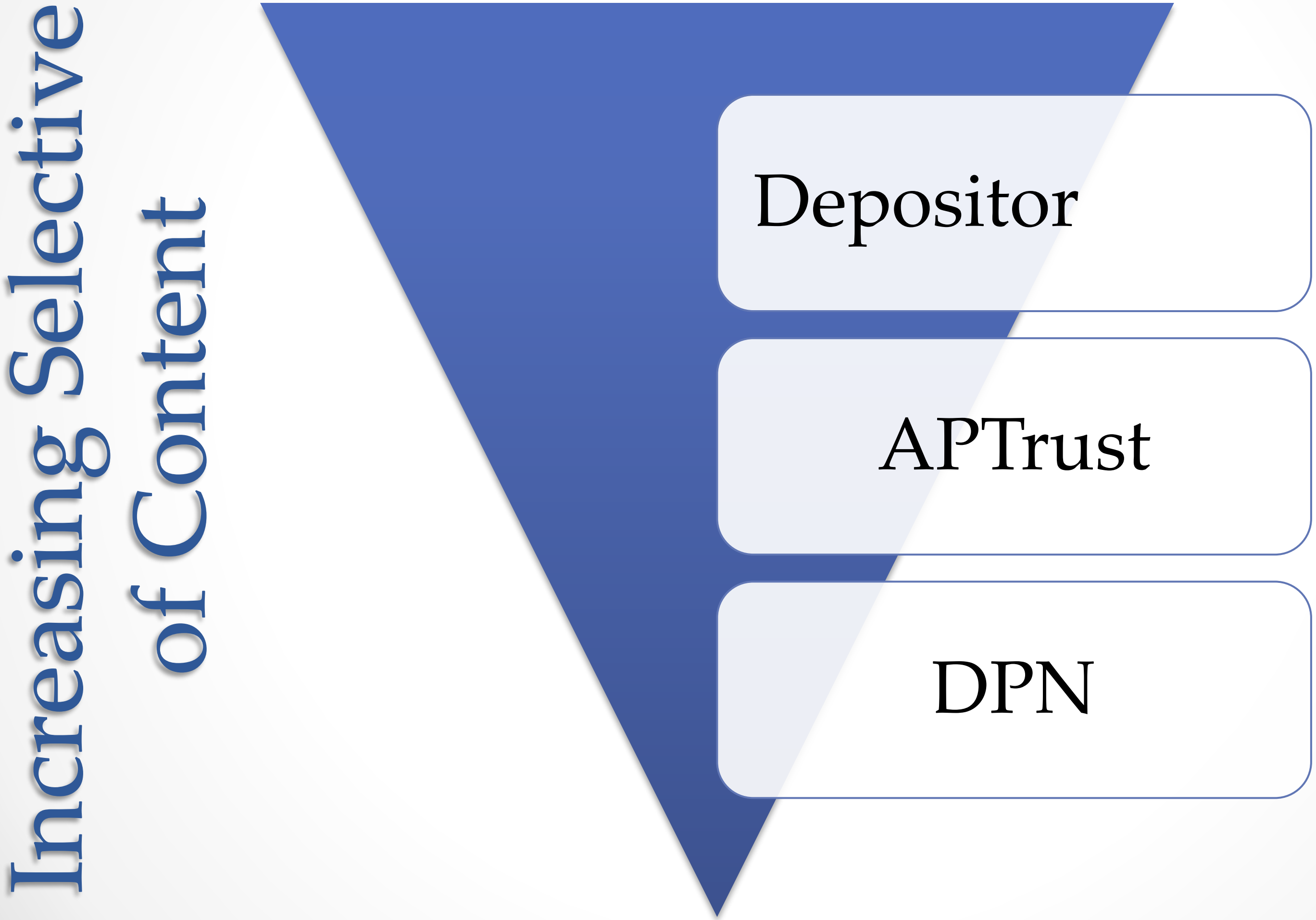
Packaging Content for Aggregated and Federated Repositories through APTrust

Scott Turnbull
University of Virginia
scott.turnbull@aptrust.org

Overview of APTrust

- Academic Preservation Trust (APTrust)
- ... incubated at the University of Virginia.
- ... aggregates content from16 member Universities.
- ... cloudbased architecture for bit level preservation.
- ... backed with Fedora with administrative access through a Hydra interface.
- Accept content regardless of format and filetype.
- Participating as First Node in the Digital Preservation Network (DPN) with 4 other Repositories.
- BagIt used common unit of exchange.

