Active Digital Preservation and Data/Metadata Migration

Nick Ruest, York University

Karen Estlund, Penn State University

Focusing on Movement

PennState
University Libraries

NDSA Preservation Activities

Table 1: Version 1 of the Levels of Digital Preservation

	Level 1 (Protect your data)	Level 2 (Know your data)	Level 3 (Monitor your data)	Level 4 (Repair your data)
Storage and Geographic Location	- Two complete copies that are not collocated - For data on heterogeneous media (optical discs, hard drives, etc.) get the content	- At least three complete copies - At least one copy in a different geographic location - Document your storage system(s) and storage media and	- At least one copy in a geographic location with a different disaster threat - Obsolescence monitoring process for your storage system(s) and media	- At least three copies in geographic locations with different disaster threats - Have a comprehensive plan in place that will keep files and metadata on
File Fixity and D Integrity				
Information Sec	individual files - Restrict who has those authorizations to individual files		and preservation actions	of logs
Metadata	- Inventory of content and its storage location - Ensure backup and non-collocation of inventory	- Store administrative metadata - Store transformative metadata and log events	- Store standard technical and descriptive metadata	- Store standard preservation metadata
File Formats	- When you can give input into the creation of digital files encourage use of a limited set of known open formats and codecs	- Inventory of file formats in use	- Monitor file format obsolescence issues	- Perform format migrations, emulation and similar activities as needed

- Check fixity of all content in response to specific events or activities
- Store standard preservation metadata
-

PennState University Libraries

Digital Preservation Storage

Lund, Ken. (2012). Timpanogos Cave National Monument. https://www.flickr.com/photos/kenlund/7187382348/ CC-BY 2.0

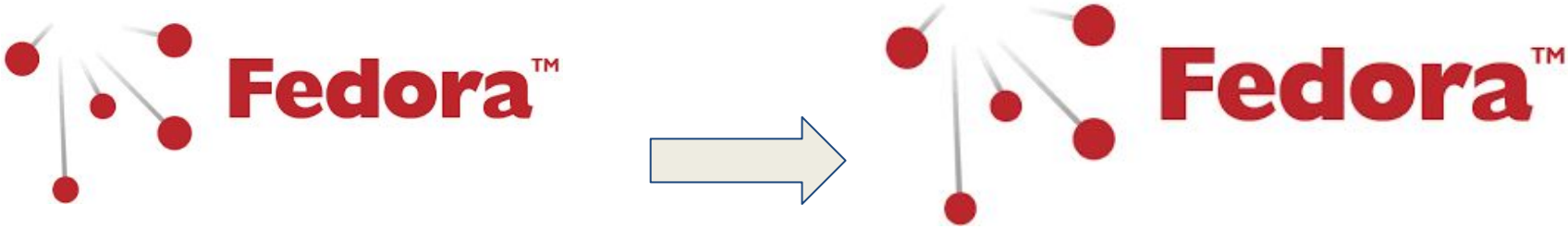
MoabAdventurer. (2010). Cataract Canyon. https://www.flickr.com/photos/tag-a-long/4818668522/ CC-BY 2.0

