April 2009

Preserving Digital Public Television: Preparing for the Broadcast Afterlife

Nan Rubin

rubinn@thirteen.org

Preserving Digital Public Television: Preparing for the Broadcast Afterlife

by **Nan Rubin** (Project Director, Preserving Digital Public Television, 450 W. 33rd St., New York, NY 10001; Ph: 212-560-2925; Fax: 212-560-2833) <rubinn@thirteen.org> *http://www.ptvdigitalarchive.org http://thirteen.org*

"Public Television is responsible for the production, broadcast and dissemination of programs which form the richest audiovisual source of cultural history in the United States." — Librarian of Congress, 1997

New Preservation Practices for Television Archives

In less than a decade, television production, distribution and preservation has undergone a radical shift. Today, programs are nearly all shot, edited, and shared as digital files. Video recording and editing systems are now well within the means of most members of the public, and the ubiquity of media on the Internet, coupled with the mass deployment of hand-held devices, have transformed not only the medium of television but the entire environment for creating and watching moving images.

Distribution and transmission have been equally transformed, as tape-based submissions to the **Public Broadcasting Service** (**PBS**) and other national program services are being replaced by digital file transfers. On-demand viewing is growing just as on-air signals become all-digital, when every analog transmitter is turned off in 2009.

What do these changes mean for television archives? Practices to conserve and protect videotape recordings are well established, and the cost for maintaining and storing physical media are easily calculated. However, in an age of digital files, the requirements for preserving television programs are far different from storing videotape. It isn't enough to close a digital file and put it on a virtual shelf. For video in particular, acceptable practices to save and access very large files, manage ever-changing file formats, and maintain rich metadata are just now emerging.

Preserving Digital Public Television, a project funded by the **National Digital Information and Infrastructure Program of the Library of Congress** (**NDIIPP**)[1] set out to solve some of these difficult problems by designing a model repository for public television. In the process, the project also determined standards for metadata, explored rights issues relating to video archives, analyzed operating costs, and brought a new consciousness about the importance of digital preservation to the public television system.

Bringing Digital Preservation to Public Television

In the *Public Broadcasting Act* of 1967, **Congress** authorized the **Corporation for Public Broadcasting** (**CPB**) "to establish and maintain, or contribute to, a library and archives of noncommercial educational and cultural radio and television programs and related materials." However, **CPB** never allocated any funds to support this charge, and no demand for system-wide preservation was implemented. Consequently, only a few stations have established formal archiving activities to preserve their own materials.

Without a preservation mandate, digitally produced programs in public television are at great risk of being lost. The rapid changes in digital technology are rendering recording and playback systems obsolete at breakneck speeds, at the same time adequate tools for managing large and complex video files are not yet perfected. This has left a very large gap in the preservation of America's public television legacy.

Public television stations **WNET** in New York and **WGBH** in Boston, which produce roughly 60% of the national prime time series including *Frontline* and *NOVA* at **WGBH**, and *American Masters* and *Great Performances* at **WNET**, recognized this challenge early. Because **WNET** and **WGBH** each maintain its own archives, the stations were already committed to long-term program preservation. Both knew that solving the demands of digital preservation would be costly and that no station could do it alone — it would take a collaborative effort.

The Preserving Digital Public Television Collaboration

When the **Library of Congress** invited proposals under **NDIIPP**, **WNET** and **WGBH** partnered with **PBS** to build a model preservation repository for "born-digital" public television programs. **PBS** operates the network that distributes public television programs to more than 300 stations, and because most national programs pass through **PBS** before they are aired, it is the principle *de facto* repository for these programs. (The **PBS** warehouse holds more than 150,000 videotapes of programs going back more than 40 years).

These institutions understood that public television had to take steps to protect its rapidly growing collection of digital assets. As broadcasters, however, they had little experience building a preservation repository. **New York University** provided the expertise that was lacking. The **NYU Digital Library** team had extensive experience designing repository systems specifically for large digital files wrapped in metadata. The project further benefited from a relationship with **NYU's Moving Image Archiving and Preservation Masters Degree Program**, whose students have provided excellent research and whose graduates have become full-time project staff.

