

Feasibility Study for the American Archive of Public Broadcasting

Linked Data Implementation for National Educational Television (NET)

By Christopher Pierce, NET Project Catalog Librarian, Library of Congress

November 2018

Thanks to: Andrea Leigh (Library of Congress), Rebecca Guenther (Library of Congress), Kirk Hess (Library of Congress), Rachel Curtis (Library of Congress), Sadie Roosa (WGBH), Casey Davis Kaufman (WGBH)

Executive Summary

Between the 1950s and the early 1970s, the National Educational Television (NET) network distributed educational programming across the nation, some of which was produced by NET and some by local affiliates (such as WGBH or KRMA) either directly for NET itself or distributed by NET after local broadcast, with additional content imported from the CBC, BBC, and other foreign broadcasting and production companies. The volatility of this historical time period and the range of international and local content makes NET content a fascinating time capsule. However, much of the material has long been considered as hidden special collections of rare and unique content due to difficulties in exposing content for access. What descriptive metadata exists remained isolated in finding aids, card catalogs, and inventories.

This feasibility study for Linked Data implementation examines the NET Collection Catalog Project, a collaboration between WGBH Educational Foundation and the Library of Congress (“the Library”) funded by the Council on Library and Information Resources (CLIR). The study’s goal is to examine the feasibility of a Linked Data strategy for exposing descriptive metadata in NET catalog records to the Web as Linked Data.

Linked Data in library and archival description is an opportunity at improving discovery, enhancing data exchange between institutions, and exposing collections to the Web. Yet implementing Linked Data in libraries and archives is a challenging proposition as it represents a difficult conversion of catalog records and finding aids designed to be interpreted and exchanged as whole records to an exchange format where the emphasis is on machine-actionable discrete pieces of data. This feasibility study examines strategies for converting a dataset of records describing NET content from XML serialization to the Resource Description Framework (RDF), the serialization used for Linked Data exchange. The XML framework for the Library’s media asset management system, MAVIS (Merged Audiovisual Information System), is examined in relation to the Public Broadcasting Metadata Dictionary (PBCore)¹, for distribution of audiovisual content, with BIBFRAME², the Linked Data model for bibliographic description and access.

The Library’s Motion Picture, Broadcasting and Recorded Sound Division (MBRS) produced four main datasets based on records curated as representative of NET content in both fiction and non-fiction works to be shared with the public. One dataset is serialized as MAVIS XML (and is basically a direct export from the Library’s media asset management system, MAVIS, in an XML serialization). One dataset is serialized as PBCore XML, which is the result of a transformation using a stylesheet designed by MAVIS developers. One dataset is serialized as MARC XML, which is the result of a transformation using a style sheet designed by MBRS for

¹ Public Broadcasting Metadata Dictionary (PBCore): <http://pbcore.org/>

² Bibliographic Framework Initiative (BIBFRAME): <https://www.loc.gov/bibframe/>

this project. One dataset is serialized as BIBFRAME-standardized RDF, which is the result of a stylesheet shared by the Library's MARC Standards and Development Office.

This feasibility study discusses four phases of this conversion process: 1) **Data Modeling**, 2) **Evaluation**, 3) **Crosswalking**, and 4) **Publishing**.

The **Data Modeling** phase compares and contrasts different standards as part of a process of examining the impact of each standard during the conversion process. Data modeling works is examined in relationship to different metadata standards and in relation to the Entertainment Identifier Registry (EIDR),³ a universal unique identifier system for movie and television assets.

The **Evaluation** phase examines the MAVIS-XML exports from the Library's media asset management system and the stylesheet used by MBRS to transform metadata for export into PBCore to facilitate exchange with WGBH and other partners in the public broadcasting community. This evaluation phase examines the functionality of various fields and the potential for Linked Data exposure in unstructured text fields (by using regular expressions to roughly match certain patterns in these fields), while also identifying potential issues with the export from the MAVIS to PBCore stylesheet.

The **Crosswalking** phase discusses mappings between different standards to prepare for the transformation of NET collection data from XML serialization to RDF serialization and to prepare for possible extensions to EBUCore properties where there are gaps between PBCore and BIBFRAME. EBUCore⁴ is a Linked Data model designed by the European Broadcasting Union to describe audiovisual resources in all the stages of their lifecycle, including production, distribution, marketing, and archiving.

The **Publishing** phase discusses a stylesheet transforming a PBCore-standardized dataset of NET records to MARC-XML that could then be transformed to BIBFRAME using stylesheets made available by the Library's Network Development & MARC Standards Office. Different publishing methodologies are also examined in the discussion of the publishing phase.

MBRS also produced a few derivative datasets of NET access points reconciled with Linked Data services to provide Uniform Resource Identifiers (URIs) that enable Linked Data to be connected to data in other datasets across the Web. These datasets include subject, genre, credits, and names as subjects. The credits and names as subjects are notable for having been reconciled with both Library of Congress Linked Data Service⁵ and Wikidata⁶; names used in these two

³ Entertainment Identifier Registry (EIDR): <https://eidr.org/>

⁴ European Broadcasting Union Metadata Set (EBUCore): <https://web.archive.org/web/20160304060240/https://tech.ebu.ch/docs/tech/tech3293.pdf>

⁵ LC Linked Data Service: <https://id.loc.gov/>

⁶ Wikidata: https://www.wikidata.org/wiki/Wikidata:Main_Page

datasets that do not have uniform resource identifiers (URIs) are marked for catalogers to establish in the Library of Congress Name Authority File (LCNAF) in the future.

The importance of transforming records to Linked Data is to provide more comprehensive coverage on important topics of high interest to scholars and the public, and to provide greater context to resources traditionally isolated from the public domain. The primary objective of this study is on discoverability and access alongside related materials of interest to maximize impact on the creation and dissemination of knowledge as a public good.

Contents

<i>Introduction.....</i>	<i>5</i>
<i>Why Linked Data?.....</i>	<i>7</i>
<i>Phases of Linked Data implementation</i>	<i>11</i>
Phase 1: Data Modeling	12
<i>Titles, works, and EIDR.....</i>	<i>14</i>
<i>BIBFRAME, PBCore, MAVIS, and MARC.....</i>	<i>19</i>
Phase 2: Evaluation	23
<i>MAVIS access points, description, and uncontrolled text</i>	<i>23</i>
<i>MAVIS to PBCore XSLT stylesheet</i>	<i>34</i>
Phase 3: Crosswalking	41
<i>Crosswalking observations.....</i>	<i>48</i>
<i>EBUCore possible extensions.....</i>	<i>50</i>
Phase 4: Publishing	52
<i>NET Collection dataset conversion process.....</i>	<i>60</i>
<i>Authorities</i>	<i>60</i>
<i>Conclusion</i>	<i>63</i>
<i>Bibliography</i>	<i>65</i>

Introduction

The National Educational Television (NET) Collection Catalog Project evolved out of the American Archive of Public Broadcasting (AAPB), a collaboration between WGBH Educational Foundation (“WGBH”) and the Library of Congress (“the Library”) to preserve and make accessible significant historical content created by public media and to coordinate a national effort to save at-risk public media before its content is lost to posterity. Funded by the Council on Library and Information Resources (CLIR), the NET Collection Catalog Project involves the creation of a national catalog of descriptive records documenting extant titles distributed by NET, public media’s first national network. The NET Collection Catalog Project aims to provide descriptive records of extant holdings of NET distributed and produced content held in public archives. As part of this effort, CLIR provided resources to hire two catalogers that resulted in the description of 7,485 NET titles held by the Library with acquisitions dating back to NET’s beginnings in the 1950s.

Between the 1950s and the early 1970s, NET distributed educational programming across the nation, some of which was produced by NET and some by local affiliates (such as WGBH or KRMA) either directly for NET itself or distributed by NET after local broadcast. Additional content was imported from the Canadian Broadcasting Corporation, British Broadcasting Corporation, and other foreign broadcasting and production companies. The scope of the Library’s collection ranges from the mid-1950s to the 1968-1971 period during which NET distribution was transitioning to the Public Broadcasting Service (PBS).

The descriptive records created at the Library’s Motion Picture, Broadcasting, and Recorded Sound Division (MBRS) relied predominantly on the existence of secondary sources that were held locally (such as paper inventories and spreadsheets) or made available by WGBH. A key resource was a document created by PBS, which records the titles, broadcast dates and summaries of NET programming, as well as instructions to affiliate stations regarding the distributed NET content and its promotion. This resource, available on microfiche, is held by WNET and the Library. During the first months of the project, staff at WNET and WGBH transcribed this document into a Microsoft Word document and made it available to the rest of the project staff. This document was key in determining the content held by the Library and its subsequent descriptive cataloging. WGBH also used this document to create their own descriptive records in the Archival Management System (AMS). The Library’s completed descriptive records were shared with WGBH as exports standardized to the PBCore XML metadata schema which was designed to facilitate the exchange of descriptive records between public broadcasting stations.

Even with a standard metadata schema to exchange descriptive records, challenges reconciling these records were evident from the start due to differences in descriptive policies and procedures across organizations. A notable example is the reconciliation of Library of Congress

established name, subject, and genre/form authorities used as access points with PBCore's data dictionary that lacks a consistent methodology in the use of controlled vocabularies for the purpose of collocation, disambiguation, and indexing. To complicate matters more, matching series and program titles was difficult due to inconsistencies resulting from a lack of agreed upon title construction procedures. It was not uncommon to name the same work by a variety of different titles dependent on both primary and secondary sources consulted. For example, a program highlighting a 1968 visit by Julia Child to the White House resulted in the following titles:

White House Red Carpet with Julia Child

NET Festival. White House Red Carpet with Julia Child

White House Red Carpet

Julia Child at the White House

It is likely that similar challenges will be replicated as affiliates and institutions share their NET holdings with WGBH. Individual institutions have their own metadata environments, policies, and institutional practices that might not readily fit under the PBCore guidelines without some work on the part of staff to crosswalk between different standards and metadata environments, reconcile authorities, clean up messy metadata, and normalize titles.

One possible solution to these challenges is to implement a method where metadata can be more easily integrated between different metadata environments, where metadata that exists on the Web or in other external environments can be re-used, and where smarter automation is possible. This is one area where a Linked Data strategy for the NET Collection Catalog can be valuable.

Linked Data has long been advertised as supporting metadata exchange where anybody can say anything about anything (commonly called the AAA standard for Linked Data) (W3C, 2002; Allemang, D. & Hendler, J. A. 2012), allowing, for example, an astrology database and an astronomy database to re-use standardized metadata about the planet Pluto but in different contexts and domains, and where metadata can be re-used through Web applications as well as by humans. Rather than publishing records that *contain* data, Linked Data implementation involves publishing data so that it can be re-purposed or re-used in other contexts.

The shift to a Linked Data environment involves a costly process of converting metadata to an exchange format substantially different from the "records-based" exchange between the Library and WGBH over the course of the NET Collection Catalog project. This report on the feasibility of Linked Data implementation for the NET Collection Catalog Project examines the transition of a dataset of NET collection records modelled in a "records-based" exchange environment to the "data-based" exchange environment of Linked Data. This report will examine challenges to this process with silos like the Library's media asset management system, Merged Audiovisual

Information System (MAVIS), aligning PBCore with the bibliographic Linked Data model, BIBFRAME, modelling differences in works between archival moving image cataloging and external domains, and possible extensions of BIBFRAME to EBUCore (the European Broadcasting Union Linked Data model) to address gaps between PBCore and BIBFRAME.

Why Linked Data?

Given the common experience of the unreliability of the information environment on the Web, it would appear that the Web-based AAA principle where anybody can say anything about anything is not enough to establish Linked Data as a structured format for data exchange. In short, links must be made smart. For example, if a hotel's website includes a link to a popular tourist attraction, then the fact that the tourist attraction is closed on a certain weekend need not be coordinated between the tourist location and the hotel, which would require the hotel to update their website separately. Smarter linking would mean that an update to the metadata about the "resource" referred to by both tourist attraction and hotel can be used to reflect such changes (Allemang, D. & Hendler, J. A. 2012).

The need for structured information for the example described above should be pretty clear to anyone who works with information. For decades, librarians have been creating authority records for subjects, names, and titles, allowing them to collocate, disambiguate, and index search results (both in the card catalog and online). All of the structure in library bibliographic and authority records depends upon data models that underpin both the encoding (MARC, XML, VRA, MODS, EAD, etc.) and the content standards that describe the values of bibliographic and authority metadata (RDA, DACS, LCSH, LCNAF, etc.); these are all carefully documented and designed to meet specific needs to organize and represent information for newer technology, diverse domains, and information seeking and retrieval behavior. In fact, the above example invites the response that librarians are already doing this work with the standards and data exchange formats central to their profession.

Additionally, the decentralization of what cataloging librarians think of as bibliographic control in favor of a system based on the AAA principle, rather than rules to aid collaborative cataloging, might seem a bridge too far. Some key differences between Linked Data and "traditional" library information standards need to be considered when assessing the value of Linked Data implementation as a departure from "records-based" data exchange in order to be better prepared to answer the question of what value Linked Data has for the NET Collection Catalog Project.

A key difference is the URI as a unique identifier that is resolvable through the standard for information exchange of the Web, the Hypertext Transfer Protocol (HTTP). This unique identifier when paired with HTTP is central to Linked Data. Together, the URI and HTTP

establish the functionality of Linked Data operating as information that is navigable through hyperlinks, as data “embedded” in the functionality of the Web. As an example, the program about Julia Child at the White House mentioned in the introduction could be better represented as a numeric identifier underpinning the name of the work no matter what title was used. Referencing a unique identifier that represents the name of the work is not unlike dialing up a friend you know by a nickname, but don’t happen to know the friend’s given name. No matter what, dialing that unique phone number will still get you to the person you want to reach.

Publishing data in this way represents the evolution of metadata from strings, which at best can be used to index, collocate, and disambiguate resources, to “things.” Another way of thinking about what it means to be a “thing not string” is to think of each URI as a node in information exchange – meaning the URI, unlike the name authority *King, Stephen, 1947-*, can, in addition to collocation, indexing, and disambiguation, be linked to a lot of different resources through any number of different relationships. This is why a URI for Stephen King (such as the VIAF URI, <http://viaf.org/viaf/97113511>) is not just a controlled string of characters that is useful for systematically ordering resources but can also be thought of as a resource itself. It’s a thing that one can say something about – a thing that can consequently be a node in a Web of relationships related to it that link it to other things.

Another key difference is the way relationships between URIs are modelled in Linked Data as RDF (Resource Description Framework) (W3C, 2002). RDF stipulates two very important things about how URIs relate to each other (and to strings) and that influences RDF-modelled information exchange.

1. URIs are related to each other through statements that take the form of SUBJECT – PREDICATE – OBJECT. An example would be, Stephen King (Subject)—is a (Predicate)—Author (Object). Such statements are called triples. Triples, in this sense, atomize the discrete metadata elements traditionally “silo-ed” in record based or hierarchical information representation (Alemu, Stevens, Ross, & Chandler, 2012). Below is an example of relationships to Stephen King delineated as triples:

Person hasName Stephen King

Person hasBook <<http://www.worldcat.org/oclc/1031918016>>

<<http://www.worldcat.org/oclc/1031918016>> hasTitle The Shining

Person hasTwitter @StephenKing

@StephenKing hasContent [pictures of his dog Molly aka Thing of Evil]

2. Another way of representing the relationships between URIs is through the graph model. Since an OBJECT can be a SUBJECT of another statement, a dataset of URIs modelled as RDF triples will look much like the Web with links between nodes that can be traversed as a user navigates from node to node through links that are in fact predicates establishing relationships between URIs. **Figure 1** is a visual representation of the linked data graph model.

