Pushing Metadata Capture Upstream into the Content Production Process: Preliminary Studies of Public Television

Howard Besser *and* Kara Van Malssen
New York University, 665 Broadway, 6th floor, New York, NY 10012, howard@nyu.edu, kara.vanmalssen@nyu.edu

Abstract

This paper examines the issue of metadata lifecycle management, highlighting the need for conscious metadata creation to become part of the digital media production workflow in order to facilitate effective digital curation. Based on a study of public television workflow conducted in 2006, it reports on ways metadata might be captured from existing resources within the television production process, and ingested into a preservation repository. The authors identify areas in the workflow where small changes may yield significant improvements for preservation and curation of assets. This report also proposes methods by which producers and custodians of cultural content can work together to ensure the longevity of digital information, and outlines further studies that might turn these preliminary studies into standard practice. It encourages digital curators in all disciplines to begin looking back up the content production process for sources of rich metadata, rather than relying primarily on that created at the archival end.

Statement of Problem

Although the importance of metadata in digital curation is undisputed, the level of detail required has become immensely complex. Unlike analog resources, digital media need to tell us what they are, how to read them, how their multiple files unite to create a “work,” and what rights restrictions may inhibit certain uses. The need for stable and consistent metadata to facilitate preservation and dissemination of materials has become crucial. Descriptive, technical, administrative, structural, and preservation metadata must accompany a resource so that it can be identified, located, and retrieved by users. Metadata standards have been developed that allow archivists and curators to capture these crucial details, so that the works can be managed and used as appropriate.

Very little of the metadata needed for sustainability accompanies a digital object when it is sent to a repository; usually the repository must create much of this metadata itself. The more metadata that needs to be created, the more work has to be done by the digital repository or collection that is the recipient and custodian of these finished works. In order to make the metadata compliant with the repository’s requirements, many manual hours must be spent going back to the creator of the resource, doing research, and other sorts of exploration to find the appropriate information. This requires significant resources -- both time and money.

One of the major inhibitors to massive input into digital repositories is the cost factor in manual assignment of metadata needed for curatorial control. In 2006 the UK's LIFE Project examined the costs to the British Library of electronic publications voluntarily deposited there and found that the metadata creation costs for an electronic serial was £20 in the first year and projected to be almost £150 per serial in the 10th year. This was the largest cost center for preservation other than storage (The LIFE Project, 2006). If more of the metadata useful for long-term stewardship could be captured during the content production process, the repository ingest and management process could be streamlined, and more content could likely be preserved more effectively.

The archival notion of “lifecycle management” is absolutely critical in the digital world, and digital curators and preservationists from non-archival environments have begun to recognize the need for digital information to be managed from its point of creation. Digital curators are now looking at new ways to generate or obtain metadata that doesn’t rely on back tracking once the files have reached the repository. They are looking back up the production stream at ways that metadata may already be used internally by content producers, and how to systematize and push that information downstream in the workflow, so that it accompanies the resource once it arrives at its long-term home. In their major Mellon-sponsored study of electronic journal preservation issues, Yale University found that “the necessary or additional metadata cannot be effectively and satisfactorily produced either as an afterthought post-production process on the publisher's side or as a pre-ingestion conversion activity at the archive's end. Approaching e-archiving in this fashion leads to distribution delays and a more complex production and distribution scenario, with all the accompanying potential to introduce production delays and errors." (Yale University, 2002) They concluded that it was necessary to find systematic ways to generate more of the metadata needed for curation as part of the regular production process.

The preservation of digital media is a shared responsibility between content creators, distributors, curators, and preservationists. Each of these is becoming aware of the growing need for lifecycle management to ensure longevity. This study looks at ways that those in different roles can work together to capture the metadata required for the preservation, reuse, and curation of digital media.

This paper will report on work being conducted by the Preserving Digital Public Television project (http://www.ptvdigitalarchive.org/), funded by the National Digital Information Infrastructure Preservation Program (NDIIPP) of the Library of Congress. Charged with finding solutions to the problem of audiovisual preservation in a digital broadcast environment, the team consists of staff members at PBS, WNET in New York City, WGBH in Boston, and New York University. One of the important challenges this project faces is finding ways to capture metadata throughout the production process, so to avoid the need for backtracking, a method which simply will not be sufficient in a digital environment. This report is based on preliminary studies of workflow conducted at television stations WNET and WGBH in 2006. By examining the workflow involved in the creation, broadcast, and archiving of a television program, this study looks at points and methods by which important metadata might be extracted as it is created.