PennState University Libraries

PROBLEM STATEMENT

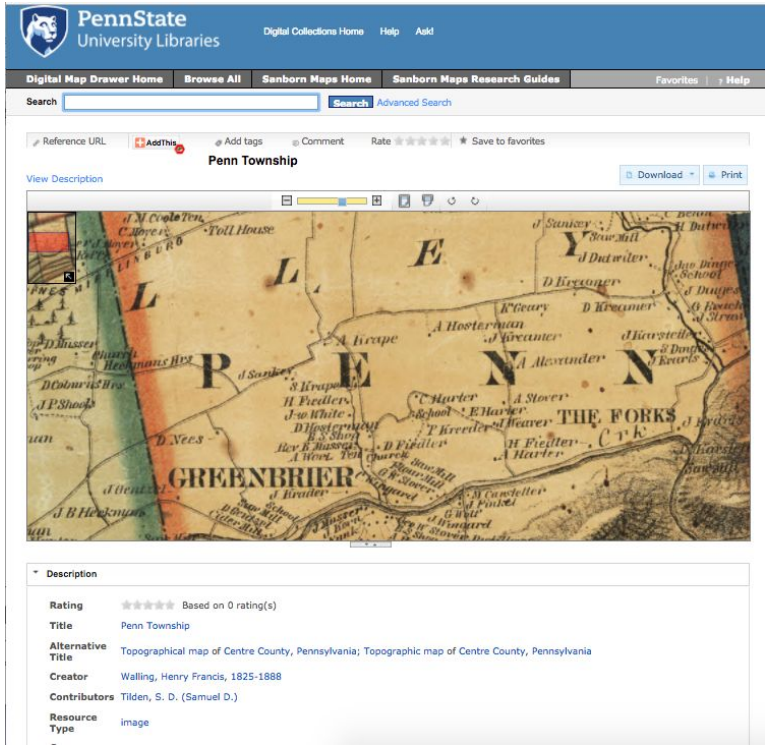
PennState
University Libraries

Repository Migrations



PennState University Libraries

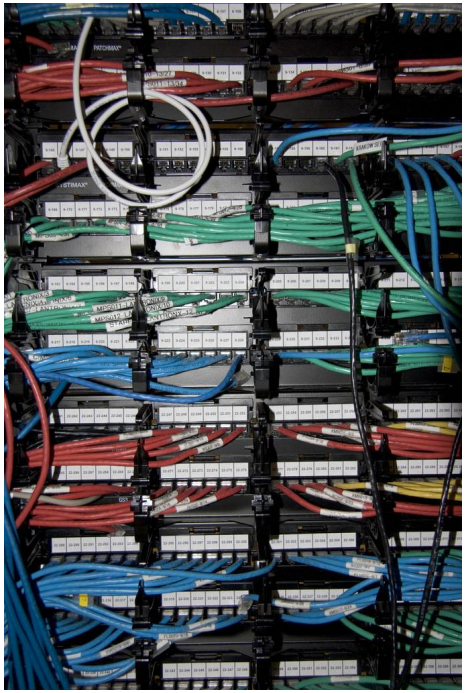
Connecting Access & Preservation Copies



```
<institutal_id.item_uid[.b###.of###]>/
|     bagit.txt
|     manifest-md5.txt
|     bag-info.txt
|     aptrust-info.txt
\----data/
     |     [payload files]
```

PennState
University Libraries

Big Data - Lack of Movement



Federal Communications Center. (2010). Data Center.
https://www.flickr.com/photos/fccdotgov/albums/72157624535788708

PennState University Libraries

ADDRESSING THE PROBLEMS

PennState
University Libraries

Staffing and Skills Required

Current Ads

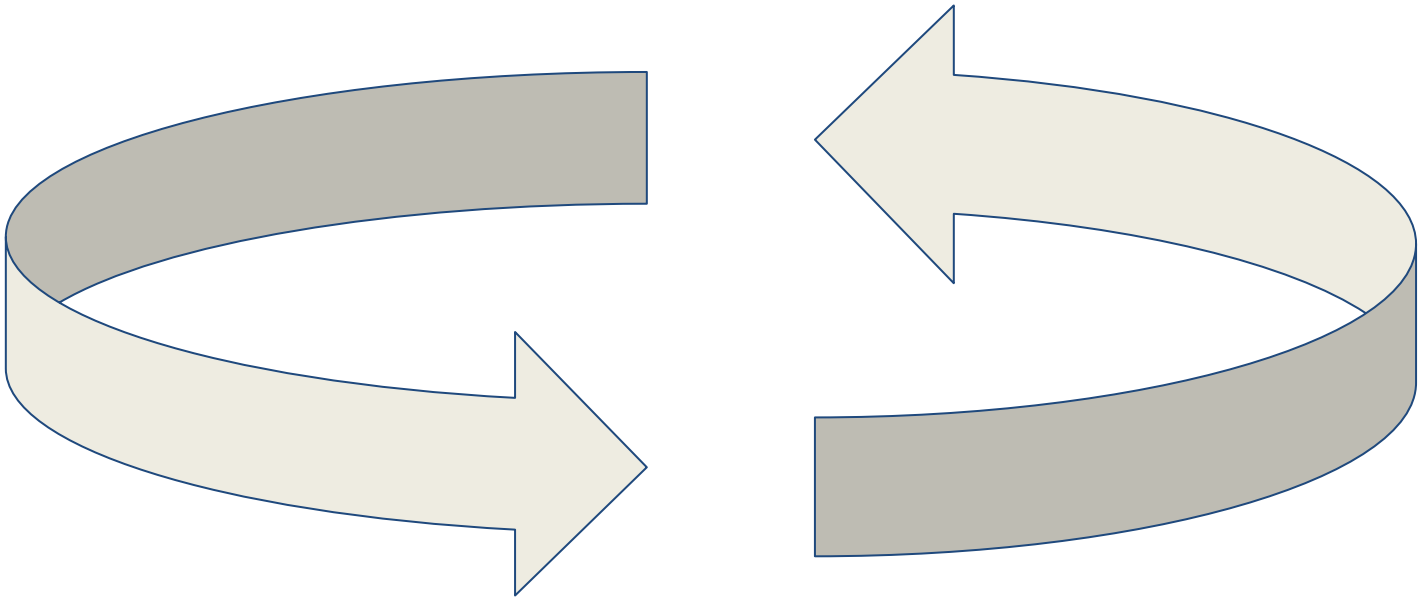
- Understanding of issues related to both digitized and born-digital formats, media, and migration is required
- METS & PREMIS (metadata standards)
- Fedora, Hydra, Islandora….
- Familiarity with national and international collaborative digital preservation efforts
- Digital preservation tools such as BitCurator, Archivematica, Preservica, BagIT and
- Application of Linked Data URIs in metadata records
- + all sorts of good standard language

