PRESERVING DIGITAL PUBLIC TELEVISION

RECOMMENDED METADATA GUIDELINES FOR DESCRIBING BORN-DIGITAL MASTER PROGRAMS FOR PRESERVATION AND DEPOSIT WITH THE LIBRARY OF CONGRESS AND OTHER DIGITAL REPOSITORIES

REPORT PREPARED BY
LEAH WEISSE AND MARY IDE, WGBH ARCHIVES
WITH KARA VAN MALSSEN, RESEARCH ASSISTANT,
MOVING IMAGE ARCHIVE & PRESERVATION PROGRAM, NEW YORK UNIVERSITY

Audience: This report is an overview of descriptive and preservation metadata necessary for the long-term access of digital television program assets. For illustration purposes, there are references to WGBH examples of WGBH in-house metadata records mapped to national metadata standards.

SUMMARY

In recent years, public broadcasting stations have found increasing value in retaining and efficiently managing their own master programs and associated production elements. There are a variety of reasons for this, including:

- New platforms for the re-broadcast of master programs;
- The potential re-use of program content for new productions;
- The use of content for websites or other educational materials;
- Research by journalists, scholars, educators, and historians;
- The sale of original footage content and a growing awareness of programs and related production elements as historical resources for scholarly and educational research.

Public television stations have tended to establish their own policies and procedures for collecting and managing production assets in analog formats. With the growth of digital editing capabilities, the advent of digital file transmission for over-the-air broadcasts and new cable broadcast opportunities, many stations have recognized the need to better organize and standardize methods for managing their digital asset collections as well.

This report will focus on the use of metadata for managing and preserving digital program assets. Metadata is considered a strategic component in structuring a station's digital asset management system. It enhances access to digital assets, and is essential to the transmission of digital program files to PBS for national distribution. It is also critical to the ability to access digital program assets over time.

This report will:

1. Discuss the development of metadata as it is utilized today;

2. Identify the different kinds of metadata;
3. Describe metadata schemas most appropriate for public television preservation;
4. Identify the points along the production process (from start up to shut down) where metadata is, or could be, generated and/or applied to digital assets;
5. Identify related points beyond shutdown where additional metadata is, or should be, generated before and at the time of transmission to PBS to stations for broadcast air;
6. Anticipated metadata required for file transmission to the Library of Congress or other digital repositories.

A **digital asset** is defined as *a computer or digital file that contains content, which includes "essence" (e.g., stills, text, audio, and/or video) plus "metadata", or data about the essence*. The digital asset also contains information about the digital/technical format (e.g., encoding format and decoding instructions).

Assets represent a production or station's investment for "depositor and an information resource for the researcher." (California Digital Library Glossary *http://www.cdlib.org/inside/diglib/glossary/?field=institution&query=CDL&action=search#D*)

We will provide specific examples of metadata schemas that we believe most appropriate for adoption by those managing public television digital collections. The metadata schemas profiled are:

- ***PBCore*** (Public Broadcasting Metadata Dictionary), which was closely based on Dublin Core;
- ***PREMIS*** *(*Preservation Metadata: Implementation Strategies), which is maintained in the Network Development and MARC Standards Office of the Library of Congress. Sample mappings to WGBH master program records are provided in the attachments.

We also include information about ***OAIS*** (*Open Archival Information System)* and ***METS*** *(Metadata Encoding & Transmission Standard).*

- The section on **OAIS** outlines how an archive prepares digital objects so they can be available for designated communities.
- **METS** is an open standard for encoding descriptive, administrative, and structural metadata regarding objects within a digital library that is necessary for both managing digital objects within a repository and exchanging objects between repositories. METS uses the *XML (eXtensible Markup Language***)** schema of the web.

These initiatives and the PREMIS schema are relevant in that they provide the structures and information for public television assets to be delivered from PBS to the Library of Congress or any other long-term repository.

Finally, the book, *Descriptive Metadata for Television* by Mike Cox, Linda Tadic and Ellen Mulder gives an excellent and through analysis of the subject of metadata and is highly recommended.

DIGITAL ASSETS: SETTING THE METADATA CONTEXT

While the Library of Congress (LC) began acquiring motion picture film in 1893, it was not until 1942 that LC established formal guidelines for acquiring, cataloging and providing access to moving image materials. *Cataloging* in its traditional sense means *classifying, assigning a location and describing a physical and immutable -- unchanging -- item*.

LC cataloging practices and standards were by and large adapted, adhered to and supported by large academic institutions and public libraries. For example, the University of California Los Angeles (UCLA) has had a long history of acquiring television materials and cataloging these according to LC/MARC record standards. (*MARC* stands for *Machine-Readable Cataloging* and was introduced in the 1960s to produce computerized cataloging records).

In 1984 the first edition of *Archival Moving Image Materials: A Cataloging Manual* was published and continues to be an important rulebook American libraries use for cataloging archival moving image materials.

