Program Service Data (PSD) and PBCore

Initial Mapping and Recommendations

WGBH National Center for Accessible Media (NCAM)

July 2006

Prepared by:
Gerry Field
Marcia Brooks
Alan Baker
abaker@visi.com
(707)-239-0200

Introduction

As part of the CPB-funded "PSD Consortium" project being managed by Public Radio International (PRI), the WGBH National Center for Accessible Media (NCAM) has been engaged to provide input, assistance and an initial mapping of the proposed PSD metadata elements to the existing structure of the Public Broadcasting Metadata Dictionary (PBCore).

This report summarizes related activities performed by NCAM from November 2005 through July 2006, and provides a mapping of preliminary PSD elements to PBCore as well as recommendations for follow-on activities.

Definitions

Program Service Data (PSD) is a text stream made possible by digital radio broadcasting technology that allows stations to broadcast text, such as artist/title information, in synch with digital audio broadcast. Listeners see this text on their HD Radio™ receiver displays. The PSD Consortium is charged with developing the basic elements of a common data structure for PSD, and to provide guidance for voluntary implementation of PSD.

PBCore is intended as a common metadata and cataloguing resource for public broadcasters and associated communities. With funding support from CPB, the Public Broadcasting Metadata Dictionary (PBCore) was created through a number of years of collaboration involving representatives of local and national television and radio organizations, numerous constituencies, and related disciplines.

PBCore Version 1.0 was published on the PBCore Web site on April 1, 2005. As part on an ongoing CPB-funded project to promote adoption and implementation of PBCore, a PBCore XML Schema was developed and published for review on April 17, 2006. The final PBCore XML schema is expected to be published Fall 2006.

NCAM (The National Center for Accessible Media) is a research and development facility at WGBH in Boston primarily dedicated to the issues of media and information technology for people with disabilities in their homes, schools, workplaces, and communities. NCAM has considerable expertise in technical standards development as well as years of experience providing technical and system support to the public broadcasting system. CPB has contracted with NCAM to provide management and support of PBCore advocacy activities through April 2007.

Related Activities

From November 2005 through July 2006, NCAM provided consultation services to the PRI PSD Consortium project, as follows:

- Weekly email, telephone, and in-person contact with PSD Consortium manager Tim Halle.

- Regular participation in PSD Consortium bi-weekly telephone conference calls.

- Weekly follow-up activities providing liaison between the PSD Consortium and PBCore, including exchange of draft documents and related information.

- Monthly planning and coordination meetings on schema and mapping strategies with PSD Consortium staff and the staff of PRX.

- Participation in the PSD Consortium workshop at the Integrated Media Association's New Media conference in Seattle, WA (March 2006).

- Participation in the PSD panel presentation at the Public Radio Engineering Conference in Las Vegas, NV (April 2006).

- Preparation and delivery of draft report. (June 2006)

- Draft semantic mapping of PBCore and PSD for PSD project team review.

- Consulted on revised PSD fields with consortium manager Tim Halle.

- Final mapping of PBCore metadata dictionary to proposed PSD.

- Preparation of final report. (July 2006)

Background

The common goal of PBCore, PSD and other related metadata standards efforts is to provide broad system integration of data collected to describe, organize, announce, and use collections of media assets.

PBCore was designed as a general-purpose metadata structure to be used across the variety of public broadcasting (television, radio, Web) distribution systems, as well as in related business and management systems.

On a basic business level, there is an obvious benefit in minimizing the number of times data entry is performed for any given asset, and to reduce or eliminate duplicate effort. Reducing the number of steps required to create, maintain and update information about individual series, programs, Web pages or other media assets has a potential cost benefit resulting from streamlined workflow.

As changing technology and business practices support and create new distribution methods, there is an increasing need to break down traditional "information silos" within organizations and industries, and enable effective and useful exchange of information among systems, people, and organizations.