Layers of Related Repositories



Increasing Selective of Content

Depositor

APTrust

DPN

Cross Repository Needs

Identifiers

- Single identifier across repositories to link content.
- Avoid complex management of multiple identifiers.
- Desire for identifiers based on information depositor already has.

Packaging

- Need for a single simple packaging format that can be easily translated.
- Need to decouple packaging from repository software and versions.
- Desire for clean content serialization that only includes important preservation metadata and content.

Overview of BagIt

- A base level specification for serializing arbitrary archival content to filesystem.
- Tag Files contain metadata.
- Preservation files are under a data/ subdirectory called the bag payload.
- Checksums for payload listed in a manifest file.
- Support optional metadata files.
- Generally wide support in many tools.
- Bag filepaths map nicely to URIs

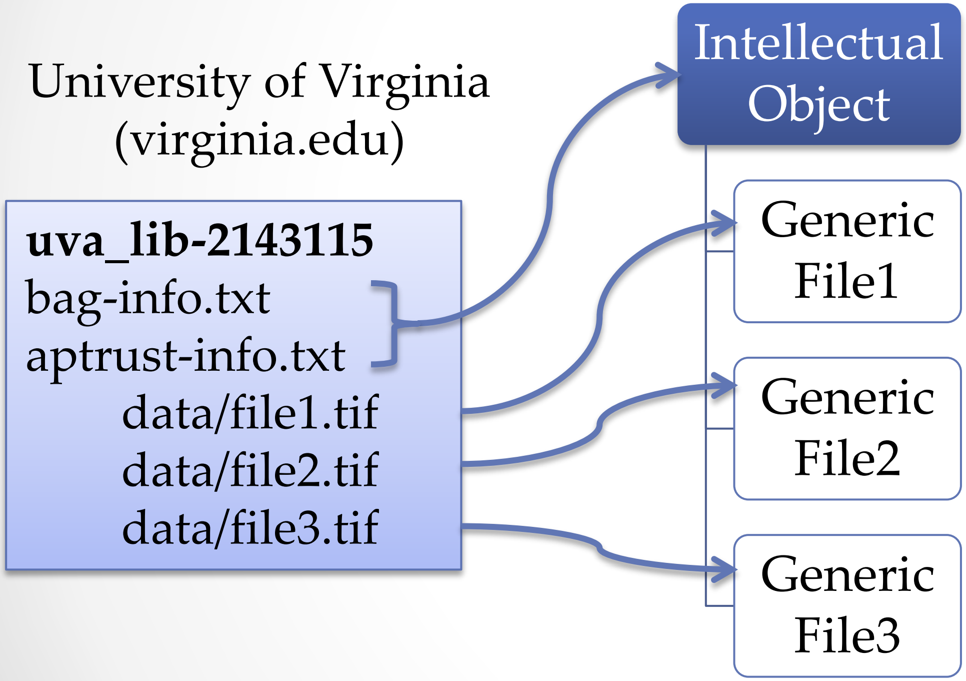
Use of BagIt at APTrust

- Submission Package:
  - Created by depositor based on shared profile.
  - Depositor determines what goes into their bags.
  - Depositor provides an institutionally unique identifier used as the bag name.
  - Supports Multi-part bags for large content (> 250 GB)
- Bags ***Not*** Used as Archival Package.
  - DPN ***DOES*** use bags as an Archival Package.
- Distribution Package (Restoration Bags)
  - Use same APTrust profile for a bag.
  - Contains exactly the same payload files as submitted to APTrust.
  - Also recoverable from DPN in “brightening” scenarios.
- Depositors give us bags, we give them back bags.