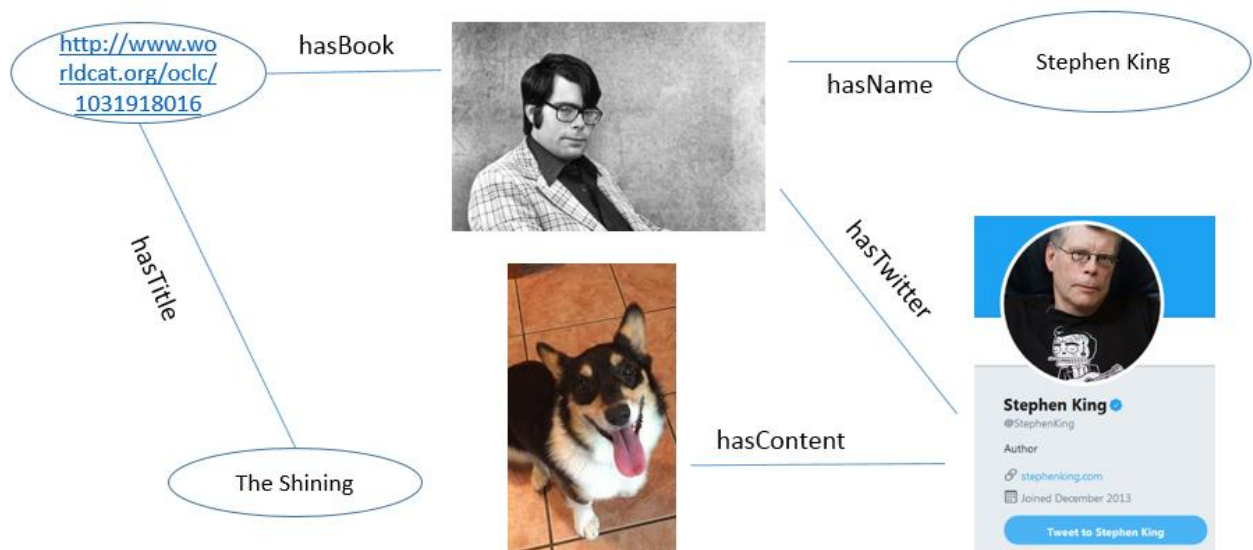


Figure 1

However, RDF only describes the structure of the exchange of Linked Data (as graphs or as triples), but not necessarily its completeness, extensibility, expressivity, or granularity. What is missing from ensuring these qualities is a way of defining this structure to fit the needs of various institutions, domains, formats, users, and other stakeholders. This is the need that various ontologies attempt to meet.

RDF Schema (RDFS) (W3C, 2014) specifies the semantics of RDF-modelled triples. It allows data modelers to create ontologies that cover whatever format, domain, topic, etc. that interests them or their users. A more expanded model called Web Ontology Language (OWL) allows for complex relationships to be expressed through richer formal semantics (Allemang, D. & Hendler, J. A. 2012). What RDFS and OWL both do is provide a RDF-standardized “language” for defining what can be said about resources and establishes the semantics of the relations between them. Ontologies use RDFS or OWL to define classes and subclasses (SUBJECTS and OBJECTS in triples) and to define properties and subproperties (PREDICATES), while also defining data types for string values (such as dates, integers, dimensions, etc.).

In this way, ontologies share structural similarities with taxonomies by categorizing concepts into specialized roles (ontologies even rely on hierarchical structures with subclasses and

subproperties), but there are also differences from such hierarchical systematization. Unlike taxonomies, the main goal of an ontology is not ordering concepts from the general to the more specific (Madsen & Erdman Thomsen, 2009) – rather one expresses relationships defined in the ontology that *may or may not* be from general to more specific. In this sense, ontologies allow one to assert arbitrarily defined relationships between enumerated classes. Ontologies are thereby designed not with systemization as a goal but with what can be said, and what can be inferred from the things that have been said, about resources.

A positive outcome of this lack of systemization is the extensibility of ontologies to include other ontologies if there is a gap in granularity or expressivity. For example, BIBFRAME, the Library’s in-development Linked Data model, provides a relationship between BIBFRAME classes and Friend of a Friend (FOAF) classes in order to represent name authorities, in that the BIBFRAME Agent class and Person subclass are subclasses to FOAF Agent and Person. The Video Resources Association (VRA) establishes relationships to classes and properties in other ontologies, including the VOID ontology that describes datasets (Mixer, 2014) in order to increase interoperability and reduce the “yet another standard” interoperability problem, while also providing support of the AAA principle and the need to be able to express any number of relationships between resources.

Based on this understanding of Linked Data and with a healthy appreciation for the challenges inherent in implementation, below are key reasons why a Linked Data implementation for the AAPB/NET Collection Catalog would be beneficial:

1. AAPB/NET metadata contains valuable and largely undiscovered relationships that, when re-used by others on the Web, can enhance the information already online about classic public broadcasting.
2. It would open AAPB/NET metadata to Web applications, making the metadata more discoverable and shareable on the Web
3. Linked data’s decentralized approach to structuring information reflects the decentralized curatorial environment of the AAPB/NET.
4. It would allow better automation through re-use of metadata in new environments.

Linked Data provides the opportunity for better interoperability through a system designed to link data inside one dataset to data in another dataset without necessarily involving centralized control. Consider the earlier example of the hotel and the tourist attraction and the smarter linking between the two entities where the linking alone is not sufficient to make the link “smart.” Of course, what makes the linking “smart” can be accomplished through collaborative cataloging standards, centralized authority files, querying centralized databases, or a number of controlled information technologies and procedures, yet publishing data in such a way that it is designed to be re-purposed and re-used in different contexts can afford great opportunity for library collections to become better connected to one another and the wider Web environment as

a whole than letting them remain constrained to indexes and records for the sake of bibliographic control.

Phases of Linked Data implementation

The focus of this feasibility report will be on the conversion of a NET collection dataset from MAVIS XML to BIBFRAME standardized RDF. Since, as discussed above, the Library exchanges NET collection records as PBCore-standardized XML with WGBH, this involves testing an alignment between PBCore and BIBFRAME and possible extensions to EBUCore. Additionally, the Library will be data modeling moving image works with Entertainment Identifier Registry IDs (EIDR IDs) and publishing a transformed BIBFRAME RDF dataset to GitHub.

This process is distributed throughout four phases: 1) Data Modeling, 2) Evaluation, 3) Crosswalking, and 4) Publishing.

The Data Modeling phase compares and contrasts different standards (BIBFRAME, EBUCore, PBCore, and MAVIS) as part of a process of examining the impact of each standard on the conversion process. Data modeling works will also be examined in relation to different standards and in relation to EIDR ids.

The Evaluation phase examines the Library's current metadata environment, including the MAVIS environment and issues with the MAVIS to PBCore stylesheet. The goal of this phase is to become familiar with the idiosyncrasies of the current environment and to examine potentialities for exposure as Linked Data. There are a variety of approaches to this step, including using named entity recognition tools (Gracy, K. F. 2015; Zeng, Gracy, & Skirvin, 2013) and crowd sourcing (Pattueli, Provo, & Thorsen, 2015). It was decided to explore the functionality of various fields and the potential for Linked Data exposure in unstructured text fields (by using regular expressions to roughly match certain patterns in these fields), while also identifying potential issues with the MAVIS to PBCore stylesheet.

The Crosswalking phase discusses a crosswalk between different standards to prepare for the transformation of NET collection data from XML serialization to RDF serialization and to prepare for possible extensions to EBUCore properties where there are gaps between PBCore and BIBFRAME.

The Publishing phase discusses a stylesheet transforming a PBCore-standardized dataset of NET records to MARC-XML that could then be transformed to BIBFRAME using stylesheets made available by the Library's Network Development & MARC Standards Office. Different publishing methodologies will also be examined in the discussion of the publishing phase.

Phase 1: Data Modeling**Figure 2**

Figure 2 is the path from MAVIS to BIBFRAME standardized RDF examined in this report. The fact that there are a possible three XML standards to consider is perhaps the more notable data modeling challenge facing a conversion of NET records to Linked Data capable RDF serialization. This path is also an artificial construct, since there are shorter paths to BIBFRAME that do not necessarily involve PBCore. This decision is based on testing an alignment with PBCore and BIBFRAME with possible EBUCore extensions. A wrinkle in this process is that there is no XSLT stylesheet developed to transform PBCore serializations into MARC. There is currently a PBCore Advisory Subcommittee grant to design a MARC to PBCore stylesheet, and a stylesheet will be produced for NET collection holdings for the purpose of this report. Fortunately, however, MAVIS developers designed a MAVIS to PBCore stylesheet. This stylesheet has allowed the exchange of MAVIS exports with WGBH who, like much of the broader moving image archival and production community, exchange records with PBCore standardized XML.

Also contributing to data modeling challenges, the Library’s Motion Picture, Broadcasting and Recorded Sound Division (MBRS), which has responsibility for the acquisition, cataloging, and preservation of moving images and recorded sound materials, operates within a differentiated metadata environment rather than one wholly or even partly governed by unitary systemization in one data model. For instance, the Library’s Integrated Library System (ILS) and MAVIS are not synchronized through an automated process. Records originating in MARC are typically manually added or updated in MAVIS as part of technical processing or cataloging workflows. This situation extends to Library of Congress name, subject and genre/form authorities, which means MAVIS authorities are local, only referencing the authority by a Library of Congress Control Number (LCCN). The ILS MARC-based cataloging environment supports the online public catalog (OPAC) that is user-directed at increasing discovery and access, while MAVIS supports the Packard Campus Workflow Application (PCWA), a local application that manages technical metadata and access to content digitized as part of MBRS’s preservation goals. The communication between MAVIS and PCWA extends only minimally into discovery – for example, searching title fields in PCWA searches preferred titles, not alternative titles.

In the following examination of the data modeling phase—challenges regarding data modeling—how works are exchanged between moving image archives and other institutions with different cataloging goals will be discussed. Entertainment Identifier Registry (EIDR) identifiers as a means to provide unique identifiers for works throughout different stages of the lifecycle of audiovisual materials will also be examined. In addition to the idiosyncrasies in the current metadata environment (particularly regarding technical characteristics of items), the conversion

process in **figure 2** will be examined for impact on data modeling and authority conversion. However, it will be helpful at first to briefly introduce the standards discussed throughout the entire report.

MAVIS is the Library's media asset management system. It is silo-ed both at the Library, where the ILS is not synchronized to it, and also, broadly, where the Library's MBRS Division is one of the few institutions using it for media asset management. There are three broad types of MAVIS records that have a hierarchical relationship to each other: 1) the title record, which describes the intellectual content of the item, including the title, credits, subjects, production and distribution dates, content summaries, etc.; 2) the component record, which describes the technical and provenance details of the item, including technical format, gauge, acquisition information, etc.; 3) the carrier record, which describes holdings, including rack numbers, preservation quality, etc. A component in MAVIS can have several carriers, and a title record can have several components. Since records in MAVIS are not shared with the public except by appointment onsite in the Library's Moving Image and Recorded Sound Research Centers, descriptive practices are more oriented to the needs of preservation workflows than what would be characteristic for discovery by the public. Even so, established Library of Congress authorities are referenced for credits, subjects, and genres to aid in exporting data. XML files standardized by a schema designed for MAVIS can be exported; such XML serializations are called MAVIS-XML throughout this report.

PBCore is a schema that originated in the public broadcasting community as a metadata standard designed for representing elements useful for the exchange of records related to the production and distribution of moving image and recorded sound content. Designed as a public broadcasting specification of Dublin Core (DC), the schema has been adopted by a broader range of institutions maintaining moving image and recorded sound collections. MBRS does not describe its moving image or recorded sound collections in PBCore but rather MAVIS-XML, which is the serialization format for the media collection management system used by MBRS, but, as mentioned above, MBRS does have the capability to convert MAVIS-XML records to PBCore through a XSLT stylesheet designed by MAVIS developers for exchange with WGBH. There are two levels of description in PBCore: 1) the asset, which describes the abstract intellectual content of a media resource, including elements like title, subject, genre, etc.; 2) the instantiation, which describes the physical or digital instance of the asset, including elements like duration, generation, file size, etc. An asset can have several instantiations, and an instantiation can have several parts (represented under instantiationPart container element).

BIBFRAME is a Linked Data model designed to transition traditional library metadata silo-ed in MARC records to RDF and to exploit the semantic capabilities of Linked Data to represent relationships between resources for metadata creation in libraries and other research and cultural heritage institutions. It is still being developed by the Library and is currently in the midst of pilot testing for the impact of the model on cataloging workflows. There are three modes of description in BIBFRAME: 1) the work, which describes the intellectual content of a resource,

including such attributes as subject, genre, creator, etc.; 2) the instance, which describes the material embodiment of the work, including attributes such as format, publisher, etc.; 3) the item, which describes the copy or the *exemplar* of the instance, including attributes such as holding institution, barcode, etc.

EBUCore is a metadata standard for the European Broadcasting Union to support archives, business, and production metadata needs. EBUCore provides the framework for descriptive and technical metadata for service oriented architectures and audiovisual ontologies for semantic Web and Linked Data environments. The European Broadcasting Union have developed an ontology alongside their schema to better support interoperability and harmonization with other standards, such as PBCore. EBUCore also complements the EBU Class Conceptual Data Model that represents a model of broadcasting in the full lifecycle of the media resource in production, in marketing, in archiving, in programming, etc. (EBU, 2016).

Titles, works, and EIDR

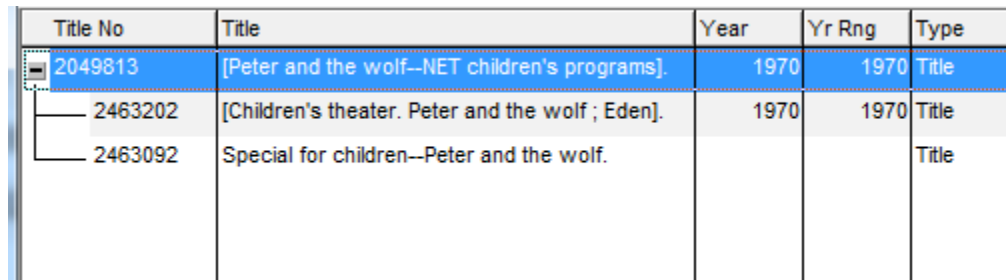
Titles play a central role in describing moving images mainly due to the collaborative nature of motion picture and television production where pointing to a single creative agent for the purpose of collocation is not typical. This reliance on title as a main collocating device is atypical in libraries and archives where collocation is dependent on naming a creator.

Although the practice of establishing preferred titles in libraries is fraught with many of the same issues that catalogers in moving image archives encounter, such as titles changing for marketing purposes, titles appearing differently in many different sources and in different contexts, e.g container label, title screen, indexes, inventories, etc., there's added difficulty in moving image archival description that will likely be affected by additional contextual information that is rooted in a film or television program's distribution history, which has a direct impact on normalizing titles. A common example throughout the distribution history of public broadcast media is distributing a title originally broadcast in one context on the BBC in another context on PBS. For instance, a limited series that originated on the BBC that is broadcast in the U.S. under the umbrella series Masterpiece.