It is safe to say that most public television stations throughout the country have designed their own systems for cataloging. Those that have affiliations with institutions of higher education may have adopted the LC/MARC cataloging rules; but most public broadcasting stations that acquire and maintain collections of their program and production assets, have developed their own idiosyncratic and "local" cataloging and tracking systems.

However, with the proliferation of digital production and distribution workflows which rely on a wide variety of equipment by different manufacturers that must 'speak' to each other, it is no longer viable for stations to utilize such individualistic systems. Instead, in order to function within the complex national public television network, it is increasingly important that stations begin to adopt common standards for cataloging and tracking so that program files can be successfully exchanged.

Over the last 20 years or so, as computers have been deeply integrated into the television production process, a myriad of electronic and digital files are being generated. These production files run the gamut from word processing text files such as transcripts; database software used for logging footage; budgets prepared in spreadsheets; to digital camera files containing footage or stills sent to digital editing workstations (e.g., AVID, Final Cut Pro).

The editing workstations in turn create files containing video and audio streams that are uncompressed or compressed, along with JPEG, TIFF and other digitally generated still visuals and digital animation files. (For a comprehensive description of file formats used in video production and distribution, see *"Survey of Digital Formatting Practices in Public Television Program Production"* written by Dave MacCarn for the NDIIPP Public Television Project, March 2007.)

It is increasingly apparent that such digital assets generated by electronic equipment require that we capture new kinds of information (data) for efficient indexing, collection management, broadcast transmission and long-term preservation of these assets. Capturing this data is done both through automated processes and traditional manual data entry procedures.

Due to the ethereal, alterable and fragile nature of digital objects, the range and type of data captured has gone beyond traditional cataloging information. What we formerly referred to as cataloging has morphed into ***metadata***, which is cryptically defined as 'data about data.'

The authors of *Descriptive Metadata for Television* define metadata as "meaningful information in its aggregate." (Cox, Tadic, Mulder 2006, 2) The *National Science Metadata Primer* says metadata "…consists of structured, standardized descriptions of resources…" to assist in the discovery and retrieval of these resources and the "…prescribed set of possible descriptive statements is known as a metadata "format" or "schema." (NSDL *http://comm.nsdl.org/viewcvs/viewcvs.cgi/*checkout*/metamanagement/metadataPrimer/overview2.html)* Metadata is also a major tool to be used in assuring that the integrity and authenticity of a digital asset remains intact over time.

As was true in the moving image cataloging world, metadata for moving image materials can be, and often is, tailored to fit individual institutional needs. But such tailored schemas can be developed with an eye to allow future Web and other access opportunities. To prepare for this expanded access, local metadata fields are "mapped" to external open metadata schema standards so they all follow the same general rules.

Types of Metadata

Metadata must be created for all digital objects. Metadata is inextricably linked in a unique fashion with the object to which it relates.

- **Embedded metadata** means that the metadata was created by the author at time of creation and is contained within the asset.
- **Associated metadata** is maintained in files related to the digital asset.

There are four major kinds of metadata:

- **Descriptive.**
- **Administrative.**
- **Structural.**
- **Preservation.**

The Cornell University Digital Imaging Tutorial is an excellent resource outlining the management of digital objects. The following definitions are based on the Cornell Tutorial, and on those in *Descriptive Metadata for Television.*

Descriptive metadata – identifies and describes the intellectual content of the asset to allow for searching and retrieval via local systems or Web. Descriptive metadata fields for program masters would include information such as –

- **Series title** (e.g. *American Experience*);
- **Program title** (e.g. Big Dream, Small Screen);
- **Program summary** information that briefly describes what the program is about;
- **Subjects** examined within the program;
- How the program fits into established **genre categories** (e.g. Popular Culture);
- **Keywords** that would provide access points for searching (e.g. Television).

Well-managed **subject** terms come from internally developed or external open source vocabulary lists or thesaurus. Ideally, subject metadata is broken into at least three major areas:

1. **Proper names** of people who appear in a program (e.g. Edward R. Murrow), and names of formal organizations (United Nations) or business entities (e.g. CBS);
2. **Geographic areas** for location of a program sequence (e.g. New York City) or intended location of program sequence;
3. **Historical events or time periods** (e.g. WWII or Cold War) covered in a program.

The Dublin Core Metadata Initiative (DCMI) is an organization dedicated to promoting the widespread adoption of interoperable metadata standards and developing specialized metadata vocabularies for describing resources. As defined in the Dublin Core Glossary, a **controlled vocabulary** is a prescribed set of consistently used and carefully defined terms. The intent of a controlled vocabulary is to normalize subjects and thus assist in subject searching. External subject source examples would be the *Library of Congress Subject Headings* or the *Art and Architecture Thesaurus*.

Genre is another key descriptive metadata field. It classifies is the type of program, such as public affairs, children's or science. PBS maintains a list of preferred genre terms and the Library of Congress' *Moving Image Materials: Genre Terms* is another major source for television genre terms.

In addition, descriptive metadata should include technical information such as format, timing, item type (original footage, etc), and footage log description.

