HydraDAM2: Building Out Preservation at Scale in Hydra

Heidi Dowding
Indiana University
1320 E. 10th St
Bloomington, IN 47401
+1 (812) 856-5295
heidowdi@indiana.edu

Michael Muraszko
WGBH
1 Guest St
Boston, MA 02135
+1 (617) 300-2000
michael_muraszko@wgbh.org

ABSTRACT
HydraDAM2, a digital preservation repository project being developed through a partnership between Indiana University and WGBH, aims to leverage new developments in the Hydra/Fedora stack in order to provide better long-term management solutions for large audiovisual files.

Keywords
Hydra; Fedora; Repositories; Digital Preservation; Audiovisual

1. INTRODUCTION AND BACKGROUND

Indiana University and WGBH, a large academic research institution and a public media organization respectively, are currently both managing large audiovisual collections. While WGBH regularly manages multimedia as part of its daily production, IU's current developments in this area are based on the Media Digitization and Preservation Initiative (MDPI), which will result in 6.5TB of digital audiovisual content. [1] As both institutions have identified similar challenges in managing large- scale audiovisual content, Indiana University and WGBH have partnered to develop a repository aimed at long-term management and preservation.

This repository project, titled HydraDAM2, will build on WGBH's original NEH-funded Hydra Digital Asset Management system (HydraDAM) [2] as well as IU's IMLS-funded Avalon Media Systems. [3] The original HydraDAM is a digital file preservation repository built for the long-term storage of media collections. Like many Hydra applications, HydraDAM is a web- based, self-deposit system that allows for the search and discovery of the files and records stored in the repository. Storage for the HydraDAM repository is limited to the server or virtual machine on which the application is installed. Avalon Media Systems is a Hydra digital access repository aimed at discoverability and use of audiovisual materials.

HydraDAM2 will leverage recent improvements to the Fedora repository. The new digital preservation repository will allow for the storage of files either online via a Hierarchical Storage Management (HSM) system or offline via LTO data tape or hard drives. Having begun work in mid-2015, the HydraDAM2 team will complete a minimum viable product to be implemented within each institution by the fall of 2016. The ultimate goal of HydraDAM2 is to create an extensible product that can be reused within any Hydra institution.

2. IDENTIFIED GAPS IN HYDRA

One of the main limitations of the current Hydra / Fedora technology stack identified by the HydraDAM2 team is the inability to store large digital files within Fedora. This has been challenging with web-based repositories because there are often limits on size when ingesting files into Fedora from a web browser. Processing large files for things like fixity and characterization is also problematic, as it can be difficult to pinpoint the problem if any processes get held up or fail.

Another identified challenge in Hydra is the favoring of self-deposit systems where a majority of the metadata describing an object is generated during the ingest process. This is a problem for many institutions dealing with years of metadata records, sometimes from legacy digital asset management systems. In moving to a new Hydra self-deposit system, an institution could immediately have a significant backlog of files that would require re-description upon ingest. Hydra self-deposit repository systems are most successful for new projects, not for migration of legacy files and metadata.

3. HYDRADAM2 STACK

The HydraDAM2 system is based on the open source Hydra repository application framework and will utilize the emerging Fedora 4.0 digital repository architecture. There has also been a recent development in data modeling in Fedora. The Portland Common Data Model (PCDM) is a flexible, extensible domain model that is intended to underlie a wide array of repository and DAM applications. By implementing PCDM in HydraDAM2, we hope that using an emerging, standardized model for our data will allow for better understanding and interoperability with current and future Hydra open source solutions.

4. MAJOR FUNCTIONALITY

4.1 Management of Large Files

One of the main aims of the HydraDAM2 project is to reconcile the challenge of large files within the Hydra/Fedora environment by building out mechanisms for connection between local storage architectures and the HydraDAM2 repository. At Indiana University, this will likely integrate an API developed for asynchronous interactions with the institution's HSM storage backend utilizing Apache Camel routes as a means of integration. This scenario will allow for better management of terabyte-size audiovisual files within HydraDAM2, as the content will be safely deposited in IU's storage backend but manageable through HydraDAM2. The implementation at WGBH will be somewhat simpler in allowing HydraDAM2 to interact with their LTO tape storage backend.

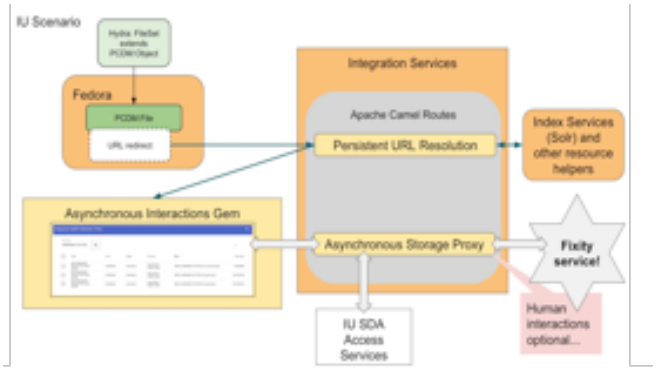

Figure 1. Indiana University Storage Interactions [4]

4.2 Reporting and Legacy Technical Metadata Management

Another goal of HydraDAM2 is to build out preservation functionality within Hydra and make it reusable. A majority of this functionality is focused on generating reports. Utilizing search functionality from the Blacklight piece of Hydra, HydraDAM2 expands capabilities in working with technical metadata for discoverability and management. This will result in the end user's ability to generate reports on things like file format, date created, and fixity events. HydraDAM2 will also include the ability for users to ingest previously created technical metadata so the system does not have to process files on ingest and generate them. As both institutions are managing collections with significant amounts of legacy metadata, this feature is crucial to scaling the repository solution.

4.3 Ongoing Curation

The final overarching goal of HydraDAM2 is to create an environment for ongoing management of digital files. Where Avalon will function as the access repository for all of IU's audiovisual content, HydraDAM2 will provide mechanisms for preservation and sustainability of content. While the first iteration of the repository focuses on basic preservation events like scheduled and triggered fixity checks, future iterations could include functionality like pathways for format migration. The main aim is to create a reusable Hydra repository with functionality for the necessary ongoing preservation functions related to audiovisual content.

5. CONCLUSION

As "an ecosystem of components" aimed at allowing individual institutions to more efficiently and effectively meet their repository needs, the Hydra project is constantly identifying gaps in infrastructures and workflows. As part of this, the HydraDAM2 digital preservation repository will fill in the gaps identified in the ongoing curation and management of large audiovisual files. By jointly developing this repository as a partnership between two very disparate institutions with two diverse storage backends, the end result will be a new set of functionality that can be utilized at a broad variety of institutions.

6. ACKNOWLEDGMENTS

This project has been developed based on a grant from the National Endowment for the Humanities. Our thanks to all staff involved in the project at Indiana University and WGBH.

7. REFERENCES

[1] https://mdpi.iu.edu/

[2] https://github.com/projecthydra-labs/hydradam

[3] http://www.avalonmediasystem.org/

[4] Floyd, R. (February 22, 2016). HydraDAM at IU. Presented at HydraDAM2 Partners Meeting: Bloomington, IN.