HydraDAM2: Extending Fedora 4 and Hydra for Media Preservation
Project Narrative

Project Significance

The Indiana University Libraries and WGBH Media Library and Archives propose to conduct a two-year project (January 1, 2015 – December 31, 2016) to extend the open source HydraDAM digital asset management system to operate on the emerging Fedora 4.0 digital repository architecture, using the Hydra repository application framework. The development of this system, tentatively known as HydraDAM2, will build on WGBH's prior NEH-supported work on the development of the original HydraDAM system and on Indiana University's IMLS-supported work on development of the Hydra/Fedora-based Avalon Media System. It will also build on both institutions' substantial experience in digital approaches to preservation of audio and video collections. The primary project goal is to develop a system that can serve as a digital preservation repository for time-based media collections at a wide range of institutions.

Radio and television broadcasting have documented in voices and images and influenced the shape of twentieth-century politics, society, culture, and economics. Broadcast recordings and related production materials are now primary sources for the history and culture of the twentieth and twenty-first centuries.

> "Television affects our lives from birth to death. Most Americans inform and entertain themselves through it… Sadly, we have not yet sought to preserve this powerful medium in anything like a serious or systematic manner. At present, chance determines what television programs survive. Future scholars will have to rely on incomplete evidence when they assess the achievements and failures of our culture." [1]

TV and radio have captured our local and national history for the past 60 years. The final broadcast programs and outtakes from public broadcasting stations hold valuable primary source materials for humanities research and study.

Recent public media discoveries range from audio recordings of the full day's events of the March on Washington in August 1963, early science education television programming produced in cooperation with MIT, outtake footage of President Barack Obama as a law student at Harvard, news coverage of the first ladies presidential debate in Iowa, and a program on the practice of "grave dowsing" in Missouri. However, the vast majority of this material is unknown and inaccessible to researchers, scholars and the public. As stated in the Corporation for Public Broadcasting (CPB)'s American Archive Project goals:

> "Massive numbers of Americans today look no further than their favorite search engines to find a huge selection of audio, music, articles and video. They have no idea what's missing – what isn't online. At the same time, hundreds of thousands of hours of public media content are currently inaccessible and undocumented in closets, on shelves, in storage boxes. Despite public broadcasting's mandate to "inform, inspire and educate," most stations' important and memorable recorded treasures, produced at significant cost, are never seen or heard again after their brief, shining moments on the air. This material is part of our American cultural heritage." [2]

In addition to broadcast materials, ethnographic audio and video recordings document cultural practices, musical styles, ceremonies, and rituals for use by scholars in anthropology, ethnomusicology, and other disciplines. Historical recordings of musical concerts, theater, dance, and other performing arts have the potential to provide access to little-performed works and the output of significant performers.

In this new paradigm, if an archive's moving image and audio content is not on the web or organized in digital form, it may as well not exist. Indeed, if it is not properly digitally preserved in the near future, it will not exist. As the current model of intensive, text-based research gives way to newer modes of cross-media, cross-platform interdisciplinary research and analysis, moving image archives must quickly find a way to make their digital materials accessible and preserved, or else face rapid obsolescence. Access is best created via the web, whether through digital preservation and online access to the content.

As these materials are exposed, scholars and educators have begun to use media materials in publications, presentations, and teaching. Scholars from as far away as Vienna, Austria, regularly visit the WGBH Archives to research program outtakes as primary source recordings of American history, and institutions, such as Columbia University, have started to integrate primary source media from series such as *Vietnam: A Television History* into their teaching environments.[3]

Time-based media collections are held by a wide range of cultural institutions, including libraries, archives, and public broadcasters, and are seeing increased use in the humanities across a wide range of disciplines for both research and teaching. Collections of born-digital materials are being created and acquired on an ongoing basis, and digitization of analog formats is increasing as institutions confront issues of degradation and media obsolescence. These institutions need flexible, reliable, manageable, and affordable solutions to store and preserve their digital audio and video files for the long term.

Open Source Approach

Media files are significantly bigger than most other digital files, and as a result, they need to be stored and handled differently. Supporting management, discovery, and preservation of large files and complex objects is difficult. Fedora 4,[4] the most recent iteration of the open source Fedora digital repository system, provides new functionality to help with the management and preservation of such large files. In addition, it provides new capabilities for storing and retrieving properties of digital objects as RDF (Resource Description Framework) data, which has not previously been taken advantage of in systems for managing media. Storage of metadata using the PBCore ontology could be used to allow greater relationships among items, helping the humanities with future discoverability of media items and allowing greater links with related items in repositories.

Commercial digital asset management systems are extremely expensive and are not focused on preservation and scholarly access needs; rather, they are largely oriented around supporting reuse of media content in a commercial production setting. In this vacuum, open source software has emerged as a flexible and affordable alternative for the management of access and preservation of large media files. Over the past fifteen years, academic and institutional repository communities have been pioneering open source architectures and services to support access to and preservation of documents and images, but most have yet to attempt media. The digital preservation community is certainly eager to deal with large media files, but work within that area is limited and is complicated by the complexity of managing and preserving these time based files.

