

**Notes from NDIIPP public television project metadata meeting  
Hosted at Thirteen/WNET, New York**

December 12, 2005

Initially drafted by Carl Fleischhauer, edited and revised by Nan Rubin

**Attendees**

Visiting experts

Grace Agnew, Rutgers University and representing Moving Image Collections (MIC)  
Lisa Carter, active in AMIA, formerly Kentucky public TV, now at University of Kentucky  
(Director of Archives)

Laura Krasnow, Kentucky ETV

Thirteen/WNET

Mike Boeglin (IT, digital asset management), Meg O'Hara (ProTrack implementation, digital  
asset management), Daisy Pommer (archivist), Nan Rubin, Winter Shanck (librarian)

WGBH

Gerry Field (National Center for Accessible Media, role with PB Core), Mary Ide, Dave  
MacCarn, Leah Weisse

PBS

Glenn Clatworthy, Bea Morse, Irene Taylor

NYU

Melitte Buchman, Rick Ochoa (programmer), Joe Pawletko (Hemispheric video project),  
Jennifer Vinopal (interim head, digital library program), Gary Shauver

Library of Congress

Carl Fleischhauer, Barbara Humphrys

**General comments.** The meeting was very productive in several ways. It highlighted metadata areas that need to be investigated or developed within the confines of the NDIIPP project. It also articulated the relationship of the NDIIPP project's preservation-oriented metadata to the development and implementation of PBCore metadata by the public television community. PBCore is a new structure intended to serve the ongoing operational needs of public television stations and PBS, the distributor of programming. In other words, the meeting connected the NDIIPP project to a larger national group. There was a big turnout (see list of attendees below), which in large measure reflects the interest in the larger context.

The meeting began with a number of presentations, giving background and context to the immediate issues --

- Dave MacCarn, followed by Gerry Field, give the history of PBCore development and current outreach activities.
- Grace Agnew presented the structure and cataloging approach of MIC (Moving Image Collections) activity and promoted its possible value in relation to PBCore. One idea that she and others floated had to do with possible MIC contributions to public television metadata. Could the MIC cataloging utility be used? Could data be put into the MIC underlying data structure/system, and then output as PBCore when needed? Might there be a public television portal at MIC? The questions were heard and will be considered as things move forward.

- Lisa Carter, who has been instrumental in assisting public television stations in planning how to digitize their collections, spoke about the current state of station projects and their frustration with not knowing how to adopt or use PBCore. Laura Krasnow reinforced Lisa's conclusions with additional station experiences.

The meeting also featured a talk about an ongoing video project at NYU. The Hemispheric Institute Digital Video Library (HIDVL) is committed to content preservation and uses the DSpace repository architecture. Although the people representing this NYU project are not officially assigned to the NDIIPP project, it seems clear that the HIDVL approach will form the basis for the model repository that is an important part of the NDIIPP activity.

The particulars of the group discussion explored some strengths and gaps in PBCore. We learned that identifiers are required by the specification, and that there are guidelines for identifiers. But we also learned that this new standard has not seen much implementation in part because there are no tools, and little guidance to stations in its value. Worse, the public television community suffers from bad habits in the area of material identification: because of the daily operating demands of stations, they were described as putting program tapes on the shelf in the back room with virtually no cataloging and little thought of later consequences or needs. So stations see little motivation or need to go back and create a new database for their tapes. Without ready-tools, PBCore is a hard sell to stations.

The group also talked about deep techy metadata, i.e., what is called "parametric" data in the SMPTE (Society of Motion Picture and Television Engineers) data dictionary. There are a bunch of such elements--maybe 40 or 50--in PBCore. Grace Agnew told about the work of a committee in AMIA (Association of Moving Image Archivists) that identified about 100 elements of interest in this category. The committee was led by Ian Gilmour (from Australia's National Film and Sound Archive) and Hannah Frost (Stanford preservation unit). Because PBCore is designed for basic use, the technical fields were pared down so there may be a shortfall here if more technical information is needed for preservation purposes.

Meanwhile, I called attention to a life cycle angle, noting that PBCore (as expanded) might work well as an OAIS-style submission information package (SIP). If additional metadata is needed for long-term preservation, as in an archival information package (AIP), I speculated, this data might be added/handled outside of the PBCore structure.