Bag Ingestion and Identifiers

University of Virginia
(virginia.edu)

APTrust Fedora Objects



uva_lib-2143115
bag-info.txt
aptrust-info.txt
data/file1.tif
data/file2.tif
data/file3.tif

Intellectual Object

Generic File1

Generic File2

Generic File3

Corresponding Identifiers

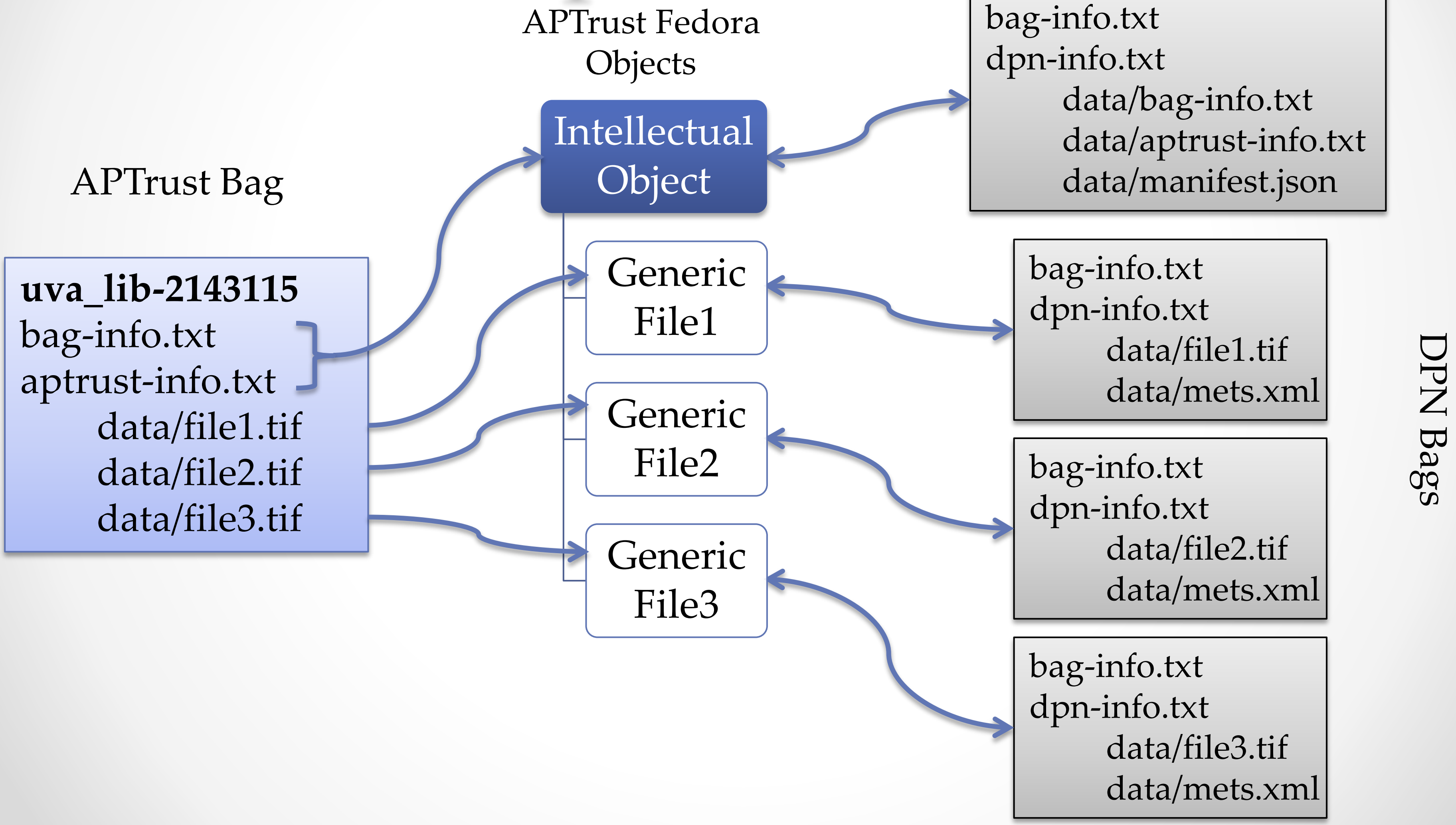
apt://virginia.edu/uva_lib-2143115

apt://virginia.edu/uva_lib-2143115/data/file1.tif

apt://virginia.edu/uva_lib-2143115/data/file2.tif