Administrative metadata – contains information about the "display, use, management and interpretation of an asset." Included in this is rights management and may include information about the asset's "…file characteristics or the capture of encoding processes used in creating the resource (aka "technical metadata); and information about the provenance of the digital resource and efforts to archive or manage the data for the long-term (aka "preservation metadata")". (Northwestern University Library, *http://staffweb.library.northwestern.edu/dl/metadata/standardsinventory/definition.html)*

Structural metadata- Structural metadata defines the digital object's internal organization and is needed for display and navigation of that object. (Dublin Core Glossary, *http://library.csun.edu/mwoodley/dublincoreglossary.html#C*) This includes pertinent data about the internal structure of the asset and any relationship to other assets.

Preservation metadata – tracks an asset's condition, documents backup procedures, dates and time; and notes any preservation work already done on an asset including migration actions.

METADATA MODELS AND SCHEMAS

Good information requires good management, and this is critically true of metadata. **Standards for data entry and terminology use are vital**. "A user 100 years from now should be able to read a metadata record and understand the programs' production intentions, contents, and whole life cycle as well as its audience and how the physical manifestations were created." (Cox, Tadic, Mulder 2006, 28)

When it comes to choosing a metadata standard or schema to follow, the multiplicity of choices can cause confusion. Often it seems that one is faced with a difficult "either/or" decision. After examining several prominent standards, schemas and conceptual models, it has become clear that no single standard or schema will suffice.

> **Instead, to meet the various needs of the NDIIPP project, a combination of standards and schemas is recommended. A combination of PBCore, METS and PREMIS used within an OAIS model appears to be the best solution.** (van Malssen, 2006, 1)
> - **The repository charged with preserving the digital public television content, should use the OAIS model.**
> - **Information Packages for the digital artifacts can be submitted to the OAIS as METS documents.**
> - **In-house databases should use or map easily to PBCore.**
> - **PREMIS should be used as the preservation metadata standard.**

The above recommendation came out of a metadata research document done for the NDIIPP project and is echoed in the Initial Design Documents report, also done for the NDIIPP project.

Developing a Trusted Digital Repository Based on the OAIS Model

OAIS (Open Archival Information System) is a conceptual model for an archive that is charged with preserving information and making it available to a designated community. In the OAIS model, a ***Submission Information Package (SIP)*** is received into the repository, within which the information is stored as an ***Archival Information Package (AIP)***. That information can then be delivered to users as a ***Dissemination or Distribution Information Package (DIP*)**.

The SIP should consist of *Content Information*, *Representation Information* and *Preservation Description Information (PDI)* in addition to the physical or digital object itself. In other words, **the SIP is the object *plus* all the metadata needed to identify, retrieve and render that object viewable.**

Additional metadata may be added to a SIP to create the AIP, such as information regarding *provenance, rights, ownership, context, relationships to other SIPs,* and *identifiers.* Thus an OAIS model repository should be able to

- Ingest material;
- Store it:
- Provide access to it;
- Manage that material;
- Incorporate administrative functions that ensure the preservation of the repository's contents.

Ideally, the OAIS model repository should integrate elements of a ***Trusted Digital Repository.*** Two of the basic concepts for a Trusted Digital Repository are **viability** and **renderability.**

- **Viability** refers to the maintenance or preservation of the digital contents over time.
- **Renderability** is the concept that those bits that make up the digital contents will be readable by humans over time.

The Trusted Digital Repository, itself, should adhere to a *principle of accountability*. That means the repository needs to have in place *clearly articulated policies and procedures that are available for review.* It must document its decisions, follow standards and be dedicated to sustaining the digital content through long-term financial, staffing and organizational commitments.

Collecting Metadata Becomes Part of the Preservation Process

The OAIS and Trusted Digital Repository concepts work together toward digital preservation. Establishing a documented system of policies and decisions must be key to developing the staff and strategies for preserving digital content.

Preservation in this sense is a collaborative effort.

- Original content and representation metadata must be gathered at time of ingest into the repository. This information then becomes the basis for defining an object's authenticity.
- Monitoring the user community's needs, changes in technology and documenting any actions that are taken on an object will result in additional metadata. This additional administrative and technical metadata works to maintain an object's fixity.

Again, it is important to document the repository's policies, procedures and actions so that over time, a digital object may be accessed and viewed by a human who has the assurance that this representation is accurate to the object originally ingested. That assurance or trust comes from the combined efforts of an object's content producers, a repository's archivists, the system's technicians and the metadata each contributes.

PBCore as a System-Wide Standard for Public Broadcasting

In the production process, metadata should be created and attached to digital objects as they are created.

> The process of ingesting, archiving, accessing, managing, and delivering digital media files is becoming an essential requirement in the everyday life of multimedia producers, and the workflow involved offers numerous opportunities for production efforts to be integrated with the needs and workflow of the archive.
> (Felix-Didier, Rubens and Van Malssen, - 2006, 2)

At WGBH, the Media Library and Archives (MLA) have developed a compliance system that seeks to collect that metadata via database templates filled out by the production units. Data entry occurs at the time of creation, or as closely as possible, with the assistance and advice of staff from the MLA.