Together, **WNET**, **WGBH**, **PBS** and **NYU** organized **Preserving Digital Public Television** (**PDPTV**)[2] as a collaboration to introduce digital preservation issues and practices to the public television system. Understandably, the priorities of public broadcasting are program production and broadcast delivery, not saving program assets. Most program preservation is handled as an afterthought. To be successful, **PDPTV** had to demonstrate that *building* a repository was technically possible, and that *operating* a repository was functionally and economically feasible.

There were two major project goals:

1. Design a model preservation repository for large digital video files, and examine operating issues related to content selection, costs, and access.

2. Build system-wide support for digital preservation.

The project formally began in September 2004 and will be completed in 2009.

Building a Model Repository

The process for building the repository was initially conceived as a series of discrete technical tasks in a lab-type environment, with the approach that identifying commonly used file formats, determining appropriate metadata requirements, and adopting technical standards would be critical to repository functionality. The project naively assumed that commercial television networks and large collecting institutions such as the **Library of Congress** (completing its **Packard Campus** of the **National Audio-Visual Conservation Center**) were already making progress solving these same problems, and that public television could simply "tag along" with work underway.

The project quickly learned, however, that this was not the case. In reality, other video producers including the networks and the Library itself were struggling with these technical issues and not making much progress.

Instead of following along, the project found itself in the unanticipated position of actually leading the effort to create a set of standards for preservation-based video file wrappers for the television industry. Likewise, little had been done to determine what meta-

Figure 1: Repository Information Packages

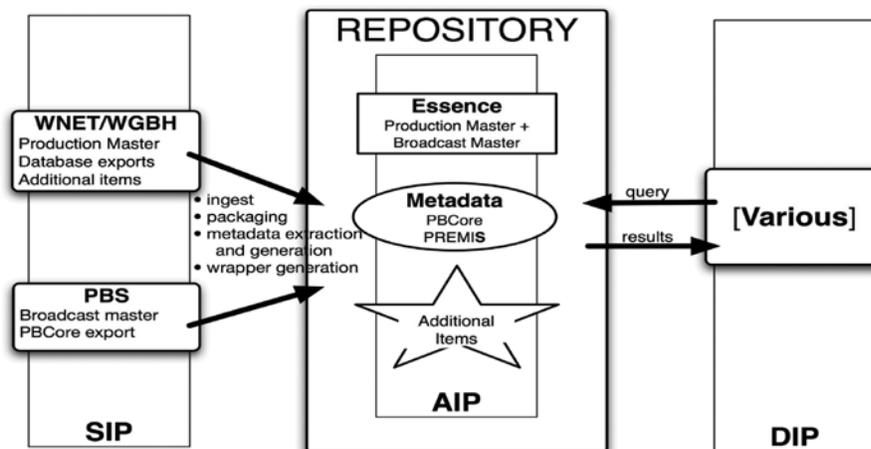*(Based on the OAIS Reference Model for Storing Files in a Digital Repository.)*

data should accompany the video files, and the project was one of the first to adopt a set of metadata schema appropriate for long-term video preservation. Both of these outcomes were unexpected.

Collecting and Analyzing Metadata

The **NYU Digital Library** team based their repository design on **DSpace**, which they had used to build other moving image archives. Technical issues rested primarily on how best to organize files and metadata to create **Submission Information Packages (SIPs)** and **Archival Information Packages (AIPs)** using test digital program files.

The team used a sample of 35 hours of program files, all standard definition, drawn from *Nature, Frontline* and *Religion and Ethics Newsweekly*, plus a local selection from *New York Voices*. The test files originated from three sources — uncompressed program masters from **WNET** and from **WGBH**, and compressed distribution versions of the same programs from **PBS**. This provided a mix of both high and low resolution program file formats, with different flavor files from each source.

A fundamental requirement was to configure **AIPs** for long term storage by aggregating content plus metadata for each program without adding anything new. The **SIPs** therefore, had to contain comprehensive program-related and technical metadata along with the program files themselves.