Hydra[5] is an open source technical framework that lets institutions deploy robust and durable repositories supporting multiple "heads" for fully featured digital asset management applications and tailored workflows. The framework is based on Fedora, Solr, Ruby on Rails, and Blacklight. Hydra has a very robust open community that shares development efforts across multiple institutions. The motto of the project partners has been, "If you want to go fast, go alone. If you want to go far, go together." Hydra has been successfully utilized to develop digital library management systems at Penn State, Stanford University, Indiana University, Northwestern University, University of Virginia, WGBH and other major institutions worldwide. There are currently 24 committed Hydra partners.

Prior Work

This project will build on previous work in applying Fedora, PBCore and Hydra technologies to time-based media, specifically, the previous WGBH Open Source HydraDAM project and Indiana University's Avalon Media System.[6] The NEH-supported WGBH HydraDAM project implemented an open source digital media preservation and DAM system, focused primarily on the needs of public media stations but relevant and applicable to all cultural institutions with moving image and audio materials. It faced head-on the challenge of handling both large and small media files in several file formats. Based on the Hydra technology stack, including the Fedora 3 repository and the Blacklight[7] discovery interface, the project built upon the work already done by Penn State through their Sufia and ScholarSphere projects[8] and showed the value of the Hydra community in tackling these large problems of digital preservation systems.

The Avalon Media System has been a joint development effort of the Indiana University Libraries and Northwestern University Library, with support from an IMLS National Leadership Grant. The Avalon project is focused on developing a software system to enable libraries and archives, particularly in an academic setting, to more easily provide online access to audio and video collections. Like WGBH's HydraDAM, Avalon is based on the Fedora 3 repository and Hydra framework, but is focused on access rather than preservation. However, Indiana University also has a strong need for a repository system that can support preservation of media files as the university embarks on a five-year project to digitize the majority of its audio and video holdings.

A challenge of the open source community generally, and the Hydra community more specifically, is how to focus and build solutions specifically for one's institutional needs and yet still develop something generic enough for others to adopt. The Avalon project confronted this issue through the involvement of multiple partners providing feedback on functional requirements and technical features, but Avalon manages derivative files for access, not preservation master files. A university's need to pull preservation master files is limited and more infrequent than, for example, WGBH, whose productions pull the preservation files for reuse in new programs. WGBH's HydraDAM project involved implementation and testing partners from other public broadcasting organizations, but did not directly address the audiovisual preservation needs of other types of organizations. Yet we are both building systems to manage digital media files. We need to find common ground to develop code and content models that can be used in the basic functionality for preservation and yet still easily adopted and tweaked for other workflows.

Fedora 4

The open source Fedora digital repository system is widely used by libraries, archives, and other organizations to manage a variety of different types of content. The current HydraDAM system was developed on version 3 of Fedora, which has a number of limitations as far as being able to handle large files such as audio and video preservation masters. As a result, the current HydraDAM system stores the files outside Fedora, using Fedora mostly to store metadata about the files, along with information about the location of the external file. HydraDAM uses Fedora 3's messaging capabilities to support this workflow.

Since the original development of HydraDAM, the Fedora community has released a major new version of Fedora. This new version, known as Fedora 4, is based on the JBoss ModeShape[9] content repository system and is much more capable of storing and managing large files and connecting to various types of storage systems beyond traditional disk-based filesystems. Fedora 4 also incorporates more sophisticated access permissions or rights security, an important feature for preservation files. However, in its current Alpha release form, it has not yet been tested for such, or specifically for media files. In addition, no one has yet tested the advantages of Fedora 4 to fully utilize RDF for storage of metadata and relationship information for media files.

Fedora 4 features a number of significant improvements over Fedora 3 that would greatly benefit a media-based repository. Multimedia files, particularly preservation masters, can be gigabytes or even terabytes in size. Fedora 4, in contrast to Fedora 3, supports large files both in terms of its REST API and from a backend storage perspective. Uploading large files to the repository and downloading them via the REST API has been successfully tested with files up to 1 terabyte in size, which demonstrates that there are no memory-bound bottlenecks in these flows. Fedora 4 also supports the continuation of interrupted downloads by allowing byte offset requests to be passed via the REST API.

Large files are often stored in an external file system outside the repository, and in Fedora 3 these files are related to objects within the repository using what Fedora calls “external datastreams.” However, files referenced in this way cannot take full advantage of Fedora’s preservation features, including fixity checks and versioning.

Fedora 4 addresses this use case in a much more robust way using a new feature known as “federation.” This feature, also known as “projection,” allows the repository to connect to a variety of backend stores and treat the files as objects and datastreams within the repository. This allows Fedora to fully manage externally stored files, which includes operation of preservation functions such as fixity checks and creating and storing relationships between objects on federated storage systems and objects managed directly within Fedora.

One of the biggest performance bottlenecks in any repository is the point at which modifications get saved to the hard disk or other storage device. In Fedora 3, every action that modifies an object or datastream results in a “save” to the hard disk; for batch operations this can result in a significant number of saves, and thus a significant performance hit. Fedora 4.0 adds support for transactions, which provide the ability to effectively wrap a series of repository actions that can be committed or aborted as an atomic unit. Not only is this useful for managing inter-dependent actions, it has also proven beneficial in terms of general performance. Bundling multiple API calls into a single transaction shows a 30% to 60% performance improvement, depending on the type of events involved.[10]

Another promising new feature is clustering, which allows Fedora 4 administrators to configure two or more instances of Fedora to work together as a single system. By contrast, Fedora 3 supports only single-node configurations. Clustering has many potential benefits, including improvements to performance, scalability, and redundancy.