This interaction provides some education and standardization within the station, but across the public television community a broader standard needs to be implemented. PBCore (the Public Broadcasting Metadata Dictionary) was developed for just such a purpose.

> **The NDIIPP project's recommendation is to adopt PBCore and encourage its use widely across the public broadcasting system.**

The following discussion is based on PBCore v.1.0. Since this writing, PBCore v.1.1 has been released and while the ideas outlined below still hold true, additional elements and sub-elements continue to be added. (For current information please visit the PBCore website: *http://www.pbcore.org*.)

Based on Dublin Core, an international standard for electronic documents, PBCore is intended for describing video, audio, text, images and interactive learning objects for television, radio and Web activities. It consists of a core set of terms and descriptors (elements) used to create information (metadata) that categorizes or describes media items (assets or resources).

There are **48 Elements and Sub-Elements** defined within PBCore (v.1.0). A Sub-Element may further clarify information provided in a related Element. The Elements and their related Sub-Elements are grouped together into 3 categories:

- **Intellectual Content**: 13 elements for describing the intellectual content of a media item.
- **Intellectual Property:** 7 elements that deal with the creation, creators and usage of a media item.
- **Instantiation:** 28 elements used to identify the media asset's existence in some format, either digital or physical.

Key to PBCore's acceptance is its relation to the already widely used Dublin Core, its creation specifically for public broadcasting and the relative ease with which it can be mapped to current in-house databases. (Van Malssen 2006, 6-7)

An example of such a mapping is included in Attachment B, which maps PBCore, Dublin Core, and PREMIS to current records in the WGBH Digital Asset Management (DAM) system, internally referred to as TEAMS, and to the WGBH Filemaker database called MARS (MARS tracks physical items only).

The "Record Level Recommendation" indicates whether the field should be used for a basic vs. a complete record for a public television program. Basic recommendations are indicated if a schema or data dictionary considers them "mandatory" or if the field is already in use at WGBH.

Many "recommended" elements in PBCore also seem appropriate for a Basic record. The recommendation level is more of an analysis of possible fields. A careful selection would need to be made in any case, so that metadata records do not become too daunting to fill or use. (Van Malssen 2006,13)

On the following pages is a sample taken from Appendix B of what may be considered a Basic Record in PBCore, mapped to a current WGBH DAM record. The Basic Record captures Series and Program or Episode Titles, Program or Episode Number, a brief summary of the content, creator and publisher information, rights information, creation date and release date, format of the original (if not born digital), various identifiers, location, duration, time code information and language. It should be noted that additional metadata does exist in the record in Appendix B that shows the fuller, more complete WGBH Digital Asset Management (DAM) system record mapped to an extended PBCore record.

You will note that some fields may be repeated (Title and Identifier), some information in TEAMS may not have a perfect equivalent in PBCore (annotations about Closed Captioning or Descriptive Video Service (DVS)), and there are basic PBCore fields with no information in the TEAMS record.

While not a perfect fit, the in-house database serves as a basis for a PBCore record that could be shared with other stations or archives. The current database and data entry procedures need not be abandoned.

Basic PBCore (v.1.0) Record Mapped to WGBH TEAMS Record

PBCORE ELEMENT/ SUB-ELEMENT	WGBH TEAMS FIELD	SAMPLE TEAMS RECORD
01.01 title	Title	Press and the People, The: An Inquiry into the Work of the American Press in Informing the American People

01.02 title Type	Title: Type	Series
01.01 title	Title	Responsibilities of Television, The; Part I
01.02 title Type	Title: Type	Program
01.01 title	Title	8
01.02 title Type	Title: Type	Episode
03.01 description	Description: Movie	Following closely on from his speech in to a professional group in Chicago where he made the following quote "Surely we shall pay for using this most powerful instrument of communications to insulate the citizenry from the hard and demanding realities that must be faced if we are to survive", Louis Lyons interviews America's foremost television journalist Edward R. Murrow, on the state of television broadcasting. Murrow believes that the program structure in imbalance and that the American television audience is being fed a diet of programming that insulates them from the realities of the world around.
10.01 creator	Creator	
10.02 creator role	Creator:Role	
12.01 publisher	Publisher	
12.02 publisherRole	Publisher:Type	
12.01 publisher	Rights:Holder	
12.02 publisherRole		
	Rights:Usage	
13.01 rightsSummary	Rights:Description	
13.01 rightsSummary	Rights:Restrictions	

	Rights:Term (Time period)	
	Rights:Credit	
20.01 dateCreated	Date:Movie Creation Date	
20.02 dateIssued	Release Date	
	Effective Date	1/24/1959
21.01 formatPhysical		
	Format:Original Moving Image	Betacam
21.02 formatDigital	Format:Digital Movie	
21.03 formatLocation	Holdings:Location	
21.04 formatMediaType	Type:Media	
	Type:Moving Image	Preservation Master
21.05 formatGenerations		
21.06 formatStandard	Format:Broadcast	NTSC
21.07 formatEncoding	Format:Video Codec	
21.08 formatFileSize		
	Desc:Time out	01:30:54;03