PBCore was designed to provide a common basis for metadata structure and exchange among public broadcasting communities. The 48 metadata elements defined in PBCore Version 1.0 represent a 'core' set of descriptors for content, intellectual property, and instantiation (format) needs most commonly found across public broadcasting systems. By providing common semantic elements, definitions, and controlled vocabularies, PBCore hopes to provide a foundation for the design of databases, applications and systems that will enable more effective exchange of information for current and emerging system needs.

But creation of common terms and language of a metadata dictionary, while critical, is only a first step toward a goal of broad system integration. To be used effectively in media distribution systems, metadata must have rules and structures defined in terms that can be understood both by people and technology. And while the core structures must be clearly defined, there must also be enough flexibility in the metadata system to allow for modification and adaptation.

XML

The use of XML (eXtensible Markup Language) is the most common method in current practice to provide data structures that are human-readable, machine-readable, and built to accommodate change. Virtually all professional broadcast and new media equipment vendors provide the ability to import and export data in the XML format.

Both PBCore and PSD have agreed to use XML as the language for data interchange. A 'PBCore record' will be an XML file providing description of a media asset in PBCore terms, in the XML language. A 'PSD record' will be an XML file providing description of Program Service Data, in PSD terms, in the XML language.

However, another structure – an XML Schema – is required to provide guidance to both people and machines on how to use the data included in these files. A PBCore XML Schema will provide the rules and structures to be followed when using a PBCore XML file.

There are specific XML rules that must be followed in the creation of a 'valid' XML Schema, and the Schema must be made readily available on the Web for reference by people and machines. In XML parlance, this 'namespace' is the location (URL) where the related XML Schema document can be found, and it must be referenced in the header of a related XML file for it to be considered valid.

During this phase of the PSD Consortium activity, the PBCore project was developing the PBCore XML Schema, which was posted for review and comment on the PBCore Web site on April 17, 2006.

A copy of the document, PBCore XML Schema: An Overview is attachment A to this report. It gives detailed narrative description of the Schema structure and its use, as well as the PBCore XML Schema itself and a sample PBCore XML file.

Once an XML Schema is created and made available, it is possible to share and use it either as a 'rulebook' in using a related XML document (file), or as a basis for creating further data structures.

For example, the PBCore XML Schema can be used as a definitive reference on how PBCore elements and terms are defined and related to each other. It is also possible to use the Schema to refer to and use individual PBCore elements within another XML Schema.

Specific rules on exchanging information based on different Schema can be defined in a number of ways.