apt://virginia.edu/uva_lib-2143115/data/file3.tif

Flow of Content Across Repositories



APTrust Bag

uva_lib-2143115
bag-info.txt
aptrust-info.txt
data/file1.tif
data/file2.tif
data/file3.tif

APTrust Fedora Objects

Intellectual Object

Generic File1

Generic File2

Generic File3

DPN Bags

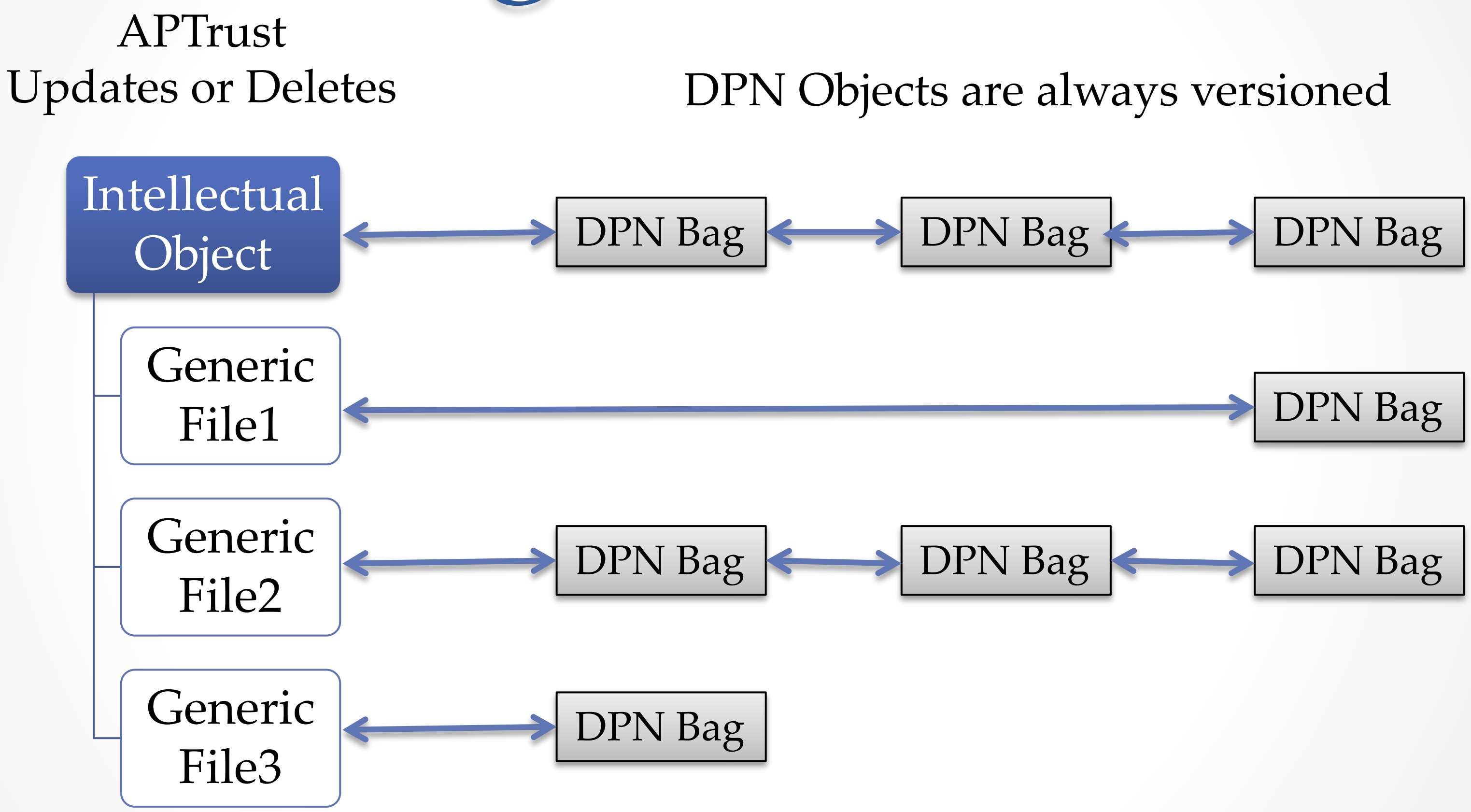
bag-info.txt
dpn-info.txt
data/bag-info.txt
data/aptrust-info.txt
data/manifest.json