Operating from the assumption that the repository should conform to the **OAIS**[3] reference model for creating a trusted repository, the project examined a broad range of metadata schema used by libraries and archives. It also looked at standards emerging in commercial television, and assessed **PBCore**,[4] a metadata dictionary based on **Dublin Core**,[5] designed specifically for public radio and television program files.

In practice, determining the appropriate sets of metadata fields was an intensive task. Individual program files were accompanied by a wide range of metadata, but because program information is not collected systematically even within **PBS**, it had to be gathered from multiple sources on a program-by-program basis. Also, because there are no uniform criteria, the quality of metadata associated with each program was idiosyncratic and inconsistent. To determine the components required for the **AIP**, the collected metadata had to be analyzed, particularly the extensive descriptive and rights metadata created by **PBS** for broadcast scheduling.

Although **PBCore** is in the early stages of adoption, the repository chose to build its descriptive metadata requirements around it, which has encouraged others to use it as well. As a result, the most important source of metadata for national programming, **PBS's Program Offer Data Service** (**PODS**), can now be exported directly into **PBCore**, making national program information much easier to access.

Incorporating technical metadata from the video files also proved to be a challenge. Because the program files were submitted to the repository in many formats (including such diverse wrappers and encoding formats as **MXF**, **Quicktime**, and various flavors of **MPEG** and **DVC Pro**), multiple tools were required to play the videos and extract information such as bitrate, file size, and frame size. Transforming this disorganized metadata into a standardized **AIP** was clearly a necessity.

Valete, from the *Economist* print edition, July 3, 2008.
www.economist.com/world/unitedstates/ displaystory.cfm?story_id=11670747

Yearbooks are a great source of photographs. If there are no more yearbooks, how will reporters and others be able to know what people looked like in the good old days? **Lolly Gasaway** has a fascinating answer to a **photograph question**. See this issue, p.58.

And since we're on **Facebook**, be sure and read **Eleanor Cook's** column this time about **social networking**. I read somewhere (was it in **Eleanor's** column?) that **Facebook is** being taken over by the older set (people in their 40s) and that many librarians are on **Facebook**. Not me! Bah humbug! See this issue, p.75.

And **Tinker Massey** asks "has technology changed you?" Is this a trick question? Has technology made us more isolated (looking at the computer all day) and is social networking the balance? See this issue, p.56.

Moving on through technology, the glamorous **Xan Arch** (who has started a new column this issue!) gives us acquisitions types some ways to implement technology in our work – **Second Life**, avatars, PODS, etc. My question is – when is something a trend and when is it just a fad? How long does a fad have to last before it becomes a trend? See **Xan's** article, this issue, p.14.

Have you been reading the *ATG* **NewsChannel**? Well, if not, shame on you! I told you already about the article "Knowledge Overload," by **Ken Coates**, that was published in *Inside Higher Ed*, March 23, 2009. **Coates** who says that with the "deluge of information" that now confronts us (specifically students, scholars, librarians and the general public), we have by necessity become more focused and so we are "reading less than in the past." Faculty, in particular, have become more interested in scholarly productivity (read publication of articles) than in keeping up with the current literature that is available. This is a thoughtful article with comments from many of the people we know – **Sandy Thatcher** (Penn State Univ Press), **Toby Green** (OECD), **Sally Morris** (Morris Consulting), among others. And it has caused a flurry of comment on **Ann Okerson's Liblicense**. You need to read it!
www.insidehieredcom/views/2009/03/23/ coates
www.against-the-grain.com/

Along these lines, **Bob Holley's** article in this issue of *ATG* focuses on the need **for fewer but higher quality publications that would deal in greater depth with more substantive issues**. Hmmm … How likely is that? See this issue, p.64.