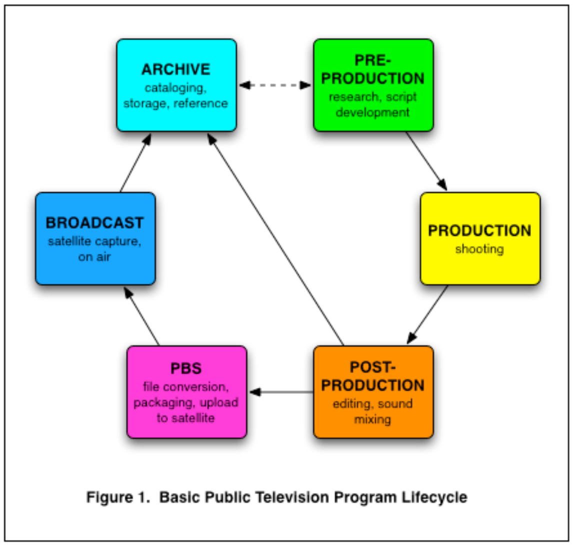

ARCHIVE
cataloging, storage, reference
PRE-PRODUCTION
research, script development
BROADCAST
satellite capture, on air
PRODUCTION
shooting
PBS
file conversion, packaging, upload to satellite
POST-PRODUCTION
editing, sound mixing

Figure 1. Basic Public Television Program Lifecycle

Public Television Workflow

When we step back and look at the broad public television program workflow (see Figure 1 above), the basic lifecycle doesn't look very different from that of many other cultural resources. The social science researcher follows a similar path, and data travels through similar stages. This may also look familiar to those acquainted with workflow involved in the creation of a magazine or e-journal. A closer look, however, will reveal public television production's unique processes.

For a typical public television documentary program, the process begins with a research phase when the script is developed, budgets projected, and initial contracts signed. The team then moves into pre-production, during which funding is sought, materials purchased, and the script is further developed with preliminary interviews recorded. During the production phase, original footage is shot, stock footage and audio clips gathered, and the script is finalized. In the post-production phase, the footage is edited, audio mixed, and credits created. The different versions are also produced: broadcast (in-house version, international, clean, etc), special resolution (16:9, 4:3, Anamorphic), and screeners. The script team creates the final transcript. The production team then conducts promotional activities, creating press kits and publicity photos. The program's website is fashioned to include attention grabbing materials such as "making of" segments, participant interviews on video and/or audio, biographical information, and downloads. Finally, the program is re-used and repackaged as needed.

Currently, the workflow between the producing stations, PBS, and broadcast is a real-time, hybrid digital/analog process. The final version of a finished program is mailed to PBS on videotape. There it is ingested and encoded, and the accompanying metadata about a program (submitted by the producer) is input into a database. Both the program and the metadata go through a number of conversions, before PBS sends the newly packaged program up to a satellite, which a local station's master control then captures in real time. The stations record the satellite feed onto tape, which is stored in the local master control area for later airing, and until the rights period for that show has expired.

In the current tape-based workflow, the production materials are turned over to the producing station's archive either after the production unit has no further immediate need for the materials, or after the production runs out of storage space as the next project begins. The producers will typically give the archive various versions of the finished program, original unedited footage and interviews, stock footage, transcripts, and stills. The archive also receives the broadcast master tapes, after rights have expired.

Typically, the most significant metadata gathering process starts either at this point, or even later, once the archivist has time to begin cataloging the program. The archives prioritize cataloging master programs (over non-masters, production elements, or promotional material), and rely on a variety of sources to populate their in-house database. These typically include the tape content itself, tape labels, and the program's website. All underlying program elements delivered to the archive (such as unedited footage and stills) are cataloged only at the box level.

Public television in the US is moving toward a file-based workflow. Already, most of the post-production process is digital, the content moving back to analog form only at the distribution stage. PBS is now installing the Next Generation Interconnection System (NGIS), which will allow for real time and non-real time transfer of digital program files between stations and PBS. Additionally, the largest producing stations (WNET in New York City and WGBH in Boston) are implementing digital asset management systems, which will manage in-house program files and accompanying metadata.

As the workflow becomes file based, the need for robust and accurate metadata will become critical. One program will likely consist of an array of files, in different formats. File relationships, video codecs used and bit-rates must be explicitly noted. Rights information will need to be accurate and immediately accessible. This will require a much deeper level of metadata than is currently captured in tape-based archives. Obtaining this very rich set of necessary metadata at the end of the production and broadcast lifecycle is simply not feasible. Thus, the metadata required will need to be systematically gathered during the production cycle and should accompany the programs as they are submitted to the preservation repository.