Consequently, dates of production, broadcast times, format, packaging and editing of content, all inform the establishment of preferred title in archival moving image cataloging. Relationships to other variations of the work are also commonly researched and documented. Moreover, describing a physical copy in hand can be difficult for content that can be the same when migrated from different formats or when combined with filler content to round out broadcast time slots.

In the NET collection, these issues are all prevalent, but there are also issues that are largely the result of local policy or local contingencies, such as legacy cataloging practices and being unable to separate components from title records in MAVIS once digital components have been added to a record. For this particular issue, a variety of solutions are used to deal with this problem,

such as making the record a parent record for two new title records that describe the individual content (an example of this is **Figure 3**).



Title No	Title	Year	Yr Rng	Type
2049813	[Peter and the wolf--NET children's programs].	1970	1970	Title
2463202	[Children's theater. Peter and the wolf ; Eden].	1970	1970	Title
2463092	Special for children--Peter and the wolf.			Title

Figure 3

These differences between archival moving image cataloging and other communities are not by themselves particularly problematic. However, issues emerge when exchanging metadata between institutions. There are two areas where this contextual aspect of archival moving image cataloging's establishment of preferred titles conflicts with the sharing of records: title matching based on unique strings and works as abstract entities related to published and distributed materials, institutional holdings, and other works. In **figure 3**, both areas of conflict can be examined.

For title matching, the preferred titles for two of these title records are supplied titles, constructed according to cataloging policy for the NET collection. The parent record's supplied title – [Peter and the wolf—NET children's programs] – was constructed because of the need for it to be a parent record for two title records describing different productions whose components were mistakenly added to the same record, and were, after digitization, now unable to be moved to new records due to constraints in MAVIS. The child record's supplied title – [Children's theater. Peter and the wolf ; Eden] – was constructed in an effort to describe the fact that the components also included short animated filler content entitled "Eden." Since supplied titles are not sourced entirely from common sources of information (they are heavily supplemented by institutional policy, content standards, and cataloger's judgment), it is highly likely that title matching based on these strings would be difficult. Additionally, the title – Special for children--Peter and the wolf – uses a title describing the program as a children's special that may not be used as a preferred title by other institutions. Of course, soft matching between similar titles during reconciliation can provide opportunity to bridge gaps between different title strings for the same content in different datasets, but this adds a human review cost to title matching that can be time consuming for large datasets.

By comparison, the second area of conflict, works, seems more systematic because it involves the structure, ordering, or semantics of a dataset's relationships. This is explicit in the standards used to describe resources, such as PBCore where works are treated as assets that are related to embodied instantiations and to other assets or MAVIS where title records are related to components, carriers, and other title records. There are also differences in the kind of metadata

used to describe works between different systems. For instance, MAVIS includes “duration” as a title record element, but for PBCore, “duration” is an instantiation element.

How works are used in datasets to systematize relationships between titles, published and distributed materials, and holdings also has an impact on the relationships established when sharing metadata between titles from different datasets. For example, **figure 3** includes a parent record whose status as an actual work is definitely questionable since at most the title represents a container record, but it nonetheless is the record attached to the components described by its child title records. It is an open question about the status of this container “work” when its metadata is exchanged.

Also, the error that is corrected in this example is an error directly connected to many of the issues with establishing works in moving image description. Two different items, both of which were part of the NET/PBS collections at the Library, were labeled as “Peter and the wolf” and incidentally had the same NOLA code (an alphabetic four letter code given to broadcasts that represents the title of a program or the series title of a series), but upon inspection, these two items were discovered to be two different productions – one was in black and white, the other in color; one contained animated filler, the other did not. It is crucial that works are reconciled when exchanged with other datasets, especially ones working off similar inventories using NOLA codes.

Moreover, the NET collection contains numerous rebroadcasts and re-brandings of different programs under new umbrella series titles. For example, a BBC production entitled “The violent universe” was rebroadcast as two different series: “What’s new” and “Public Broadcasting Laboratory” (PBL). Additionally, the Library’s NET collection does not utilize series or season records to collocate records, and instead handles series and season through compound strings, such as “NET journal. Huelga!” or “Spectrum. [Series 4], The jet train is here.” Such differences described in the above examples can alter relationships between titles from different datasets enough that it is critical to pay close attention to dissimilarity and, if necessary, document differences.

These are all challenges that are beyond simply mapping to title-work authority records in the LCNAF that can then be re-used as URIs in a Linked Data Environment, as is currently the process for libraries where the focus is on transforming bibliographic MARC records to BIBFRAME standardized RDF. Library cataloging, as discussed above, is not necessarily concerned with the types of title relationships and title descriptions that can influence the establishment of a work for moving image archives, and, currently, the establishment of relationships between works for libraries and external sources, such as Wikipedia or Internet Movie Database (IMDb), is currently only being handled as a “same as” relationship, when moving image archival description needs to be able to express robust relationships between works throughout production, distribution, marketing, and programming lifecycles.

A promising alternative is the Entertainment Identifier Registry (EIDR). EIDR creates unique DOIs using hierarchical and semantic notation to represent unique content in relation to broader context, such as series, seasons, composites, edits, broadcasts, rehearsals, etc. “EIDR IDs are globally unique, externally resolvable; applicable to works in the abstract (title records), versions of works (edits), and representations of works (encodings or manifestations); and are able to support multiple types of asset groupings and relationships and to store multiple alternate titles and alternate identifiers per asset. This last feature starts to alleviate the point-to-point translation problem: EIDR acts as a bridge between multiple systems” (Kroon, Drewry, Leigh, & McConnachie, 2015). Such a system can help mitigate issues with establishing works in archival moving image cataloging where context is central and can facilitate title reconciliation with less reliance on matching strings, instead utilizing resolvable unique ids.

The screenshot shows the EIDR website interface. At the top is a green header with the EIDR logo and navigation links: Home, Search, Register, and Help. A search bar is also present. Below the header, the 'VIEW' section is active, showing a hierarchy of identifiers for a specific episode. The hierarchy is as follows:

- PARENT #1:** Series | Saturday Night Live | 10.5240/52F4-B813-D880-CD5B-9414-P
Referent Type: Series | Structural Type: Abstraction | Publication Status: valid | Release Date: 1975
- PARENT #2:** Season | Saturday Night Live: Season 41 | 10.5240/4198-B778-1019-B384-22A6-4
Referent Type: Season | Structural Type: Abstraction | Publication Status: valid | Release Date: 2015
- CURRENT:** Episode | Donald Trump/Sia | 10.5240/85CF-AE5C-26B9-F0EA-04AC-N
Referent Type: TV | Structural Type: Abstraction | Publication Status: valid | Release Date: 2015
- Edit:** Donald Trump/Sia | 10.5240/02D3-19F7-50EB-D47C-6582-W
Referent Type: TV | Structural Type: Performance | Publication Status: valid | Release Date: 2015
- Edit:** Donald Trump/Sia | 10.5240/1112-228A-C434-BDA7-E1D6-Q
Referent Type: TV | Structural Type: Performance | Publication Status: valid | Release Date: 2015
- Edit:** Donald Trump/Sia | 10.5240/802E-5C5C-7E51-59D2-98AC-C
Referent Type: TV | Structural Type: Performance | Publication Status: valid | Release Date: 2015

At the bottom of the page, there is a footer with copyright information: © 2018 Entertainment ID Registry Association, the DOI International DOI Foundation logo, and links to Sandbox, 29e8262, sandbox1.eidr.org, and v2.1.

Figure 4

Some of the work that MBRS has already done with EIDR includes born digital ingest of *Saturday Night Live* episodes that link to already existing EIDR IDs. A smaller project creating new EIDR ids is that for the Watergate Hearings broadcast and distributed by PBS. In **figure 4**, a representation of an episode from *Saturday Night Live*’s 41st season – with host Donald Trump and musical guest Sia – is displayed with its relationships to identifiers at the series, season, and edits (Original broadcast, shortened, and Web versions of the episode) levels. Each node in the hierarchy has its own identifier to aid identification of content across a broad spectrum of user and institutional needs.

Even though EIDR IDs are DOIs, MBRS evaluated creating EIDR ids for a small dataset of titles as part of its Linked Data report because of EIDR's support of complex and granular relationships between works. Appendix A is a table representing the dataset that was curated to test a number of works with titles that range from relatively straightforward to complex. A summary of the discoveries from testing EIDR IDs for the titles in Appendix A is below.

- EIDR is dependent on a hierarchical model of relationships between titles, and support for cross-referenced or horizontal relationships (such as edits) sometimes seems constrained and should be engaged very cautiously
- EIDR has rigid rules governing value entries that likely requires training and/or learned familiarity to navigate
- EIDR's de-duplication process tends to overflag content, which further requires administration by EIDR staff

Additionally, the following observations are relevant: EIDR handles parent records such as "series" as a general container that encapsulates title records for disparate entities like *Saturday Night Live*, the Senate Hearings on Campaign Activities (Watergate Hearings), and oral history collections. Functionally, these can be individually coded so as to represent something close to how these "collective" runs are represented in the Library's database, but how close is sufficient for title matching based on EIDR ids? What would be the impact on the data modeling works for library information systems where the Senate Hearings on Campaign Activities are being tagged collectively as a series when in fact they represent one continuous multi-day broadcast event? Controlling the functionality of mapping these container relationships together by specifying seasons, ordered or unordered sequencing of parts (episodes) or allowing for open and unordered sequencing as EIDR ameliorates these challenges. However, it will be necessary to build a nuanced and documented understanding of granularity gaps going forward if EIDR is to play a role in encoding for Linked Data implementation for titles because those gaps would need to be traced somewhere in library discovery systems.

BIBFRAME, PBCore, MAVIS, and MARC

BIBFRAME is a data model (represented in Figure 5) that has been designed as a means of

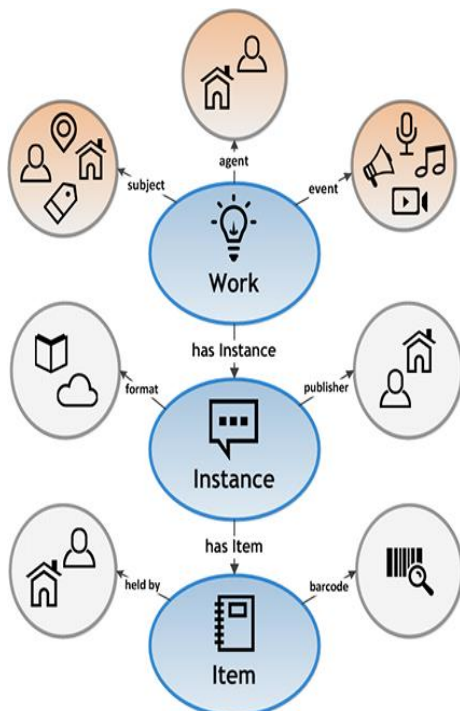


Figure 5

transitioning out of a MARC-based environment to a Linked Data environment. The reasons for this transition were advocated by Roy Tennant in 2002, who argued that the rigidity of MARC's flat structure that is based on outdated concepts in information architecture and the outsized priority display has in regards to MARC's functionality (Tennant, 2002). Since library metadata has historically been silo-ed in MARC records, BIBFRAME development is primarily focused on converting existing MARC records to the Linked Data exchange environment enabled by RDF encoding.

In addition to the modernization of library and cultural heritage and research institution metadata outlined by Tennant (2002), MARC records do not easily represent the Functional Requirements for Bibliographic Records (FRBR) Work, Expression, Manifestation, and Item (WEMI) relationships. **Figure 6** is a MARC record where WEMI fields have been highlighted in the following legend: **green** for work fields, **yellow** for expression fields, **orange** for manifestation fields, and **blue** for item fields. There are also

three items circled in **green** to highlight the fact that within individual fields or subfields there is variation in FRBR. Work fields are elements that represent metadata describing intellectual content; expression fields are elements that represent metadata describing the form that each particular work takes each time it's realized; manifestation fields are elements that represent metadata describing the embodiment of an expression or work; item fields are elements that represent metadata describing individual holdings of manifestations.

In **Figure 6**, a single MARC record can be said to be about four different resources:

- Work: *Mal d'Archive* by Jacques Derrida that was originally presented as a lecture on June 5, 1994
- Edit: the English translation with the parallel title *Archive Fever* by Eric Prenowitz that was released as part of a series called "Religion and postmodernism."
- Manifestation: the 1996 University of Chicago Press publication that libraries assign classification numbers to,
- Item: the item described by holdings information.

Tag	11	12	Subfield Data
035			\$9 (DLC) 96018568
906			\$a 7 \$b cbc \$c orignew \$d 1 \$e ocip \$f 19 \$g y-gencatlg
955			\$a pc14 to sa00 04-23-96; sh20 04-24-96; rec'd 05-13-96 sh07; sh07 05-13-96; sh15 05-14-96;aa05 05-16-96; CIP ver. ta06; to SL 01-09-97
010			\$a 96018568
020			\$a 0226143368 (cloth)
040			\$a DLC \$c DLC \$d DLC
041	1		\$a eng \$h fre
050	0	0	\$a BD181.7 \$b .D4713 1996
082	0	0	\$a 153.1/2 \$2 20
100	1		\$a Derrida, Jacques.
240	1	0	\$a Mal d'archive. \$l English
245	1	0	\$a Archive fever : \$b a Freudian Impression / \$c Jacques Derrida ; translated by Eric Prenowitz.
260			\$a Chicago : \$b University of Chicago Press, \$c 1996.
300			\$a 113 p. ; \$c 22 cm.
440		0	\$a Religion and postmodernism
500			\$a Originally presented as a lecture June 5, 1994, at an International colloquium entitled: Memory : the Question of Archives in London, England.
504			\$a Includes bibliographical references (p. 113).
650		0	\$a Memory (Philosophy)
650		0	\$a Psychoanalysis.
600	1	0	\$a Freud, Sigmund, \$d 1856-1939.
856	4	2	\$3 Publisher description \$u http://www.loc.gov/catdir/description/uchi051/96018568.html
856	4	1	\$3 Table of contents \$u http://www.loc.gov/catdir/toc/uchi051/96018568.html
920			\$a **LC HAS REQ'D # OF SHELF COPIES**
991			\$b c-GenColl \$h BD181.7 \$l .D4713 1996 \$t Copy 1 \$w BOOKS

Figure 6

BIBFRAME should be understood as not just about converting bibliographic data out of a flat record based structure in terms of concrete separate statements as required in a Linked Data model but also about exploiting Linked Data's capabilities for representing relationships to better accommodate common bibliographic data modeling such as FRBR's WEMI relationships with its own multi-level relationships between resources.

The NET Collection Catalog Project at the Library has chosen this data model as its target ontology for the following reasons:

- BIBFRAME 2.0 is a complete model, with available documentation and OWL and RDFs data modeling plus a working XSLT that converts MARC-XML to BIBFRAME.
- As part of the Library, the NET project was able to leverage human and other institutional resources available dedicated to BIBFRAME implementation and analysis.
- At the moment, there is no corresponding transformation of PBCore into another Linked Data standard, so it makes sense to work with readily available tools and infrastructure involved with BIBFRAME at the Library.
- This is also an opportunity to test BIBFRAME's applicability to cataloging in moving image archives.
- LC has already commissioned two studies analyzing BIBFRAME in terms of the needs of moving images: BIBFRAME AV Modeling Study (<http://www.loc.gov/bibframe/docs/pdf/bibframe-avmodelingstudy-may15-2014.pdf>) and

BIBFRAME AV Assessment (<http://www.loc.gov/bibframe/docs/bibframe-avassessment.html>).