Missing Abilities:

- Ability to communicate and understand networking and data center environments
- Ability to identify metadata necessary to preserve and track data in changing technical environments
- Ability to identify and manage copyright-related concerns for data preservation and metadata management
- Ability to forecast access needs and identify content preservation standards
- Ability to discern preservation vs. ephemeral content

PennState University Libraries

Export & Import



PennState
University Libraries

Infrastructure & Local Collaborations Required

Federal Communications Center. (2010). Data Center. https://www.flickr.com/photos/fccdotgov/albums/72157624535788708

PennState University Libraries

Open Source Commitment

NASA. A Precocious Black Hole. (2015).
https://www.nasa.gov/image-feature/a-precocious-black-hole

PennState University Libraries

Decision Trees – Where to Preserve & How to Access*

Preservation Strategy	Back-up	Local Preservation Strategies**	MetaArchive	APTrust	Digital Preservation Network
Research Data	X			X	
ETDs	X		X		
PA unique CHO materials	X	X		X	
Purchased content	X				
High value, at risk unique materials	X	X			X

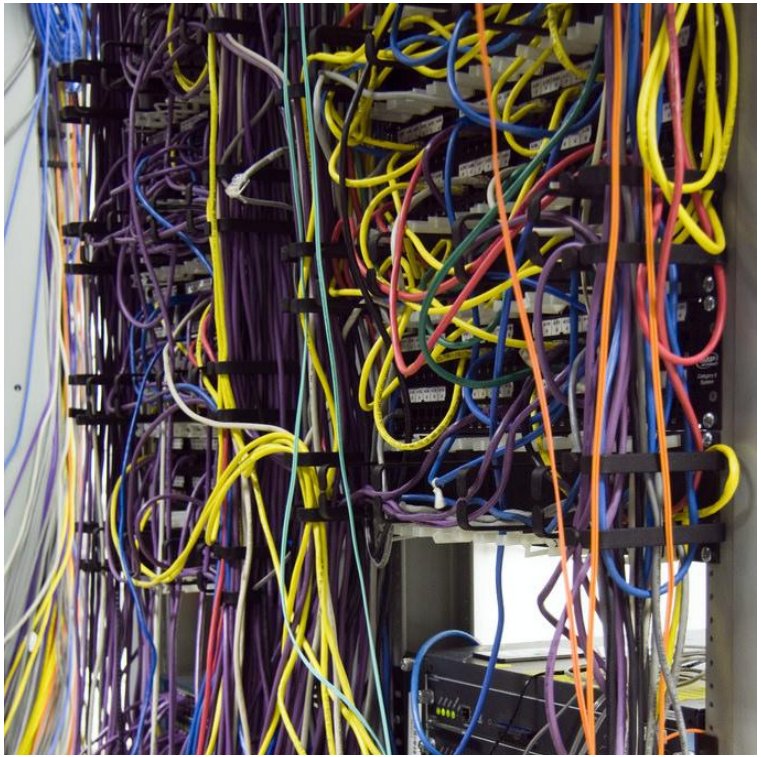
* Draft decision tree
** May include off-site storage

PennState University Libraries

NEXT STEPS

PennState
University Libraries

What Does Digital Preservation Look Like Now



Federal Communications Commission. (2010). Data Center
https://www.flickr.com/photos/fccdotgov/4808782554/