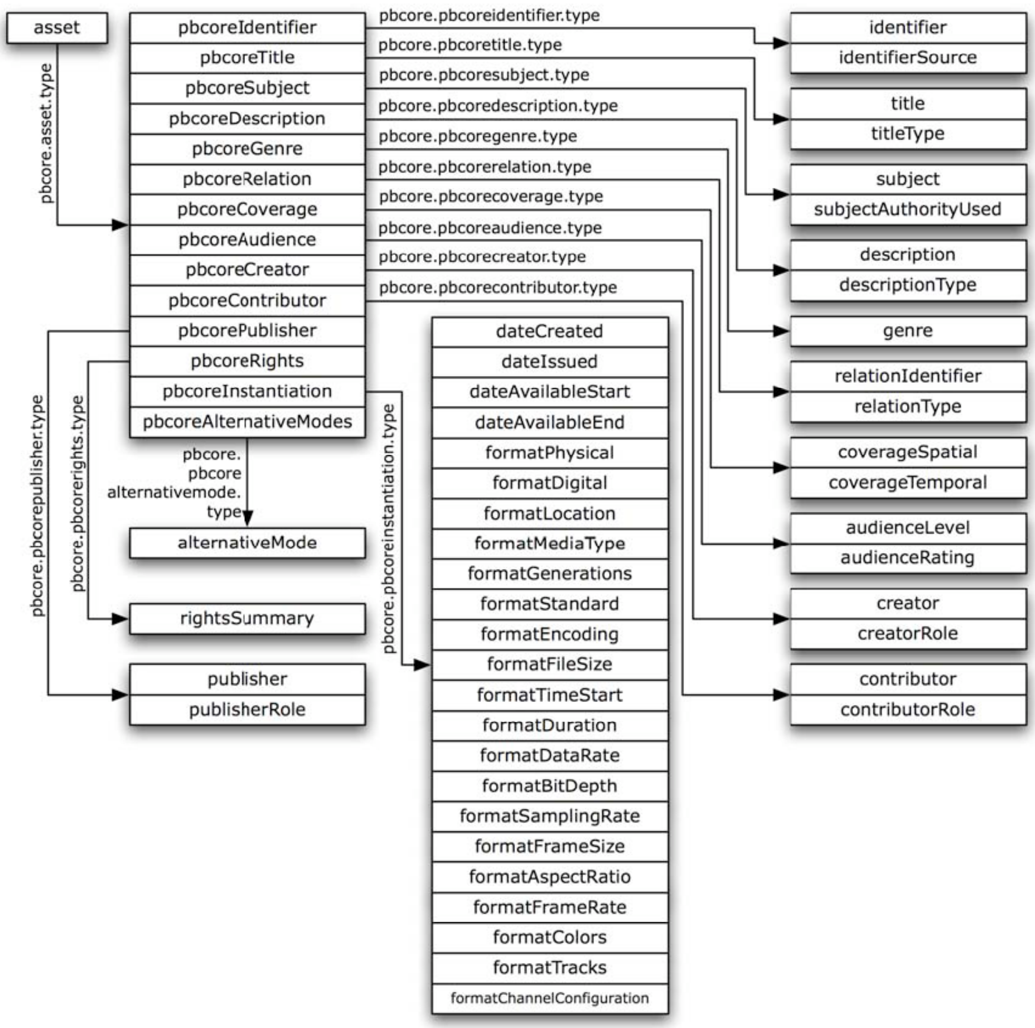
One might be to provide a simple 'mapping' from one Schema to another. This is somewhat similar to what you have done if you've ever transferred your email address book from one application to another. Typically, you're presented with two lists of terms – one for each application – and asked which term from the first list matches which term from the second. This can be a tedious process, and typically involves some trial and error. But once completed, you can transfer the data knowing that what started out as a contact's work phone number shows up in the correct place in the new application.

XML provides specific structures to support these kinds of mappings, called XSLT (eXtensible Stylesheet Language Transformations) a language for transforming XML documents into other XML documents. So, once valid XML Schema are created, enabling the creation of valid XML documents, an XSLT can be created that will provide the rules and styles needed to transform one type of document to the other successfully. These Schema and XSLT can be made available to database designers and managers, and to equipment and software manufacturers, to provide a common reference for adoption of PBCore and PSD. Another approach to utilizing XML Schema is through RDF (Resource Description Framework) structures. Simply put, a variety of related XML Schema and associated namespaces can be referenced in a 'stacked' structure within a document, allowing terms and elements from a variety of data structures to be used.

The RDF approach provides a distinct advantage if it will allow "extensions" to a known data structure or specification without requiring modification or revision to it. For example, use of RDF could allow a metadata document to be authored referencing specific elements from PBCore Version 1.0, as well as extended (new, modified) elements that might be required for a specific application such as PSD, without requiring changes to PBCore 1.0.

PBCore Data Structure Overview

The PBCore Metadata Dictionary elements and the PBCore XML Schema are posted on the PBCore Web site for your reference, along with substantial detailed information and tutorials. Included here for reference is a graphical view of the PBCore XML Schema, indicating the current data model of the PBCore elements:



Suggested PSD Field Descriptions

A document describing the suggested PSD field descriptions was posted on the PSD Consortium Web site in May 2006. The revised, and final suggested field description documentation is available on the PSD Web site. This resource should be referenced for the full description and explanation of the intention of each field.

