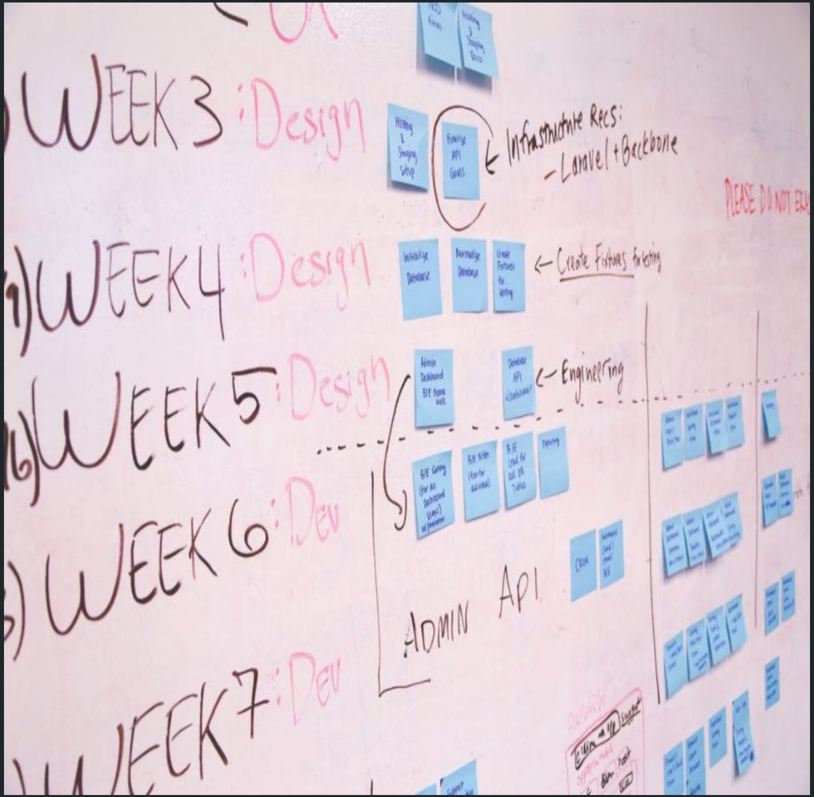
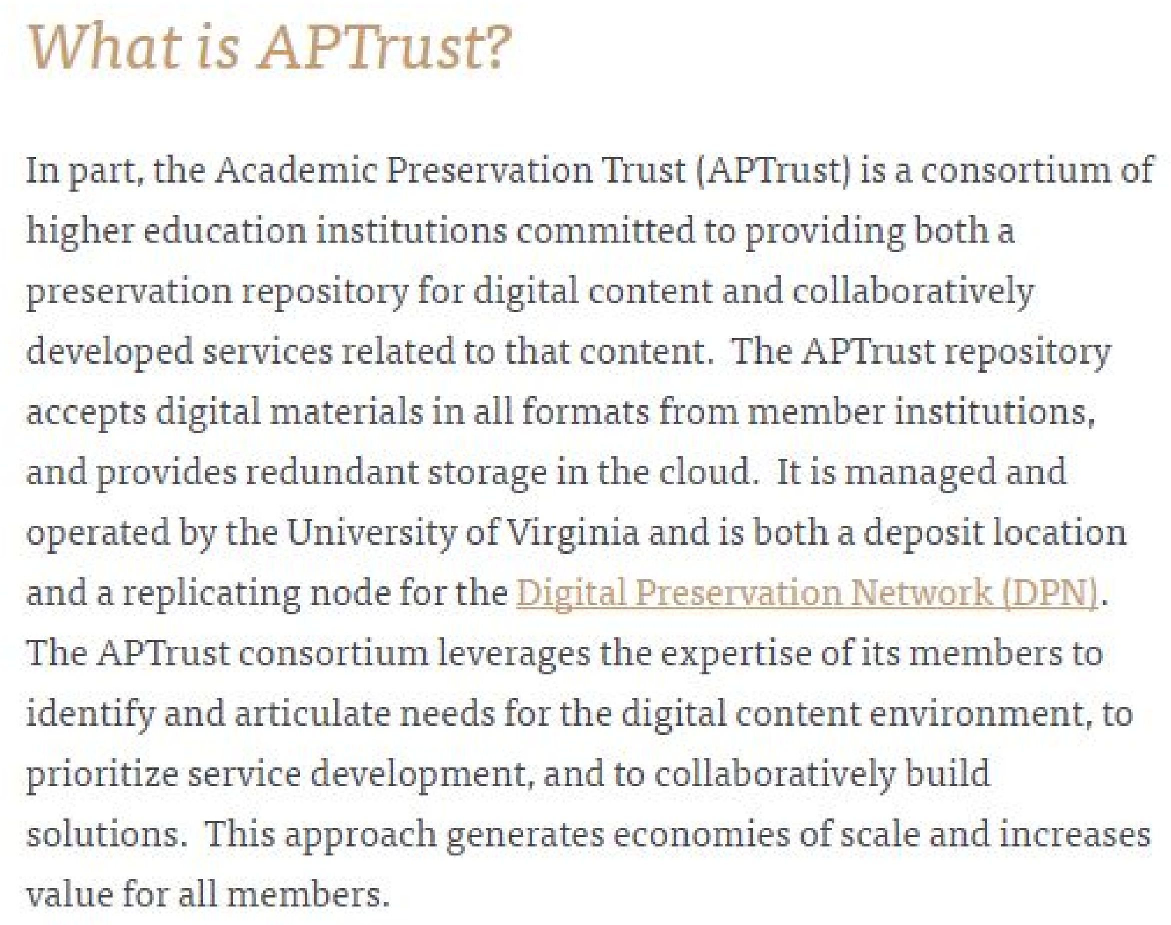


Image: Pexels.com, CC0 license, no attribution required.

Academic Preservation Trust (APTrust)



What is APTrust?

In part, the Academic Preservation Trust (APTrust) is a consortium of higher education institutions committed to providing both a preservation repository for digital content and collaboratively developed services related to that content. The APTrust repository accepts digital materials in all formats from member institutions, and provides redundant storage in the cloud. It is managed and operated by the University of Virginia and is both a deposit location and a replicating node for the Digital Preservation Network (DPN). The APTrust consortium leverages the expertise of its members to identify and articulate needs for the digital content environment, to prioritize service development, and to collaboratively build solutions. This approach generates economies of scale and increases value for all members.

Academic Preservation Trust

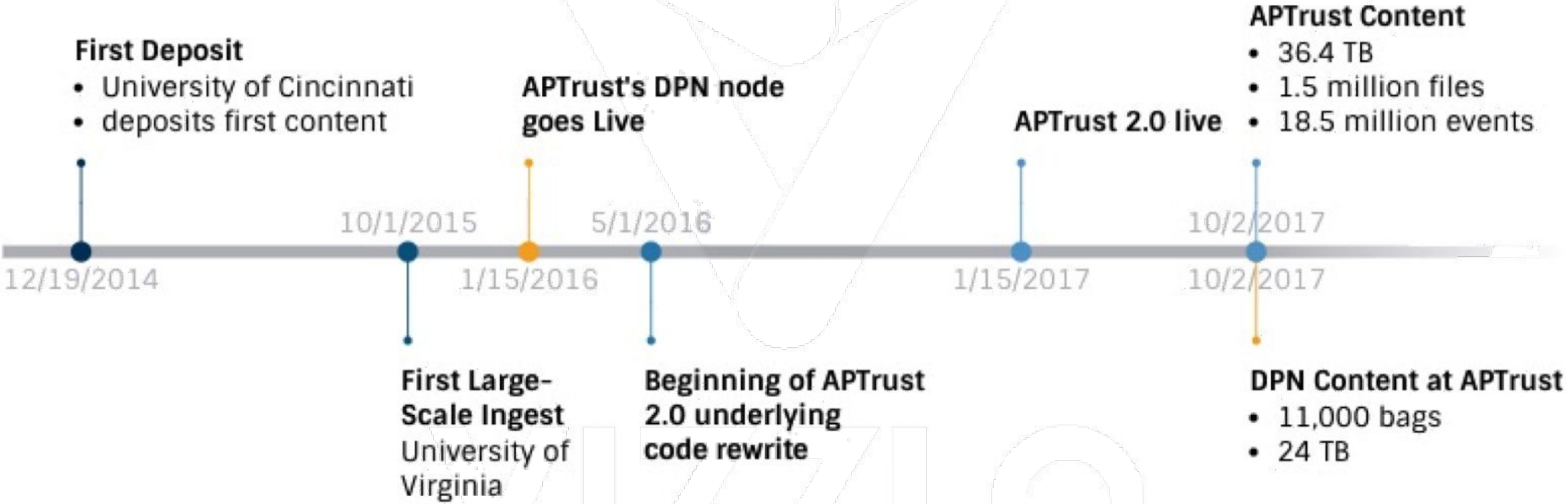
http://aptrust.org/

APTrust

Academic Preservation Trust

- 5 years since formation, 3 years with preservation repository in production
- Member meetings twice a year, often hosted by member institutions
- Built on Amazon Web Services, exploring additional storage options
- 6 copies (3 in S3, 3 in Glacier) in two regions (Virginia, Oregon)
- 36+ TB of unique content (216 TB total content)
  - 1.5+ million files, 18.5+ million PREMIS events
- Fixity checks every 90-days
- Tiers of service in development
  - Sustaining members
    - High assurance storage
    - Low assurance storage
  - Subscription memberships, consortial memberships
- Command line tools, drag and drop bagging web tool in development