At the end of the meeting, Leah Weisse from WGBH, was identified as the key coordinator of metadata for the purpose of the NDIIPP project. Leah, along with Nan and others, will follow up on some of the ideas expressed on this occasion.

## **Minutes of the meeting as jotted down by Carl, edited by Nan.**

The meeting started with introductions and an opening set of comments by Nan Rubin. She explained that Ken Devine (the CTO for Thirteen) [Ken IS the CTO here!!!] was ill and would not attend. She went on to introduce PBCore and to sketch its status: developed by CPB, convoluted path, some in the room were on the development committee or advised the process (Bea, Dave, Grace). PBCore was officially released in April 2005, as a data dictionary with no initiative at the time to promote its value or assist stations with using it. The question for this project is -- will it fit asset management and preservation needs? The discussion of that topic is the purpose for this meeting. Nan noted that there had been a good meeting at AMIA a week ago, there were several other public TV folks there, to talk about PBCore in general and how it relates to preservation. Implementation of PB Core may be manageable at big outfits like WGBH and Thirteen, but this will be more difficult at the many smaller stations. For PB Core to succeed, we need to make sure it works for them.

PBCore not designed for archival work, was developed for operations not preservation. We want to explore what is missing and where it might come from. Nan notes that one purpose is to make the programs digestible/ingestible by LC for its collections (via the PBS agreement, not via the copyright process), we want to make sure that we structure our content in a way that works for LC. Mentions NYU MIAP graduates who are doing an inventory of LC public TV program holdings, which are mostly analog, so that they can be considered in the overall planning effort.

Nan mentioned that public television has a tech conference in April, just before NAB. This is a deadline for our metadata examination; we want to have something ready for that.

MacCarn: Outlined history of PBCore: in 1989, WGBH was working on DAM development, this raised metadata issues. We were starting to get some headway in two departments: (1) Mary Ide and the archive, and (2) film and video resources, a department that does research on program content for future reuse, also for footage sales. They want to describe tape holdings down to the scene level. These two departments were developing their own databases, we wanted to connect the two. By 2000 this asset management effort came into focus, and we approached CPB for sponsorship. We proposed to share ideas and our approach with the public TV community, a kind of homegrown standardization effort. We got a little money, brought together folks in public TV, various stations and other producers, looked at field (data element) structures. Initially the thrust was for program exchange and distribution. What do I need to know about this program for those purposes?

The effort was aimed at public broadcasting, including radio. We trolled for interested parties from public TV organizations. Who was working on DAM? We brought those experts together, from their organizations. We took a look at what public TV people were doing, what the industry was doing. We took field (data element) lists from many existing sources and put them into spreadsheets. For example: Dublin Core, MPEG-7, ViDE, SMEF, SMPTE's metadata dictionary, etc. We had 1200 fields, then whittled these down. Grace Agnew helped, as did others. We looked at other standards, like Dublin Core, MPEG-7, ViDE, SMEF, SMPTE's

metadata dictionary, etc., and found that none were ready “off the shelf.” Most of these were too complex for the organizations we were working with. MPEG-7 for example, there was just too much there, people were not going to “get it.” What we work on is called PBCore for a reason, it is a *core set*, can be expanded.

Issues: once we finished version 1, the funding from CPB was used up. The group asked for more funds to sustain the core set. CPB is coming around for more help. We have to teach people how to use it, we need a maintaining org. WGBH has standards experience, we knew what was entailed.

Nan: not sure there are real advocates at CPB in the current political climate.

Dave: CPB support will grow out of stations saying that they need this.

Grace: Notes that PBCore maps well to other schemas, like MPEG-7.

Gerry Field from WGBH picks up Dave’s story and continues. We do have more funding now, a contract with CPB has come thru. We call this a sustainability effort, meaning that we are looking for ways to maintain and sustain the PBCore metadata standard. Our group included folks from other stations, including Alan Baker from Minnesota Public Radio [I know Gerry said this, but I don’t think it should be put down on paper like this!] PBCore 1.0 went public on April 1, 2005. Web site in Utah is very rich, lots of info.