The introduction of NGIS as well as custom digital asset management (DAM) systems at WGBH and WNET are indications that public television is moving away from a tape-based environment. Public television is also working closely with the broadcast industry on the development of two wrapper formats for post-production, distribution, and preservation of born digital television material. This is an important moment in the broadcast world, which has traditionally approached preservation as an afterthought to the more lucrative arena of distribution. These open source formats will wrap metadata, including edit decision lists (EDLs), important for future reuse of digital video files, with the program essence as it travels through the post-production path. Once production is finished, the packaged file will be converted to a storage format, complete with all related production files, synchronization information, and the necessary video and audio codecs. These efforts are a significant step toward a "universal preservation format"[1] for audiovisual material.

Possible Points of Capture

Metadata generated at the time of creation has the invaluable advantage of being fresh and more accurate than that reconstituted at the end of the line in the archives. "Those in the production unit are the creators and have first hand knowledge of who, what, where, when, and why the content was created," note Mary Ide and Leah Weisse, Archivists at WGBH (Ide and Weisse, 2006). Before promotional activities, other projects, and general distance from the program erodes the production team's memory and their local personal notes disappear, it is critical to capture the (often informal or ephemeral) records they generate during the production workflow.

Already technical metadata tends to be automatically embedded in the header of digital files, which can easily be extracted and mapped to a repository's internal database. This can eliminate the need for one level of metadata creation. Elucidating descriptive, administrative, structural, and preservation metadata, however, is not yet a simplified or systematic process.

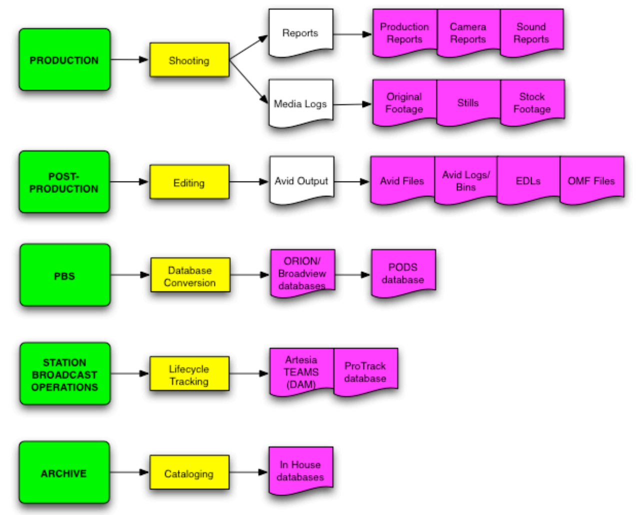
[1] The "Universal Preservation Format" (UPF) concept was developed in the late 1990s at WGBH by archivists and technologists who were concerned with the growing issue of digital media preservation (see http://info.wgbh.org/upf/). The format was conceived as an open source wrapper that incorporates metadata, including all technical specifications required to build applications to read the contained materials. The Preserving Public Television project is exploring ways to incorporate UPF features into the formats currently used in most digital video production: the Advanced Authoring Format (AAF) and Material eXchange Format (MXF).

Currently, much of the metadata created during the production process has only a short lifespan, as detailed local records are prepared, utilized and then discarded. Media logs, for example, which provide clip-level details, including the clip name, a description, and the type of element (interview, archival footage, etc.), are only passed from the producer to the editor. This type of dataset could prove invaluable down the line as digital curators look to find detailed descriptive elements for archives public television programs.

Figure 2 below highlights points in the current production workflow where rich datasets might be captured. The diagram draws out just a few activities of each unit within the process. These particular actions consistently yield reports, spreadsheets, files, or rely on databases. Each document represented in the diagram could potentially be carried along the production process with the program itself, or extracted and packaged with the final asset for ingest in a preservation repository.

Between March and September 2007, the Preserving Public Television team at NYU, WNET, WGBH, and PBS will begin gathering test programs and related metadata for ingest in the model repository at NYU. Each of the examples in this diagram will be tested, along with other relevant production records. As the initial tests are conducted, the value of each dataset will be analyzed, and will allow us to define an ideal Submission Information Package (SIP) for the preservation repository.

Figure 2. Possible Points of Metadata Capture within Current Public Television Workflow



Creating Partnerships

Convincing creators of digital cultural resources to change their workflow to accommodate the needs of archivists is difficult for any creator group, whether it be producers of e-journals, e-prints, or web-based art. Content creators are often pressed for time, money, and materials. They focus their attention on balancing a quality product with pressing delivery requirements. In the broadcast industry, many creators are independent producers, freelancers, or third parties, and each has traditionally performed tasks in his or her own way. The challenge is to implement a standardized process that won't disrupt the current workflow too much.