APTrust



12/19/2014

First Deposit

- University of Cincinnati
- deposits first content

10/1/2015

First Large-Scale Ingest
University of Virginia

1/15/2016

APTrust's DPN node goes Live

5/1/2016

Beginning of APTrust 2.0 underlying code rewrite

1/15/2017

APTrust 2.0 live

10/2/2017

APTrust Content

- 36.4 TB
- 1.5 million files
- 18.5 million events

10/2/2017

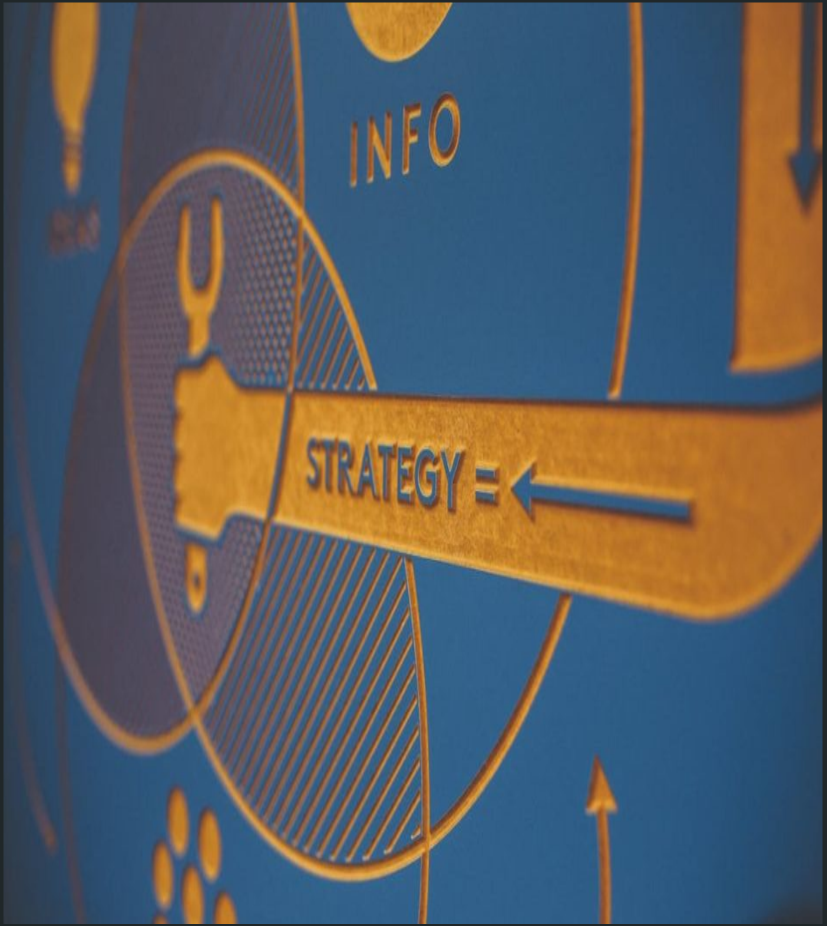
DPN Content at APTrust

- 11,000 bags
- 24 TB

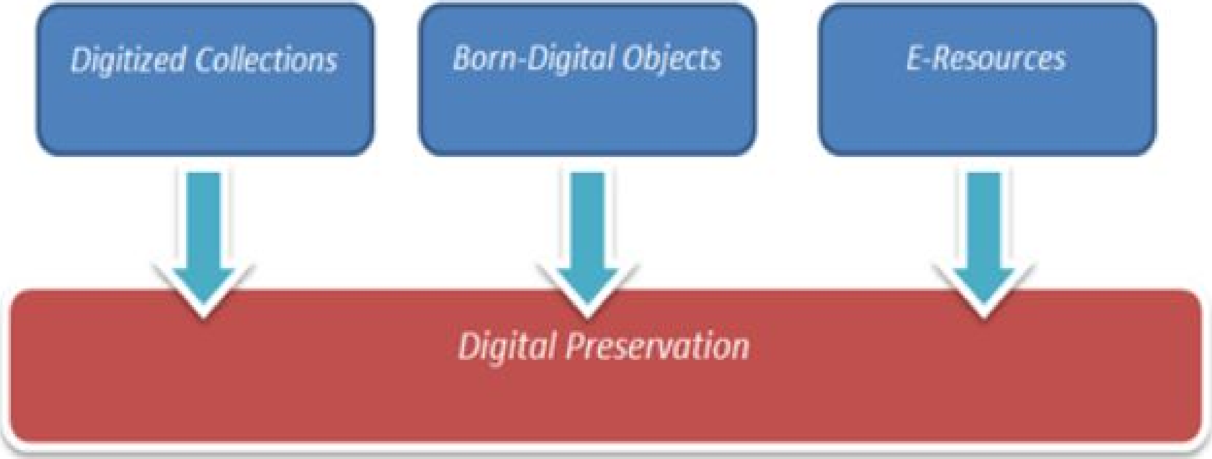
Georgetown University Library

Digital Preservation Strategy

- 2013 (Fall) || Library Leadership Council
- Costs || Platforms
- e-journals/e-books/licensed content || e-resources
- Scope || Objectives
- Challenges
- Priorities
- **ROLES**



Digital Preservation Strategy



Digitized Collections

Born-Digital Objects

E-Resources

Digital Preservation

Choosing Academic Preservation Trust (APTrust)

- Assessed Preservica, DPN, Chronopolis, APTrust
- Had phone calls with pre-determined questions
  - Access
  - Ease of getting to preservation content
  - Update content and metadata
  - Development Phase
- Assessed their offerings against our needs and situation
- Internal discussions with pros and cons and why
- Recommended APTrust
- Lucky for us - APTrust was accepting new members

Choosing the platform/service was the easy part.

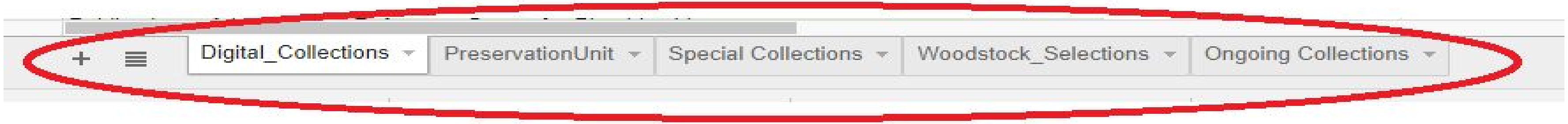
Then came the difficult part...

Division of Labor

- Right stakeholders involved
- Ensure the same starting point
- Understand the scope and responsibility, and timeline with deliverables
- Vested interests
- Develop workflows, automate them, learn and test them

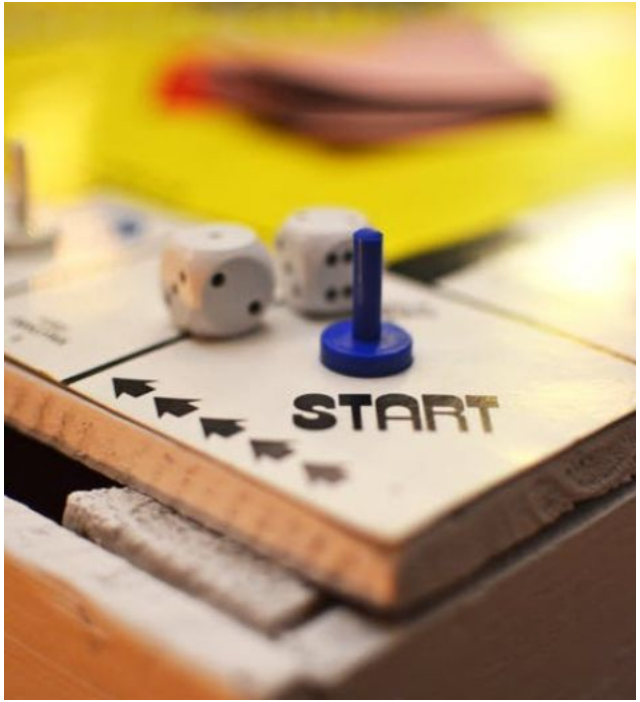
Digital Preservation Assessment (Where Art Thou?)