21.10 formatDuration	Desc:Time total	00:29:17;19
21.14 formatFrameSize		
	Format:Audio Playback Rate	
	Format:Video Playback Rate	
22.01 language	Language	en
22.02 alternativeModes	Format:Closed Captioning?	
	Format:DVS?	
	Format:Digitizing Hardware	
	Format:Movie Creation Software	
	Format:Software Version	
23.01 identifier		
	Asset:Asset ID	d66be05b25078ed88b935760c9989b02c2173f54
	Identifier:Movie	
	Import ID	
	Identifier:Reference ID	
	Tracking Number	143750

	Creation Order	1156882
	Date:Review Date	
23.02 identifierSource	Date:Movie Creation Date	
	Model	MOVIES
	Content Editor	karen_colbron
	Asset:Date:Imported	4/12/2006
	Asset:Imported By	
	Start Frame	00:00:00:00
	End Frame	00:29:18;05

PREMIS Elements

While PBCore contains elements for basic title, program, file format information and includes elements for rights and ownership, it does not contain appropriate elements for preservation information. Adopting elements from *PREMIS (Preservation Metadata: Implementation Strategies)* becomes necessary in order to track preservation actions as part of a Trusted Digital Repository.

PREMIS was developed to fill the preservation metadata gap found in most schemas and standards. It includes a data dictionary but does not focus on descriptive information. Instead, PREMIS concentrates on administrative, technical and structural metadata used to support the digital preservation process.

In PREMIS, elements are described as semantic units that are broken down into semantic components. These semantic units are used to describe five entity types:

- **Objects:** discrete units of information in digital form
- **Intellectual Entity**: a coherent set of content that is described as a unity
- **Event:** an action that involves at least one object or agency
- **Agent:** a person, organization or software program associated with preservation events
- **Rights:** assertions of one or more rights or permissions pertaining to an object or agent

Three types of Object Entities are defined in PREMIS:

1. **Files** are a named and ordered sequence of bytes that can be read, written and copied.
2. **Bitstreams** are a set of bits embedded within a file.
3. **Representations** are a set of files, including structural metadata, needed for a complete and reasonable rendition of an intellectual entity.

Versions, as defined by PREMIS, become distinct objects themselves. Their relationship to the original object is described by an event.

Each of the entity types, except Intellectual Entities, has semantic units defined for them. PREMIS does not include semantic units for Intellectual Entities because it considers such information to be descriptive and covered by other metadata schemas. Appendix C provides an explanation of the top-level PREMIS fields.

General descriptions for each entity are given below.

- **Object Entity:** 85 Semantic Units and Semantic Components available. Object entities are associated with rights statements and events and can be related to agents.
- **Event Entity:** 16 Semantic Units and Semantic Components available. These record information about an action that involves one or more objects. Events must relate to an object and may relate to an agent.
- **Agent Entity:** 5 Semantic Units and Semantic Components available. An Agent may hold or grant rights, carry out, authorize or compel an event, and may create or act upon an object.
- **Rights Entity:** 16 Semantic Units and Semantic Components available. Rights must be related to an object and an agent.

While PREMIS is well defined, there is little experience using it for audiovisual items, such as television programs. There are no examples in the PREMIS documentation for moving image or sound materials, and guidance for implementing PREMIS is sparse.

Although these drawbacks have the potential to mean extra work, it is still recommended that PREMIS be implemented to record and manage digital preservation metadata. (Van Malssen 2006, 10-12)

METS Documents

PBCore and PREMIS provide standardization across institutions by defining language and terms through their data dictionaries, and by defining the type and organizational structure of what information to gather. However, they are not, necessarily, easily read across systems. Thus, it would be like having cataloging information standardized but in different languages at different institutions.

METS (Metadata Encoding and Transmission Standard) is an open standard XML encoding format that can be used to manage digital objects both within a repository and between repositories. METS encodes descriptive, administrative and structural metadata for both text and images. It is a product of the Digital Library Federation and fits well into the OAIS repository model.

METS, which is maintained in the MARC Standards Office of the Library of Congress, recognizes *extension schemas* that allow additional metadata that may be specific to an object. For example, PREMIS has a METS approved extension schema. A PBCore XML schema is currently being developed with the hope that it will obtain METS approval, thereby allowing for easier encoding of PBCore records into METS documents.

There are seven sections to a METS document, briefly described below:

- **Header**: metadata about the METS object itself and its creation (who, what, where and when).
- **Descriptive Metadata**: this can point to an external metadata document or contain the embedded descriptive metadata or both (for instance, this can be used to point to an in-house database document).
- **Administrative Metadata:** this section provides information on how the files were created and stored, along with information on intellectual property rights, provenance and source. As with Descriptive Metadata, this section can point to an external document or contain the embedded information.
- **File Section:** this allows related files to be grouped together to comprise a single version of the digital object.
- **Structural Map:** this is an outline of the hierarchical structure of the digital object. It maps relationships (links) between elements of the structure to content files and to relevant metadata.
- **Structural Links:** this records the hyperlinks between the nodes in the structural map.
- **Behavior:** is used to associate executable behaviors with the content of the METS object.

