Packaging Content for Aggregated and Federated Repositories through APTrust

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

Abstract

The Academic Preservation Trust (APTrust) provides an aggregated cloud-based bit level preservation repository for institutional content. It also serves as a first node in the Digital Preservation Network (DPN). By providing a format and repository agnostic submission packaging strategy, it is able to accept content of any type for digital preservation and helps to establish a layer of abstraction for digital objects that maintain references, allow for distributed health checks, and set the stage for relating content across other preservation systems. Digital Object health checks involve confirmation fixity against hashes generated both by the original contributor and by APTrust itself to provide a strong chain of custody. As a format agnostic repository, a larger burden is placed on content packaging and preparation to ensure interoperability and easy restoration, but once a lightweight method is established, it has great utility for maintaining the portability of objects across the preservation ecosystem as well as for restoration.

Audience

Repository Managers, Content Specialists and Developers who design systems for serializing content or identity management systems for preservation content.

Background

The Academic Preservation Trust aggregates content from 16 American partner universities by providing an abstraction and packaging policy allowing for basic file level preservation services. It also sets the stage to serve as an entry-point for even deeper levels of preservation in the DPN federated repository. As a loosely coupled service, its focus is to create portable formats for digital object exchange that preserve the integrity of files and metadata, satisfy a strong chain of custody requirements, and allows linking objects across a number of related services while keeping coupling to a minimum. Checksums are managed both by the original owner and APTrust and auditing provided to ensure that the results of object health checks are confirmed by the aggregator and by the original owner. Having the submission packaging create persistent identifiers for unstructured content helps APTrust enhance preservation with its own services and keeps objects linked across repository ecosystems.

Submission packages to APTrust are created by serializing local repository content using the Library of Congress BagIt open standard. Metadata about objects to be preserved in APTrust are supplied as part of BagIt tag files and the actual files to be preserved grouped under the bag

data directory in whatever format makes sense, given the original local implementation. Bags are decomposed upon receipt and the tag metadata registered with a top level Intellectual Object in the APTrust repository. Child objects called Generic Files represent the digital content provided in the bag data directory and store pointers to the related file in the APTrust preservation space. Intellectual Objects are linked to the original bag by the bag name, which is a composite of the institutional identifier and a unique identifier for the object provided by the owner during packaging. Related Generic Files for each Intellectual Object are referenced by a filehandle formed by a composite of the intellectual object identifier and relative path of that file in the bag. Using this scheme of unique identifiers APTrust is able to relate content in the APTrust repository back to the content from the original repository and form a common identifier to link references to files preserved in other systems like DPN. Linking these identifiers across systems provides a straight-forward way to version items by identifier and date and provide reporting on content across different repository solutions.

APTrust's submission packaging and identity management also ensures that both the original content owners and APTrust are actively participating in the long term preservation of each item. As part of content preparation and bagging for submissions, an MD5 checksum is generated on packaging and stored with the Bag manifest. This checksum is primarily used to ensure the good transfer and chain of ownership of items as it moves from the original repository through each stage of transfer and processing into APTrust. As part of the final ingestion progress of content into APTrust, a SHA256 checksum value is generated on the content to ensure cryptographically security. Fixity is provided for long term preservation and to mitigate against malicious tampering as well as bit loss. As part of the full lifecycle of digital object preservation, both the original MD5 checksum generated during content packaging and the SHA256 checksum generated during ingest are matched and reported for fixity providing collaborative management of object integrity between APTrust and the original owner. This value is further confirmed as part of reporting of content that may have been sent onto the DPN federated repository.

Presentation content

This presentation will express ways of linking materials through common identifier schemes across several tiers of repositories. Methods for maintaining proper provenance as content changes hands across a linked repository ecosystem and creating reporting frameworks back to content originators. It will describe methods for scaling repository services horizontally and experiences with bottlenecks in current common repository architectures. It will also describe how mixing localized work around into preservation metadata is counter-productive to long-term preservation goals and how the repository ecosystems exacerbate the problem.

Conclusion

As libraries move toward decoupled preservation systems that need to communicate, we need to develop better abstractions of our content that are repository system agnostic. Maintaining the identity of these items as they move across systems is critical for proper management, and

good local practices need to be developed for persistent identifiers that allow the explicit reference of not only digital objects across related repository services, but individual components of a digital object as well. We should consider the format of a digital object as much in need of migration as the binary filetype used for preservation. Frameworks for decoupling services are critical for developing a better preservation ecosystem. Reducing the artifacts of local application work-arounds are critical to increase the portability and utility of digital objects across systems.

References

BagIt Speicification v0.97-10, Network Working Group Internet Draft
APTrust BagIt Profile
APTrust Bag Receiving and Processing
APTrust Bag Ingest Scenarios (Linking Content for Replacement and Management)