- Several WRLC inventories -- done in a vacuum
- 2 hr meeting -- everything on the table -- remote items
- Divided it into physical ownership of preservation
- Homework -- note the full location, preservation format, and ownership
- Not as disperse as it could have been
- Responsibilities
- Full assessment of what we have, where, why, and should we?



Getting jump started

- NDSR residency host institution proposals were coming up
- Dedicated person who could focus on just APTrust
- Wrote the detailed proposal and submitted
- Phone interviews, in-person site visit, feedback on the proposal
- Matched ... Yay!
- 1 - yr residency started ... reality hit!
- Automated workflows put to use, new workflows developed
- Testing and in-production
- Best part
  - Focused on the project
  - Issues with collections, workflows, processes -- reconcile collections and objects
  - Save time on back and forth (APT and internally)
  - Work with Digital Services Team to fix issues



Georgetown’s Digital Preservation Landscape

- Content and metadata is stored in DigitalGeorgetown, our DSpace digital collections and institutional repository.
- Metadata is stored in ArchivesSpace, our publicly available finding aids database.
- Content is on network and external media.
- In the future, we may ingest content stored in other systems, such as EmbARK, our art collection management system.

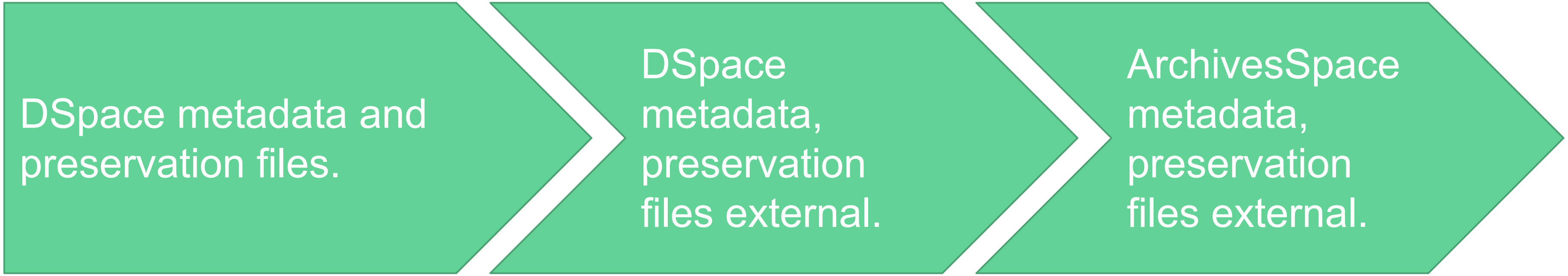
ACADEMIC PRESERVATION TRUST

DG DIGITALGEORGETOWN

ArchivesSpace

DSPACE

Implementing APTrust: Content and Workflows



DSpace metadata and preservation files.

DSpace metadata, preservation files external.

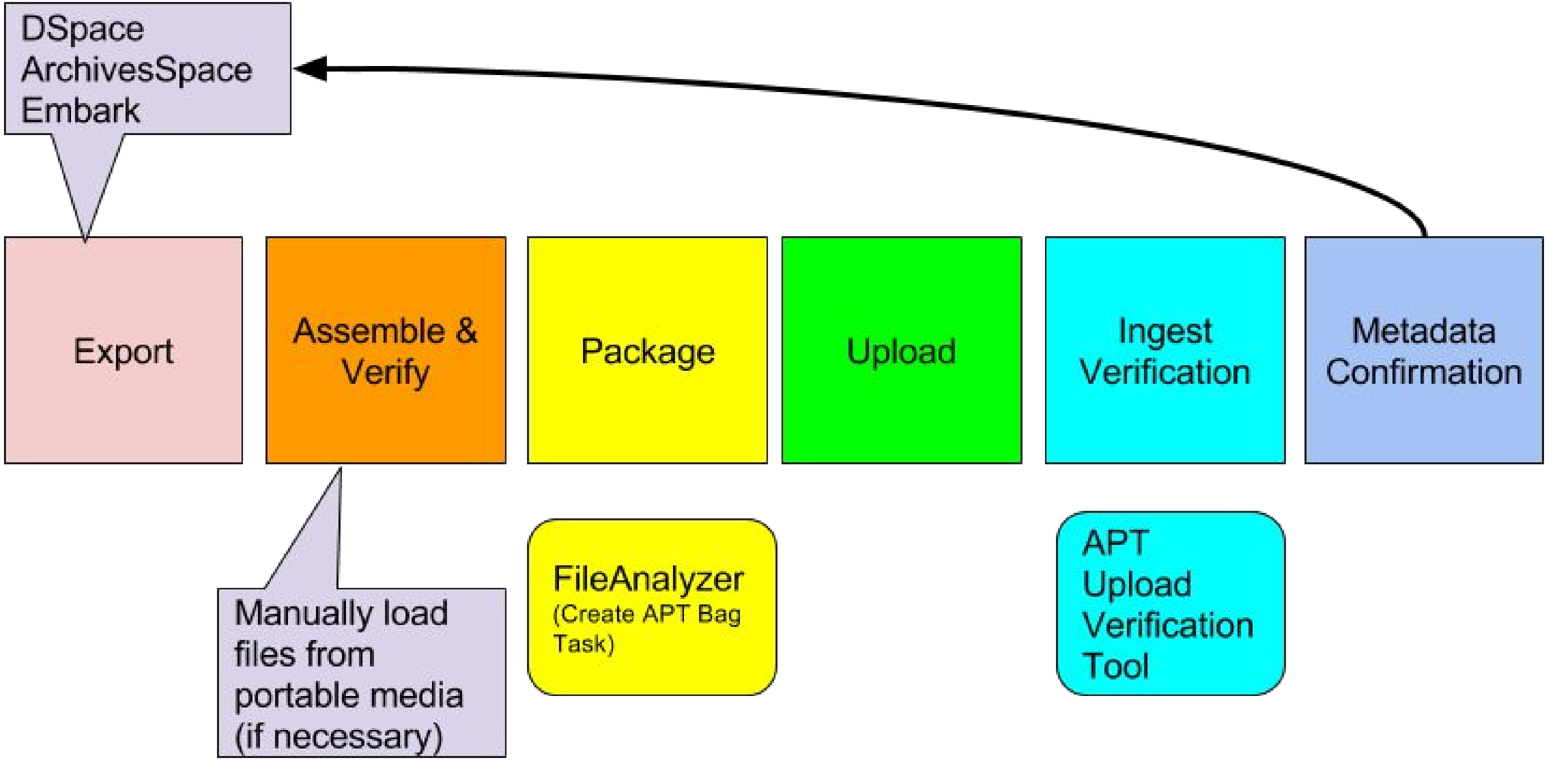
ArchivesSpace metadata, preservation files external.

“Low hanging fruit 1”

“Low hanging fruit 2”

“Low hanging fruit 3”

Workflows



DSpace
ArchivesSpace
Embark

Export

Assemble & Verify

Package

Upload

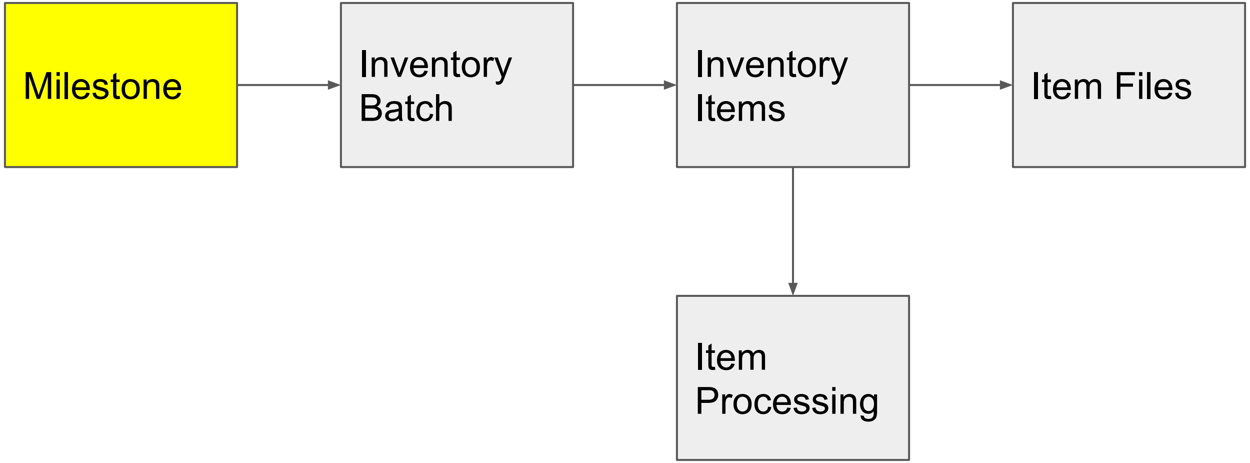
Ingest Verification

Metadata Confirmation

Manually load files from portable media (if necessary)