Metadata and RDF

More and more scholars are looking to media as primary source of recorded history for the 20th century. This content is complicated and needs to be adequately described to enhance its discoverability and context, and it needs to well described and organized for the web for easy access. One of the challenges with media materials is their relationship to each other and the many elements and pieces of content that are created to create a final product, and the many different copies or instances of the final product.

For example, the series *War and Peace in the Nuclear Age* was originally broadcast as a 13 part series. Yet as part of these final 13 programs, interviews were shot with world leaders, of which only small portions ended up in the final programs. Scholars would want to know that these interviews exist, and on their own merit are important, but also that they were created as part of the series. It gives them context, and it gives scholars primary source materials. In current cataloging standards it is very hard to do this elegantly and intuitively. PBCore is an emerging metadata schema being used by archives and institutions with large media collections. PBCore has been found to more adequately describe media collections and relationships among items than other standards that exist, and preservation of both content and context is required in order to support use by future scholars.

Work is currently underway to develop a PBCore ontology for use in RDF, and this matches well with Fedora 4's increased support for RDF and current work in the Hydra community to make greater use of RDF. The use of RDF in Fedora could greatly enhance exposing these relationships to discovery systems and to users. It could also help scholars find content that was otherwise unknown.

The implemented work of PBCore RDF data in Fedora 4 will help model the potential of relational data for media collections enhancing discoverability of these items. Although this project's main focus is the capabilities of Fedora 4 to manage media preservation needs, testing the full use of RDF streams in Fedora 4 will inform other preservation and systems considering a migration to Fedora 4 and demonstrate the advantages.

Workflows are also needed to move selected assets, including metadata, from a preservation repository (HydraDAM) to an access repository (Avalon, Open Vault) for public consumption. Creating a single repository with both a preservation and access application user interfaces would require substantial work to build multi-tiered security to assure preservation files are not exposed via the public access Hydra head. Such a security layer would require an effort much more significant than the workflows to publish an asset to a separate access repository.

Background of Applicant

Indiana University

Indiana University has been a pioneer in the field of digital libraries, in the development and application of open source software, and in the use of digital and Internet technologies for preservation of and access to audio and video media.

Indiana University's Variations project[11] was one of the world's first streaming media digital library systems, focused on music teaching and learning. The current version of Variations, developed by the IU Libraries with support from a Digital Libraries Initiative – Phase 2 grant from the National Science Foundation and National Endowment for the Humanities in 2000-2005[12] and a National Leadership Grant from the Institute of Museum and Library Services in 2005-2009,[13] was released as open source software under a BSD license in 2009 and is presently in production use at over a dozen colleges and universities beyond IU.

IU's experiences with Variations led in part to its current work on development of the Avalon Media System, intended to allow libraries and archives to more easily provide online access to audio and video collections. Avalon is being co-developed by the IU Libraries and Northwestern University Library using the approach of a single virtual development team split across two institutions, with development work managed via an Agile Scrum[14] methodology. We intend to use a similar approach to the development of HydraDAM2 on this project. Avalon is based on the Fedora 3 repository and Hydra framework.

Beyond Variations and Avalon, the IU Libraries have been involved in a wide variety of efforts both within and outside the university related to digital audio and video. The EVIA Digital Archive Project,[15] a joint effort of Indiana University and the University of Michigan, with support from the Andrew W. Mellon Foundation, has developed workflows for scholarly contribution, annotation, and editing of video, along with software tools for video segmentation, annotation, and searching. The IU Libraries were also a key partner in the Digital Audio Archives Project, an IMLS-supported effort led by Johns Hopkins University that worked on developing efficient workflows for preservation-level digitization of audio collections and established a preservation audio digitization lab within the Cook Music Library at IU. In addition, Sound Directions,[16] a series of NEH-supported projects in partnership with Harvard University worked to define and execute best practices in the use of digital technologies for audio preservation.

A major survey of media collections on the IU Bloomington campus in 2009[17] identified over 80 campus units with collections of over 560,000 audio and video recordings and film reels or cores, 41% of which are either unique or rare. Efforts over the past five years to develop a plan[18] for systematic digitization of these materials for preservation and improved access led to the announcement in October 2013 of the IU Media Digitization and Preservation Initiative (MDPI).[19] MDPI is a five-year comprehensive effort to digitally preserve unique and rare time-based media collections from across the university, supported by $15 million in funding from the IU Office of the President and university administration.

Digitization work for the IU MDPI is being carried out in partnership with Memnon Archiving Services of Brussels, Belgium, who will be setting up a digitization facility in Bloomington that is expected to produce as much as 12 terabytes per day in new digital content for archiving.

Output of this digitization will be stored in IU’s Scholarly Data Archive (SDA). SDA is a disk and tape-based hierarchical storage management system utilizing consortially-developed HPSS (High Performance Storage System) software, with storage mirrored between IU’s Bloomington and Indianapolis campuses, approximately 50 miles apart. IU intends to use HydraDAM2 as a preservation layer in managing this content.

IU’s experience with Fedora dates back to 2003, when the IU Libraries served as one of the initial implementation partners on the Fedora project led by the University of Virginia and Cornell University. IU is currently a financial sponsor of the Fedora 4 development project. IU is a member of the Association of Research Libraries, Digital Library Federation, Academic Preservation Trust, Digital Preservation Network, National Digital Stewardship Alliance, and together with the University of Michigan is co-host for the HathiTrust Digital Library.