In order for the transition to digital public television to take place, the various groups responsible for the creation, distribution, and preservation of content must work together. It is in everyone's interest to see that public television content is preserved, and that includes the production team. As Ide and Weisse have found, "They [Producers] do care that the content is saved, whether it is because they understand it could be reused or because they want their work saved for posterity." (Ide and Weisse, 2006)

Because other administrative imperatives are forcing a redesign of all production workflow, now is the perfect time for some of the necessary changes to begin. With the implementation the Artesia TEAMS digital asset management system into the post-production process, and file-based delivery to and from PBS, production units will need to re-design their workflow to become compliant with station-wide procedures for asset tracking and distribution. As these changes take place, the requirements for metadata can also be put into practice. Adjusting to the increase in metadata capture will be relatively simple in respect to the overall re-structuring of workflow that must take place, and producers will likely not differentiate between workflow changes driven by implementation of the asset management system and those driven by the need to capture more curation metadata. Additionally, production units will likely view the implementation of the asset management system positively, as it will improve their ability to find the information and assets they need during the production process. Producer resistance to workflow redesign is likely to be less because the asset management system will provide immediate concrete benefits to their daily work.

WGBH currently has a system in place that requires production units to work closely with the Archive and Media Library on delivery requirements. When a production unit contracts with the station to create a show for broadcast, it also agrees to deliver supporting production documents and media to the station's Archive before receiving final payment. At the point a WGBH program is green-lighted, the Archive's Media Library is notified and initiates contact with the unit. The Library sends a start-up letter, File Maker Pro database templates and a *Production Handbook*. The *Handbook* includes information about the archival and library resources available to producers and instructions on how to track information and organize media and information for eventual delivery to the Archive.

The production unit is asked to designate one crewmember to be responsible for eventual delivery of the program elements, and the supplied databases. Once a program shuts down, the production team delivers completed databases for original footage, still images, stock footage, and materials used in the final edit.

Because this system was designed for analog content, the metadata collected is not extensive enough for digital curation. But two important lessons can be learned from this prior experience. (1) Gathering metadata on one system and trying to use it for curatorial maintenance on another introduces both conversion and quality control problems. Even though both current systems are Filemaker Pro, they are different versions and re-formatting and other conversion is sometimes required. In addition, the inability of anyone outside the individual production to see the metadata before it is handed over to the library affects quality control. (2) Extensive quality control is necessary both because most metadata is gathered at the end of the production (rather than early on in the production) and because producers are not specialists in metadata creation. As the Archivists at WBGH have noted, "[Producers] are not trained in metadata standards and data entry nor do they have the time to do the data entry or fully understand why others just can't get the information from their labels." (Ide and Weisse, 2006) So, in order to be effective, the movement of metadata gathering into the production stage needs to be accompanied by training of production staff, and by convincing them of why it is important for others to be able to understand the labeling that they put on their material. This is a challenge that most asset management implementations face.

Another potential way to reinforce the importance of accurate and consistent metadata entry would be a mandate from the Corporation for Public Broadcasting (CBP) and other funders that preservation planning must be part of the initial proposal. This has already been seen in the sciences, where national funding agencies now require that long-term data sharing and preservation planning be part of research proposals (United States Department of Health and Human Services, 2003 and United States National Science Foundation, 2001). This may become a reality in public broadcasting as the industry develops its plan for an "American Archive," of public television and radio content (Behrens and Egner, 2007). The idea was proposed to Congress in February 2007, and already has the unwavering support of the Association of Public Television Stations (APTS), the lobbying arm of the public broadcasting system. If it does pass, the CBP will likely require that everyone involved in the program lifecycle participate in an effort to ensure that the content will be preserved for future use.

Finally, creating clear financial incentive for producers will certainly influence the amount of effort they put into entering metadata. Currently, producers do not take full advantage of the potentially money saving opportunities they have to use the station's own content at no charge. Producers and production assistants have said that the time it takes to search through the archive, find potential content, then screen to see if it is actually useful, will be much longer than the time it takes to contact a stock footage house and just purchase footage, or simply shoot new footage. Though re-use can save money, under current conditions it is seen as a waste of the production team's precious time. The advantages of current tape-based archival environment are far outweighed by the convenience of the online access options offered by large stock footage houses such as Getty Images and Corbis. However, once the digital asset management system is in place, new and legacy content will be searchable from the producer's desk. In the future, a fully functional preservation repository will likely be user friendly and convenient, which will encourage re-use rather than the current practice of creation of original materials, or purchase of stock footage.