FileAnalyzer
(Create APT Bag Task)

APT Upload Verification Tool



Milestone

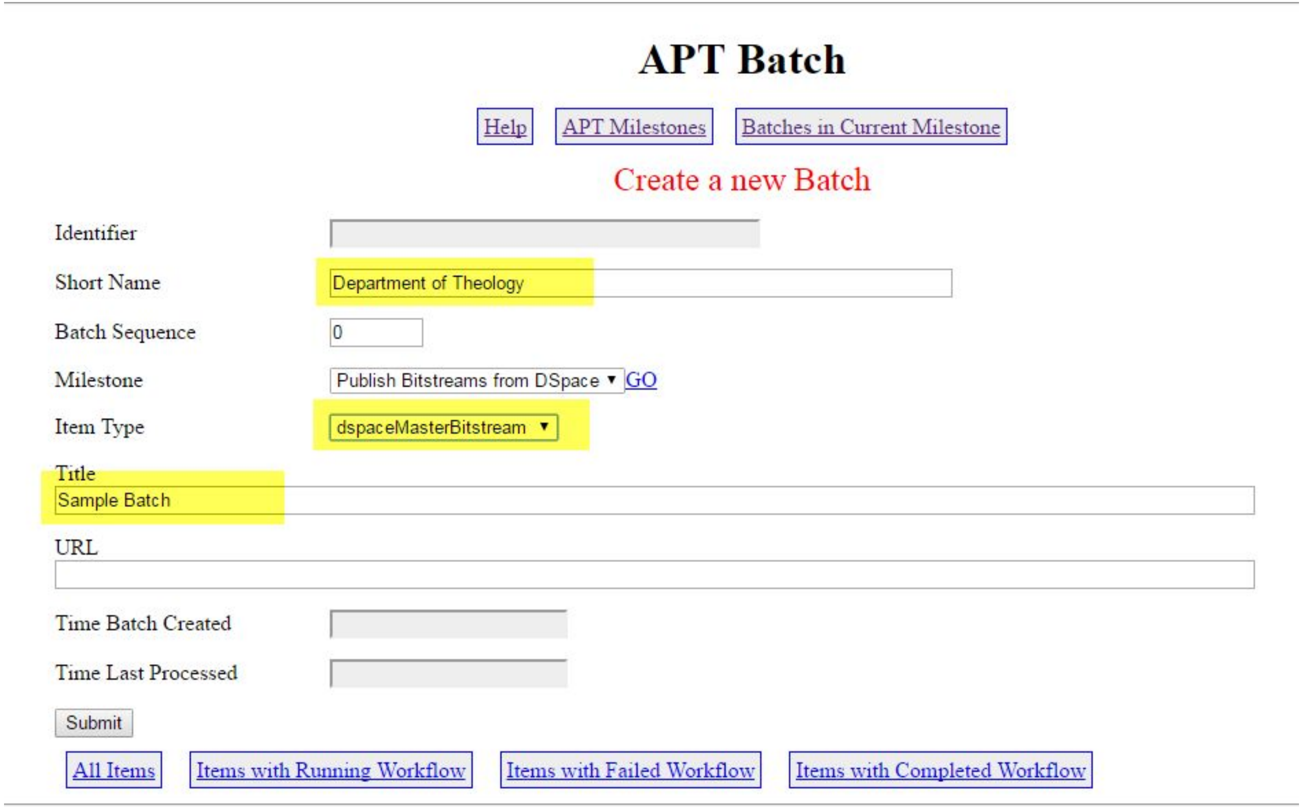
Inventory Batch

Inventory Items

Item Files

Item Processing

DSpace preservation bitstreams



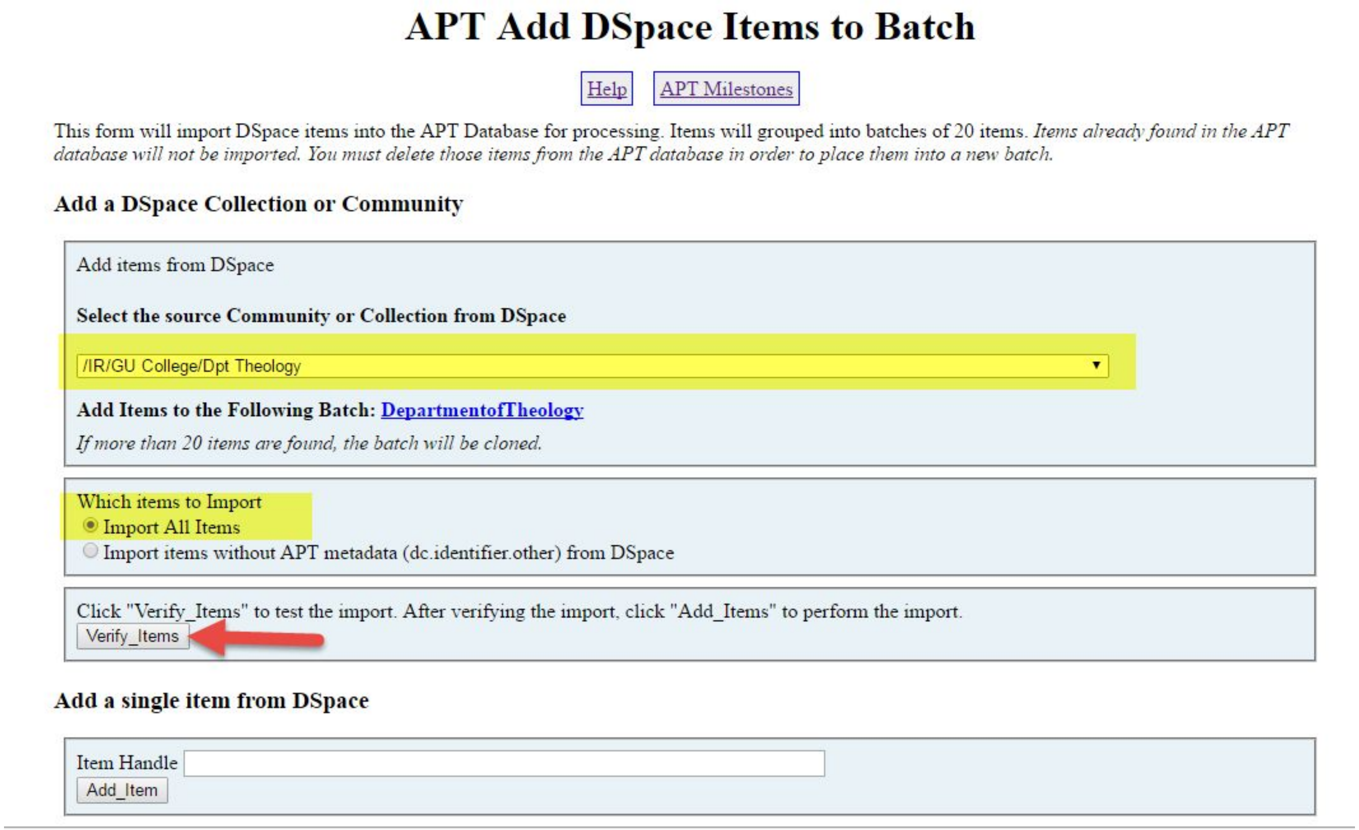
APT Batch

Create a new Batch

Identifier	
Short Name	Department of Theology
Batch Sequence	0
Milestone	Publish Bitstreams from DSpace ▾
Item Type	dspaceMasterBitstream ▾
Title	Sample Batch
URL	
Time Batch Created	
Time Last Processed	

Help
APT Milestones
Batches in Current Milestone
GO
Submit
All Items
Items with Running Workflow
Items with Failed Workflow
Items with Completed Workflow

Can select a community or collection (bagged on item level)



APT Add DSpace Items to Batch

This form will import DSpace items into the APT Database for processing. Items will grouped into batches of 20 items. *Items already found in the APT database will not be imported. You must delete those items from the APT database in order to place them into a new batch.*

Add a DSpace Collection or Community

Add items from DSpace

Select the source Community or Collection from DSpace

/IR/GU College/Dpt Theology ▾

Add Items to the Following Batch: DepartmentofTheology

If more than 20 items are found, the batch will be cloned.