About WGBH and taking up the challenge of continuing their lead role with PBCore. Semi-joke from the accessibility specialist about his role in this activity: there is an accessibility angle here, you know, i.e., how to find content. New CPB funds are for (a) advocacy, (b) training, and (c) sustainability. Will PBCore stay where it is? Will it roll into some other activity?

(a) Advocacy: we want to help potential users understand the need and the standard, we need to prepare info in “short speak” form, i.e., draft compelling talking points aimed at station managers. And we need to do some liaison with other standards organizations, e.g., SMPTE, NCAM (National Center for Accessible Media), and other standards organizations. At the same time, advocacy is needed with system manufacturers, e.g., ProTrack. (ProTrack is a traffic and scheduling tool used by a majority of public TV stations.) A lot of stations are building systems, want to get PBCore built into those. Regarding standards harmonization efforts, we have gotten good feedback, including from SMPTE, related to DMS-1 (digital metadata version one).

(b) Training: we proposed creation of tools to support PBCore. PBCore 1.0 is just a data dictionary on HTML pages, we would like to see an XML schema for it. Want to see implementation in various systems. Want to make PBCore “presentable” on the Web. Want to build a community of interest on PBCore. Will also make a pitch at the NETA (National Educational Telecommunications Association) conference.

(c) Sustainability is in a development process; see notes four paragraphs back.

Carl: what classes/chunks of data are there in PBCore?

Dave MacCarn:

- \$ Dublin Core (descriptive) plus what we need to exchange content
- \$ added timecode, encoding, techy stuff
- \$ also admin metadata, about program rights (?) and availability

Grace Agnew adds:

- \$ admin metadata, there is “tiny bit of rights”
- \$ it is like METS except no structural map
- \$ very strong on tech [parametric in SMPTE terms] metadata, I used a lot in MIC

MacCarn: PBCore’s element set relates to our metadata at WGBH, in our production/DAM systems, but our data set is bigger, it is a superset, or rather PBCore is a subset of what we have. We don’t actually use PBCore per se at this time.

Lisa Carter: KET (Kentucky Educational Television) sees it this way for sure, it isn’t what you use for operations or production, it is an input or output (Carl’s terms). There is a perception problem of people thinking “this is it.” Stations around the country do not have people who think about organizing their assets. It is quite a leap to get them to think about the title, the director, and to put this in place in a systematic way. Stations have a big learning curve.

Leah Weisse: with WGBH producers, it is battle to get people to fill in (as few as ) four fields. (meaning more is a HUGE obstacle.)

Group discussion: about the problem of getting tv stations people to understand the importance, how to put PBCore (or any archiving approach) into use.

Grace: a lesson I have learned is: “don’t confuse data model (could be complex) with tools (streamlined to seem simple to users).” A lot of detailed metadata will be yanked from your assets.

[Gerry Field mentions Ann Wilkins in Wisconsin, who has developed a clever demo of PBCore using all open source tools.]

Grace Agnew delivers a slide-based talk on Moving Image Collections (MIC). MIC will be hosted at LC, may be a public face for the Culpeper center. MIC is a union catalog, with a content tilt just now toward science due to NSF funding. It all started years ago, at first considering the use of MARC. But there was so much resistance to MARC that it was dropped. So we worked to come up with a flexible approach that would accommodate various kinds of metadata produced by a variety of players. We are now developing a cataloging utility for participants to use that fits the project.

We have asked about funding from LC, to continue development. Greg Lukow at LC says this is two years away. But we may have additional funding from a third party with an interest in rights management.

Grace's slide shows a diagram with the various MIC elements, including a data registry, element tables, an OAI table, and other boxes. The structure we use for our storage of the underlying data means that MIC supports display in MARC, Dublin Core (XML and thru OAI gateway), and MPEG-7. We also maintain the participating archive's own local metadata formats. Next will be the finalization of export in these formats.

One original motive for the development of MIC was the 1997 LC film preservation survey, this led to a push to determine priority items to preserve. Another motive: move moving image content into education. When we talked to educators, they all said, "We have to be able to get it, I have to be able to show this in my classroom." This demand led us to create a directory database (for archives-as-archives) to provide info to users on how to get resources. This info is at the "archive" and not the item level. Now there are about 14 archives participating. The directory database also tells about preservation roles played by these archives.