It is important to note some of the expected difficulties with BIBFRAME that transforming NET records to BIBFRAME-standardized RDF involves. These difficulties are discussed in individual sections below.

Technical characteristics

Both MAVIS and PBCore XML standards are designed with a shared understanding that not only are the technical characteristics of physical carriers essential to the preservation, production, and distribution of AV material but also that the intellectual content “captured” by the carriers need not – and most often does not – have an essential relationship to the carrier itself. This is not an unusual idea in BIBFRAME either, which posits the *Work* class as a resource that has a non-dependent relationship to the various forms and methods of expressing and manifesting the intellectual content as a concrete entity.

In this sense, MAVIS/PBCore use of “title records” (MAVIS) and “assets” (PBCore) aligns with the BIBFRAME model, where works are understood as independent, more so than with MARC, which produces a record intermingling all aspects of the bibliographic data model (with the exception of holdings records). However, MAVIS and PBCore contain more technical elements because of the mediated nature of AV material, a fact that involves descriptions of technical characteristics to aid preservation and to determine the nature of and the dependencies required by the type of media, while BIBFRAME is intended for all kinds of resources.

Also, since BIBFRAME is MARC-targeted, another difficulty is the fact that many technical characteristics are handled through content standards rather than encoded directly as attributes (Lyons & Malssen, 2015). An example of this is the MARC 300 element that in AACR2 and AMIM2 combines multiple technical characteristics in single subfields, as below (where duration and item count are recorded in one subfield, sound and color characteristics in another):

‡a 6 videoreels of 6 (383 min.) : ‡b sd., col. ; ‡c 2 in.

The LC AV Assessment for BIBFRAME by Lyons and Van Malssen in 2015 examined these issues in thorough detail and was part of the recommendations that BIBFRAME used to expand its model to include audiovisual technical characteristics in BIBFRAME 2.0. BIBFRAME 2.0 is the version of BIBFRAME to which MBRS is transforming this dataset to be shared through GitHub publication. MBRS also traced gaps in the NET collection data between PBCore and BIBFRAME 2.0 that can be addressed with specific EBUCore properties, which are examined in the Crosswalking section.

However, it is important to share a word of caution. As noted by Lyons and Malssen (2015), the EBUCore “track” object type and the BIBFRAME “instance” object type are not necessarily

equivalent classes, so mapping properties whose domains in EBUCore are the track class to properties asserted of BIBFRAME instance classes would be problematic for machine querying, discovery, and semantic inferencing. Likewise, the instance class in BIBFRAME might pose difficulties in mapping from PBCore instantiation metadata that has been transformed from MAVIS because the hierarchy in MAVIS between title records, component records, and carrier records is more complex than PBCore's simpler hierarchy between asset and instantiation. Overall, while conceptualizing "item" and "manifestation" metadata for MAVIS, PBCore, BIBFRAME, and EBUCore need to be tested more fully, there needs to be considerable caution.

Authority conversion and reconciliation workflow

The BIBFLOW project at Stanford (Smith et. al, 2017) and most library communities recommend adding URIs to records as soon as possible rather than attempting reconciliation after transitioning to BIBFRAME. This is good advice for transitioning to BIBFRAME, especially when the data being transitioned is MARC data for which there are few, if any, intermediary steps between the original dataset and BIBFRAME-standardized RDF. Unfortunately, for the purposes of this project and the need to align MAVIS with BIBFRAME through PBCore this approach has a number of different drawbacks, examined briefly below.

MAVIS authorities include 1) local authorities not authorized through LCNAF, which can be either treated as literal values (strings rather than URIs) or reconciled with Linked Data identity sources other than LC, such as VIAF, 2) name authority metadata silo-ed in MAVIS, including information such as LC identifiers and elements describing identities only recorded locally, and 3) other controlled vocabularies, which are useful but silo-ed, such as roles (production company, etc.) and title types.

At what stage in the transformation process should URIs be inserted into records? As early as possible is a good rule of thumb, but it is important to consider the availability of tools that might make it easier to insert URI's into MARC-XML than PBCore or MAVIS.

How should this reconciliation proceed? Is there a way to pull LCNAF identifiers from MAVIS and insert them into PBCore, MAVIS, or MARC-XML or should NET collection data be reconciled with LC Linked Data services externally?

Should local MAVIS authorities be isolated and matched to non-LC Linked Datasets, such as VIAF or Wikidata?

Is it possible to share local name authority metadata (variant names, etc.) for identities not recorded in LCNAF or other Linked Datasets?

These questions will be explored through the Crosswalking and Publishing phases.

Phase 2: Evaluation

The evaluation of the MAVIS-XML serialization was performed on the NET dataset that was later converted into BIBFRAME. This dataset was curated by focusing on two anthology series – NET Playhouse and NET Journal. The former is a fictional drama anthology series and the latter a non-fiction anthology series of documentaries. These titles were chosen because they were numerous and were central to NET programming, but also because together they could represent both fictional and non-fictional content metadata for which there would be different patterns (such as more subjects for non-fiction and more credits for fiction).

For the Evaluation phase, the MBRS's current metadata environment was examined in order to ascertain potentialities for Linked Data exposure of access points and descriptive metadata in the NET collection as cataloged in MAVIS. Component and carrier metadata were mostly excluded from this review because their values are treated differently by individual schemas and standards throughout the conversion process from MAVIS to BIBFRAME to the point that reviewing them systematically seemed less useful than individually mapping item metadata through the different standards in the Crosswalking section. For the Evaluation phase, it was desired to focus less on technical metadata in MAVIS that is stable and consistently identifiable – such as format, file size, duration, etc. – and more on the metadata describing content characteristics that is more variable between institutions and cataloger judgment– such as subjects, titles, genres, notes, etc. Additionally, BIBFRAME 2.0 accommodates Lyons and Malssen's 2015 recommendations for technical characteristics, and any gaps still remaining can be handled by extension to EBUCORE (or other Linked Data models, such as PREMIS). Nonetheless, technical characteristics metadata are still mapped in the Crosswalk and are included the published preliminary dataset.

Also for the Evaluation phase, potential challenges in how the MAVIS to PBCore stylesheet transforms MAVIS XML exports to share with institutions using PBCore to exchange XML records were also reviewed by examining a sample record transformed from MAVIS into PBCore.

MAVIS access points, description, and uncontrolled text

The MAVIS metadata environment was examined by first creating an ad hoc “data dictionary” for MAVIS with additional evaluative information about what kinds of values are used to populate MAVIS fields. This has been provided for reference as Appendix B. This data dictionary focused only on descriptive metadata and not on holdings or technical metadata, although the dataset includes both holdings and technical metadata. Technical metadata is nonetheless still mapped in the Crosswalking section and published in the preliminary BIBFRAME-standardized dataset.

Although the aim was primarily the descriptive values that catalogers use to populate MAVIS title records, representations of the XML hierarchy for each element and references to the relative paths of codes MAVIS uses to uniquely identify local controlled values are included. For each element, there are nine categories of information:

DESCRIPTION—a statement regarding the purpose and/or meaning of the field

EXAMPLE—a representative example from NET catalog MAVIS records

REQUIRED—whether the element is core or not

HREF/RELATIVE ADDRESS—relative path to identifiers and codes MAVIS uses to uniquely identify local controlled values

XML HIERARCHY—a representation of the XML hierarchy of the element [retrieved through an export of MAVIS XML that is transformed into a spreadsheet using OpenRefine]; this was to track the structure of the MAVIS metadata environment

ALTERNATIVE LABELS—alternative names of the element described; useful for thinking through crosswalks to PBCore or RDA or MARC

CONTROL—information regarding the type of control catalogers use to populate values for the elements described, including controlled vocabularies, uncontrolled fields, and content standards

CODE—abbreviated term representing a controlled vocabulary.

Following from the data dictionary of MAVIS metadata elements, it was decided that the key point of variation in metadata values in MAVIS was the type of control that is declared for each element value. The following are the types of control for metadata values in MAVIS:

- | | |
|----------------------------|-----------------|
| • Controlled vocabulary | • Binary |
| • Automatic generation | • Uncontrolled |
| • Authorized access points | • Transcription |

The dataset was then divided into three separate, smaller datasets focusing on the potential for element values to be exploited for Linked Data: controlled vocabulary, transcribed field, and uncontrolled field values. Like the data dictionary above, these datasets were imported into OpenRefine—a tool for working with messy data—and then converted into.csv files.

Binary and automatically generated values were excluded from this analysis since they are machine oriented values that lack flexibility in semantic repurposing. These fields are definitely useful for creating facets in front-facing discovery systems, but the controlled vocabulary field, transcription field, and uncontrolled field values would seem to provide the most potential for developing the sorts of Linked Data relationships RDF should support.

It should also be noted that authorized access points were included with the controlled vocabulary dataset. Since authorized access points are associated specifically with Library of Congress authorities such as subjects, names, and genre/form, it's useful to distinguish between the broader controlled vocabulary concept of controlled field values and the construction of

authorized access points that are regularly updated under the auspices of the Program for Cooperative Cataloging (PCC), wherein members contribute bibliographic records and related data under a common set of standards and conventions. Authorized access points under the NACO and SACO programs have already been made Linked Data capable and in use broadly by the library and archive communities, so examining them separately for potential Linked Data exposure does not seem useful, especially since many descriptive controlled vocabulary values themselves also have URIs that can be re-used.

In MAVIS XML, modifying attributes, such as controlled vocabulary terms specifying the type of data being modified, are treated as sibling elements to elements that they are modifying, a situation that does not translate very well into flat formats like spreadsheets. They were listed in these datasets alongside the values they are modifying for clarity, for example, “Original language: English” or “BROADCAST DATE: 1969.” A factor to consider about this decision is that this effectively combines several controlled vocabulary values with transcribed values, such that they could be in either dataset, but the methodology of review and analysis makes this approach more feasible than if MBRS were using a more controlled or quantitative methodology.

The controlled vocabulary and transcription fields datasets are reviewed in the following discussion, following which the uncontrolled fields dataset is analyzed for potential Linked Data exposure using regular expression matching in XSLT.

Controlled vocabulary and transcribed fields

1	TitleNumber - id	Medium	Credit Role	Country	Language	Genre	Object Identifier	Color	Subject	Name as subject
			Ritchard, Cyril, Reader/Reciter; Carroll, Lewis, Author/Based on; National Educational Television and Radio Center, Broadcaster; National Educational Television and Radio Center, Presenter; Verdon, Gwen, Reader/Reciter; Jordan, Glenn, Producer; Jordan, Glenn, Director; WNET, Production Company; Novak, David., Scriptwriter;				NET/PBS NOLA: NYAC; NET/PBS number: 0; Televised NET/PBS number: 215;			
2	/TitleWork/key/208686	Moving Image	Holbrook, Hal, Host;	United States;	Original Language: English;	Nonfiction television programs (lcgft); literary readings (lcgft); Televised performances (lcgft);		Color		
3		Moving Image								
			National Educational Television and Radio Center, Broadcaster; Australian Broadcasting Commission, Production Company; Intertel, Undetermined;			Documentary television programs (lcgft); Educational television programs (lcgft);	NET/PBS NOLA: AFMI; NET/PBS number: 41;		Israel-- History-- 1948- 1967;	
4	/TitleWork/key/232969	Moving Image	Holtzman, Boris, Producer;	Australia;	Original Language: English;	Nonfiction television programs (lcgft);		Black & White		Israel;
5		Moving Image								

Figure 7

Controlled vocabularies control element values for access points and descriptive elements. These are managed in MAVIS separately from bibliographic records through local controlled vocabulary lists and authority records (sometimes with references to external authority files, such as LCNAF). **Figure 7** is a screenshot of the controlled vocabulary dataset in spreadsheet format.

Reviewing controlled vocabularies in MAVIS brings up some observations regarding the scope of some elements and vocabularies. Access point controlled vocabularies, or value standards, such as subjects, names, and genre/form conceptually are independent from the records they modify, since they facilitate relationships to other concepts through collocation and disambiguation.

Descriptive elements are more embedded in the records they modify. They rely on smaller vocabularies, or they represent more empirical properties of the objects they describe. Examples of these are “medium” and “colour.” These still have value for collocation or disambiguation, but they are inseparable from the records they define, even when they can be identified in controlled

vocabularies and re-used. This is why these latter elements seem to work better as facets or delimiters rather than functioning as authorized access points, which help the user to narrow down, rather than expand, more specific needs as discovery progresses.

	A	B	C	D
1	MAVIS id	Preferred title	Alternative title	Dates
2	/TitleWork/key/2086861	Actor's choice. Lewis Carroll.	Lewis Carroll.; Actors choice 1.; NET playhouse. [1970-11-26]. Actor's choice. Lewis Carroll.; NET playhouse. [No. 215]. Actor's choice. Lewis Carroll.;	Broadcast: 1970/ 11/ 26-0
3	/TitleWork/key/2329697	After the miracle.	Intertel. After the miracle.; NET Journal. After the miracle.; After the miracle. NET Journal/Intertel.;	Broadcast: 1967/ 2/ 20-0;
4	/TitleWork/key/1162460	Cathy come home.	NET playhouse. [No. 259], Cathy come home.; NET playhouse. [No. 187], Cathy come home.; NET playhouse. [No. 130], Cathy come home.; Two off the cuff.; The Wednesday play. Cathy come home. [Great Britain release title].; NET playhouse. Cathy come home. [U.S. release title].;	Broadcast: 1969/ 3/ 27-0, 11/ 16-0/ 0/ 0;
5	/TitleWork/key/2048799	The ceremony of innocence.	NET playhouse. [No. 211], The ceremony of innocence.; NET playhouse. [No. 191], The ceremony of innocence.; Ceremony of innocence.; NET playhouse. The ceremony of innocence.;	Broadcast: 1970/ 10/ 29-0
6	/TitleWork/key/2432362	Children in balance.	NET journal. Children in balance.; Children in the balance.; NET journal. [No. 221], Children in balance.;	Broadcast: 1969/ 1/ 6-0/
7	/TitleWork/key/1167394	A conversation with Earl Warren.	A conversation with Earl Warren.; NET journal. A conversation with Earl Warren.; NET journal. [No. 247], A conversation with Earl Warren.; Warren, Earl; Earl Warren.;	Broadcast: 1969/ 9/ 8-0/
8	/TitleWork/key/2352561	Conversation with Milovan Djilas.	Milovan Djilas.; [Conversations with Milovan Djilas].; NET journal--conversation with Milovan Djilas.;	Broadcast: 1968/ 12/ 2-0;
9	/TitleWork/key/2047578	Culloden.	NET playhouse. The Battle of Culloden.; NET playhouse. [No. 160], The Battle of Culloden.; NET playhouse. [No. 23], The battle of Culloden.; NET playhouse. [No. 86], The battle of Culloden.;	Broadcast: 1969/ 10/ 23-0
			NET journal. Dick Gregory is alive and well.; NET journal. [No. 260], Dick	

Figure 8

Figure 8 is a screenshot of the transcribed fields dataset. Transcription fields contain information that is recorded exactly as it is represented in sources of information. These may include titles that appear on screen, labels, slates, or secondary sources such as collection documentation (NET microfiche and inventory lists), and they also include statements about creative responsibility, dates of distribution, production, or broadcast, etc. Transcribed fields thus follow similar patterns of empirical descriptive properties whose metadata is contingent more on the material or embodied aspect of the resources being described than on separate, abstract indexical access points. Like controlled vocabulary, transcription includes both access points and description. Access points include *preferred titles*, *related titles*, *alternative titles*, and *contents*. Description includes *statements of responsibility*, *object identifiers*, and *dates*.