Which items to Import

● Import All Items

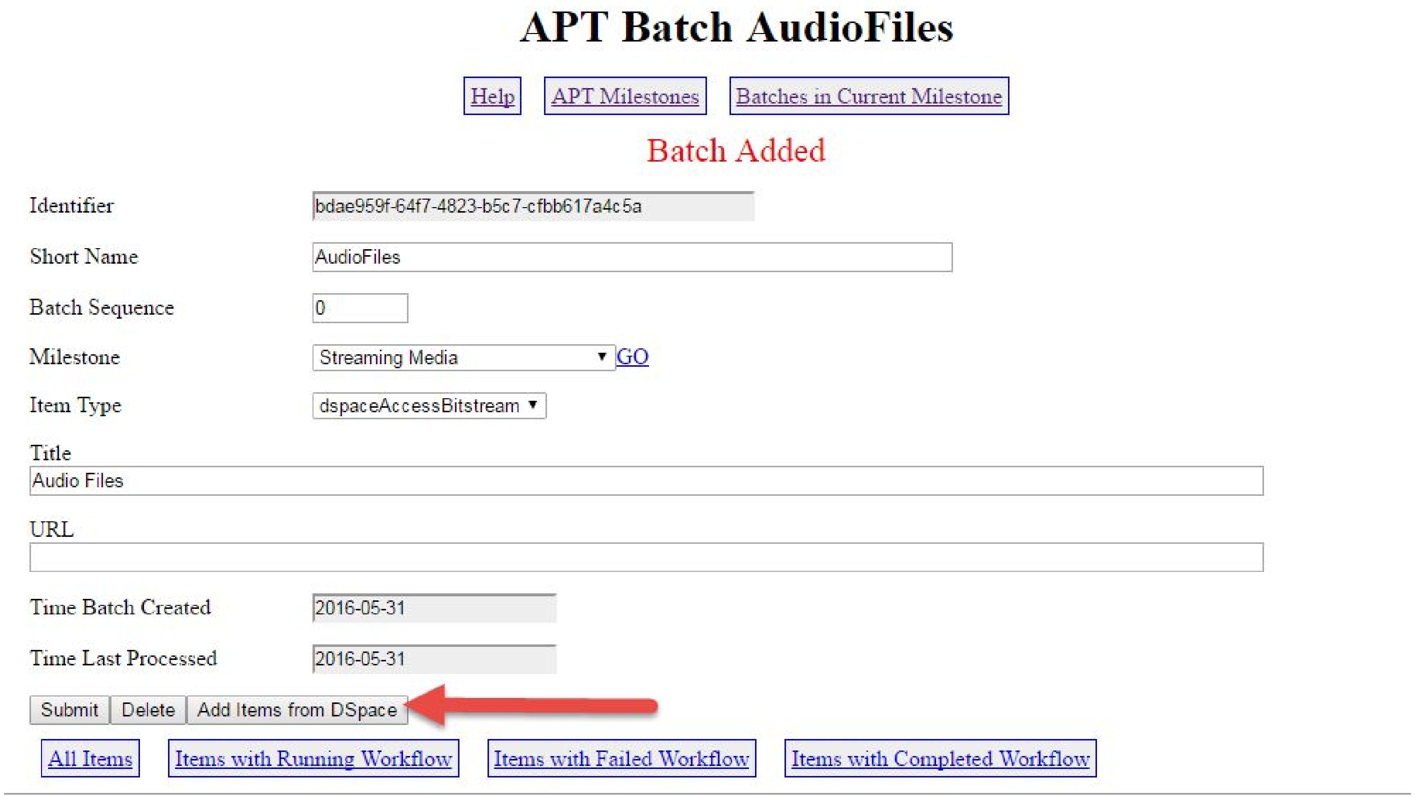
○ Import items without APT metadata (dc.identifier.other) from DSpace

Click "Verify_Items" to test the import. After verifying the import, click "Add_Items" to perform the import.

Add a single item from DSpace

Item Handle

DSpace Access Bitstreams (External Preservation Files)



Help
APT Milestones
Batches in Current Milestone
Submit
Delete
Add Items from DSpace
GO
All Items
Items with Running Workflow
Items with Failed Workflow
Items with Completed Workflow

APT Batch AudioFiles

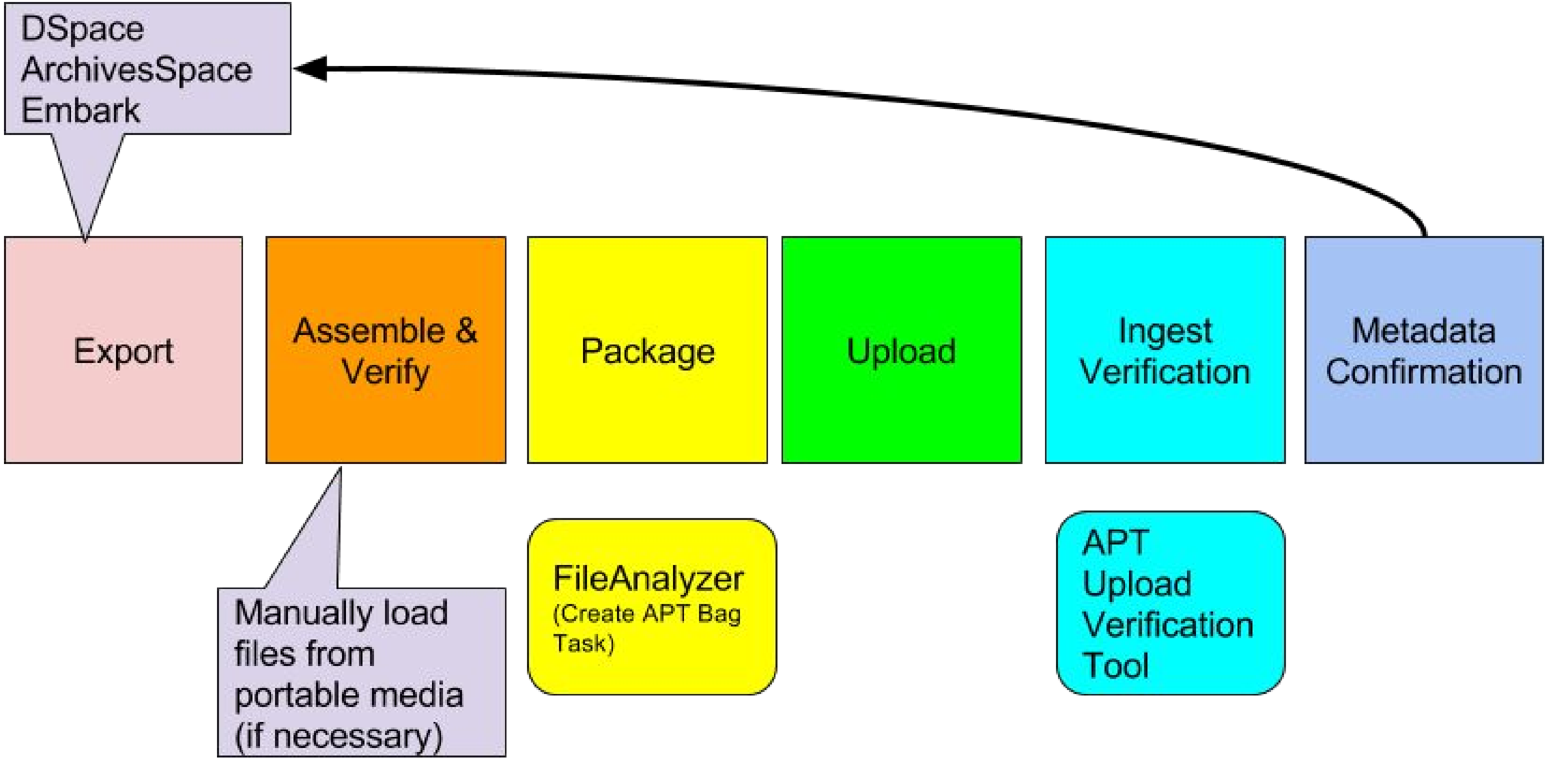
Batch Added

Identifier	bdae959f-64f7-4823-b5c7-cfbb617a4c5a
Short Name	AudioFiles
Batch Sequence	0
Milestone	Streaming Media
Item Type	dspaceAccessBitstream