About educational materials, lots are being created by MIC users, we disseminate these using Creative Commons licenses. Lots of educators use them.

Current and future initiatives: working on refining our own underlying-data data model. And sorting out our next developments vis-a-vis our relationship to LC. Moving our technology to the Library is made difficult by security concerns.

Grace Agnew's view of a data model:

- \$ identifies primary entities associated with resource
- \$ identifies durable relationships between entities
- \$ dynamic event-based relationships between entities
- \$ situates data in place and time

About what Grace learned from her work with an earthquake data model. Lots of changes led to headaches. Learners that data ought to be *context independent*, and that it ought to relate various events or actions to a core set of data. Events or actions in terms of object, agent, place. Grace talked about the influence of FRBR (Functional Requirements for Bibliographic Records) on the MIC data model.

Agnew: Drawbacks of Dublin Core and PBCore: no recognition of events and no recognition of FRBR hierarchy of work/expression/manifestation/item.

About developing a METS model, I didn't find good pre-existing model. Then there was a diagram with four levels: Origin of info (owner), Source, Preservation master, and Access copies.

About the MIC data model: diagram with object, agent, place. If these come together you have an event, e.g., a "film showing."

Agnew slide: Descriptive metadata

Ideal: DMD describes unchanging intellectual content'

real: DMD tends to mix descriptive and tech metadata  
DMD should be user centered, enable discovery

Another slide: Digital preservation enabling strategy

- \$ tech metadata documents object creation, tech characteristics, mediation space
- \$ continuous checking for integrity
- \$ tech metadata
- \$ source metadata

Rights metadata, in MIC it is not in the shape it should be in

- \$ should durably link to rights holder
- \$ identify permissions
- \$ identify citation for attribution
- \$ identify copyright, contracts
- \$ may involve data elements across metadata classes

About MIC's Mapping Utility. One motive was that LC wants to go between MAVIS and MARC. Then Grace listed a group of mapping projects, Jane Johnson oversees this. Mapping has been a quick process, the actions are in the hands of the collection owner.

The MIC cataloging utility is related to FEDORA repository. The utility offers a full METS implementation, and this encompasses MODS, MIX, and PREMIS metadata. The data captured there maps readily to MPEG-7 and PBCore. AMIA support is very active in this development. But the utility needs to implement some additional elements. And the utility is not yet sufficiently user friendly. Grace shows a screen grab of the input screen, which looked rather like the AV prototype project's METSmaker.

Grace on the importance of events in inputting, screens allow for this. Provenance data is event based, e.g., "create," "exchange," "loan," etc. Grace showed the export utility, permits mapping as part of its setup.

At the end of her slide show, Grace Agnew asked: how to relate MIC to PBCore? There is an export-as or display-as possibility, MIC could build a public broadcast portal.

Lisa Carter: endorses the MIC model for community building. Also notes the value of the resources at the MIC Web site for answering user's how-to questions and for finding service providers.

Lisa Carter talks about stations that are PBCore leaders: Wisconsin, Iowa, some stations in Indiana, New Jersey. We have been talking to the people at the stations, the challenge is putting these still-theoretical standards into practice. Station people hoped PBCore would be a tool, like a database, but there is an implementation shortfall. The stations use ProTrack, some use DOS-based ProTrack. They see PBCore and related ideas as things to support educational outreach, getting stuff to their users. How can PBCore help them? How can a DAM help them? Their ethos is to focus on the next production, what do we put on the air tomorrow, not about

archiving. They are used to the model of put a tape on the shelf and then later you can play it back. Looking at their metadata work, there is no consistency and it is unreliable -- we see series titles in the show title slot, show titles in the series titles slot, etc.

Laura Krasnow: Lots of these folks use FileMaker and want something that will work in that setting. But one bright spot is that once they actually start using a common database, we have stations “discover” multiple uses for their data: “we need this same information for our Web site, there is value to standardized data.”

Nan: part of this pertains to the relationship to PBS, to the handoff from the local show producer to the national distributor. The station systems need now to be integrated with BroadView (the PBS system to support the distribution interconnection).