Further Studies

As noted in the introduction, this paper is based on preliminary research. Thus far, the Preserving Digital Public Television research team has conducted studies on general workflow at WGBH and WNET. The observations and suggestions in this report are the result of the workflow study, along with the input of archivists, broadcast operators, producers, and IT staff at the two stations. However much more work needs to be done to test the effectiveness of the suggested methods for metadata capture. One of the more important tests will involve capturing all existing metadata documents and databases during the sample ingest to the model repository at NYU. These tests are currently underway and will be reported on later in 2007.

Additional studies will also need to be conducted as initial ingest test results are returned, and as the workflow transitions to a digital environment. One of these is a more detailed study of the metadata throughout the program lifecycle, the mapping of databases to one another, and the value of captured metadata to end users. This future study will look at which organization (and which unit within that organization) should enter what type of metadata, and at what point. The findings will hopefully help focus on how to get reliable metadata from each unit through the lifecycle, by showing, for example, that certain types of descriptive and some technical metadata will come from the production units, technical and structural from the distribution (PBS or broadcast operations) arm, and administrative and preservation metadata will come from the archives.

Finally, there are questions of accuracy that concern digital curators in any field. How can we be sure the metadata is reliable? What methods are there to check the accuracy of metadata capture? How many hours should be spent on quality control without completely defeating the purpose of pushing the metadata upstream? Is metadata from one station more reliable than that of another? Is the metadata that comes from PBS more accurate than what a producer entered for the same program?

Conclusions

Developing systems to capture more metadata during the production phase of a digital work can ease the burden of ingest and long-term management. But inserting metadata capture into the production stream requires close cooperation between producers and curators. For the metadata to be useful, curators need to train producers in creating metadata according to standards, and producers need to see the utility of producing metadata that can be understood by others. Because content producers are frequently driven by tight production deadlines, they must be able to see concrete benefits from any changes made to their normal workflow. Curator analysis of production workflows can locate the points of metadata capture that will yield the best metadata yet might be least disruptive to the production process. In the next phase of this project, we will see whether incentives were sufficient to convince line-level producers to create metadata useful for curation, and whether the metadata collection points derived from curator analysis of production workflows resulted in clean and useful metadata.

This desire to push the metadata capture process further upstream is not limited to the television industry. Many other fields are working on this problem. Social science data repositories, for example, are looking at ways to work with researchers to capture metadata throughout the research lifecycle. Their aim is to better serve users and guarantee that the goals of reuse, dissemination, and long-term management can be achieved. In a digital environment, this is nearly impossible without established partnerships between the researcher/creator, his/her host institution, and the domain repository.

As Sarah Higgins of the UK Digital Curation Center states at the beginning of her work on metadata standards, “Metadata is the backbone of digital curation.” (Higgens, 2007) Without robust, accurate, and consistent metadata, digital curators cannot effectively do their job. Yet today’s digital curators shoulder most of the weight in both verifying the sparse metadata that comes from content producers, and in creating new metadata needed for digital curation. Finding ways to encourage content producers to generate and systematically record more of the metadata needed for long-term retention can lessen the burden on digital curators and lead to richer and more robust digital repositories.

References

Behrens, S. and Egner, J. (2007) APTS Preps Proposals for ‘American Archive,’ Copyright Legislation. Interview with APTS President John Lawson. Washington, DC: Current Publishing Committee. Accessed 14 March 2007 at http://www.current.org/federal/fed0702apts.shtml

Higgins, S. (2007). Using Metadata Standards. Digital Curation Center. Accessed 14 March 2007 at http://www.dcc.ac.uk/resource/standards-watch/using-metadata-standards/

Ide, M. and Weisse, L. (2006). Recommended Metadata Guidelines for Describing Born-Digital Master Programs for Preservation and Deposit with the Library of Congress. Preserving Digital Public Television, an NDIIPP Project.

The LIFE Project (2006). Lifecycle Information for E-Literature. A summary from the LIFE project Report Produced for the LIFE conference 20 April 2006. Accessed 14 March 2007 at http://eprints.ucl.ac.uk/archive/00001855/01/LifeProjSummary.pdf

United States Department of Health and Human Services (2003). NIH Data Sharing Policy and Implementation Guidance. National Institutes of Health Office of Extramural Research. Accessed 14 March 2007 at http://grants2.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm

United States National Science Foundation (2001). NSF's Grant General Conditions (GC-1). Accessed 14 March 2007 at http://www.nsf.gov/pubs/2001/gc101/gc101rev1.pdf

Yale University (2002). YEA: The Yale Electronic Archive, One Year of Progress. New Haven, CT: Yale University Library and Elsevier Science. Accessed 14 March 2007 at http://www.diglib.org/preserve/yalefinal.html