Indiana University is a leader in the academic open source community, having served as a cofounder of the Sakai open source learning management system and Kuali community source administrative application suite. The IU Libraries are playing a lead role in the design and development of the Kuali Open Library Environment[20] (OLE) integrated library system.

WGBH

With a repository of over 50 years of public broadcasting history, both TV and radio, the WGBH Media Library and Archives is committed to developing a usable digital media preservation system based on open source solutions. As an educational foundation, creator and steward of a valuable collection of media resources, WGBH has embraced new developments in online media in its efforts to bring its archived materials to a broader audience and to serve the needs of the academic community. As a consequence, WGBH has already taken steps to identify what these needs are and to act upon them. We plan to build upon that expertise.

WGBH began developing its digital asset management (DAM) system in 2000 at a time when the open source community was fledgling and still unstable. WGBH, by necessity as a media creator, needed to resolve the issue of digital asset management. WGBH turned to a commercial vendor as a partner for a sustainable business solution. Working with Sun Microsystems and Artesia (now OpenText), WGBH developed a DAM architecture for media access and published reference architecture documentation for other media organizations to replicate the work.[21]

WGBH has been actively involved in on-going digital management and preservation projects in various communities. WGBH was a partner on the *Preserving Digital Public Television* NDIIPP project,[22] which set out to design a model repository for public television and focused on defining archival submission packages for ingest of large video files into a repository. Recently WGBH reinvigorated the PBCore community and plans to release a new schema next spring. WGBH has been an active participant in the Library of Congress' National Digital Stewardship Alliance working groups. And the WGBH Media Library and Archives (MLA) in collaboration with the Library of Congress, is the new home for the American Archive of Public Broadcasting[23] (AAPB), an initial collection of 40,000 hours of digital media from over 100 public media stations across the country. The AAPB is a project initially funded by the Corporation for Public Broadcasting to preserve and make accessible the many hours of content created by the American public media community.

WGBH has embraced new developments in online media in its efforts to bring its archived materials to a broader audience and to serve the needs of the community. WGBH is successfully creating value from archival content within various web portals like Open Vault, Stock Sales, Lab Sandbox, Teachers' Domain, and National Productions websites.

The digital preservation community has acknowledged WGBH as a media expert and has an interest in the knowledge base it has developed, from rights issues, to digital media file formats, to access interfaces. WGBH's Open Vault digital library project[24] has also fueled new relationships with the open source community and the digital library community, learning from their approaches to shared challenges and contributing solutions. Originally Open Vault was based on Blacklight, a discovery interface jointly developed by the University of Virginia, Stanford University, WGBH, and others, which focuses on providing access to materials through customized, lightweight interfaces on top of the Apache Solr indexing system. The Hydra project grew from this previous collaboration with Blacklight and is currently a collaboration of over 20 partner institutions, of which WGBH is a member. WGBH is active in the Hydra community and continues to engage with other Hydra partners. Inspired by possibilities of open source software for archivists, WGBH has been active in developing an Open Source Committee within the Association of Moving Image Archivists (AMIA), and are working to bring institutional repository and digital library communities together with AMIA to share ideas and solutions.

WGBH has already built a robust open source web access system for Open Vault using a Fedora repository. Recently Open Vault was moved to a fully functioning Hydra stack. The content model and source code have been shared through a Github account. The WGBH team has spoken to many organizations interested in building upon the work and has helped them as much as possible to understand what has been developed in order to improve upon it and replicate it. The WGBH team understands the needs of an active public media creator and the importance of following preservation standards. Building upon past work and experience, WGBH is uniquely qualified to continue work on an open source preservation system that will serve this community. WGBH offers deep knowledge of video and audio technology, management and web access. WGBH will focus this effort on the needs to manage digital media (audio and video files), an area of work lacking in both the academic and commercial systems.

The WGBH Media Library and Archive department (MLA) manages a collection of over 750,000 items dating back to the late 1940's. The MLA is committed to making its collection accessible for research and scholarly use and thus launched the Open Vault website for on-line access. It is currently actively implementing the current version of the WGBH HydraDAM system to capture born digital assets throughout WGBH, and to digitize analog formats as funds permit.

Project History, Scope, and Duration

Project History

As noted before, WGBH pioneered the development of an early DAM system for media in 2000 in partnership with a vendor and learned many invaluable lessons from the experience. The preservation DAM system is based on a proprietary system from the publishing and creative industries with limitations for metadata structure and interface. The vendor tended to develop the system toward what they saw as market trends and viable business sales. WGBH has found that although the system works, it is not flexible to the changing needs of the media industry, and the vendor is unable to tailor the software to our particular user needs without significant additional investment. In addition, upgrades are costly and time consuming, and all of the site-specific customizations built around the software need simultaneous upgrading by internal teams (e.g. extensive customizations to support media ingestions of large video files requiring limited technical knowledge). The customization links often break and need to be rewritten with every upgrade.