Lisa: the high price of DAM systems is a deterrent for a lot of stations. In the absence of such things, there is no base system that connects to or demands metadata making.

Bea Morse: About AVID, problem of proprietary data. PBS has gotten AVID to output for BroadView, mostly timing data

MacCarn: the AVID tool is a post-production tool, not made for archiving, we have looked at what you can get out of AVID, and it isn't much. At WGBH, we don't do shot-listing at that stage.

Leah: we log tapes, but this is before the editing, before they get to editing.

Carl: is PBCore just for finished programs? Dave MacCarn: PBCore is for both program material and source material. We often go back to source material. Carl: will you describe at the scene or shot level? What about identifiers for scenes, segments, and programs?

Agnew: MPEG-7 can get to shots or segments.

MacCarn: Well, the identifier is a required field in PBCore, and there are some recommended practices. There is an identifier [value] element and another for the identifier source [type]. In our system, we use the Artesia number, prefixed by WGBH. We sometimes use PBS's NOLA (Network Operations Log Application) codes. Lisa Carter: at Kentucky, they also use shelf numbers. Carl: I can see that identifiers are a need in this metadata context. Agnew: we have an identifier for archives, and MIC then assigns an additional element for the individual work (item) within their own system.

Barbara: you need to have data that tracks relationships, this “is part of” or “is child of.”

Leah: at WGBH, we found that we have items with multiple relationships to multiple other items, one problem with Artesia was that they had a less complex notion about relationships, could not express the relationships we wanted to.

Barbara talked about MAVIS and its use at LC. All of this concerned the handoff of public TV content to the Library in the future. She explained that MAVIS is a proprietary system, with no relationships to any other systems, but it was designed specifically for media cataloging (it is not an asset management system) and is very rich in that respect. It is only used for media holdings, and has to link to other databases used throughout the Library.

Carl: How far does PBCore go regarding parametric techy data?

Grace: We had an AMIA committee led by Hannah Frost and Ian Gilmour. This list includes both analog and digital elements, and they found about 100 elements with this kind of data. The full set is not represented in any of these: MPEG-7, PBCore (about one half are in this), [others?]. It is an open question is what data elements are needed to preserve content.

Carl: Nan, please note that the NYU model repository in the NDIIPP project provides an opportunity to test the metadata in working environment. The NYU model repository will provide a platform not available at LC at this time.

Grace: I will note that MIC and NDIIPP may be competing for funding on this front, we want to work on this category of metadata too, need to get money for such work.

Joe Pawletko then presented a slide show on the Hemispheric Institute Digital Video Library (HIDVL). The focus was on mapping HIDVL metadata to PBCore. The HIDVL content is content that we commit to preserve in perpetuity. We get (a) our encoding from METS, (b) our descriptive metadata from MARC, (c) our tech metadata is strongly influenced by the schemas from the LC audiovisual prototyping project, and (d) structural metadata from METS. Regarding (c), we used the LC AV data structure mostly as-is, and extended it some, creating some additional elements. There were pointers to the xsd files at the LC avprot website.

Joe said that a crosswalk from their (i) videoTechMD plus (ii) MARC did map successfully to PBCore. All 9 mandatory PBCore elements are covered. Some optional elements are not covered or may not be covered, e.g., titleType, descriptionType, and three others. Grace said that these optional ones are probably covered, if you know how to extract everything from MARC.

Melitte Buchman: in the HIDVL project we also save edit decision lists, which we generate ourself so they are under our control.

Nan: how's the HIDVL system working?

Joe: well, it has been defined, not yet put to work.

Nan: what volume of content are going to deal with?

Melitte: 250 hours

Barbara: what wrapper?

Joe: Well, we will maintain metadata separate from essences, METS will track where things go, physical and cyber. Not going to use a wrapper in the sense of actual encapsulation.

Grace: Notes that all of PREMIS is incorporated in the MIC cataloging utility.

Joe: Our identifiers for cyber items come via DSpace, assigned by the system. For the physical items, you would go to the MARC records and see what the call number/shelf number is.