And speaking of *Inside Higher Education*, got a tip from **Mark Stengel** of **Cuesta College** <mark_stengel@cuesta.edu> that **Sandy Thatcher's** column "**The Hidden Digital Revolution in Scholarly Publishing: POD, SRDP, the 'Long Tail,' and Open Access**" in this issue of *ATG (v.21#2, p.60)* was mentioned (actually more than mentioned, it was quoted heavily) in *Inside Higher Education*, April 8, 2009. See – "Print or Byte?" by **Scott McLemee**.
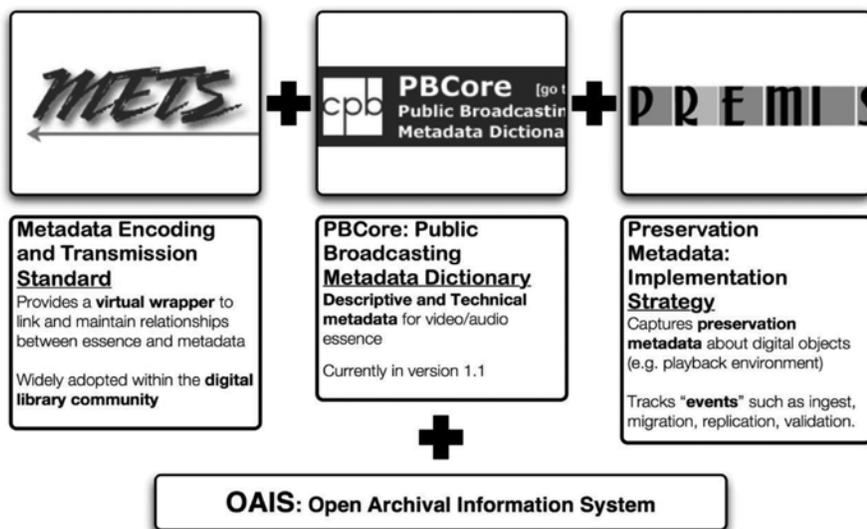*www.insidehighered.com/views/mclemee/ mclemee237*

The solution was to select schema from several data dictionary standards that encompass descriptive and technical metadata, while maintaining information unique to public television programming. This **AIP** structure uses not only **PBCore**, but also **PREMIS** (**Preservation Metadata Implementation Strategies**)[6] and **METSRights** (**Metadata Encoding and Transmission Standards Rights Declaration Extension Schema**). Appropriate fields from these standards, along with virtual links to the program files themselves, are all contained within a **METS** wrapper.

this issue, the project convened a "Wrapper Roundtable" of technologists, digital archivists and industry leaders.

The group was surprised to learn that the lack of consistent wrapper standards was also a major problem for the commercial networks. Any initiative to create technical standards for public television must dovetail with the needs of the commercial broadcasting industry, because public television on its own does not carry enough economic clout to influence hardware vendors. The wrapper standard remains to be solved, but due in part to the "Wrapper Roundtable," the **Advanced Media Workflow Association** which represents vendors, has made a commitment to develop an appropriate standard that vendors will support.

operating costs are being documented by such institutions as **The National Science Foundation**, which commissioned the **Blue Ribbon Task Force on Sustainable Digital Preservation and Access** (**BRTF-SDPA**) in 2007 specifically to study cost models for large database repositories. The contribution of **PDPTV** has been to focus on the particular concerns of maintaining very large digital video files with a manageable scale of operations. **PDPTV** is also closely monitoring the growing body of research being published on this topic.

Promoting System-wide Support

From the beginning, project partners promoted a position within public broadcasting that planning for digital preservation was no longer optional — it was a necessity. The explosion of online broadcast content, coupled with a constantly changing array of viewing devices, have created a fundamentally altered video environment which requires programming to be viewable on everything from the very smallest iPod screen to giant wall-size flat panels.

Amid such extremely fluid technology, the project emphasized the importance of adopting *standards* for technical operations, plus consistently collecting critical *metadata*. Because these are the very same factors necessary for successful multi-platform digital distribution, the project was able to tie digital preservation to effective reuse of program content. The concept of digital preservation thus became highly relevant to stations, elevating it from a marginal concern to a major subject in the public television debate on how to make content available to reach more viewers.