On the access side, based on earlier work supported by the Andrew W. Mellon Foundation, WGBH developed the Open Vault website to deliver digital media materials to scholars for research in ways that would address their needs. As part of this project, WGBH implemented a Fedora repository for access and utilized other open source tools to enhance use of archival moving image material on the web through features such as annotation, tagging, and citation. Through the use of Fedora and the dissemination of this project, WGBH became involved in open source communities and gained experience working with the Fedora repository architecture, Blacklight and other open source media management solutions. Having one of the few instantiations of a digital repository in active use as a media repository, WGBH staff are looked to as experts in managing, preserving and disseminating digital media.

WGBH has had great success with the Open Vault technologies allowing easy web access to archival materials. And basic preservation needs are more or less addressed by the vendor based DAM system. However, the integration and workflow of materials between the vendor-based DAM system and Open Vault is not smooth. The vendor-based system continues to be expensive and hard to customize at a time when resources are short, and needs are rapidly growing, and this led to WGBH's desire to build on its knowledge and expertise with managing web based media and open source repository technologies to apply those technologies to digital preservation and storage needs.

WGBH is in the process of completing a previous NEH Preservation and Access Research and Development grant to build an initial open source DAM system using Hydra technology. WGBH successfully created HydraDAM building off the work of the Sufia "gem," developed by Penn State as part of its ScholarSphere Hydra-based institutional repository application. HydraDAM has been demonstrated to successfully met the needs of ingest, cataloguing, discoverability, and storage of media files. HydraDAM can ingest any file format, create access proxies for 3-4 media file formats, and batch ingest files and metadata. It can also be indirectly linked to an HSM storage system to store large media preservation files.

The HydraDAM project built a replicable model with clear implementation documentation[25] so that other stations can use the system. It was tested at WNYC and SCETV, and WGBH is currently laying plans to implement this system for its DAM needs. Workflows for large media files preservation and storage still need to be customized depending on the storage being used at any institution.

The IU Libraries, over the course of the past ten years, have developed an extensive repository infrastructure for digital text and images based on Fedora 3, and more recently, Hydra. However, integration of large-scale cost effective storage for preservation has always been a challenge due to

Fedora 3's limitations in interfacing with multiple storage systems and with high-latency storage environments such as HSM systems. On the access side, the Avalon Media System project has demonstrated the value of using Fedora and Hydra technologies to provide online access to audio and video for scholarly, teaching, and learning use and the potential for using these technologies to deal with media in other ways.

Project Scope and Duration

We plan to accomplish the following goals over the two year period of the grant:

1. Extend the HydraDAM digital asset management system to operate on the Fedora 4 repository system

The existing HydraDAM system was developed using version 3 of the Fedora repository system, which is not optimized to handle large files. As a result, HydraDAM stores media files outside of Fedora, using Fedora mostly to store metadata about the files, along with information about the external file storage location. IU's existing in-house Fedora ingest processes use similar methods. In addition, Fedora 3 relies largely on XML datastreams for metadata storage, with minimal capability for dealing with RDF and linked data. The Fedora 4 development project, which began in 2013 and plans to release its first beta release in summer 2014, represents a complete re-architecting of Fedora to address many of the shortcomings of Fedora 3.

The project team will adapt HydraDAM to make use of Fedora 4 for storage of metadata and, optionally, content files themselves. This will include development of a Fedora 4 content model for audio and video preservation objects, which defines the content and metadata components of a preserved piece of media, as well as how those components are represented and related within the data model of Fedora. This data model take advantage of Fedora 4's enhanced RDF capabilities and will make use of RDF for representation of metadata when appropriate, to enable greater interoperability with other objects and metadata and increase its potential for sharing and reuse as linked data.

The project will build upon work currently underway within the Hydra and Fedora development communities to adapt the Hydra framework to Fedora 4 and to develop a standard means of using RDF with Hydra and Fedora 4.

The team will implement the new content model within the HydraDAM2 system, to allow ingest and dissemination of audio and video objects to and from the repository. The team will also explore the use of other new Fedora 4 features to enhance the functionality and scalability of HydraDAM2, including use of clustering and transactions.

2. Develop Fedora 4 content models for audio and video preservation objects, including descriptive, structural, and digital provenance metadata, based on current standards and best practices and utilizing new features in Fedora 4 for storage and indexing of RDF

One of the most critical steps in digital preservation is deciding what information and context about an object needs to be preserved, and then designing and implementing a consistent way of representing that content and context within a storage system or repository.

The project team plans to develop a Hydra/Fedora 4 content model for preserved digital audio and video that incorporates essential descriptive, technical, administrative, and digital provenance metadata components along with the content files. This work will be based on prior research and best practice documentation, including the Indiana/Harvard Sound Directions project and IASA TC-04[26] document for audio; practices implemented by institutions such as the Library of Congress, National Archives and Records Administration, Stanford, WGBH, and IU for video; and metadata standards including PBcore, AES audio technical and process history metadata standards,[27] NARA reVTMD,[28] and Library of Congress videoMD.[29]

Of particular note is the fact that the PBCore schema group plans to have an RDF-based ontology completed by Jan 2015, which could then be used tested with Fedora 4.