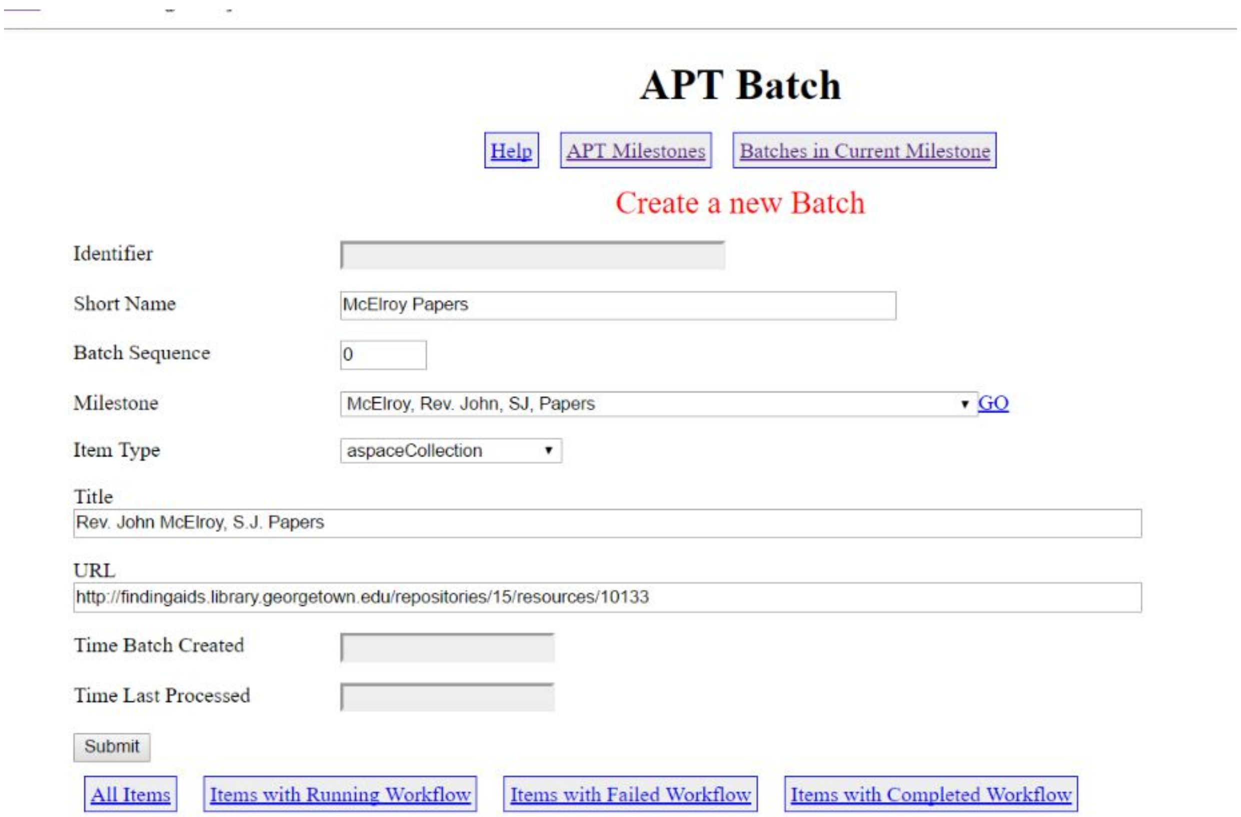
Title
Audio Files

URL

Time Batch Created	2016-05-31
Time Last Processed	2016-05-31


DSpace
ArchivesSpace
Embark
Export
Assemble & Verify
Package
Upload
Ingest Verification
Metadata Confirmation
Manually load files from portable media (if necessary)
FileAnalyzer
(Create APT Bag Task)
APT Upload Verification Tool

ArchivesSpace Metadata (External Preservation Files)



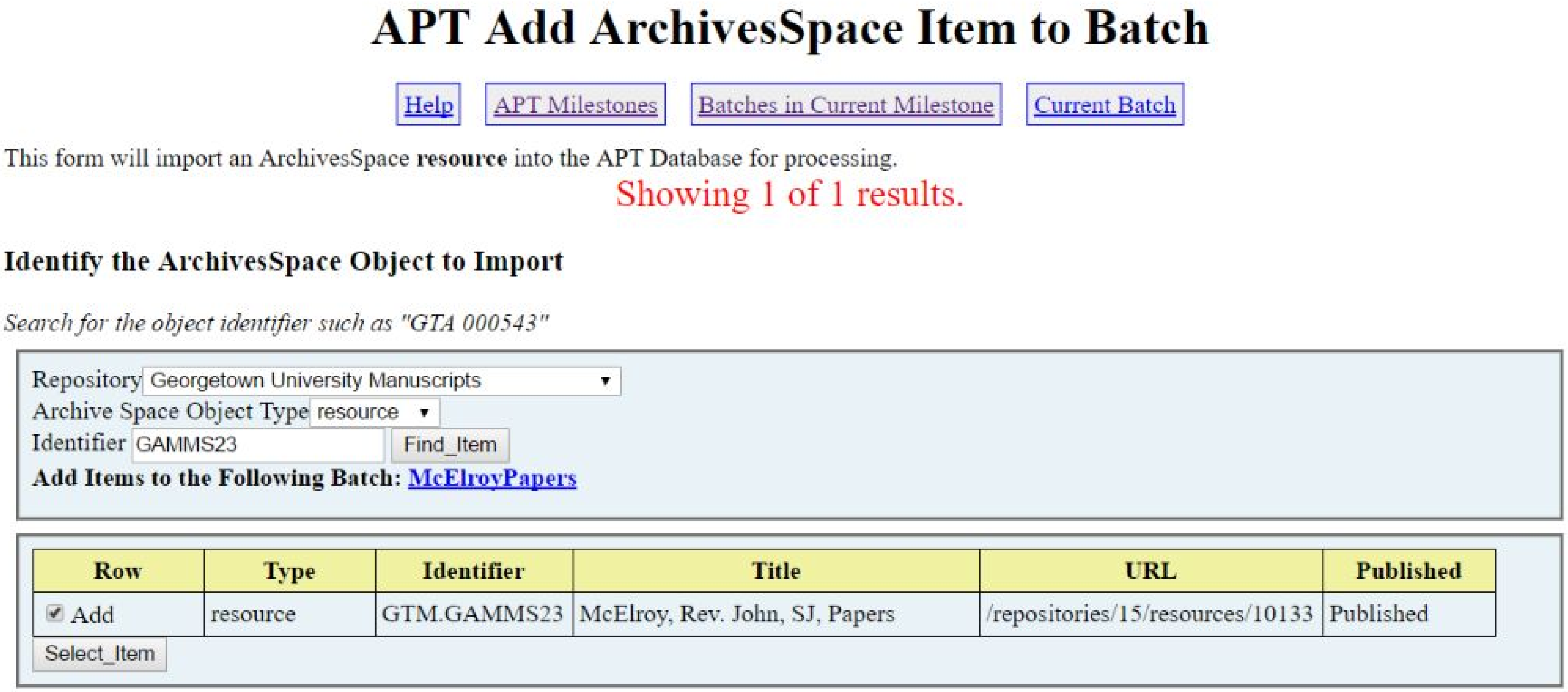
Help
APT Milestones
Batches in Current Milestone
GO
Submit
All Items
Items with Running Workflow
Items with Failed Workflow
Items with Completed Workflow

APT Batch

Create a new Batch

Identifier	
Short Name	McElroy Papers
Batch Sequence	0
Milestone	McElroy, Rev. John, SJ, Papers
Item Type	aspaceCollection
Title	Rev. John McElroy, S.J. Papers
URL	http://findingaids.library.georgetown.edu/repositories/15/resources/10133
Time Batch Created	
Time Last Processed	

Can select an accession or collection (bagged on collection level)



APT Add ArchivesSpace Item to Batch

This form will import an ArchivesSpace **resource** into the APT Database for processing.

Showing 1 of 1 results.