It may in fact be that certain controlled vocabularies or transcribed strings are useful as strings in RDF rather than URIs, especially in metadata environments where some content management systems or discovery systems might only validate strings or otherwise not make use of linking. In fact, the OWL RDF data model distinguishes between these two types of data with the categories of properties: *data type properties* (properties describing resource attributes that are expressed as literal “string” values, in essence connecting resources to literal values) and *object properties*

(properties describing classes or individuals, in essence connecting resources to other resources). In the Crosswalk, domains and ranges of BIBFRAME and EBUCore properties were traced during mapping to track how each individual data element will be utilized in the Linked Data environment, and the Crosswalk also maps controlled vocabularies relevant to each element.

Uncontrolled fields

B	C	D	E	F	G	H	I	J
Summary	Note	Sources	Date Notes	Credit Notes	Administrative Notes	Additional credits	Related title notes	
Free Press July 27, 1969. Page 68: "The life and work of the great artist-naturalist and a glimpse of what is left of Audubon's America. Filmed throughout the North American continent; written and narrated by Lister Sinclair, with Albert Millaire as the voice of Audubon." (taken from newspaper.com search RPAR 3/8/2017). From NET microfiche: Suggested Newspaper	americanarchive.org : "Intertel is an anthology series that features documentaries on world issues. The episodes come from many producers, and some aired as individual programs before airing on Intertel." This means that this content may match MAVIS 211342 but I have no way of verifying that (RPAR 3/8/2017). Updated by SBOO. 2017-3-20. MAVIS 211342 is a match; this program was originally aired under NET Journal in 1968. Sources used: NET microfiche; PBS-1994.xls Excel spreadsheet (on MBRS shared drive).	Sources used: NET microfiche; PBS-1994.xls Excel spreadsheet (on MBRS shared drive).	Under Intertel ; date from NET microfiche.; Under NET Journal; date from NET microfiche.;	Photographed by Irving Saraf.;	Updated by SBOO. 2017-3-20.	Additional credits: Dialogues by Jerzy Skolimowski. Music by Krzysztof T. Komeda. Cameraman: Andrzej Gronau. English subtitles: Cecylia Wojewoda and Michael Elster.	anthology series that features documentaries on world issues. The episodes come from many producers, and some aired as individual programs before airing on Intertel." This means that this content may match MAVIS 211342 but I have no way of verifying that (RPAR 3/8/2017). Updated by SBOO. 2017-3-20. MAVIS 211342 is a match; this	
Summary from NET microfiche: The film shows a three-year-old boy, walking through the thick undergrowth of	record, LCCN 76707749; video was viewed; Copyright catalog, Motion Pictures, 1960-1969; NET microfiche (NET playhouse. Dublin one). NET microfiche						1. MAVIS 2332444, video copied	

Figure 9

Uncontrolled fields do not necessarily use standardization for values – particularly controlled vocabularies or transcription rules - although there is nothing preventing institutions from devising standardized approaches that work best for their collections. Uncontrolled fields in MAVIS are fields that are generally used to clarify elements of the title, component, and carrier

records, for instance credit notes and date notes. However, the summary and note fields in title records individually play other roles analyzed below for potential Linked Data contributions.

The unstructured text is mostly apparent when reviewing the note field, which is entered under notes tabs in MAVIS – this is a field that encompasses a wide range of administrative and contextual notes that help clarify the entire record or provide notes for elements that cannot be provided elsewhere in the record. **Figure 10** is an NET title record that has been updated a number of times, each time adding more detail and more clarification and context to the rest of the NET collection and KRMA broadcasting in general, but also tracing changes made and points of contact between catalogers/technicians and the record over the years (such administrative tracing is particularly important since MAVIS only logs the last point of contact).

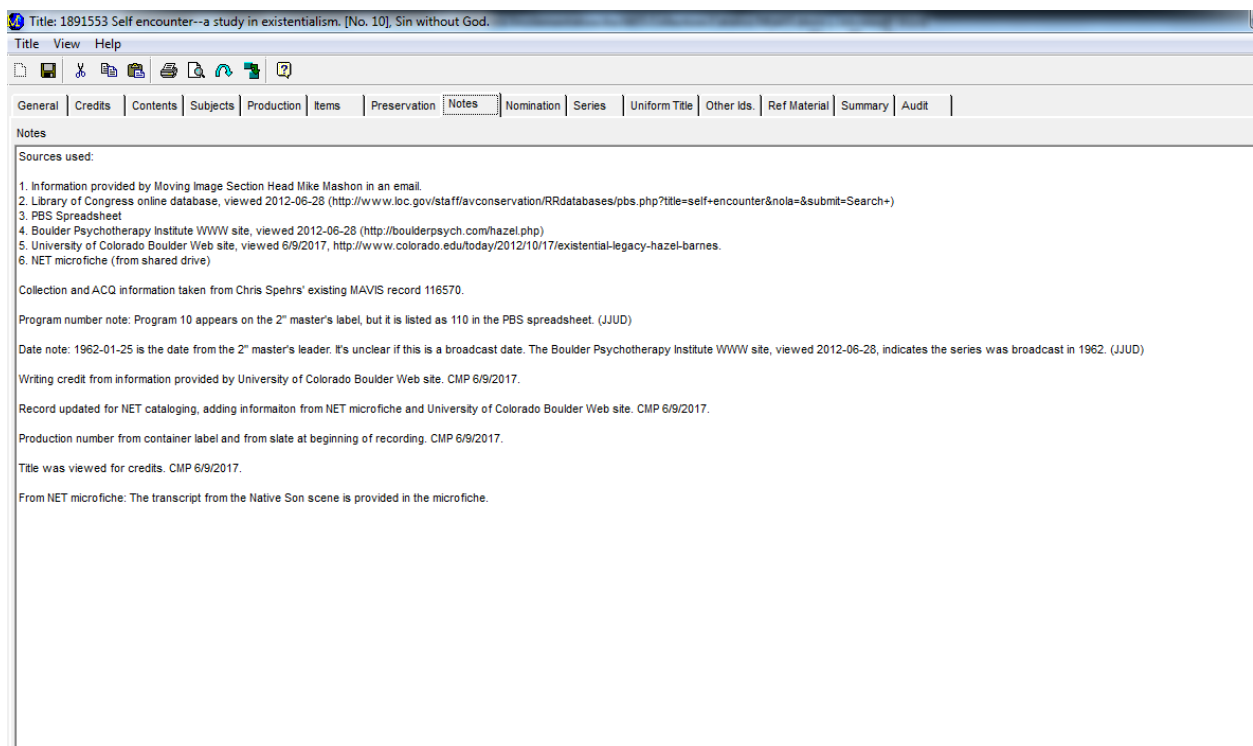


Figure 10

In MAVIS XML, these notes are all entered under the element *note*, as illustrated below:

<note>Sources used:

1. Information provided by Moving Image Section Head Mike Mashon in an email.
2. Library of Congress online database, viewed 2012-06-28
(<http://www.loc.gov/staff/avconservation/RRdatabases/pbs.php?title==self+encounter&nola==&submit==Search+>)
3. PBS Spreadsheet
4. Boulder Psychotherapy Institute WWW site, viewed 2012-06-28 (<http://boulderpsych.com/hazel.php>)
5. University of Colorado Boulder Web site, viewed 6/9/2017,

<http://www.colorado.edu/today/2012/10/17/existential-legacy-hazel-barnes>.

6. NET microfiche (from shared drive)

Collection and ACQ information taken from Chris Spehrs's existing MAVIS record 116570.

Program number note: Program 10 appears on the 2" master's label, but it is listed as 110 in the PBS spreadsheet. (JJUD)

Date note: 1962-01-25 is the date from the 2" master's leader. It's unclear if this is a broadcast date. The Boulder Psychotherapy Institute WWW site, viewed 2012-06-28, indicates the series was broadcast in 1962. (JJUD)

Writing credit from information provided by University of Colorado Boulder Web site. CMP 6/9/2017.

Record updated for NET cataloging, adding information from NET microfiche and University of Colorado Boulder Web site. CMP 6/9/2017.

Production number from container label and from slate at beginning of recording. CMP 6/9/2017.

Title was viewed for credits. CMP 6/9/2017.

From NET microfiche: The transcript from the Native Son scene is provided in the microfiche. </note>

Since this information is entered without qualifying elements providing context and meaning, these individual notes within the *note* element rely solely on the context of the record as a whole and moreover, without individually being structured by element or attribute, resist querying by machines individually (instead reading as collectively one string of the node *note*). These notes are mostly useful for keyword searching.

Similarly, the summary element is unstructured while providing synopsis or other descriptive details of the content being described in the record. However, it is not generally composed of discrete individualized notes but instead text that provides an uncontrolled description of the content. Sometimes, NET title records might boast two or more descriptions from different sources (usually entered on different occasions), but these descriptions can be thought of as serving the goal of providing a collective summary of the content. Converting to a metadata environment where summaries are repeatable, such as MARC, would make this much simpler, but the difficulty with NET collection records is that NET summaries are frequently long and complex and devising a strategy to separate NET summaries from other summaries could prove challenging, especially since NET summaries often have subsections describing series information, episode information, and other miscellanea. As an example of a summary with two different sources, below is the summary from the same record described above in **figure 10**:

[summary](#)>Hosted and presented by Hazel Barnes and broadcast over PBS station KRMA in Boulder, CO.

From NET microfiche:

General Description of Series:

SELF ENCOUNTER is a series designed to explain and illustrate the most important principles of existential philosophy, and the implications of their application to everyday life and problems. The title suggests the two themes of the series: one, an explanation of the existential thesis that man must meet and recognize himself honestly, without recourse to myths or vain or supernatural hopes; two, the attempt to draw each viewer of the series into a closer and more careful understanding of himself. The technique used to clarify these themes is a combination of lecture and drama. Dr. Hazel E. Barnes, professor of classics at the University of Colorado and a noted student of existential philosophy, is the host for the series. She describes, in a direct, almost lecture style, the themes and topics most important to an understanding of existentialism. Her comments alternate with scenes from plays or novels by noted authors whose work reflect, or explain, existentialism; these dramatizations, performed by students at the University of Colorado, do much to clarify the material Dr. Barnes has been discussing.

The series was produced by KRMA-TV, Denver. Director: James Case. Producer: John Parkinson.

Featured Personality:

Dr. Hazel E. Barnes, who received her PhD from Yale University, has taught or studied classics and philosophy in North Carolina, Ohio, New York City, Hawaii and Athens, Greece. Currently she holds the title of professor of classics at the University of Colorado; however, she teaches not only classics, but also humanities and an occasional course in philosophy. And she has found time to write, edit or translate four books and a considerable number of articles, and to prepare a series for radio called ["Philosophy of You."](#)

Program 10: Sin without God

This program deals with the way the religious existentialist and the humanistic existentialist look at the subject of sin. To those who believe in God, the concept of sin makes sense. But what becomes of good and evil, right and wrong, if there is not God to refer to? Can there be sin without God? Some say there can be, that sin takes place in the gap between what man is and what he aspires to be. Dr. Barnes relates this view of sin to the existentialist idea of freedom. Pointing up these philosophical questions are scenes from Camus's [The Plague](#) and the play [Native Son](#) by Paul Green and Richard Wright. There is considerable dramatic impact in the excerpt from [Native Son](#), which shows a Negro on trial for murder. This is its theme: This man committed a crime, but society made him what he is, and thus society itself is on trial. The scene illustrates quite well a point that Dr. Barnes is seeking to make.</summary>

MBRS decided to treat the note fields and summary fields as providing different forms of unstructured text. Consequently, MBRS used different approaches for evaluating note fields and summary fields. It was decided, at least for this initial study, to treat summaries collectively as a single summary, in essence a transcription of the MAVIS summary element for the purposes of the conversion process.

Notes fields, which have a variety of different patterns for potential Linked Data exposure (including additional named entities not enumerated as an additional entry credit), were analyzed by disclosing broad patterns and assessing their frequency and variation.

Figure 8 below illustrates a grouping of different types of note metadata under three categories – *Sources*, *Administration*, and *Relating*.

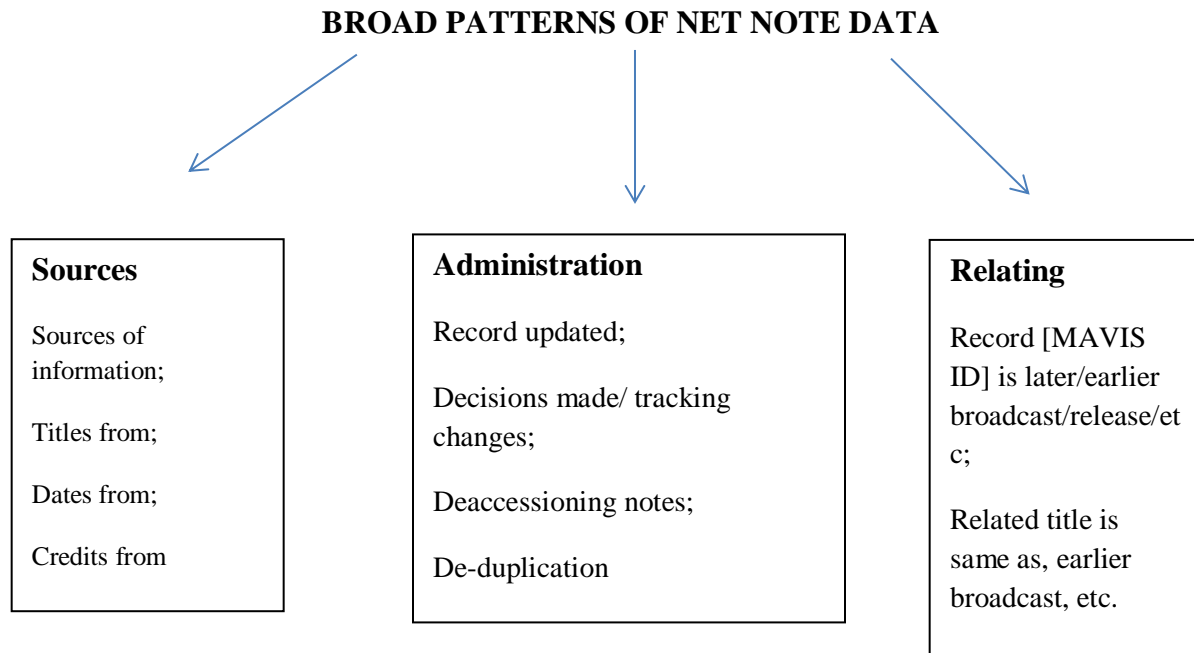


Figure 11

Using an XSLT transformation of an XML dataset of uncontrolled elements and values from the original dataset of NET playhouse/NET journal records, MBRS used a number of different general expressions targeting certain patterns of text in the notes tab in order to group these under these categories. For instance, for sources, the phrase “Sources used” was targeted and for administration cataloger initials and the word “note” were targeted and for relating the phrase “Related title” and the use of MAVIS ids were targeted. These are admittedly rough and very broad categories whose instances overlap with each other frequently. See **Figure 12** for a chart outlining the frequency of use for each category in comparison to the use of summary and note tabs on their own.