3. Implement support in HydraDAM for two different storage models, appropriate to different types of institutions:

- *direct management of media files stored on spinning disk or on tape in a hierarchical storage management (HSM) system; and*
- *indirect management and tracking of media files stored offline on LTO tapes*

Audio and video media files are much larger than most other files that libraries and archives are managing in digital repository systems, and the output of video preservation projects in particular frequently overwhelm the traditional disk-based filesystems used by most repositories. Some larger institutions have the technical capability to utilize hierarchical storage management (HSM) systems that typically integrate a small amount of expensive disk with much less costly automated tape storage, but many institutions do not have access to such resources or the technical or financial means to make use of them. Cloud-based storage options for preservation are starting to emerge, such as the Digital Preservation Network and Academic Preservation Trust, along with more bit storage-focused services such as DuraCloud and Amazon Glacier. However, these services are still very costly and often are designed or priced to serve more as backups for local preservation systems rather than as primary preservation storage, particularly in settings where preservation master files or larger-sized derivatives may need to be retrieved somewhat frequently.

As a result, we propose to develop HydraDAM2 to support two different approaches to storage. The first approach is direct Fedora management of digital content files stored on a filesystem or in a hierarchical storage management (HSM) system. Indiana University will develop a storage connector for Fedora to allow it to store and retrieve files from the consortially-developed HPSS (High Performance Storage System) HSM system,[30] which is used by IU to manage its Scholarly Data Archive system,[31] a 20 petabyte+ HPSS instance mirrored between its Bloomington and Indianapolis data centers and supported by IU's central IT organization with base funding as a university resource. In the direct management scenario, files can either be ingested into the repository through Hydra and Fedora or can be transferred directly to the underlying storage (filesystem or HSM) and ingested into the repository using Fedora 4's new federation capability. The second approach, which will be developed by WGBH, will implement support in HydraDAM2 for management of files stored offline on LTO (Linear Tape Open) tapes, with Fedora and Hydra serving as a metadata registry and tracking system with pointers to the identifiers and shelf locations of these LTO tapes.

4. Integrate HydraDAM into preservation workflows that feed access systems at IU (Avalon) and WGBH (OpenVault) and conduct testing of large files and high-throughput workflows

Preservation of media content is only valuable if it can be discovered and used by researchers, teachers, students, and other users. Both IU and WGBH have robust access systems, and HydraDAM2 will be built to be able to feed content and metadata into Avalon and OpenVault for discovery and delivery. IU and WGBH will implement HydraDAM2 as part of their preservation and access environments in order to support large scale testing.

5. Document and disseminate information about our implementation and experience to the library, archive, digital repository, and audiovisual preservation communities

One of the major goals of this project is to share both our code and experiences widely to encourage adoption by others and to allow others to learn from our experiences. Activities to support this goal are described in more detail in the Dissemination section below, but will include both traditional conference presentations and articles as well as a “hackathon” for developers at the beginning of year 2 to allow exchange of lessons and ideas at a code level with the larger repository developer community. The project will clearly document its activities, findings and installation instructions.

The more institutions that adopt these open source systems, the more support for development and evolution of the systems. There are numerous institutions looking for solutions in media preservation. With the strong partnership of Indiana University and WGBH, we will disseminate and promote the project to academic institutions, public media stations, libraries, and cultural institutions.

Methodology and standards

Software

Through use and research, both Indiana and WGBH have found that the Fedora repository architecture and structure is very well designed to fit the needs for the management of digital media. Fedora stores and manages data in native formats, rather than using application-specific data normalization and conversion, which makes it especially attractive for the development of enterprise-level repository ecosystems with large, heterogeneous datasets including various levels for metadata and relational information and media assets. In addition, it is flexible, extendable, and easily accommodates complex relationships between objects, which are all necessary for media management within the public media community. However, Fedora based repository solutions are difficult to implement from scratch, as while Fedora provides core functionality for storage, management, and retrieval of digital content and metadata, it does not itself have user interfaces, ingest, and access methods that are designed for end-users. Two technology ecosystems have emerged to ease development of Fedora applications: Hydra and Islandora, both of which have robust and growing communities of adopters, developers, and service providers. Indiana and WGBH are both members of the Hydra community and heavily involved in both the technical development of Hydra’s core components, based on the Ruby on Rails language and framework, as well as in helping to guide the community as a whole as it matures and brings on more partners and users.

Hydra makes use of Blacklight, a Ruby on Rails based discovery interface using the Apache Solr indexing system, that has a growing community of implementers and developers. Both Indiana and WGBH have successfully used Blacklight for several projects to provide search and discovery over large volumes of aggregated metadata records, including WGBH’s Open Vault digital library for public access, the American Archive Content Inventory, and the Boston Local Television News project, and Indiana’s IUCAT online library catalog, Digital Collections Search, and Avalon Media System.

Metadata Standards

PBCore will be utilized as a key component of the metadata structure. PBCore is an extended Dublin Core schema that was originally developed to help in the exchange of digital program files within the public media system. Public media stations found that other metadata standards don’t adequately describe time based media materials such as radio and television programs. It has quickly become a useful schema to describe digital media and is being adopted by numerous other institutions both within the public media system and among other cultural institutions with media collections, including the UCLA Film &

Television Archive, Illinois Public Media, the Alliance for Community Media and the Academy of Motion Picture Arts and Sciences.

As a key part of the American Archive of Public Broadcasting, PBCore has recently entered into a new phase of its development. Over 40 participants from institutions including public media organizations, media libraries and archives, public libraries, and university special collections are involved in various aspects of future development and dissemination of PBCore. Schema development plans include a reassessment of the current schema, refinement of controlled vocabularies, harmonization with EBUCore, and the creation of an RDF ontology. Additionally, participants are also working to improve the usability and navigability of the PBCore website, develop educational opportunities and resources for adopters and potential adopters, and enhance outreach to the community.