Mapping PBCore to draft PSD Fields

Summary of Approach

In November 2005, the PSD Consortium began to determine a proposed set of PSD elements, based on a survey of user needs and the constraints of the technical systems involved.

Representatives from the PBCore development team participated in the IMA and PREC conferences, which provided significant opportunity for comment and guidance from end users and vendors.

The semantic mapping of proposed PSD fields to PBCore V1.0 fields began in late May 2006. After an initial period of review, and based on feedback from the PBCore team, PRX, and vendors, a revised set of fields was proposed on July 10th, and the mapping that is attached to this document is based on the revised set of Proposed PSD fields V0.15. This source PSD data is included as the second tab in attachment B to this report.

Mapping Documentation Guide

Attachment B, Initial Mapping of PSD Field Descriptions to PBCore, shows the most likely matches of equivalent fields between PSD and PBCore 1.0. There are three categories of mapping: Direct, Partial, and None.

- **Direct** = there is a direct and obvious mapping to PBCore, and 64 characters should be sufficient space to accommodate direct exchange of information between these two fields
- **Partial** = possible semantic mapping but there are data limitations, or other exceptions
- **None** = no mapping possible in current PBCore. Recommendations will appear in notes field

Comments are included in the mapping document, which explain any conditions that would have to be met for the data to be exchanged freely between PSD and PBCore fields. This mapping is as comprehensive as possible with the PSD fields in their proposed state.

Summary Comments on Proposed PSD Fields

While the mapping of PSD from/to PBCore is the primary directive for the representative from the NCAM/PBCore project, it was also requested that the NCAM/PBCore project representatives give some general observations, and recommendations, to help ensure the successful development of the PSD specification.

The recommended PSD fields form a rich set of data that can be presented to listeners via a number of emerging channels. The most obvious of these channels is digital receivers for terrestrial digital broadcast signals. Along with the first generation digital receivers ability to display Program Associated Data (PAD,) there are other existing outlets for this very basic text information, such as RDS on some analog FM receivers. The generation, delivery, routing, and display of PAD/PSD data will become ever more complicated as devices and new channels proliferate. As part of the PBCore project we have seen how demands for this type of metadata has continued to outpace the industry's ability to accommodate the emerging needs. It is important to consider a mechanism by which the PSD specification can be managed and updated so that it can grow and change as needs dictate. This will greatly enhance the PSD specification's usability, and likelihood that it be adopted by a broad user base.

Mapping Overview

Of the 54 proposed PSD fields, 22 map directly to PBCore with some data limitations, 19 have a possible partial mappings, and the remaining 13 have no obvious mapping. Attachment B gives detailed mapping, and comments on each field.

Direct

<SeriesTitle>
<SeriesHost>
<SeriesUnderwriterName1>
<SeriesUnderwriterName2>
<SeriesUnderwriterName3>
<PgmID>
<PgmHost>
<PgmTitle>
<PgmUnderwriterName1>
<PgmUnderwriterName2>
<PgmUnderwriterName3>
<PgmAudioFileLocation>
<SegType>
<SegTitle>
<SegGuest>
<SegHost>
<SegUnderwriterName>
<SongArtist>
<SongConductor>
<SongComposer>
<SongSoloist>
<SegAudioFileLocation

Partial

<SeriesDescription>
<SeriesGenre>
<SeriesComment1>
<SeriesComment2>
<PgmGenre>
<PgmDescription>

<PgmComment1>
<PgmComment2>
<PgmAudioFile>
<SegID>
<SegDescription>
<SegComment1>
<SegComment1>
<SegComment2>
<SongTitle>
<SongTitleLong1>
<SongTitleLong2>
<SongAlbum>
<SegAudioFile>

None

<SeriesIsSeries>
<SeriesID>
<SeriesUnderwriterTag1>
<SeriesUnderwriterTag2>
<SeriesUnderwriterTag3>
<SeriesHasPgmInfo>
<PgmNumber>
<PgmUnderwriterTag1>
<PgmUnderwriterTag2>
<PgmUnderwriterTag3>
<PgmHasSegInfo>
<SegNumber>
<SegUnderwriterTag>

For a discussion of which fields, if any should be included in the PBCore dictionary, see Recommendations, Next Steps section below.