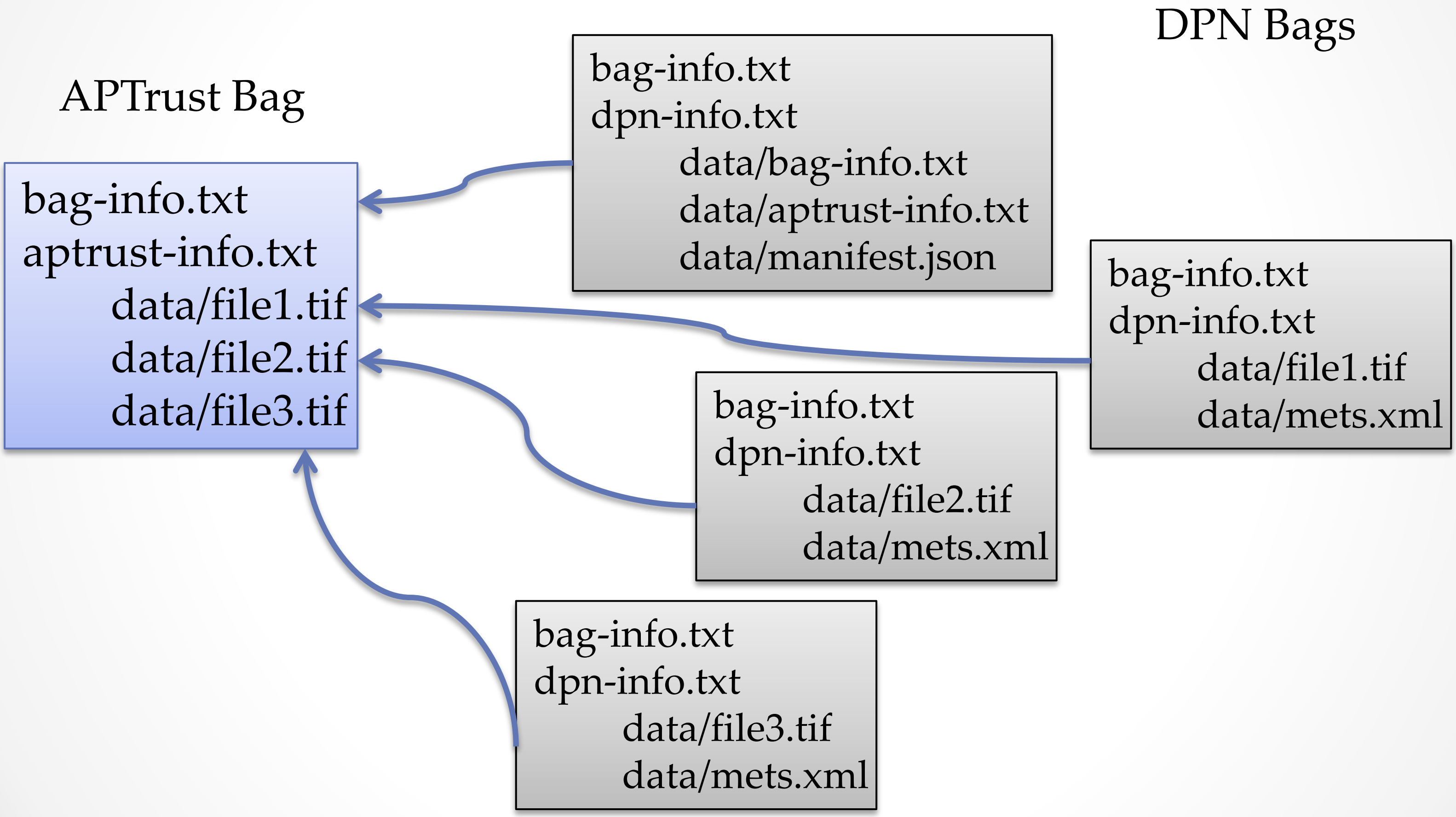
bag-info.txt
dpn-info.txt
data/file1.tif
data/mets.xml