PennState University Libraries

Fedora Import/Export

Background

Stakeholders

Esmé Cowles, *Princeton University*
Ben Armintor, *Columbia University*
Mike Durbin, *University of Virginia*
Josh Westgard, *University of Maryland*
Youn Noh, *Yale University*
Mike Giarlo, *Stanford University*
Jon Stroop, *Princeton University*
Karen Estlund, *Penn State University*
Jim Tuttle, *Duke University*

Planning

Use Cases

Use cases

1. Transfer between Fedora and external preservation systems, such as APTrust, MetaArchive, LOCKSS, DPN, Archivematica, etc
2. ~~Package~~ [Export] the content of a single Fedora container and all its descendant resources
3. Transfer between fedora instances or (more generally) from Fedora to an LDP archive
4. ~~load~~ [Import] ~~the contents of a package~~ into a specified container.
5. Round-tripping resources in Fedora in support of backup/restore
    a. A start has been made on this in FCREPO-1990;
    b. The implementation referenced in the above ticket is not dead, though not actively being worked on at the moment; pull requests welcomed (though others may well wish to take it in a different direction).
    c. A rebuilder that:
        i. Is not solely dependent on a intact backup of the repository index
        ii. Works off shredded serializations that can be supported with file preservation techniques
        iii. Can recover as much as possible of a repository in the face of integrity issues (supports partial recovery)
        iv. Supports gathering copies of the shreds (serializations) from multiple sources to recover a repository
6. Round-tripping resources in Fedora in support of Fedora repository version upgrades
7. ~~Batch loading arbitrary sets of resources from metadata spreadsheet and binaries (may well be difficult – or not worth it – to try to generalize such a feature).~~
8. Import or export containers or binaries using add, overwrite, or delete operations. Configure the data model and the source and the target for each resource that will be updated. Allow target containers to be non-empty before import and source containers to be non-empty after export. Maintain ordering, etc. Support versioning. Examples: add issues to a publication; add fragments to a manuscript; add data sets to a longitudinal study; add time-series images from telescopes; remove resources determined to be under copyright; release resources after restrictions on access have expired.

Use cases yet to be rolled into requirements

1. Import objects from an external system (such as Figshare, where a research data object might be prepared) into a Fedora preservation repository with either Hydra or Islandora on top. (Implies compliance with Hydra and/or Islandora object models)

Design

Requirements

External Systems

1. PHASE 2 Support import from and export to a TBD list of external systems.
   a. APTrust - University of Maryland (@Joshua Westgard)
   b. Archivematica - Artefactual Systems (@Justin Simpson)
   c. MetaArchive - Penn State (@Ben Goldman)
   d. Perseids - Tufts - @Bridget Almas

General

1. PHASE 1 Support transacting in RDF
2. PHASE 1 Support allowing the option to include Binaries
3. PHASE 1 Support references from exported resources to other exported resources
4. PHASE 2 Support transacting in BagIt bags
5. PHASE 1 Support import into a non-existing Fedora container
6. PHASE 2 Support import into an existing, empty Fedora container
7. PHASE 3 Support import into an existing, non-empty Fedora container with various policies: add, overwrite, delete, version, skip
8. PHASE 3 Support export of resource versions
9. PHASE 3 Support import of resource versions
10. PHASE 1 Support export of resource and its "members" based on the ldp:contains predicate
11. PHASE 2 Support export of resource and its "members" based on a user-provided membership predicate
12. ~~Support recursive RDF insert/updates with LDP Indirect Container specified POST (and PUT / PATCH?) (ref: FCREPO-2042)~~

Requirements

Round-tripping

Defined as: Export all or a subset of a Fedora repository and importing the export artifacts into a Fedora repository.

1. PHASE 3 Support preservation of dates during round-tripping
2. PHASE 3 Support preservation of version snapshots during round-tripping
3. PHASE 1 The URIs of the round-tripped resources must be the same as the original URIs
4. PHASE 3 Support lossless round-tripping. (ie, if you export a resource, delete that resource and import there is no difference from if you had never performed any of those operations).