Gap Analysis

Intended Audience, and Use

There is potentially a lot of shared information between PSD and PBCore. However, it is important to note that the intended use of PBCore metadata and PSD data is very different. PSD is specifically aimed at displaying messages to listeners to, or users of, produced content, while PBCore was designed to describe rich media assets for content exchange.

Another key distinction to note is that, although there are exceptions, there is almost always a one-to-one relationship between a PBCore metadata record and a media item. Each PBCore record typically describes a single instance of a complete program. The splitting of a single program into multiple segments creates a challenge for PBCore. There is currently no obvious method for describing multiple program segments within a single PBCore record. Potentially each discreet segment of a program could have its own PBCore record that could then be concatenated to produce a master data set that would describe a single complete program, including all of its discreet segments, in sufficient detail. This is a known deficiency in PBCore, and is not a problem that is easily solved.

Terminology

It is important for producers, broadcasters, and vendors to understand the terminology used. There are conflicting terms in use within our existing system that will carry forward into the future. Future documentation should clarify these terms for the audience.

They key distinctions that need to be defined for PSD user in the public radio sphere are the terms Series, Program, Episode, and Segment.

PBCore and PSD use the following hierarchical structure:

- Series
    - Program
        - Segment
        - Song (In PSD is same as segment, doesn't exist in PBCore)

A Segment is the smallest discreet unit of audio. This may belong to a program or stand on its own. And a program may be part of a larger series or not. These terms are used in a similar fashion in several key metadata standards including MPEG-7 and SMPTE.

The reason this is important for radio personnel, is that the soon-to-launch PRSS Content Depot uses slightly different terminology. The semantics are the same but their hierarchical structure is as follows:

- Program
    - Episode
        - Segment

Also it is important to note that the term segment can be used to indicate a single feature, or piece, but in the generic terms of radio, including PRSS terminology, it is simply the most granular unit of time in a program. Segment 3 of ATC may well have three features, all with different reporters/hosts, and multiple guests/interviewees. This also affects how the Host and Guest fields are used in PSD, and how they might be displayed at the proper times within the playback of a segment is an open question.

Structure of PSD

The current structure of PSD information appears to be fairly flat and not hierarchical. Depending on how these records are structured, and how this data might be stored and used, other than a straight pass-through to a receiver display, there may be a need to introduce a hierarchical structure to the PSD record. Keeping underwriter names, and their associated tag text bound together might be a necessity in the future.

Field Labels and Implementation Guidelines

In the process of training people to use the PBCore metadata dictionary it became clear that ambiguity is the enemy of quality metadata being gathered. Within PSD there are several fields that have multiple uses, or are intended to be flexible in what kind of data they are intended to capture/hold. This may be an acceptable approach with proper documentation and training. It is not safe to assume, however, that even the more motivated users will spend enough time with the application profile, or users guide, to implement these fields consistently across the system.

Specifically the transfer of data between the Description fields in PSD and PBCore becomes difficult because both metadata dictionaries allow for some flexibility in what is captured in these fields. There are observations in the mapping document about this as well but in the long term it could be wise to update the description element in PBCore to include a mechanism for storing PSD descriptions (see recommendations below).

In order to create a system that enables data to flow through the system to consumer devices without editorial oversight it is of vital importance that the originator of the PSD apply it properly. And because of that, it is important to create a set of fields that leave as little room for improvisation as possible on the part of the producers who will be populating the PSD information.

Repeatable, and Segmented Entries

The desire to support more detailed or lengthier entries than will fit in a single 64 character field leads to a fairly cumbersome set of field labels, such as the three fields defined for song titles.