bag-info.txt
dpn-info.txt
data/file2.tif
data/mets.xml

bag-info.txt
dpn-info.txt
data/file3.tif
data/mets.xml

Versioning Across to DPN





APTrust Bag

bag-info.txt
aptrust-info.txt
data/file1.tif
data/file2.tif
data/file3.tif

DPN Bags

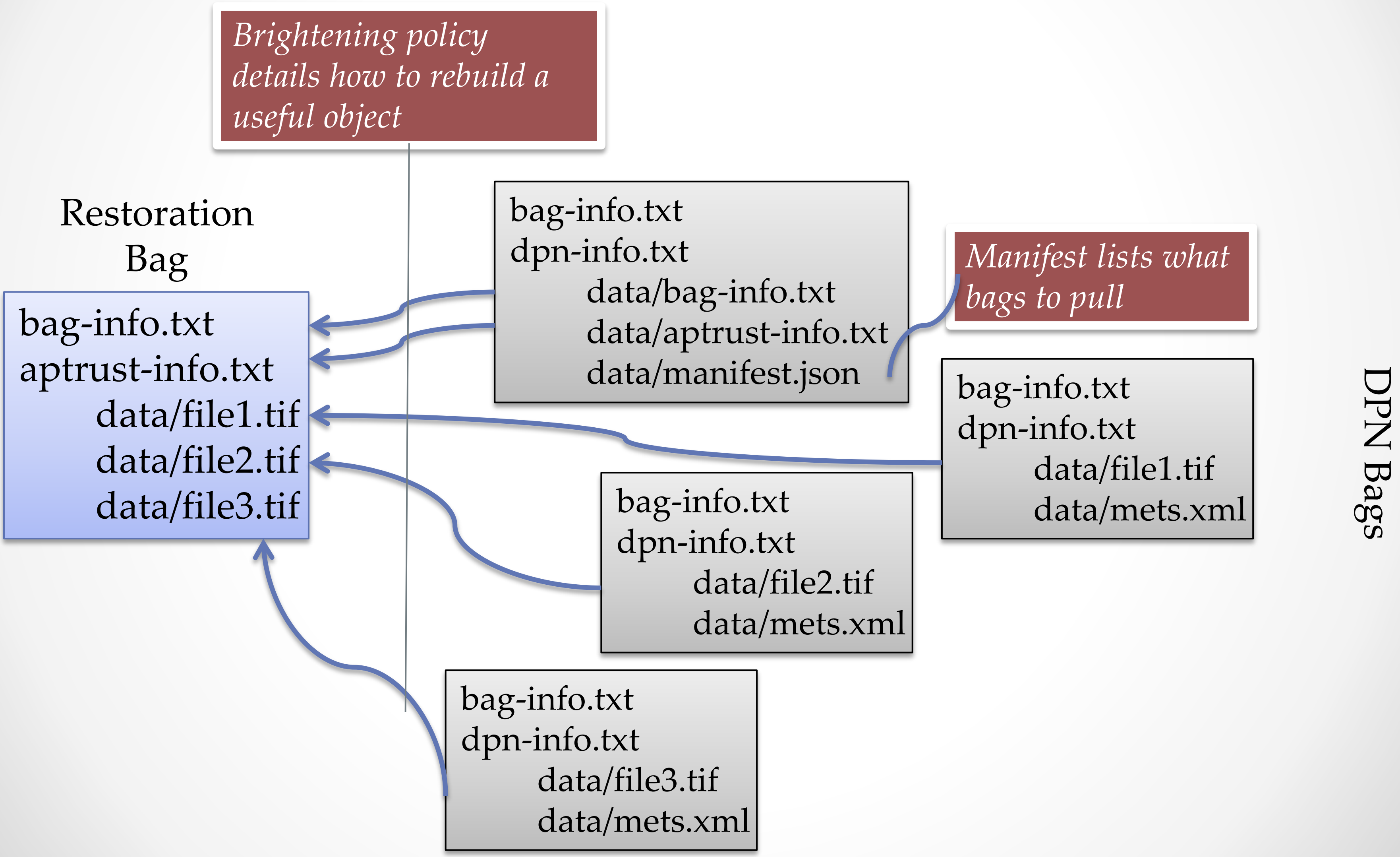
bag-info.txt
dpn-info.txt
data/bag-info.txt
data/aptrust-info.txt
data/manifest.json