BagIt

1. PHASE 2 Single resource bags
2. PHASE 2 The structure and scope of accepted and produced BagIt bags must be configurable (resource)
   a. Clarification: *structure* relates to required and optional tagfiles in the bag
   b. Clarification: *scope* relates to contents of the bag, e.g. single object or object and all members based on specific membership predicate
3. PHASE 3 Multi-resource bags
4. PHASE 3 Unambiguously support linking between resources within a bag, and from resources in the bag to resources outside the bag
   a. e.g. for bagged resources A and B, if A contains statement <A> myns:rel <B>, then it is unambiguous that B is a resource in the bag. Suppose some archive ingests the bag and exposes its contents as web resources with URIs P and Q. If the archive preserves intra-bag links, resource P will have statement <P> myns:rel <Q>. Likewise, if A contains external link <A> myns:rel2 <http://example.org/outside/the/bag>, then an archive that preserves links will have <P> myns:rel2 <http://example.org/outside/the/bag>

Requirements

Verification Tool

1. PHASE 2 Verify same number of resources on disk as in fcrepo
2. PHASE 2 Verify same number of resources in fcrepo as on disk
3. PHASE 2 Verify same checksum for binaries
4. PHASE 2 Verify same triples for containers
5. PHASE 2 Record which resources have been verified (Include checksum for binary resources)
6. PHASE 2 Verify subset of repository resources
7. PHASE 3 Verify fcrepo to fcrepo
8. PHASE 3 Verify disk to disk
9. PHASE 3 Use generated config file as sole input

Considerations

- Import/export performance as is possible under the assumption that this work is done via the REST interface

Sprints

August, 2016

Devs

Esmé Cowles
Ben Arminator
Nick Ruest
Mike Durbin

Testing & Validation

Mike Durbin
Josh Westgard
Justin Simpson
Youn Noh
Yinlin Chen
Bethany Seegar

Documentation

Youn Noh
Josh Westgard

September, 2016
(Penn State)

Devs

Esmé Cowles
Nick Ruest
Andrew Woods

Testing & Validation

Adam Wead
Nick Ruest
Andrew Woods
Karen Estlund

Documentation

Adam Wead
Nick Ruest

December, 2016

Devs

Esmé Cowles
Nick Ruest
Jared Whiklo
Danny Bernstein

Testing & Validation

Josh Westgard
Nick Ruest
Kieran Etienne
Adam Wead
Justin Simpson

Documentation

Josh Westgard
Nick Ruest

Stakeholder evaluation

Sprint 3 Stakeholder Feedback

Created by Andrew Woods, last modified on Mar 14, 2017

With the completion of the third Import/Export sprint we are now in a position to ask our stakeholders to test and verify the Phase 2 requirements.

Specifically, the Import/Export utility is capable of round-tripping Fedora resources from one repository into a second, initially empty repository, or into an existing repository to an empty container, based on the ldp:contains predicate or based on a user-provided membership predicate. It is also capable of exporting and importing BagIt bags based on a default BagIt Profile, or an APTrust BagIt Profile. BagIt tag files are configurable with a user supplied configuration. Therefore, you should be able to create resources in a Fedora repository, export those resources as RDF and Binaries to the filesystem, then re-import them.

This foundational work supports use cases including:

- Migration from one version of Fedora to another
- Persistence of Fedora resources to disk as standard RDF and Binaries for preservation
- Repository and Object-level disaster recovery
- Transfer between Fedora and external preservation systems, such as APTrust

Testing and verification of the Import/Export utility requires three elements:

- The Import/Export utility
- A running Fedora repository (4.7.0 one-click-run download)
- Data in the repository to export then re-import

Instructions on how to use the utility are outlined in the utility README, and the Import and Export Tools Administrator Guide.

Although we encourage you to create your own data to test the utility; you can also use one of the datasets created by the sprint team. If you have questions on how to use these datasets, please do not hesitate to ask.

We would like to finalize sign-off on the Phase 2 requirements from Import/Export stakeholders prior to moving further on additional Phase 2 external systems, and moving on to Phase 3. It would be helpful if you **completed your testing by January 31, 2017**. Feedback can be provided by creating a child page here using the example template. Andrew Woods has created the first one, which can also be used as a template.

In the course of your testing, please let us know if there are any issues with the documentation or the utility by creating a JIRA ticket with the "Component" field set to "f4-import-export". Existing issues can be found here. If you need help or a hand with creating issues, please do not hesitate to ask.

Feedback

- Sprint 3 Feedback - A Woods
- Sprint 3 Feedback - Esmé Cowles
- Sprint 3 Feedback - Example Template
- Sprint 3 Feedback - Joe Atzberger
- Sprint 3 Feedback - Kieran Etienne

Where we’re at now

Next steps