Bea Morse talked about PBS distribution. We are changing from tape acceptance to file acceptance (via I2, etc.), this is for completed programs. We are developing an integrated database, one piece of this is BroadView, but there is also Orion as a front end (data from producers, e.g., contact information, program description, some tech data). We are setting up [continuing to use] 16 distribution satellites for distribution to stations. We are distributing both NTSC and ATSC forms of the programs. We also have ScheduAll system for scheduling. We hope to have detail on within-program locations to permit local broadcasters to use segments or substitute elements, but it is not clear how often we will we get that level of detail.

Bea: about segment substitution for underwriting credits, etc.

Q: could this be supported by PBCore data?

Bea: no not at this time.

Leah: will Orion be replaced by BroadView?

Bea: no, both will exist.

**BREAK**

Nan: what are next steps? Who can take some of these activities on?

Dave MacCarn: At this meeting, there have been questions raised and it would be useful to go back over this.

Nan: We heard about some possible relationships with MIC, can this support this?

And about the need for identifiers.

Nan: regarding who will do the work: Leah Weisse is the metadata lead person for the NDIIPP project.

Leah: I was glad to learn about the work reported here, we can follow up on these things.

Especially the HIDVL work, and the MIC work.

MacCarn: today's conversation was focused on the NDIIPP project, but it also rolls back to the larger PBCore development.

Nan: what about PBCore and the interaction with the stations?

Gerry Field: meetings like these give us guidance for talking to stations. Remember that PBCore is a separate project from the NDIIPP project. If we conceive of extensions to PBCore, we need to figure who is the authority here.

MacCarn: if the NDIIPP project identifies some gaps in PBCore, what do we do with this?

Field: go back to the committee . . . but in large part, that is me, on a part time basis.

MacCarn: no funding from CPB to look at extension issues, CPB funding is about advocacy and training and maintenance, not extension development.

Field: interested in working with vendors about implementation, to be able to demonstrate PBCore in action. But there is very limited availability of funds.

Laura Krasnow: there are 200 people on the PBCore listserv. They are waiting for leadership and direction.

Lisa Carter: People on the listserv have broader interest than PBCore, they care about DAMs and related issues.

Group: Discussion of relationship with MIC, mutually beneficial development. Possible working connection to New Jersey and PBCore development.

Nan and MacCarn: Discussion of extended PBCore rather than a new structure. And it needs a schema.

Carl: Comment: I can see PBCore as an SIP metadata block, uncertain about PBCore as an AIP metadata block.

Grace: reminds of the importance of expressing FRBR relationships; I see a need for this in PBCore. This relates to the problem of breaking the immediate user context from long term needs. May need PBCore records for all levels: work/expression/manifestation – in order to express the relationships.

Issues raised for follow-up and/or next steps:

- Analysis of Pamela’s report on metadata fields (report is forthcoming.) and gaps in PBCore. (Leah)
- How does this feed back into localization needs to PBCore? Where do we take changes?
  - PBCore Governance – who is in charge? (Gerry to reconstitute prior project group plus others; clarify current decision-making.)
  - what might we do to set up new governance structure tied to repository?
- Demo of PBCore to be presented at PBS Tech conf (Gerry)
- Post discussions and substantive issues to PBCore list (Nan; Laura; Lisa)
  - Better communications about what PBCore is and is not
  - Start to reframe the discussion to help stations understand it
- Look at PBCore website for how to expand
  - What is PBCore?
  - FAQ
  - Resources, links, experiments etc,
  - Check status of web site – who is maintaining it, how to expand it, etc.
- Consider idea of Portal for PBS and Local Broadcasters as a collaborative space, with links to other useful resources, etc. (Lisa has initial list; follow up by MIAPP interns to scour web for resources.)
  - to be linked to MIC, complement each other; post sample records on MIC, etc.
  - what does a basic record look like?
- Data model behind PBCore – not just a data dictionary, does this need to be developed?
  - Workshop to “FRBRize” PBCore?
  - Structure map to break out components and their relationships?
- Map PBCore top MAVIS (MIC? Barbara?)
- Organize a session at PBS Tech conference for Media Asset Managers, both radio and tee vee (Lisa has list)
  - Discussion of “What do you need?”

Adjourn.