It is possible to convert in-house database records (such as from WGBH's TEAMS DAM system) into METS documents. There are tools and DAM software available that can generate XML METS documents, and for public television, this will be easier once the PBCore XML schema is completed. As stated above, the PBCore XML schema is still being developed, although a draft version does exist.

Who Supplies Metadata and When

Descriptive Metadata for Television provides a rather extensive program production process workflow, indicating the times and types of metadata that can be generated. We have also analyzed the workflows of our own program production units.

The following discussion identifies points along the WGBH production workflow where metadata is currently created and captured -- from the production start up and shut down, to Master Control ingest into the local digital on-air library, to scheduling on the broadcast system, to final delivery of files to PBS for national distribution. The main assumption in the discussion is that there are institutional standards for identifying a program series and title as the unique identifier of an asset.

In the television production workflow, a significant volume of metadata comes from the production unit, but some is added by the archivist, some is captured by the system upon ingest and still more information is added by the system technician. Any metadata recommendation needs to take into account these various metadata creators and their roles in the production workflow -- each creator has a different level of understanding regarding metadata, its purpose and its relation to them.

Content creators --The production unit is *the content creator*. Their needs and outlook are temporal. They are focused on getting a program completed and delivered to either a producing station or PBS. When this task is completed, the production team is disbanded.

Nevertheless, those in the production unit are the creators and have first hand knowledge of who, what, where, when and why the content was created. They 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.

But they do not think about the life-cycle of the program past their immediate deadlines. Often their labeling is unclear or includes abbreviations that make sense for the production team but relay little coherent information to those outside the team. They 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.

Even so, their information is critical. It has been the experience at WGBH that production teams are amenable to filling in a template with some basic information that can then be ingested into an in-house database. *This captures metadata at the time of creation, a crucial step.* This is done at the production element level, for originally-shot footage, stock footage, stills and graphics, both those used in the final program, and that material which is not used.

Specifics of what metadata is captured at the element level includes:

- program or episode title;
- format;

- originating production unit or office;
- brief content description;
- content description at shot level with reference to time code;
- rights information;
- date of creation;
- location;
- participants;
- contacts.

Programs going to PBS --Delivering a completed program to PBS follows a similar routine. Prior to delivery, the producers are required to fill out several forms that provide information about the completed program.

The primary form to be completed for PBS is the **Program Acceptance Agreement** (PAA) which includes a significant amount of metadata information:

- Series and program or episode title;
- Contact names;
- Copyright ownership information;
- Broadcast rights;
- Ancillary rights (home video distribution);
- Format information (closed captioned, aspect ratio, audio configuration, duration);
- A brief summary of the program's content.

Once PBS has accepted the program, the producers send a **Format Sheet.** This details the program's structure with time code references --

- Where the tease goes;
- Underwriting credits;
- Promos;
- Production credits;
- PBS credits, etc.

If necessary, a separate **Interactive Web Tag** form is also submitted. This information populates the PBS database.

Again, it should be possible to map these fields to PBCore elements.

The Archivist -- The archivist's role at this point is to provide advice on standardizing the data entry (language) and oversight to make sure it is done. **The goal is to enter information once and have it "live" with the content throughout its lifecycle**.

How the template looks to the production team does not matter, as long as it conforms to PBCore elements or easily maps to them. Thus, using such a template, *an in-house database could be populated with metadata that meets standards from the point of creation onward.* A set of WGBH shutdown templates is appended to this paper.

The archivist's role upon delivery to the archives, at both the program and element levels, is to review and clean the information so it meets set standards. In most cases cleaning will probably involve editing language and choices from drop down or pick lists.

The archivist will also *add descriptive metadata* as needed to complete the record (including rights and ownership information) and *administrative metadata*, such as time and date received into the archives, location of the object, any action taken upon the object.

It may be that some *additional technical information* also needs to be provided by the archivist, for instance, if the format or aspect ratio was not noted earlier.

Finally, the archivist will, most likely, *add preservation metadata or information as placeholders for future preservation events.*

The archivist is responsible for making sure that the metadata meets accepted standards. PBCore and PREMIS have data dictionaries to assist with language choices. These may need to be supplemented with other approved sources, such as Library of Congress authority lists.

> **Public television would benefit from an ongoing group that would recommend standards regarding metadata terminology, similar to the technical standards outlined in the PBS 'Red Book'.** (The PBS Red Book specifies the comprehensive packaging and delivery guidelines for PBS programs. It is used in the preparation of programs accepted for broadcast by the PBS National Program Service and other national distribution services. *http://www.pbs.org/producers/redbook)*

METADATA AND THE DIGITAL REPOSITORY

File Configurations

Up to this point, the object and its related metadata have been tracked in local or in-house systems. The elements could be stored and tracked in a producing station's archives; while the completed program could be tracked and stored either at PBS or the producing station, or both.