- Song Title
- SongTitleLong1
- SongTitleLong2

This necessitates the need to build logic into the software that manages the broadcast/playback of this data to support the reassembling of these fields into a sequential set of displayed text strings. And by defining three fields you are limited to those three. In addition to being cumbersome for users to populate these fields easily, this also builds on top of a standard that assumes a 64 character limit, which likely will disappear soon enough as PSD services mature. Why not propose that these fields can contain any reasonable amount of data, and define the proper implementation for software is to carve the text up into chunks that are 64 characters, or less, and to display them in sequence? This way the receiver manufacturers do not have to support an arbitrary, but limited, number of data fields. They can support funneling all of this data in to a single Song Title field.

By doing this you have, to some extent, future-proofed the design but also not locked the standard into accommodating limitations that won't be around for long. This approach also enables presenters to use the same data to feed different channels this same information custom formatted for that channel's display capabilities.

This same challenge exists for the comments field, which was split into two fields each for Series, Program, and Segment levels.

Ideally, the following fields would allow more than 64 characters, and rely on software to parse the text out in appropriate sized chunks for the display mechanism/channel involved.

- SeriesComment
- PgmComment
- SegComment
- SongTitle

A related but different challenge comes from fields that may need to include more than one discreet bit of information. In general it is best to have a single field contain one discreet piece of information. Most obvious here is the situation where you have multiple hosts, guests, or other contributors to a program. With a panel discussion of just a few people, you could easily exceed the sixty-four character limitation. There are two obvious solutions to this problem. You could allow a larger amount of data to be stored and have the software carve it into smaller chunks, as proposed above. The better solution is to allow certain fields to be repeatable. With classical music, the SongArtist, and SongSoloist fields can easily exceed 64 characters in fields containing multiple names. There are workarounds to this that users will invent on their own, like tossing the information into the Description or Comment fields, which could possibly hinder future functionality.

Although you could put some or all of the artists/performers involved into the two comment fields, you then lose the ability to have machines be smart enough to know when a name is a performer, or possibly mentioned as a topic of discussion.

Either way, the user guide should dictate how names are to be stored. The common- Last, First -format works well if there is a single entry per field. There is a need to have a known, consistent, separator between names for fields that contain more than one name.

Ideally the following fields would be repeatable and contain only a single entry per field:

- SeriesHost

- SeriesGenre
- SeriesUnderwriterName
- SeriesUnderwriterTag
- PgmHost
- PgmGenre
- PgmUnderwriterName
- PgmUnderwriterTag
- SegGuest
- SegHost
- SegUnderwriterName
- SegUnderwriterTag
- SongArtist
- SongConductor
- SongComposer
- SongSoloist

Part of the thinking behind this recommendation to allow these fields to be repeatable is that at some point in the future, you should be able to tell your receiver to grab content in a certain genre, or with certain contributors involved. It is less likely that this functionality will be available with musical selections due to restrictions on store-forward scenarios with music. But certainly we should see the ability to grab traffic and weather reports for retrieval. Storing the hosts, guests, artists, etc as discreet pieces of information should enable this "advance search" functionality, which will be much more difficult to facilitate if these names are stored in a block of freely formatted text.

Recommendations, Next Steps

PBCore, PSD Description Fields

One deficiency that has become obvious as we look at efforts to implement PBCore, and to create similar semantic mappings and crosswalks with other metadata standards, is that the Description element was too loosely defined. While the limited number of characters available for PSD data poses a hurdle to sharing data between these fields, there would also be problems mapping this data because PBCore does not differentiate on a granular enough level the type of descriptive data to facilitate the mapping of this data to PSD.

The current Description field in PBCore has the following controlled vocabulary choices available for the sub-element descriptionType:

- Abstract
- Table of Contents
- Other

There would be no structural change necessary to PBCore to accommodate a larger set of choices under descriptionType. Depending on the final release version of the PSD fields, we could recommend the addition of the following description types:

- PSDSeries
- PSDProgram
- PSDSegment

In this way we could impose the strict data limitations on these fields to ensure the data would appear on various devices as entered. These same descriptions are often needed, by Web and other outlets, for displaying metadata related to programming.