Figure 2: Metadata Schema

Thinking it was desirable to capture some types of metadata during the production process, the project analyzed the workflows used to produce the sample programs. The intent was to identify points where key metadata and file types were created, and track them through the program lifecycle. However, in a complex production process that involves many stages, the need to plan for file preservation had little bearing on production deadlines. Consequently, there was no opportunity to test preservation practices in the production workflows.

The problems encountered testing these various file formats, combined with the time-consuming efforts to collect metadata, demonstrated the high priority for setting uniform metadata and technical standards for any future operation. Without them, automating the functions for extracting and managing the metadata and file integrity of large collections is simply not feasible.

Video File Wrappers

The use of a standardized video file wrapper is considered a requirement for successfully exchanging digital files, particularly to support future file migration. A number of so-called video wrapper "standards" exist, but despite vendor claims, the files do not all actually interoperate with many equipment configurations used by public broadcasters. To examine

Assessing Rights and Operating Costs

Television programs are multi-layered productions comprised of original and licensed elements from a myriad of sources and subject to a wide array of contract obligations, rights restrictions, and other encumbrances. Typically, rights to use this material in a noncommercial broadcast are granted for a finite period, for example five to ten years. When these rights expire, the system can no longer use the program without "re-upping" the rights by getting new permissions from each individual rights holder. It can be very expensive to find all the rights holders, renegotiate and pay for new use agreements. Consequently, unless a program is in great demand, rights are rarely renewed.

Specific authorization to preserve public television programs after the broadcast window expires is largely absent, and making older programs available for anything but the most narrow uses is fraught with risk of copyright infringement. There are a number of efforts in the U.S. and Europe working to improve the situation, but under current conditions, the **PDPTV** model repository is "dark" to the public until viewing and use rights become more favorable.

As existing digital repositories mature,

Lessons Learned

Over the course of the project, a number of important lessons became evident.

- Technical problems will eventually be solved and standards will be adopted when private industry agrees to collaborate. But this is a slow and bumpy process.

- With producers beginning to use all-digital production workflows, now is the moment to introduce preservation compliant metadata requirements into the process. This should be done quickly or the opportunity might be lost.

- Prompted by the preservation message, stations around the country are actively exploring partnerships with other local cultural heritage institutions to share resources for preserving their respective digital collections.

- Despite a great deal of progress, a system-based commitment to preservation must be reinforced as an important national investment. Instead of being seen as overwhelming, costs need to be presented as feasible and manageable.

- Although there are some aggressive efforts to tackle the thicket of rights issues, especially for educational use, overall public television seems unwilling to push boundaries for wider access to archival content. Much more can be done in this area.

Preserving Digital Public Television ...
from page 20

Maintaining Momentum

Since **Preserving Digital Public Television** began, broadcasting has shed its analog systems and moved completely into a digital universe. This project has been able to impress on the public television system the message that digital preservation is not an optional "add-on" cost, but a requirement for any future use of the materials. In this, the project has been instrumental in transforming an attitude of indifference to one that acknowledges the value of properly managing our collective archival holdings.

In a further indication of support, for the very first time **CPB** allocated preservation funding to pilot **The American Archive**. **The American Archive** will develop a repository for public radio and television, and **PDPTV** anticipates making a significant contribution to this initiative.

Viewers keep reminding us that public television programming is precious and has made an indelible imprint. What remains is to continue building commitment across the entire system, so the critical responsibility for saving this American media legacy will be shared, sustained and nurtured over time. 🐾

Endnotes

1. *http://digitalpreservation.gov.*

2. *http://www.ptvdigitalarchive.org.*

3. "From **Wikipedia: An Open Archival Information System** (or **OAIS**) is an archive, consisting of an organization of people and systems, that has accepted the responsibility to preserve information and make it available for a Designated Community. The information being maintained has been deemed to need 'long term preservation,' even if the **OAIS** itself is not permanent."

4. *http://www.pbcore.org.*

5. From **Wikipedia**: The **Dublin Core** metadata element set is a standard for cross-domain information resource description. It provides a simple and standardized set of conventions for describing things online in ways that make them easier to find. **Dublin Core** is widely used to describe digital materials such as video, sound, image, text, and composite media like Web pages.