Identify the ArchivesSpace Object to Import

Search for the object identifier such as "GTA 000543"

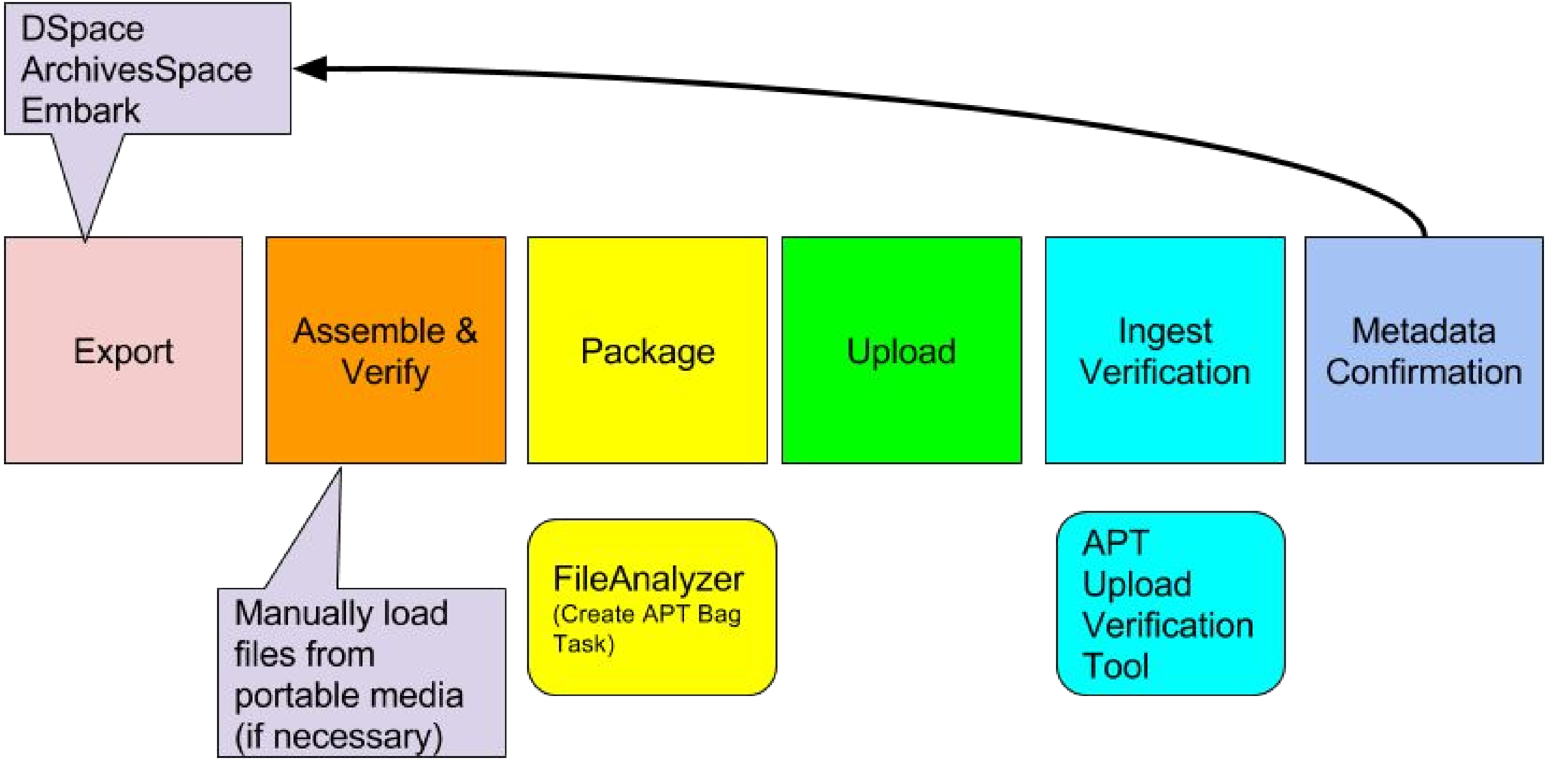
Repository: Georgetown University Manuscripts

Archive Space Object Type: resource

Identifier: GAMMS23

Add Items to the Following Batch: McElroyPapers

Row	Type	Identifier	Title	URL	Published
☑ Add	resource	GTM.GAMMS23	McElroy, Rev. John, SJ, Papers	/repositories/15/resources/10133	Published


DSpace
ArchivesSpace
Embark
Export
Assemble & Verify
Package
Upload
Ingest Verification
Metadata Confirmation
Manually load files from portable media (if necessary)
FileAnalyzer
(Create APT Bag Task)
APT Upload Verification Tool

Open Source Tools from Georgetown University Library

- FileAnalyzer (Bagging)
- ArchivesSpace Export (command line interface to ArchivesSpace API)
- APT Ingest Verification (command line interface to the APT API)

GitHub

http://georgetown-university-libraries.github.io/

Georgetown University Library Information Technology

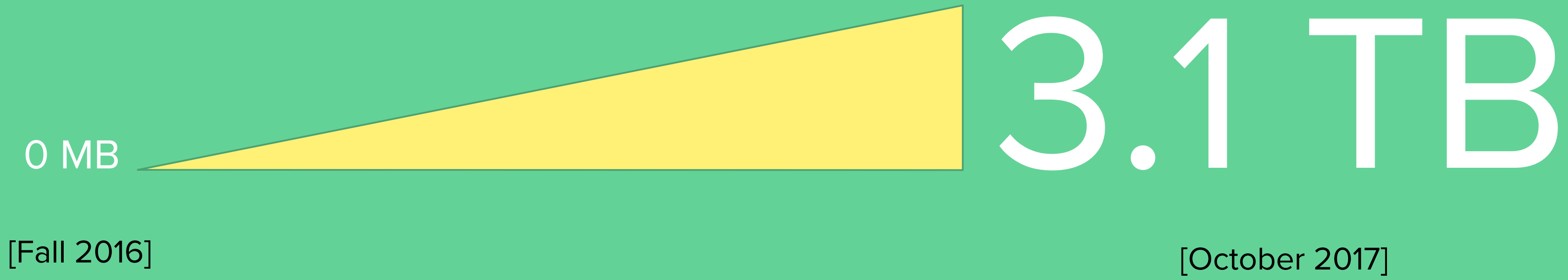
Washington, DC http://georgetown-university-libraries.github.io/

More details on our bagging process

Georgetown, APTrust bagging webinar

□ □ □ □ □ □ □ □ □ □

How do you measure a year?



Georgetown can now meet their digital preservation needs, grow, and take on challenges as they come.

Training library staff for digital preservation

- Post NDSR, we needed to distribute the work to multiple people
- No one position solely dedicated to digital preservation
- Digital preservation activities are centralized in the Library IT department
- Staff in other departments have been trained on tools and workflows, and contribute to digital preservation efforts

Roles and Responsibilities

- Head, Library Information Technology (Salwa Ismail)
    - Head, Digital Services Unit (Suzanne Chase)
        - Digital Services Specialist
    - National Digital Stewardship Resident (Joe Carrano)
    - Senior Programmer Analyst
    - Senior Systems Administrator
    - Preservation Coordinator
    - Assistant University Archivist

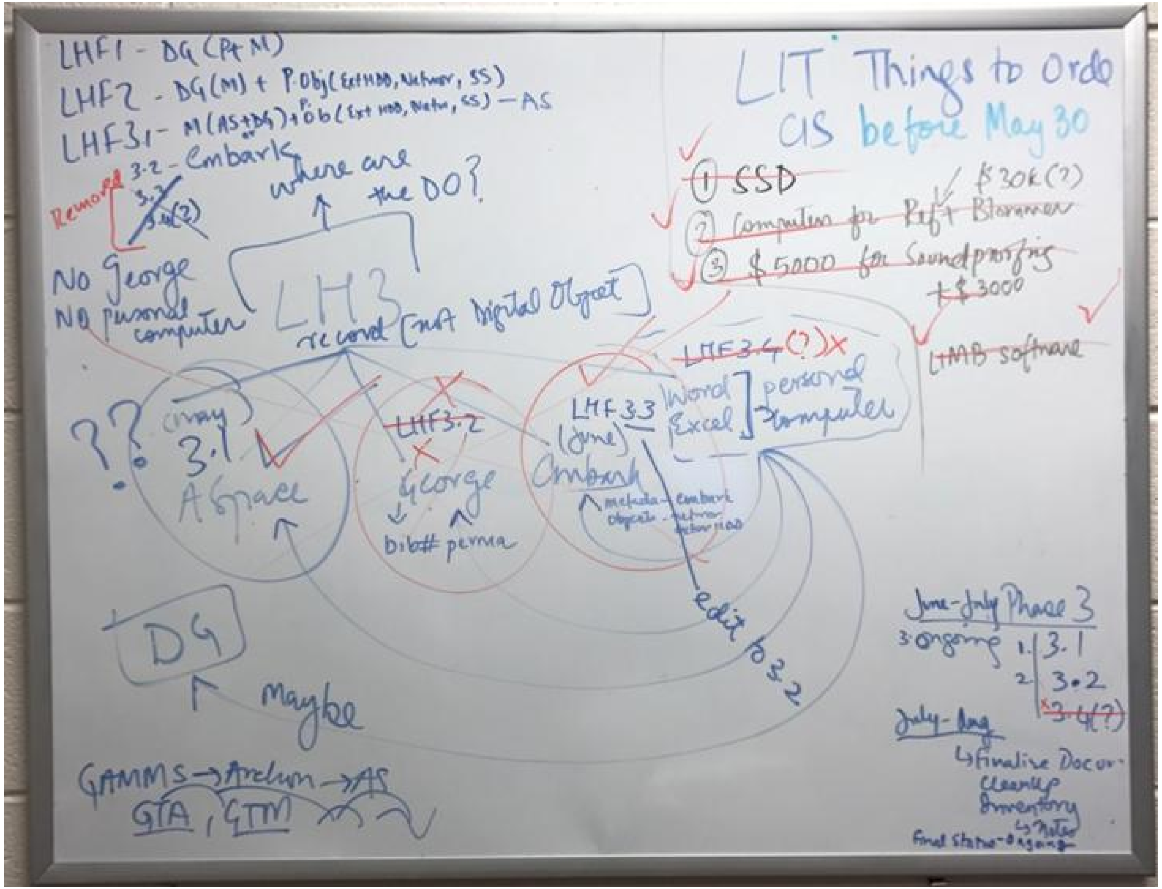
What We Learned Along the Way

Basically, the importance of:

- Selection and appraisal
- Iteration
- Testing
- Documentation
- Training
- Remaining flexible

*http://aptrust.org/post/2017-06-13-post-from-the-front-by-joe-carrano

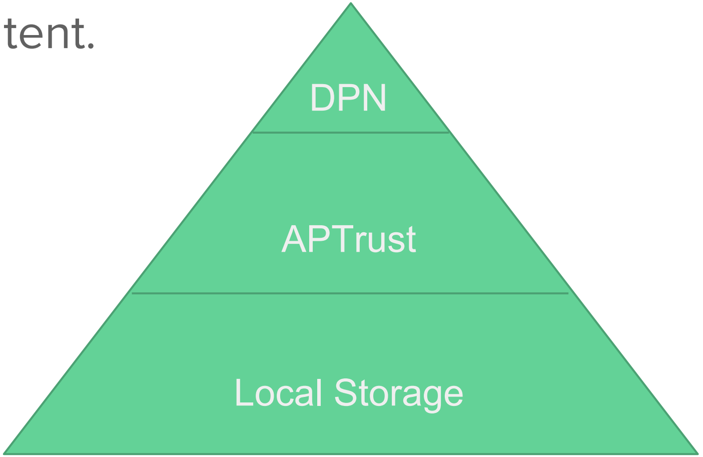
Developing digital preservation workflows is a process



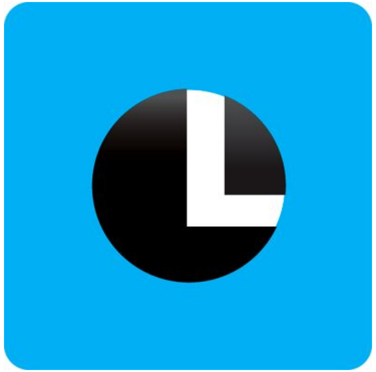
University of Cincinnati Libraries

Choosing Academic Preservation Trust (APTrust)

- Academic consortium working as a community to create a digital preservation environment for all types of content.
- Active members, dynamic meetings, sustained progress.
- APTrust is designed to stand on its own, sufficient basic preservation storage.
- If you're a member, you can also selectively send content to DPN.
- Easy deletion of content.



The Content Landscape



OJS

Open Journal Systems

DSPACE

samvera

Implementing APTrust: Toolkit

- Bagit-python
- APTrust Partner Tools and Amazon AWS CLI
- DSpace Packager
- jq (JSON processor)
- XSLT
- Bash
- Red Hat Enterprise Linux VM
- Github <https://github.com/uclibs/tricerabagger>

Implementing APTrust: DSpace Workflows

1. Export AIP packages from DSpace, by collection
2. Copy to staging storage
3. Extract AIPs zips, keeping all collections together in subdirectories
4. Bagging in place
5. Add APTrust specific tag files to bag
6. Re-generate tag manifests
7. Create uncompressed TAR balls for each collection
8. Push to APTrust receiving bucket
9. Query APTrust with API to track progress and remove staging files once ingested

Implementing APTrust: Samvera Workflows

Scholar@UC (at the time) uses Fedora 3

1. Copy Fedora directory tree to staging storage
2. Bag in place
3. Add APTrust tag files
4. Create uncompressed TAR balls for each collection
5. Push to APTrust receiving bucket
6. Query APTrust with API to track progress and remove staging files once ingested

Implementing APTrust: Filesystem Workflows

Used for LUNA, OJS, Website, and archival masters in Isilon storage

1. Export databases and metadata from content system.
2. Copy data and content files to staging storage.
3. Bag in place
4. Add APTrust tag files
5. Create uncompressed TAR balls for each collection
6. Push to APTrust receiving bucket
7. Query APTrust with API to track progress and remove staging files once ingested

Implementing APTrust: Roles and Responsibilities

- Head, Digital Collections and Repositories (Linda Newman)
    - Digital Content Strategist (was Nathan Tallman)
    - Repository Developers [2]
    - Developer Librarians [2]
- Head, Library IT
    - Systems Administrator
- Digital Archivist, Archives and Rare Books Library (Eira Tansey)

What We Learned Along the Way

- Your sysadmin is your friend. Make them cookies.
- Storage requests can take a long time, plan strategically.
- Permissions (particularly with Isilon) can be a blocker and complex to overcome if sysadmin is stingy with sudo.
- Firewalls do their job, which can prevent you from doing yours.
- Do you know where all your archival masters are? Metadata?
    - They are probably across multiple systems in multiple pieces.
- Even with scripts, documentation and cross training is invaluable.
- Character encoding is the devil.
- Size matters.
- Test restoration! Test, test, test!

Thank You!

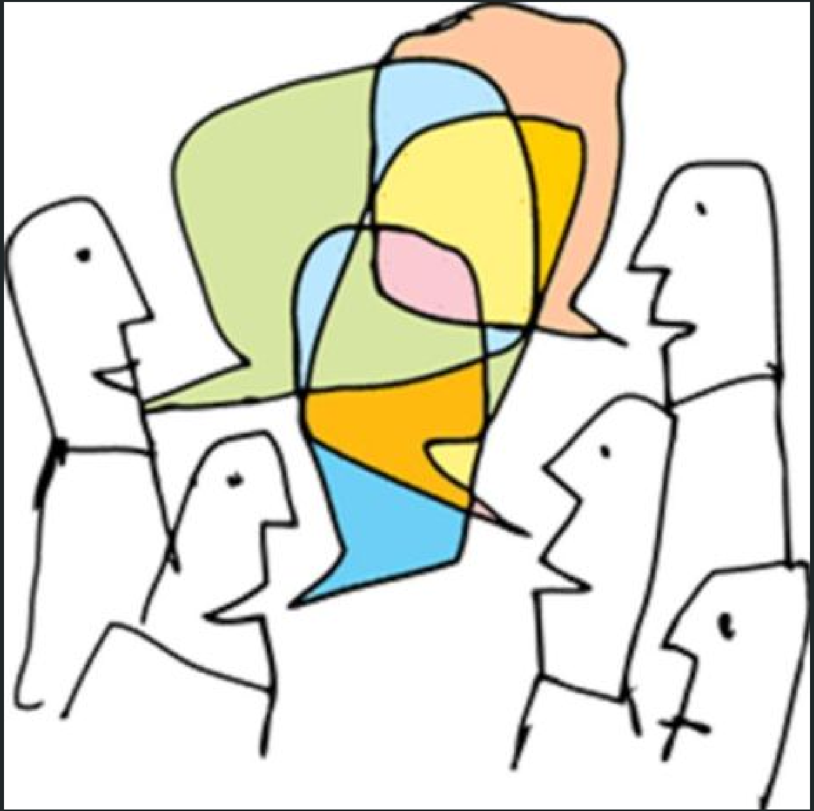
Questions?

CONTACT US:

- CARRANO@MIT.EDU
- DIGITALSCHOLARSHIP@GEORGETOWN.EDU
- NTT7@PSU.EDU
- NEWMANLD@UCMAIL.UC.EDU

Image Credits:
Icons made by Baianat and is licensed by Creative Commons BY 3.0
Images from Pexel.com, CC0 license, no attribution required

Discussion



Discussion

Q: What are the major barriers to developing and implementing digital preservation workflows at your institution?

Some common pain points:

- Content appraisal and selection
- Funding
- Administrative or inter-institutional blockages
- Confusion about who should “own” digital preservation responsibilities