The NDIIPP project is developing a digital repository for sharing these elements and programs on a broader level for preservation purposes, and eventually for deposit with the Library of Congress. The model repository builds upon existing standards, technology and workflow but with an eye to the future. Thus, this discussion of metadata has been prepared with this longer-term goal in mind.

The "Contribution File" -- Currently PBS receives standard definition (SD) master videotapes on Digital BetaCam that are ingested into a **50 Mbps IMX (D-10) file.** This becomes the *"Contribution File."*

The Contribution File may be ready for further processing or it may undergo additional editing for various purposes, such as to add or replace underwriting credits. The file then goes through a "Flattening" process before being distributed where features are added, including

- Closed captioning;
- Descriptive video service;
- Bugs;
- Nielsen information;
- V-chip information.

The "Distribution File" -- Flattening also downsizes the Contribution File from a 50 Mbps file to an 8 Mbps file. The 8 Mbps file becomes the *"Distribution File"*, which is the completed program ready for broadcast. The Distribution File is sent out and delivered to public television stations for local broadcast via satellite. (Davila, p. 9.)

Currently, this process of transferring the final broadcast package occurs in real time; however planning is underway for the additional option of non-real time file delivery. Called the NGIS (Next Generation Interconnection System) project, this will soon put in place a national system to transfer digital program files in non-real time to the broadcast stations. NGIS will also allow stations the option of delivering digital files to PBS. Thus, the NGIS project is a start at creating an OAIS model repository.

Submission Information Packages -- Based on these technical requirements for broadcast file distribution, *Submission Information Packages (SIPs)* are being defined at PBS, as well as *Distribution Information Packages (DIPs).* Working with New York University, the NDIIPP project's model repository is defining the *Archival Information Package (AIP)* that will complete the model.

File Wrappers

As mentioned earlier, digital television program files are composed of *essence* and *metadata*, or data about that essence. The essence and metadata need to be wrapped in a computer code that acts as an interface to facilitate transfer between systems.

Several wrappers already exist, such as the Advanced Authoring Format (AAF). A subset of AAF is MXF (Materials eXchange Format). MXF supports metadata exchange and recomposing separate from the essence. It also supports essence independent of the codec. (The term *codec* is a concatenation of the terms *coder* and *decoder*. *Codecs* are a method of encoding data, in this case, specifically video. Information about the codec of a file must also be maintained to guarantee that the essence can be decoded in the future.)

PBS saw MXF as a valuable existing wrapper that could be adapted for use in the NGIS system to support aspects of the file distribution process. The result is a specific application of MXF developed especially for PBS known as *AS-PBS.*

AS-PBS is based on many existing reference standards:

- The MXF file format specification and subsequent container mapping;
- Operation pattern;
- Metadata (including PBCore);
- Stream partition and digital television standards; mostly from SMPTE. (Davila 2006, 56)

It is hoped that, eventually, AS-PBS will be adopted at producing stations. Programs could be delivered as essence and metadata files, wrapped in AS-PBS. This would streamline the NGIS system.

Preparing Program Files for the Repository

Ideally, a *SIP* for program delivery to PBS would be a program composed of essence (50 Mbps IMX (D-10) MPEG 2 video and BWF or Broadcast Wave Files) plus the associated metadata, all wrapped in AS-PBS. The metadata would follow PBCore standards.

A *DIP* coming from PBS would be a program composed of essence that has gone through the feature and editing process, been downsized to 8 Mbps, had updated metadata added to it and then the package wrapped in AS-PBS.

For submission to the NDIIPP repository as an *AIP*, there are several possibilities.

- One would be to capture the 50 Mbps program file from PBS after it has undergone the feature and edit stage but prior to downsizing. At this time, however, this option by itself is not very viable due to workflow and cost issues.

- Another would be to submit the 8Mbps DIP file as the AIP. This, however, does not capture the higher resolution file.

- Alternatively, the AIP could capture *both* files -- the 50 Mbps SIP and the 8 Mbps DIP could *both* be ingested as related AIP files. This is the approach we are currently considering.

The AIP file would then have PREMIS metadata added to it in a METS wrapped MXF file that may also contain PBCore metadata. (Davila 2006, 103) Updates to the metadata, for instance when a preservation event occurs or a new version is submitted to the repository, would require changes to the METS object which then serves as a key to the digital object itself.

Conclusion: Only Metadata Makes Access and Retrieval Possible

> Without metadata, the archive could have the perfect storage strategy and would still be meaningless, because there would be no retrieval and hence no need to store the bits. With appropriate metadata, the archive becomes accessible.
> (Wactlar and. Christel 2002, 93)

The ideal outlined above is predicated on several core conditions:

- The availability of mapping or crosswalk tools that allow in-house databases to conform to PBCore and PREMIS schemas;
- The completed development of a PBCore XML schema and its acceptance as a METS extension;
- Further testing of AS-PBS and its performance as a functional file wrapper for public television program files.

Decisions also need to be made regarding conformance with UPF (Universal Preservation Format). The UPF is characterized as "self-described" because it includes within its metadata all the technical specifications required to build and rebuild appropriate media browsers to access its contained material throughout time. (See "UPF Glossary" http://info.wgbh.org/upf/glossary.html)
It may be possible to include in the AIP the metadata and codec to playback the digital file or store that information separately with a pointer in the AIP's METS object.