Potential Additions to PBCore

Other potential additions to controlled vocabularies to ensure direct mapping from PBCore to PSD

- Add to sub element TitleType the choices of Song and Album.
- Add to sub element ContributorType the choice of Soloist.

Recommendations for the process for updating version 1.0 of the PBCore Metadata Dictionary, and sustainability, will be included in NCAM's final report to the CPB, due May 2007.

XML Schema, XSD

Once the PSD fields are published in an initial official release version, it will be easier to determine the ease of data exchange between PBCore and PSD data structures. When the PSD fields are in their initial release version, there will be another development step required to facilitate the easy sharing of relevant data from PSD to PBCore, and back. A valid PSD XML Schema needs to be created. Both the PSD and PBCore project then need to draft, and agree upon, an XSLT document for the interchange of PSD and PBCore data. This will provide vendor and producers the ability to validate their published PSD files.

It is important that test implementations be conducted and required to meet satisfactory proof of "round-tripping", i.e. from one structure to the other, and back again. Resulting data must be valid and match the original source.

Federated Entity and Content ID Registry

As we move forward in the multi-channel, non-real-time mediascape we are increasingly finding ourselves lacking a good technology to match metadata with content. And making that content, or the metadata that describes the content, searchable, or viewable, by partners is increasingly necessary. Having a federated system that assigns producers, distributors, broadcasters, and our partners a unique ID, and also provides a mechanism by which producers can assign globally unique identifiers to all the content they produce, would enable a whole new set of functionality to our increasingly interwoven production and distribution network. Because a program or segment ID stored one system will not likely match the ID assigned to that content if it is moved to a partners storage infrastructure, it becomes impossible to track versions, determine redundancies, and to match metadata to essence, if they are not wrapped within the same package, and stored on the same system.

Support and Future Development

Although the long-term development and support mechanism for maintaining PSD and PBCore dictionaries is yet to be determined, it is assumed that there will need to be ongoing support and development for these dictionaries.

As part of the current PBCore advocacy project, (final report due to CPB May 2007) there is an opportunity to document recommendations for future changes to PBCore. It is still too early to make recommendations for precisely which, if any, of the PSD fields might actually be considered part of the core metadata set that would facilitate exchange but there is a case to be made for including at least a basic set of PSD data in the PBCore metadata dictionary. However, it should be taken into consideration how these two metadata dictionaries might interact. This might determine what if any fields are included in the PBCore set. It must serve the users needs to have a subset of this information in PBCore to warrant the inclusion.

The fragmenting of PSD to a supported sub-set within PBCore, with the rest existing outside of PBCore, might be more difficult to support than if we build a system where by PBCore, and PSD remain discreet, and PSD is seen as a defined, and supported, extension of PBCore. The mechanism by which we might do this has yet to be determined but it is assumed that there will be a mechanism developed to extend the PBCore metadata set to enable industry, or channel, specific functionality.

The fact that there are over twenty fields that map easily to PBCore, and provide the ability to harvest series titles, program titles, segment titles, host names, guests and other contributors, as well as underwriter names, means that a good deal of the most important data is already available to PSD from PBCore. With some simple updates to PBCore, like the addition of new options in controlled vocabularies, it could support the storage of series/program/segment descriptions, song titles, album titles, and other relevant information. At that point the most important information would be shared between the two dictionaries.

It is also recommended that the PSD consortium work with stations who are

engaged in developing their PSD services to test the assumptions made in this phase of the development, to further refine what data is needed for display, and how to support its delivery.
ATTACHMENTS

Attachment A

Draft PBCore XML Schema and documentation

Attachment B

Initial Mapping of PSD Field Descriptions to PBCore