Standards for technical and digital provenance metadata will also be examined for potential use in the implementation of HydraDAM2's content model, including audio technical and process history metadata standards from the Audio Engineering Society.

Digital Preservation

HydraDAM2 is intended to be a preservation repository system that can be deployed as part of a trustworthy digital repository environment, per the emerging set of standards in this area, including OAIS (Open Archival Information System),[32] TRAC (Trusted Repository Audit and Certification),[33] and ISO 16363 (Audit and Certification of Trustworthy Digital Repositories).[34] Given that a software system on its own cannot constitute a trustworthy repository, the HydraDAM2 system will not itself be certified as a TDR, but we intend to pay close attention to these standards and evolving best practice in the digital preservation community as we design and implement the system.

Development Metholology

IU and WGBH will develop HydraDAM2 through a Scrum agile process using the three standard roles of Scrum Master (facilitates process), Product Owner (defines product features) and Scrum team (cross-functional team of developers and other experts). IU and Northwestern have used this process successfully in the development of Avalon, and similar processes are used in other cross-institutional projects in the Hydra and Fedora communities. The Scrum agile process has a value driven focus, and assumes that requirements will evolve over time; work is scheduled instead of estimated and feedback is integrated throughout the project instead of happening at the end. The IU and WGBH team members, though located in two separate locations, will function as a single Scrum team.

Work Plan

Year One (2015)

1. Hire Programmer/Analyst (Dunn, Cowan)
2. Develop high-level functional requirements document (Cowan, Muraszko, and team)
3. Conduct technical assessment of current HydraDAM code base (Cowan, Muraszko, Programmer/Analyst, Myers)
4. Conduct project team kick-off meeting in Boston to refine architecture and requirements (team)
5. Conduct two Advisory Board meetings via phone/Internet to obtain input on system requirements and design (Dunn, Cariani, team)
6. Develop high-level technical architecture and design document for HydraDAM2 (Cowan, Muraszko, Carter, Programmer/Analyst)

7. Develop Fedora 4 content models for audio and video preservation objects (Hardesty, Ma, Floyd, Programmer/Analyst, Myers, Carter)
8. Train development team in Agile Scrum methodology (Cowan)
9. Develop initial backlog of user stories (Cowan, Muraszko, team)
10. Carry out development of HydraDAM2 release 1 (Programmer/Analyst, Ma, Myers, Carter), including:
    a. Develop Fedora 4 connector for HPSS hierarchical storage management system (Programmer/Analyst)
    b. Develop support for offline tape storage of media (Carter)
11. Conduct initial test and evaluation of HydraDAM2 release 1 at IU and WGBH (Dunn, Cariani, Cowan, Wheeler, Floyd,Myers, Carter, Muraszko)
12. Participate in Hydra partner meetings (Cowan, Programmer/Analyst, Dunn, Cariani, Myers, Muraszko)
13. Promote the project via presentations at conferences (Dunn, Cariani, Programmer/Analyst, Muraszko)

Year Two (2016)

1. Continue to develop user story backlog (Cowan, Muraszko, team)
2. Carry out development of HydraDAM2 release 2 (Programmer/Analyst, Ma, Myers, Carter)
3. Host Hydra/Fedora Hackfest at IU or WGBH or in conjunction with another major library technology meeting (Cariani, team)
4. Conduct two Advisory Board meeting via phone/Internet to discuss sustainability options and future development (Dunn, Cariani, team)
5. Conduct testing at IU and WGBH (Dunn, Wheeler, Ma, Hardesty Carter, Myers, Muraszko)
6. Promote the project and HydraDAM2 software via presentations at conferences (Dunn, Cariani, Programmer/Analyst, Muraszko)
7. Participate in Hydra partner meetings (Cowan, Programmer/Analyst, Dunn, Cariani, Muraszko, Myers)
8. Conduct project team meeting in Bloomington to review final release and testing requirements and develop sustainability plan (team)
9. Document and release HydraDAM2 software (Cowan, Programmer/Analyst, Muraszko, Myers)
10. Write final project report and white paper (Dunn, Cariani, team)

Staff

Key Project Staff

Jon Dunn, Director of Library Technologies at Indiana University Bloomington, will serve as Principal Investigator and will oversee IU’s contributions to the project. He has served as Project Director or Project Manager for numerous grant projects funded by IMLS, NSF, NEH, and the Andrew W. Mellon Foundation, including the IMLS-funded *Variations on Video* (Avalon Media System) implementation and planning grants and *Variations3* project (both as Project Director), NSF/NEH-funded *Variations2* project (Project Manager), and Mellon-funded *Integrating Licensed Library Resources with Sakai* project (Project Director). He also served as Lead Technical Investigator for the Mellon-funded *Ethnomusicological Video for Instructional Analysis Digital Archive* project. Jon is a member of the Open Repositories conference steering committee and served as co-program chair for Open Repositories 2013. He is serving as the IU Libraries’ technical lead for the IU Media Digitization and Preservation Initiative.