Figure 12 also points out some exceptions to the broad patterns outlined in **Figure 11**. One exception is that date notes were actually counted not from the note element but from the date note element, which is an element that was used for nearly every single NET record to record the source of the dates provided in the title record where there are extra patterns within these date notes, such as tracing whether a date is a re-broadcast date. Credit notes are also included in **Figure 12**, even though these are external to the note tab, for reasons similar to the date notes.

It should also be remarked that the practice for NET cataloging that even though credits are listed both in the summary element and are also added as access points by catalogers, it was also

important to trace the frequency of additional credits notes in the note field because these might be sources for future Linked Data exposure, although it should be noted that NET cataloging was rigorous in adding most identifiable entities as access points, even if there were no established name authority records for the entity in LCNAF. Additional credits notes are thus notes for credits that have not been established in LCNAF or were otherwise judged as not useful as an access point and are included only in the note field, not as access point or in the summary.

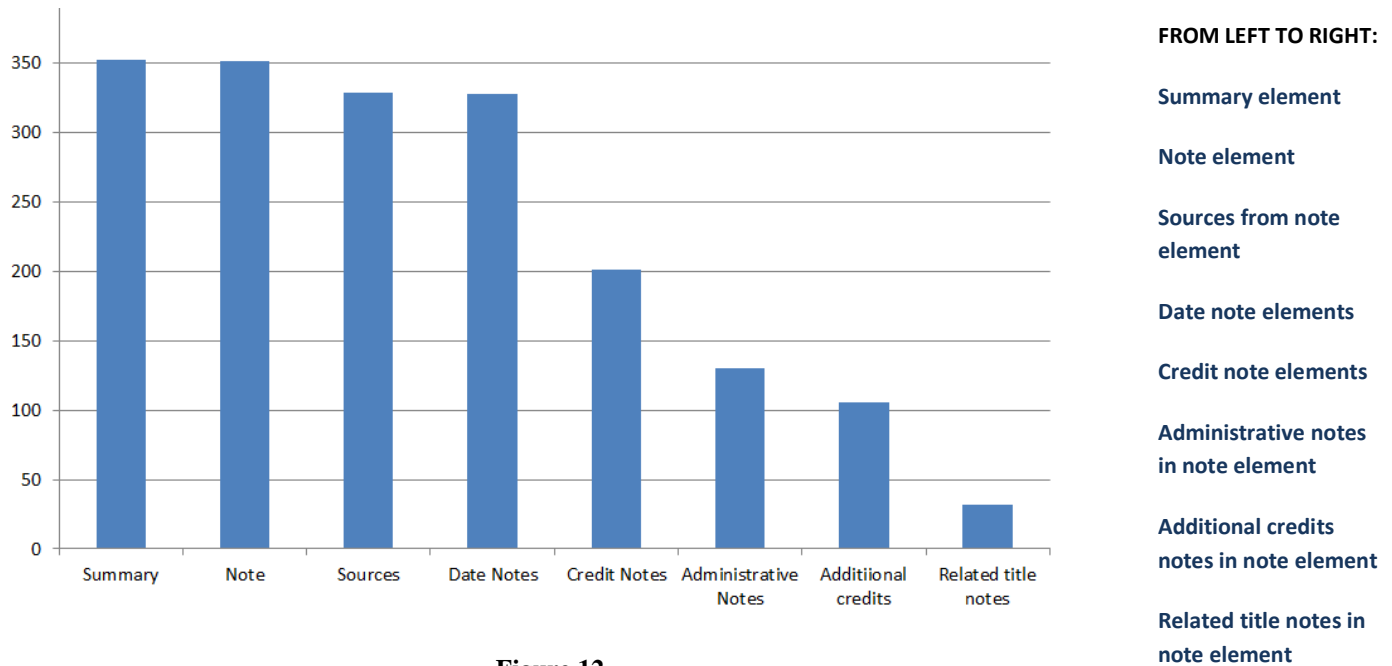


Figure 12

How can any of the above be leveraged into Linked Data?

The first, and perhaps most important aspect of leveraging uncontrolled metadata into Linked Data, is to incorporate the patterns into the Crosswalking phase. One opportunity would be to map them into MARC or PBCore before mapping to BIBFRAME or EBUCore extensions. Of course, this would depend as well upon availability of transformations and the degree to which potential mapping is reliable, and it should also be understood that since uncontrolled fields are unstructured there is a great deal of fuzziness in matching patterns during transformation so it should be expected that some amount of data cleaning would be necessary after conversion from MAVIS into another data model. It should be noted that relation notes might be problematic in mapping to MARC since MARC's representation of relations between titles might not necessarily map to the unstructured related title notes catalogers entered while cataloging NET titles, as illustrated by the "more researched needed" type of relationship in the example below:

Not viewed. More research needed to determine the relationship between this program and MAVIS 2322080, Platform. Dr. Thomas Dooley (linked as a related title).

A major issue, however, in leveraging any potential mapping between unstructured text in MAVIS and standards downstream of the conversion process to BIBFRAME is the fact that

MBRS uses the MAVIS to PBCore XSLT provided by MAVIS developers to transform MAVIS datasets to PBCore. This and similar challenges are discussed in the following section.

MAVIS to PBCore XSLT stylesheet

PBCore is a data model that originated in the public broadcasting community as a metadata standard designed for representing elements useful for the exchange of records related to the production and distribution of moving image and recorded sound content. The XML schema has been adopted by a broader range of institutions managing moving image and recorded sound collections. MBRS, as explained above, does not describe its moving image or recorded sound collections in PBCore, but, rather, in MAVIS-XML, which is the serialization format for MBRS's collection management system. However, MBRS does convert MAVIS-XML records to PBCore through a XSLT stylesheet designed by MAVIS developers.

Since MBRS was working with a previously existing stylesheet, data modeling decisions and crosswalking formalizations had already been determined. Even so, it is important to be aware of idiosyncrasies in the implementation of the transformation that could affect conversion downstream. Additionally, it is necessary to trace areas where possible Linked Data exploits as discussed above are minimized by this transformation and also how potential Linked Data exploits identified in the examination of uncontrolled fields were preserved.

Below, the handling of broad patterns of NET note data identified in the uncontrolled fields section by the MAVIS to PBCore stylesheet are listed.

1. *Sources*

```
<pbcoreAnnotation annotationType="Note">Sources used: Lost UK TV shows search
engine WWW site,
viewed June 28, 2010; Public Broadcasting Services videos research database; NET
microfiche; PBS-1994.xls Excel spreadsheet (on MBRS shared drive). Additional
credit notes:
Death -- Robin Chapman Cousin -- Patricia English Goods -- Arthur Pentelow
Confession --
Ralph Michael Discretion -- Derek Birch Five-Wits -- Sean
Lynch</pbcoreAnnotation>
```

Source notes were preserved only if they were present in the MAVIS XML note element; source notes – such as dates notes or credits notes – were not preserved in the MAVIS to PBCore transformation.

2. *Administration*

```
<pbcoreAnnotation annotationType="Note">NOTE: VUB 3028, 2" video, was
erroneously added to
MAVIS 2117330. Component was deleted from this record after it was moved to
```

MAVIS 2413283,

per CPM 16-01 instruction that AAPB Collection records should include only digital assets.

LSLee. 2017-07-14.</pbcoreAnnotation>

Administration notes recorded at the title record level were preserved as pbcoreAnnotation notes since they were entered in the note element in MAVIS XML. It should be noted as well that instantiationAnnotation was also used to preserve MAVIS component and carrier notes.

3. *Related titles*

<pbcoreAnnotation annotationType="Note">Sources used: NET microfiche; PBS-1994.xls Excel

spreadsheet (on MBRS shared drive). Title from collection documentation is from

NET

microfiche. Related title is the original broadcast of this title. This title aired without the original's filler and under the umbrella of NET playhouse biography rather than

NET

playhouse.</pbcoreAnnotation>

Related titles notes were generally recorded in the MAVIS XML note element and were thus preserved during the transformation from MAVIS to PBCore.

4. *Additional credits*

<pbcoreAnnotation annotationType="Note">Sources used: NET microfiche; PBS-1994.xls Excel

spreadsheet (on MBRS shared drive); IMDB Web site, viewed 4/25/2018,

<https://www.imdb.com/title/tt0060860/> Record updated for NET cataloging.

CMPierce

4/25/2018. Title from collection documentation is from NET microfiche. Additional

credits

note: Anne of Austria - Katharina Renn Fouquet - Pierre Barrat Mme. Du Plessis -

Dominique

Vincent Louise de la Valliere - Francoise Ponty Marie-Therese - Joelle Laugeois

D'Artagnan

- Maurice Barrier Father Joly - Andre Dumas Photography by: Georges LeClerc Art

direction

by: Maurice Valey Costumes by: Christiane Coste Sound by: Jacques Gayet From

NET

microfiche: Acquisition of this program was made possible by grants from the

National

Endowment for the Humanities and the Andrew W. Mellon
Foundation.</pbcoreAnnotation>
<pbcoreCreator>

Additional credits notes were preserved during the conversion from MAVIS to PBCore since they were entered as notes in the note element in MAVIS XML.

Below is a NET collection record that has been transformed from MAVIS XML to PBCore followed by a sampling of a spreadsheet of the same record in **Figure 13**. With this record it is possible to identify some data modeling challenges this XSLT transformation of MAVIS to PBCore brings to the path to Linked Data. Some of these issues will be enumerated below with reference to the NET record for the program below, *Onion Johnnie*.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<PBCoreDescriptionDocument xmlns:xl="http://www.w3.org/TR/xlink"
  xmlns:mv="http://www.wizardis.com.au/2005/12/MAVIS"
  xmlns="http://www.pbcore.org/PBCore/PBCoreNamespace.html"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://www.pbcore.org/PBCore/PBCoreNamespace.html
http://www.pbcore.org/PBCore/PBCoreSchema.xsd">
  <pbcoreAssetType>Moving Image</pbcoreAssetType>
  <pbcoreAssetDate dateType="BROADCAST">1963</pbcoreAssetDate>
  <pbcoreIdentifier source="MAVIS Title Number">2321420</pbcoreIdentifier>
  <pbcoreIdentifier source="NET/PBS NOLA">ONJO</pbcoreIdentifier>
  <pbcoreIdentifier source="NET/PBS number">1</pbcoreIdentifier>
  <pbcoreTitle titleType="Preferred Title">Onion Johnnie.</pbcoreTitle>
  <pbcoreSubject source="MAVIS Subject Authority List Library of Congress" subjectType="Topic"
    >Farmers</pbcoreSubject>
  <pbcoreSubject source="MAVIS Name Authority List Library of Congress" subjectType="About"
    >Brittany (France)</pbcoreSubject>
  <pbcoreDescription descriptionType="Original Summary">From NET microfiche: This excellent
    documentary won first prize for television films at the 1956 Vancouver International Film
    Festival. The judges commented, "The narration is charming, the interest of the viewers is
    sustained throughout, the treatment of the subject shows understanding, gentleness and
    tenderess." The on-location film follows the picturesque life of a Breton onion farmer.
    Viewers see "Onion Johnnie" as he plants and harvests his crops, ship it from Roscoff in
    Brittany across the channel to England, take leave of his family and goes to England for about
    six months to sell his onions. When his crop is sold, his year has come full cycle and he
    returns to his home to plant onions for the next year. This production's "Onion Johnnie" is M.
    Francois Mazeas, deputy mayor of Roscoff in Brittany and an onion farmer for thirty
    years.</pbcoreDescription>
  <pbcoreGenre source="MAVIS Genre Authority List Library of Congress">Documentary television
    programs (lcgft)</pbcoreGenre>
  <pbcoreGenre source="MAVIS Genre Authority List Library of Congress">Educational television
    programs (lcgft)</pbcoreGenre>
```

```

<pbcoreGenre source="MAVIS Genre Authority List Library of Congress">Nonfiction television
  programs (lcgft)</pbcoreGenre>
<pbcoreAnnotation annotationType="Note">Sources used: NET microfiche; PBS-1994.xls Excel
  spreadsheet (on MBRS shared drive).</pbcoreAnnotation>
<pbcoreAnnotation annotationType="Nomination / Award Notes">First prize at 1956 Vancouver
  International Film Festival</pbcoreAnnotation>
<pbcoreCreator>
  <creator>British Broadcasting Corporation ; producer, Richard Cawston ; writers, Richard
    Cawston, Stephen Hearst.</creator>
  <creatorRole>Statement of Responsibility</creatorRole>
</pbcoreCreator>
<pbcoreContributor>
  <contributor>National Educational Television and Radio Center</contributor>
  <contributorRole>Broadcaster</contributorRole>
</pbcoreContributor>
<pbcoreCreator>
  <creator>British Broadcasting Corporation</creator>
  <creatorRole>Production Company</creatorRole>
</pbcoreCreator>
<pbcoreCreator>
  <creator>Cawston, Richard</creator>
  <creatorRole>Producer</creatorRole>
</pbcoreCreator>
<pbcoreCreator>
  <creator>Cawston, Richard</creator>
  <creatorRole>Writer</creatorRole>
</pbcoreCreator>
<pbcoreContributor>
  <contributor>Mazeas, Francois</contributor>
  <contributorRole>Appearing</contributorRole>
</pbcoreContributor>
<pbcoreCreator>
  <creator>Hearst, Stephen</creator>
  <creatorRole>Writer</creatorRole>
</pbcoreCreator>
<pbcoreContributor>
  <contributor>Brunius, Jacques-B.</contributor>
  <contributorRole>Narrator</contributorRole>
</pbcoreContributor>
<pbcoreInstantiation>
  <instantiationIdentifier source="MAVIS Item ID">2321420-1</instantiationIdentifier>
  <instantiationLocation>Unknown</instantiationLocation>
  <instantiationStandard>Film</instantiationStandard>
  <instantiationMediaType>Moving Image</instantiationMediaType>
  <instantiationEssenceTrack>
    <essenceTrackType>Film</essenceTrackType>
  </instantiationEssenceTrack>
</pbcoreInstantiation>
</PBCoreDescriptionDocument>

```

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	
Creator	Creator Role	Genre	Genre Source	PBCORE identifier	PBCORE identifier source	Contributor	Contributor role	Subject	Subject source	Subject type	Annotation	Annotation type	Title	Title type	Asset Date	Asset date type	Asset type	Description
British Broadcasting Corporation ; producer, Richard Cawston ; writers, Richard Cawston, Stephen Hearst.	Statement of Responsibility	Documentary television programs (lcgft)	MAVIS Genre Authority List Library of Congress	2321420	MAVIS Title Number	National Educational Television and Radio Center	Broadcaster	Farmers	MAVIS Subject Authority List Library of Congress	Topic	Sources used: NET microfiche; PBS-1994.xls Excel spreadsheet (on MBRS shared drive).	Note	Onion Johnnie.	Preferred Title	1963	BROADCAST AST	Moving Image	From NET micro excellent documentary television films at Vancouver International Festival. The commented, "The charming, the in viewers is sustained the treatment of the understanding, the tenderness." film follows the Breton onion farmer. Viewers see as he plants and crops, ship it from Brittany across England, take it and goes to England six months. When his crop comes full cycle returns to his onions for the next production's "Or Francois M. mayor of Rosco onion farmer for years.
British Broadcasting Corporation	Production Company	Educational television programs (lcgft)	MAVIS Genre Authority List Library of Congress	ONJO	NET/PBS NOLA	Mazeas, Francois	Appearing	Brittany (France)	MAVIS Name Authority List Library of Congress	About	First price at 1956 Vancouver International Film Festival	Nomination / Award Notes						

Figure 13

1. Subjects lack clear source and identifier attributes