6. *http://www.oclc.org/research/projects/pmwg.*

Rumors
from page 18

Just got word that **The Informed Librarian Online** has selected an article from *Against the Grain* as **Editor's Picks**. Each month a few journal articles are highlighted for readers. The April 2009 issue of **The Informed Librarian** picked **Bryan Carson's** article in the December 2008/January 2009 issue (v.20#6,p.62) – "Is Open Source Software a

---

Nan Rubin

Project Director, Preserving Digital Public Television
450 W. 33rd Street, New York City, NY 10001
Phone: (212) 560-2925 • Fax: (212) 560-2833
<rubinn@thirteen.org> • *www.ptvdigitalarchive.org*
Thirteen/WNET.Org

BORN AND LIVED: 1949, Newton, MA. Have lived in Ohio, Colorado, Washington DC, New York City.

EARLY LIFE: Hippie, Folkie, Lefty, Techie.

PROFESSIONAL CAREER AND ACTIVITIES: 30+ years building community radio stations and community media with special focus on facilities, technical planning and creating infrastructure. Put two community radio stations on the air, in Cincinnati and Denver. Long-time supporter of ethnic public media, particularly Native American projects. A founder of the **National Federation of Community Broadcasters** and the **World Association of Community Radio Broadcasters**. Producer of **The Hidden Jews of New Mexico** radio series, one of the most the most popular programs ever aired on **NPR**. Organizer of the **Highlander Media Justice Gathering**, which helped launch the modern media reform movement. A primary team member involved with restoring the **WNET** broadcast signal after all analog and digital transmitters were destroyed at the **World Trade Center**. Last five years, have been Project Director of **Library of Congress NDIIPP** project **Preserving Digital Public Television**.

IN MY SPARE TIME: I make Jewish papercuts in non-traditional designs [*www. nanrubin.com*]. I produce segments for a weekly radio program on progressive Jewish politics and culture aired on **WBAI**, the **NYC Pacifica** station. [*http://www. beyondthepale.org*]. I dabble in handwriting analysis.

FAVORITE BOOKS: *A Distant Mirror*, **Barbara Tuchman**. *The Lymond Chronicles*, **Dorothy Dunnett**. *Mass Communications and the American Empire*, **Herbert Schiller**. *The Rabbi's Cat*, **Joann Sfar**.

PET PEEVES: Lima beans.

PHILOSOPHY: "You don't need to be Jewish to love Levy's." (real Jewish Rye bread...). "Any sufficiently advanced technology is indistinguishable from magic" – **Arthur C. Clarke**.

MOST MEMORABLE CAREER ACHIEVEMENT: Signing my first community radio station on the air in Cincinnati. Being invited on the **Martha Stewart Show** to demonstrate papercutting techniques. Being invited to make presentations on digital preservation to the **Blue Ribbon Task Force for Digital Sustainability**, and to the **National Library of Medicine**.

GOAL I HOPE TO ACHIEVE FIVE YEARS FROM NOW: Help community media win significant access to the digital spectrum. Help bust open the copyright stranglehold on access to archival video. Develop activities to train volunteers to take on distributed cataloging. Create a new **Foundation for the Preservation of Television**, alongside the existing **Foundation for Film Preservation** and **Foundation for Preservation of Recorded Sound** chartered by **Congress** ten years ago.

HOW/WHERE DO I SEE THE INDUSTRY IN FIVE YEARS: I'd like to see a better understanding of the importance of digital preservation, an ongoing commitment from the system to support preservation and access, improved open source tools to manage files, and more robust outlets for archival video. But given the very uncertain future of funding for public media, at this moment it's very hard to know what direction things will actually take. 🐾

Violation of Antitrust Law: Considering the Case of Wallace v. IBM." This article is fulltext on the *ATG* Website.

www.against-the-grain.com
www.infosourcespub.com/

Speaking of preserving, it is important to read all the articles in this issue of *ATG*, but I would like to especially point you to **James J. Jacobs'** article about government information.