bag-info.txt
dpn-info.txt
data/file1.tif
data/mets.xml

bag-info.txt
dpn-info.txt
data/file2.tif
data/mets.xml

bag-info.txt
dpn-info.txt
data/file3.tif
data/mets.xml

DPN Brightening



Brightening policy details how to rebuild a useful object

Restoration Bag

bag-info.txt
aptrust-info.txt
data/file1.tif
data/file2.tif
data/file3.tif

bag-info.txt
dpn-info.txt
data/bag-info.txt
data/aptrust-info.txt
data/manifest.json

Manifest lists what bags to pull

bag-info.txt
dpn-info.txt
data/file1.tif
data/mets.xml

bag-info.txt
dpn-info.txt
data/file2.tif
data/mets.xml

bag-info.txt
dpn-info.txt
data/file3.tif
data/mets.xml

DPN Bags

Cross Repository References

Scenario: University of Virginia wants to rebuild object uva_lib-2143115 from DPN:

- apt://virginia.edu/uva_lib-2143115
- Read manifest.json
- Follow Instructions

```
{
  "Institution": "University of Virginia",
  "RestorationBagName": "uva_lib-2143115",
  "instructions": "Directions for building a 
restore bag.",
  "manifest": [
    {
    "DPNObjectID": "aptrust-6c84fb90-12c4-
11e1-840d-7b25c5ee775a",
    "src": "data/bag-info.txt",
    "dst": "bag-info.txt"
    },
    {
    "DPNObjectID": "aptrust-6c84fb90-12c4-
11e1-840d-7b25c5ee775a",
    "src": "data/aptrust-info.txt",
    "dst": "aptrust-info.txt"
    },{
    "DPNObjectID": "aptrust-110ec58a-a0f2-
4ac4-8393-c866d813b8d1",
    "src": "data/file1.tif",
    "dst": "data/file1.tif"
    }
  ]
}
```

Packaging Concerns

- Significant responsibility on the original depositor.
- Balancing flexibility against enforcing standards.
- Straight export presents problems for a true preservation package:
- Extraneous data can build up in some records and lead to confusion:
    - Solr Doc data
    - Aggregation Data
    - Internal administrative data
    - Preservation Package .neq. Application Backup

Summary

- Convention based identifiers allow for easy reference.
- Having a common format of exchange is useful to cut down on complexity.
- Strong need to evolve how we are serializing content from our repositories.

Questions?

- Public Website: http://www.aptrust.org
- Technical Documentation: https://sites.google.com/a/aptrust.org/aptrust-wiki/
- Email: scott.turnbull@aptrust.org