It is most likely that no single institution could host a repository that contains completed programs plus all the production elements associated with it. A network of repositories may need to be developed, along with accompanying deposit agreements. Such a network could be based on the existing network infrastructure already in place among public television broadcasters, and the resources being installed locally at each station to manage and playback their digital broadcast files.

Despite the work that needs to be accomplished before the NDIIPP model repository becomes a working reality within the PBS system, the biggest step has already been taken.

There is common recognition that digital television files need to be preserved, which cannot happen just at the end of the production process but must be integrated into the overall production workflow.

Moreover, preservation means assuring that the essence survives along with the metadata used to identify and access it. If a genuine preservation repository is going to be successful, it will depend on collaboration between content producers, archivists and systems technicians, between producing stations and other producers, PBS and other distributors, broadcast stations and the Library of Congress, to agree on the minimum standards required to create and/or capture necessary metadata, and to demonstrate that such requirements can be adopted and implemented.

Without such a broad effort, long-term access to digital program materials will be lost. With it, there is the potential that programs produced in digital formats can be preserved and accessible well into the future.

* * * * *

REFERENCES

[Some of these documents and reports are available on the project website: http://www.ptvdigitalarchive.org]

Balkansky, Arlene, Laurie Duncan, Pearline Hardy, Stephen Kharfen, Marzella Rhodes, Betty Wilson. Archival Moving Image Materials: A Cataloging Manual, 2nd edition. Washington, DC, 2000.

Broadcast Wave Format. Available at: http://www.ebu.ch/en/technical/trev/trev_274-chalmers.pdf

California Digital Library Glossary. Available at: http://www.cdlib.org/inside/diglib/glossary/?field=institution&query=CDL&action=search

Cox, Mike, Linda Tadic, Ellen Mulder. *Descriptive Metadata for Television.* New York: Focal Press, 2006.

Davila, R. Justin. "New York University Preserving Digital Public Television Project: Initial Design Documents." Report prepared for the NDIIPP Public Television Project, October 2006.

Dublin Core Metadata Initiative (DCMI): http://library.csun.edu/mwoodley/dublincoreglossary.html#C

Felix-Didier, Paula, Caroline Rubens and Kara Van Malssen: "The Role of Workflow and Metadata Management in the Preservation of Public Television: NYU, EBC, WGBH and PBS." Report prepared for the NDIIPP Public Television Project, September 2006.

Getty Art & Architecture Thesaurus Online, The. Available at: http://www.getty.edu/research/conducting_research/vocabularies/aat/

Library of Congress Authorities. Available at: http://authorities.loc.gov/

MacCarn, Dave: "Survey of Digital Formatting Practices in Public Television Program Production." Report prepared for the NDIIPP Public Television Project, March 2007.

Metadata Encoding & Transmission Standard (METS). Available at: http://www.loc.gov/standards/mets/

Moving Theory into Practice: Digital Imaging Tutorial. Cornell Digital University Library Research Department. Available at: http://www.library.cornell.edu/preservation/tutorial/metadata/metadata-01.htm.

NSDL Metadata Primer: Metadata Overview – What is Metadata. Available at http://comm.nsdl.org/viewcvs/viewcvs.cgi/*checkout*/metamanagement/metadataPrimer/overview2.html

Northwestern University Library Staff Web, Digital Library Committee. Joint Committee on Metadata. Available at: http://staffweb.library.northwestern.edu/dl/metadata/standardsinventory/definition.html

OAIS (Open Archival Information System). Links from http://nost.gsfc.nasa.gov/isoas/

PBCore Public Broadcasting Metadat Dictionary Project. Available at: http://www.utah.edu/cpbmetadata/

PREMIS (PREservation Metadata: Implementation Strategies. Available at: http://www.oclc.org/research/projects/pmwg/

Van Malssen, Kara: "Preserving Digital Public Television: Metadata Research." Report prepared for the NDIIPP Public Television Project August 8, 2006.

Wactlar, Howard D. and Michael G. Christel. "Digital Video Archives: Managing Through Metadata." In *Building a National Strategy for Digital Preservation: Issues in Digital Media Archiving.* Washington, DC: Council on Library and Information Resources and Library of Congress, 2002.

Universal Preservation Format (UPF). Available at http://info.wgbh.org/upf

LIST OF ATTACHMENTS

Attachment A: "Preserving Digital Public Television: Metadata Research", by Kara Van Malssen, August 8, 2006.

Attachment B: Metadata Map: PBCore, Dublin Core and Premis to WGBH TEAMS and MARS, Kara Van Malssen, August 8, 2006.

Attachment C: PREMIS Elements, Kara Van Malssen, August 8, 2006.

Attachment D: WGBH Original Footage Shutdown Template.

Attachment E: WGBH Stock Footage Shutdown Template.

Attachment F: WGBH Materials Used Shutdown Template.