Karen Cariani, Director of the WGBH Media Library and Archives, will co-lead the project and oversee WGBH's portion of the work. Karen has 20 years experience as Director of the WGBH Media Library. She has developed standards and procedures for production deliverables of original and stock footage tape logs and databases. She is currently overseeing implementation of the DAM system. Karen has been project director of a number of digital library projects at WGBH including Teachers' Domain, The Evolution Digital Library, Open Vault, the Vietnam collection, the Mellon Digital Library Project and PBCore development.

Will Cowan, Head of Software Development in the IU Libraries, will serve as Project Manager for IU, managing the software developers, coordinating other IU staff and librarian participation in the project, and serving as Scrum Master for the development process. Cowan served as development manager for the EVIA Digital Archive project.

Julie Hardesty, Metadata Analyst in the IU Libraries, will provide metadata expertise in the design of the Fedora 4 content model for the project. She currently provides metadata support for the Avalon Media System project and is involved in the design of preservation packages for IU's Media Digitization and Preservation Initiative.

Nianli Ma, **Randall Floyd**, and **Brian Wheeler** at IU will provide Fedora and Hydra expertise and consulting to the project, on issues of content model design and storage integration design and implementation.

Michael Muraszko, Digital Archive Manager at WGBH, will serve as the WGBH project manager managing the software developers, coordinating other WGBH staff and participation in the project. Muraszko recently managed various Open Vault projects and is overseeing WGBH DAM system development and migration.

Andrew Myers is Supervising Developer at WGBH Media Library and Archives. He has worked as a web developer across a broad range of domains since 2004. He was first introduced to the Hydra ecosystem in January of 2013. He has been working closely with Hydra components since then, and has become an active member in the development community, working to improve code and implementation practices for the benefit of all institutions interested in using Hydra.

Kevin Carter is a Programmer Analyst at WGBH. He has been instrumental in customizing DAM-related workflows for its staff, implementing an asset-level security schema within the current Artesia DAM system and migrating data between databases and applications, including OpenVault and HydraDAM. He also manages the interaction with WGBH's current HSM system.

The team will also consult closely with staff from the Research Storage group and Council of Enterprise Architects in IU's University Information Technology Services on the integration of Fedora 4 with the HPSS hierarchical storage management system and on issues of overall technical architecture, and with media archivists from the IU Libraries and other IU campus units on preservation content model design and implementation.

Project Advisors

We have identified a small unpaid project advisory board, whose members will advise the project directors and project team on issues of Hydra/Fedora 4 development and audio/video digital preservation needs. The advisory board members are:

- **Hannah Frost**, Services Manager, Digital Library Systems and Services, Stanford University Library
- **Adam Wead**, Analyst and Programmer, Information Technology Services, Penn State University (formerly Systems and Digital Collections Librarian, Rock and Roll Hall of Fame).
- **Andrew Woods**, Fedora Repository Technical Director, DuraSpace

Each of these advisors is actively working in the field of digital media preservation, digital asset management, and/or repository technical development. They are also actively involved in the Hydra or Fedora communities.

Indiana and WGBH will engage the advisors via four teleconference meetings over the course of the two year project, as well as via an email list, to review and provide feedback project plans, technical designs, system functionality, testing plans, and test results.

Sustainability of project deliverables and datasets

Project technical documents and reports will be deposited into Indiana University's institutional repository system, IUScholarWorks Repository,[35] which is operated by the IU Libraries in conjunction with IU University Information Technology Services to provide long-term preservation and access to research output of the university community.

The HydraDAM2 system will be released as open source software under an Apache 2.0 or equivalent non-viral open source license that allows for open reuse and future adaptation of the software. The source code will be developed in an open GitHub repository on github.com and will remain accessible beyond the end of the grant. Both IU and WGBH intend to put the HydraDAM2 software into production at their own institutions and to continue to maintain and develop the software to meet changing local needs. The use of existing open source technologies with strong communities around them such as Fedora and Hydra will help ensure the sustainability of HydraDAM2, and funding for participation in Hydra community meetings is included in the project budget.

Of course, the real key to sustainability is to create something so valued that people insist on sustaining it. Broad community engagement and adoption are the most important means of ensuring the ongoing availability and maintenance of any open source system. The dissemination efforts of the project, along with engagement of the project advisors and broader Fedora and Hydra communities, will be critical to achieving this goal.

Dissemination

The dissemination plans include distributing the report, software, and project findings as widely as possible. A website, wiki, and email list will be created for the project, and project documents, planning notes, and the final white paper will be posted on that website, and announcements of significant project milestones will be posted to the project email list and to relevant public lists targeting the repository, library, archive, and media communities. The project will report and promote the project at conferences that most target the audiences interested in the findings, such as: Open Repositories, International Conference on Preservation of Digital Objects (iPres), Preservation and Archiving Special Interest Group (PASIG), Joint Conference on Digital Libraries, Association of Moving Image Archivists, Digital Library Federation, the Library of Congress Digital Preservation conference, Hydra Connect, and Code4Lib. Placeholders for conference participation are included in the project budget. The team also plans to submit articles and publicize the project in venues such as D-Lib Magazine, Code4Lib Journal, AMIA newsletter, and NDIIPP newsletter.

As noted earlier, all software code will be developed in an open GitHub repository on github.com and will be released under an Apache 2.0 or comparable non-viral open source license, and all documentation will be made available under an appropriate Creative Commons license.