```

<pbcoreSubject source="MAVIS Subject Authority List Library of Congress" subjectType="Topic"
>Farmers</pbcoreSubject>
<pbcoreSubject source="MAVIS Name Authority List Library of Congress" subjectType="About"
>Brittany (France)</pbcoreSubject>

```

Subjects in MAVIS can be local or identified as a Library of Congress Subject Heading (LCSH) with the applicable LCCN identifier, as illustrated in **Figure 14**.

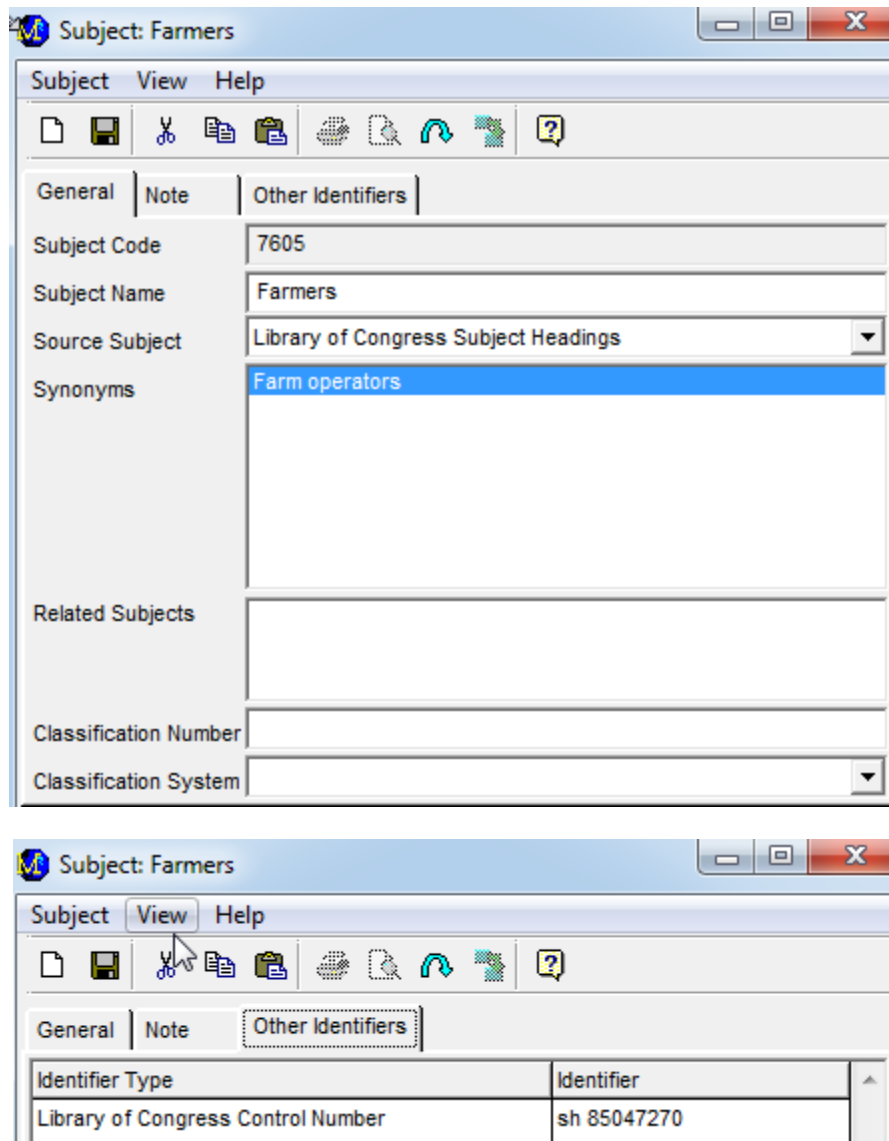


Figure 14

A clearer mapping would have identified the source as LCSH, followed by a reference attribute with the URI taken from a Linked Data authority service, in this case LC's Linked Data Service (id.loc.gov):

```
<pbcoreSubject source="Library of Congress Subject Headings"
ref="http://id.loc.gov/authorities/sh85047270#concept" subjectType="Topic" >Farmers</pbcoreSubject>
```

A strategy needs to be established to reconcile subject authorities with Linked Data authority services already available. Even if it is possible to reconcile subjects in MAVIS to a Linked Data Service, work will remain on how to handle post-coordinated subject strings and identities that are local in MAVIS and do not correspond to an existing Linked Data authority service.

2. *Statement of responsibility*

```
<pbcoreCreator>
  <creator>British Broadcasting Corporation ; producer, Richard Cawston ; writers, Richard
    Cawston, Stephen Hearst.</creator>
  <creatorRole>Statement of Responsibility</creatorRole>
</pbcoreCreator>
```

The transcribed *statement of responsibility* element in MAVIS-XML does not have a clear map to PBCore. The decision made in the MAVIS XSLT was to target the *creator* element in PBCore. The problem, however, is that the statement of responsibility is a transcribed field that copies statements of creative responsibility from chosen sources of information that generally name multiple identities. The *statement of responsibility* is useful in that it records how entities originally were identified, then used as the basis for establishing authorized access points. For moving image materials, the *statement of responsibility* lists a grouping of entities with primary creative responsibility due to the collaborative nature of motion picture and television production, while PBCore requires a single named entity that can be referenced to a single role. With the latter in mind, it may be preferable to map the *statement of responsibility* to a Note element with source equaling statement of responsibility. Doing so will retain the *statement of responsibility* in those instances when no creative agents in MAVIS are established as authorized access points. This will be traced in the Crosswalk.

3. *Contributor and creator and the question of transcription*

```
<pbcoreCreator>
  <creator>Cawston, Richard</creator>
  <creatorRole>Writer</creatorRole>
</pbcoreCreator>
<pbcoreContributor>
  <contributor>Mazeas, Francois</contributor>
  <contributorRole>Appearing</contributorRole>
</pbcoreContributor>
```

PBCore creator elements are ambiguously defined as far as how values are constructed. The XSLT pulls from authorized access points without referencing identifiers or even the LCNAF. Without using identifiers, it would appear that the creator fields resulting from the transformation are not taken from an established authority list. This matters particularly when one considers the difference between the following values: “Richard Cawston” and “Cawston, Richard.” For the purpose of collocation, disambiguation, and indexing it is recommended that an authorized list is used that is referenced using PBCore’s *source* and *ref* attributes to clearly identify creative agents. In order to do this effectively, all names in MAVIS that were created locally (i.e. names lacking a LCCN identifier referencing the LCNAF) will require the formal establishment of names as authorized access points in the LCNAF.

It is important to note that some of the issues identified are not necessarily symptomatic of PBCore in general, but, rather, representative of the stylesheet's mapping between PBCore and MAVIS. One of the challenges in this conversion process will be identifying gaps between stylesheet transformation of MAVIS to PBCore and more ideal transformations between the two standards. Some of these problems can lead to confusing or messy metadata such as the confusion between authorized access points established in the LCNAF and transcribed names in the *statement of responsibility*.

Phase 3: Crosswalking

As discussed in the Data Modeling phase, the Crosswalking phase was focused on aligning MAVIS and PBCore with BIBFRAME and addressing any gaps in this alignment with possible extensions to EBUCore. This required mapping MARC21, MAVIS, PBCore, MARCXML, BIBFRAME, and EBUCore elements and properties into a systematic presentation, while determining the degree of match between the MARC/BIBFRAME standards and the MAVIS/PBCore standards in order to recommend EBUCore extensions.

Granularity was maintained as much as possible with the “left anchored nature” of the crosswalk's focus on MARC fields, and where possible, mappings were provided for MAVIS/PBCore fields that do not easily map to MARC, such as elements related to documenting migration between different formats. This means that mapping between attributes, indicators, and subfields were often traced in the same row, often through notes explaining the reasoning.

There were also concerns with differences between an ideal mapping between MAVIS and PBCore and how the MAVIS to PBCore stylesheet handled the mapping. In a separate crosswalk, a core XSLT mapping was designed based on a NET descriptive record that had been converted to PBCore using the stylesheet. The record was transformed with OpenRefine from XML to a spreadsheet and the titles of the columns were normalized to reflect PBCore names (rather than the hierarchy of elements resulting from the conversion from XML) and these columns were transposed into rows, with new columns representing the following information: MAVIS source, MAVIS XML path, Match, Control type, Data type, PBCore scope notes, MAVIS example, PBCore example, MAVIS object type (title, component, carrier), PBCore object type (asset, instantiation, essence track), Notes, XSLT examples (images from the XML documenting complex points in the transformation). An example from this spreadsheet is provided below in **Figure 15**, but it should be noted that this spreadsheet was limited to the elements that were in the record, so its value is a bit limited; MBRS also made use of mapping documentation provided by the developers who designed the MAVIS to PBCore stylesheet as reference.

1	PBCORE Target	MAVIS source	MAVIS path	Match	Control type	Data type	PBCore scope notes	MAVIS example
2	pbcoreCreator	Name-Role	<Name-Role> <party xl:href="[local identifier" xl:title="Authorized heading"/>	narrower than	Authorized name heading	String	The element creator identifies the primary person, people, or organization(s) responsible for creating the asset. Note that non- primary names and roles should be included within the pbcoreContributor container.	<Name-Role> <party xl:href="/Organisation/key/241 0-1" xl:title="National Educational Television and Rad Center"/>

Figure 15

Additionally, the following were also consulted as sources for the Crosswalk: documentation of the mapping done by MAVIS developers who designed the XSLT transformation to PBCore (as mentioned above), work being done currently by the Association of Moving Image Archivists' (AMIA) PBCore Advisory Sub-Committee to map from PBCore to MARC, and conversion specifications provided by the Library of Congress's Network Standards and MARC Development Office regarding the transformation of MARC to BIBFRAME in addition to the ontologies themselves and the scope notes and comments contained therein..

The main crosswalk is divided into five sections, outlined below:

XML elements					
MARC FIELD	MAVIS element	PBCore element	MAVIS to PBCore XSLT element	MARCXML tag, indicator, and subfields	Notes
010, \$a (LC identifier)	objectIdentifier + identifierType="Library of Congress control number"	pbcoreIdentifier @source="Library of Congress Control Number"	pbcoreIdentifier @source="Library of Congress Control Number"	tag=010, ind1="" ind 2="",subfield a	
024, \$a (Other Standard identifier)	objectIdentifier + identifierType=	pbcoreIdentifier @source	pbcoreIdentifier @source	tag=024, ind=7 ind2=0, subfields "a" and "2"	Common to NET collection is NOLA code and NET/PBS number, which also repeat as component and carrier identifiers (and which can also be mapped to instantiationIdentifier in pbcore).
033 \$a (Date/Time of place of event)	objectDate	pbcoreAssetDate @source="BROADCAST" [or] "RECORDED"	pbcoreAssetDate @source="BROADCAST" [or] "RECORDED"	TAG=033 ind=0 ind2=0 [or] 1 subfields"a"	broadcast, recording as opposed to distribution dates; see 260 \$c

Figure 16

A. XML elements – This section covers the crosswalk between MAVIS, PBCore, and MARCXML.

1. MAVIS element – the source element in MAVIS
2. PBCore element – the target element in PBCore
3. MAVIS to PBCore XSLT element – how the XSLT handled this transformation
4. MARCXML tag, indicator, and subfields – the target element in MARCXML; mirrors the left anchored MARC fields
5. Notes – general notes related to the XML crosswalk

G	H	I	J	K
XML data properties				
MARC match	Control type	Controlled vocabulary list references	Data type	Notes
Match	Machine generated	LCNAF	String	
Match	Transcription		String	Can be machine generated depending upon source and type of identifier
Match	Transcription		Date	

Figure 17

B. XML data properties – properties describing the XML crosswalk data

1. Match – the relative reliability of the match to MARC
2. Control type – how values for the elements are controlled
3. Controlled vocabulary list references – references to controlled vocabulary relevant for the element
4. Data types – what types of data are used for the elements (string, date, etc.)
5. Notes – general notes about the XML properties

BIBFRAME crosswalking							
BIBFRAME pattern	BIBFRAME properties	BIBFRAME domains	BIBFRAME ranges [classes]	Data types	URI services	BIBFRAME term match	Notes
IdentifiedBy-lccn--rdf:value	IdentifiedBy <http://id.loc.gov/ontologies/bibframe.html#p_IdentifiedBy>		LCCN <http://id.loc.gov/ontologies/bibframe.html#c_Lccn>		http://id.loc.gov/authorities/names.html	Match	LCCN is sub-class of Identifier; inverse of Identifies <http://id.loc.gov/ontologies/bibframe/identifies>
IdentifiedBy - Identifier rdfs:label "content of \$2"	IdentifiedBy <http://id.loc.gov/ontologies/bibframe.html#p_IdentifiedBy>		Identifier <http://id.loc.gov/ontologies/bibframe.html#c_Identifier>			Match	Identifiers usually identified through ind1 are identified through the following specifications: IdentifitedBy--"Isrc, Upc, Ismn, Ean, Sici" (International Standard Recording Code, Universal Product Code, International Standard Music Nujmber, International Article Number, Serial Item and Contribution Identifier)
capture - Capture	capture <http://id.loc.gov/ontologies/bibframe/capture>		Capture <http://id.loc.gov/ontologies/bibframe/capture>			Match	did not check EBU CORE mapping

Figure 18

C. BIBFRAME crosswalking

1. BIBFRAME pattern – a representation of the pattern of RDF statements the properties would appear in, sometimes copied from the BIBFRAME conversion specs (in this section defaulting to a RDF-XML serialization format).
2. BIBFRAME properties – properties mapped to the elements.
3. BIBFRAME domain – the domain the property defines.
4. BIBFRAME ranges [classes] – classes that can be the object of the property
5. Datatypes – datatypes
6. URI services – links to lists for URIs that can be re-used for these values
7. Match – how reliable the match in BIBFRAME is to the elements
8. Notes – general notes about the BIBFRAME crosswalking

	T	U	V	W	X	Y	Z
1	EBUCORE crosswalking						
2	EBUCORE pattern	EBUCORE properties	EBUCORE domains	EBUCORE ranges [classes]	Data types	EBUCORE term extension	Notes
3	identifier-identifier-identifierType="LCCN"	identifier <http://www.ebu.ch/metadata/ontologies/ebucore/ebucore#identifierValue> [and] hasIdentifierType <http://www.ebu.ch/metadata/ontologies/ebucore/ebucore#hasIdentifierType>	Identifier <http://www.ebu.ch/metadata/ontologies/ebucore/ebucore#Identifier>		URI or string	No	annotation property
4	identifier-identifier-identifierType	identifier <http://www.ebu.ch/metadata/ontologies/ebucore/ebucore#identifierValue> [and] hasIdentifierType <http://www.ebu.ch/metadata/ontologies/ebucore/ebucore#hasIdentifierType>	Identifier <http://www.ebu.ch/metadata/ontologies/ebucore/ebucore#Identifier>		URI or string	No	annotation property
5							

Figure 19

D. EBUCore crosswalking

1. EBUCore pattern – a representation of the pattern of RDF statements the properties would appear in
2. EBUCore properties – properties mapped to the elements.
3. EBUCore domain – the domain the property defines.
4. EBUCore ranges [classes] – classes that can be the object of the property
5. Data types – datatypes often defined as ranges or as annotation types.
6. EBUCore term extensions – recommendations to extend to EBUCore
7. Notes – general notes about the EBUCore crosswalking

AA	AB	AC	AD
Object types			
MAVIS object types	PBCore object types	BIBFRAME object types	Object types notes
TitleWork Component Carrier	Asset	Instance Work Item	BIBFRAME conversion notes specify Instance, but comments in ontology record identifiedBy as "used with unspecified"
TitleWork Component Carrier	Asset	Work Instance Item	BIBFRAME conversion notes specify Instance, but comments in ontology record identifiedBy as "used with unspecified"
	Asset		
			BIBFRAME conversion notes specify

Figure 20

E. Object types – the types of objects the elements described for MAVIS, PBCore, and BIBFRAME and notes regarding the object types.

Crosswalking observations

Below, observations that arose from the crosswalking are discussed, many of which are connected to observations made during the Data Modeling and the Evaluation phases.

Not everything is transformed from MAVIS to PBCore:

A number of elements are not imported from MAVIS to PBCore that have not been noted yet.

For instance, acquisition information is not mapped – neither the identifier for specific acquisition records associated with items from the NET and PBS collections at the Library nor strings representing the identity of donors, dates of receipt, etc. However, since PBCore is focused on distribution of media and exchange of records, acquisition metadata would not need to be managed with the same standard or even in the same database, and acquisition information is rarely shared broadly between institutions.

Also, a statement about the country of origin does not appear to be mapped to PBCore. Country of origin statement is valuable because it traces the production origin of AV material and thus allows users to narrow or broaden searches based on this particular factor. In the NET collection, there's a large amount of international content, so it is particularly important to provide support for such searches. The transformation does import MAVIS' objectPlace element, but this is a transcription element describing specific recording or location information that was not controlled through local or external codes or vocabulary lists. An example of objectPlace is provided below:

```
<ObjectPlace>
  <place>Los Angeles, California--Ambassador Hotel ballroom, Good Samaritan Hospital,
Watts neighborhood</place>
  <placeType xl:href="/Code/key/OBJECT_PLACE_TYPE/TITLE/L"
xl:title="Location">L</placeType>
</ObjectPlace>
```

The discussion of pbcoreCoverage below covers this in more detail.

Also, the Evaluation phase indicated some note fields that were not imported using the MAVIS to PBCore stylesheet, the date and credit notes particularly, and the component and carrier notes were not imported to instantiationAnnotation.

Dissimilarities between ideal MAVIS to PBCore mapping and the XSLT:

In the Evaluation phase, some discrepancies in the XSLT mapping were discussed, including the values of pbcoreCreator and pbcoreContributor, the lack of identifiers, and problems with statements of responsibility. The statement of responsibility element discrepancy was noted by the difference between the mapping for the ideal MAVIS to PBCore crosswalk and the XSLT and also through a note. Identifiers were mapped for the most part as if they had been included, but the conversion process will need to identify when and how mapping to Linked Data

authorities for names, subjects, and genres will be handled. Additionally, some new discrepancies were discovered, including the mapping of pbcoreAssetType to medium in MAVIS whose values better match instantiationMediationType.

pbcoreCoverage:

Statements about countries of origin and statements about locations related to the recording or performance of an event are challenging, since the XSLT mapped statements about the former to statements about the latter. Instead of mapping MAVIS objectPlace values to pbcoreCoverage, which normally maps to MARC 522 (Geographic coverage), MBRS instead mapped it to 370 (Associated place). The MAVIS element country (a child of workCountry) is mapped now directly to MARC elements dealing with country of origin, including 044 and 257. However, the MAVIS to PBCore XSLT does not map workCountry.

Miscellaneous issues:

A number of peculiarities are noted in note fields, usually in the XML elements section, but also throughout. For situations where there are, for instance, two elements mapped to the same MARC field, notes were provided to aid in future conversion; these were frequently the result of the MAVIS to PBCore stylesheet mapping two elements to the same values in MAVIS, for instance essenceTrackStandard and essenceTrackEncoding, which were mapped to MARC 347 \$b; examples of the PBCore mapping is provided below:

```
<essenceTrackStandard>Video: M-JPEG 2000 Audio:
WAV</essenceTrackStandard>
<essenceTrackEncoding>Video: M-JPEG 2000 Audio:
WAV</essenceTrackEncoding>
```

```
<essenceTrackStandard>Video: MPEG2 Audio:
MPEG2</essenceTrackStandard>
<essenceTrackEncoding>Video: MPEG2 Audio:
MPEG2</essenceTrackEncoding>
```

The PBCore Advisory Sub-Committee made more extensive use of MARC 887 (Non-MARC information) mainly because MBRS leaned more extensively on an existing MAVIS to PBCore XSLT whose values are not necessarily ideal values and so were mapped as found and because MBRS was interested in seeing this metadata populate fields that are targeted for conversion to BIBFRAME in MARC. Consequently, some of the mappings to MARC might be secondary or tertiary choices by the Sub-Committee. MBRS has also tried to limit the use of 887 to migrations between different formats, which don't appear to have reliable mapping to MARC.

It should also be noted that some PBCore and MAVIS elements were excluded from this crosswalk. Below is a list of PBCore elements that were not mapped, either because they were

not part of the NET collection or because they had not been mapped to PBCore from MAVIS elements:

pbcoreExtension
extensionWrap
extensionElement
extensionValue
extensionAuthorityUsed
extensionEmbedded
instantiationTimeStart
instantiationTracks
instantiationEssenceTrack
essenceTrackTimeStart
essenceTrackDuration
essenceTrackLanguage
essenceTrackDuration
essenceTrackBitDepth
essenceTrackIdentifier
essenceTrackExtension
instantiationExtension
instnatiationAnnotation*
instantiatonRights
rightsLink
rightsEmbedded
*Matching elements in MAVIS that were not imported to PBCore

EBUCore possible extensions

MBRS decided on possible EBUCore extensions based on the reliability of matches to MARC and BIBFRAME. These extensions are listed below, including notes about issues relating to the extension.

It should be noted that implementing BIBFRAME has not been entirely explored as of yet for the NET collection beyond providing a preliminary dataset and crosswalk. Implementation would need to be examined for the effects on discovery, access, and inferential querying of semantic technologies that could exploit the semantic relationships and URIs provided in the dataset. An example of a problem that is currently notable in the dataset (and because of which an extension to EBUCore has been included) is mapping component identifiers from MAVIS to instantiationIdentifiers in PBCore and to 776 in MARC, which is transformed to BIBFRAME as extra items not connected to the item metadata otherwise referenced; this was a temporary measure to preserve the component identifiers, but it's not ideal. EBUCore extensions could likely resolve this problem.

MAVIS	PBCore	MARC	EBUCore
Video tape component, audio tape component, disc component, nitrate component, acetate component=strings="Film, video, audio"	essenceTrackType	337, \$c (Media type, specified materials)	trackType <http://www.ebu.ch/metadata/ontologies/EBUCore/EBUCore#trackType>
sampleRate	essenceTrackPlaybackSpeed	345, \$b (Projection speed)	playbackSpeed <http://www.ebu.ch/metadata/ontologies/EBUCore/EBUCore#playbackSpeed>
XSLT duplicates another mapping for this one	essenceTrackEncoding	347, \$b (Encoding format)	hasEncodingFormat <http://www.ebu.ch/metadata/ontologies/EBUCore/EBUCore#hasEncodingFormat>
Video height; Video width + Video Frame Layout	essenceTrackFrameSize	347, \$d (Resolution) ; this does not exactly map to MARC and another pbcore/mavis element is also mapped here.	width <http://www.ebu.ch/metadata/ontologies/EBUCore/EBUCore#width> ; height <http://www.ebu.ch/metadata/ontologies/EBUCore/EBUCore#height>
Digital component - Date copied	instantiationDate +dateType="digitized"	887 (Non-MARC information)	dateDigitized <http://www.ebu.ch/metadata/ontologies/EBUCore/EBUCore#dateDigitised>
objectPlaces + objectPlace + place + type	pbcoreCoverage + @type="Spatial" (XSLT); pbcoreAnnotation @annotationType="location"	370, \$i, \$c, \$g (Associated place)	hasEventRelatedLocation <http://www.ebu.ch/metadata/ontologies/EBUCore/EBUCore#hasEventRelatedLocation>; can be used with eventType <http://www.ebu.ch/metadata/ontologies/EBUCore/EBUCore#eventType> with range of string or class Event <http://www.bbc.co.uk/ontologies/coreconcepts/Event>
	pbcoreCoverage (not event related but coverage of content)	522 (Geographic coverage) [or] 500 note [or] 651 [or] Geographic subdivision	hasCoverage <http://www.ebu.ch/metadata/ontologies/EBUCore/EBUCore#hasCoverage>
language @title [title level]; language @title [component level]	instantiationAlternative Mode	546 (Language)	hasDubbedLanguage <http://www.ebu.ch/metadata/ontologies/EBUCore/EBUCore#hasDubbedLanguage> ; subtitles can be signaled via the hasSubtitling property <http://www.ebu.ch/metadata/ontologies/EBUCore/EBUCore#hasDubbedLanguage>
Copy history [from technical document]	instantiationRelation + instantiationRelationType + instantiationRelationIdentifier	887 (Non-MARC information)	clonedTo <http://www.ebu.ch/metadata/ontologies/EBUCore/EBUCore#clonedTo> clonedFrom <http://www.ebu.ch/metadata/ontologies/EBUCore/EBUCore#isClonedFrom>
Component href key	instantiationIdentifier	776 (Other physical item)	identifierValue <<http://www.ebu.ch/metadata/ontologies/EBUCore/EBUCore#identifierValue>

Phase 4: Publishing

Much of the project-based literature encompassing the exploration of Linked Data by cultural heritage institutions is light on the topic of publishing RDF triples, beyond the idea of transforming datasets into RDF. There is plenty of literature about the different types of services that can be offered by different projects for implementing linked data, but the publishing of linked data itself is often regarded as a future question.

One reason for this is project goals – a good many of the larger projects, while working on the conversion side themselves, have seen the issue of metadata creation as a final goal. For example, projects on BIBFRAME tend to publicly spend more time addressing pilot studies regarding the BIBFRAME editor, authority work, and cataloging workflows, which are important in gaining an understanding of the enormity of the transition towards a Linked Data environment since this is a significant shift. (Wiggins, McCallum, Frank, & Hess, 2016; Smith, Stahmer, Li, & Gonzalez, 2017).

Another issue is that publishing is often conflated with the entire process of conversion to Linked Data, such as W3C's own guidelines for publishing Linked Data (W3C Consortium & others, 2014) or the UC Davis and Zepheira study on Linked Data implementation (Smith, Stahmer, Li, & Gonzalez, 2017), both of which list roughly analogous steps outlined here under the umbrella of publishing Linked Data.

The problem is that often issues like serialization formats, data dumps, SPARQL endpoints, resolving into HTML displays, etc. are not frequently discussed (Morgan, 2014). Some of these gaps are covered simply by discussing the choices made without investigating why these choices were made and how this might have impacted the workflow (S. Bechhofer, K. Page, & D. De Roure, 2013), though there have been some developments in projects like Linked Jazz and Hardesty's work with the Avalon Media transition to Fedora 4 digital repository using RDF where reasoning is examined in regards to publishing methodology (Hardesty, 2016; Hardesty & Young, 2017; Pattuelli, Provo, & Thorsen, 2015). One of the goals of this feasibility report is to examine publishing methodology to help guide decision-making processes for other institutions.

The Linked Data implementation strategy analyzed so far has been the process of transforming NET collection metadata in LC's MAVIS-XML standard to BIBFRAME standardized RDF (with possible EBUCore extensions), while testing alignments between EBUCore and PBCore. Following this approach would appear to be committing to a particular set of Linked Data publishing strategies that are discussed below, but some alternative strategies are also considered, their strengths and weaknesses for the project overall. Finally, particular challenges with NET collection metadata, including challenges with a final conversion of Linked Data capable RDF dataset and challenges with reconciling authorities, both local and established (LCNAF, LCSH, and LCGFT) will be discussed. The goal of this section is to explore publishing methodologies broadly and not specifically to recommend any one method although, ultimately,

the focus will be on conversion and publishing a sample NET collection dataset as RDF files shared at a data repository such as GitHub or through a server.

As noted above, the conversion process analyzed commits the NET collection dataset to a data dump method of publishing. Data dumps are a method of backing up or transferring datasets and databases that can also involve sharing data more broadly. Web sites, government agencies, and other services often share their data through this method. This seems like a simple strategy, and for the purposes of many projects, this strategy of dumping of a dataset into a repository or a server to share it represents an effective low cost method of dataset publication.

However, there are a number of factors to consider for ensuring access and retrievability of information published as Linked Data, particularly in the interests of contributing to the Semantic Web, exposing previously silo-ed information, or establishing authoritative relationships on the Internet more broadly. The data dump method works particularly well for a dataset that is small and static, i.e. has a small file size and is not regularly altered. However, a dataset that is stored in a relational database or critical legacy systems that there is no intent to transition might need alternative methods to provide access to the dataset as Linked Data (Heath & Bizer, 2011).

Currently, WGBH and the Library are planning on hosting a finished dataset of NET Collection metadata on GitHub as a static file. For our purposes, this is acceptable currently as an initial step to deeper implementation of Linked Data, but processing through triplestore databases, enabled with a SPARQL endpoint, or through a Linked Data interface (such as Pubby) (Heath & Bizer, 2011) could be a future goal.

Another factor to consider is the serialization format of the static RDF dataset. Serialization refers to the format in which a dataset is “streamed” for storage, for machine processing, or for aiding access to human readability of data. Unfortunately, RDF serialization formats are numerous and bewildering in scope, goals, and value in comparison to XML. It should also be clarified that, like XML, RDF serialization formats are agnostic to ontologies and standards used to describe resources in RDF, but, in contrast to XML, RDF serialization formats have multiplied as RDF technology has adapted to the changing climate of the Internet and the shifting priorities of various stakeholders in information exchange on the Web. This has led to a bewildering number of RDF serialization formats, including RDFXML, Turtles, N3, JSON-JD, Microformats, etc. It is sometimes recommended to include publications of Linked Datasets in multiple formats for this reason. Below is some commentary and examples of some of these formats.

RDFXML is a RDF format that expresses RDF statements in the hierarchical structure of an XML document. It is the earliest RDF format, designed alongside other RDF specifications by W3C’s original efforts to provide a model for the Semantic Web. The XSLT that is used to transform MARCXML to BIBFRAME 2.0 transforms the XML records to RDFXML. RDFXML

is a common standard due to its status as an official W3C specification, but its use has fallen into some disfavor over the years, with issues with XML restrictions and with the hierarchical structure of XML distorting the graph structure of the RDF data model. The following is a snapshot of RDFXML serialization of a BIBFRAME 2.0 record:

```
<bf:Work rdf:about="http://example.org/16614942#Work">
  <bf:adminMetadata>
    <bf:AdminMetadata>
      <bf:generationProcess>
        <bf:GenerationProcess>
          <rdfls:label>DLC marc2bibframe2 v1.4.0-SNAPSHOT: 2018-06-07T10:31:46-04:00</rdfls:label>
        </bf:GenerationProcess>
      </bf:generationProcess>
    <bf:status>
      <bf:Status>
        <bf:code>c</bf:code>
      </bf:Status>
    </bf:status>
    <bf:encodingLevel>
      <bf:EncodingLevel>
        <bf:code>5</bf:code>
      </bf:EncodingLevel>
    </bf:encodingLevel>
    <bf:descriptionConventions>
      <bf:DescriptionConventions>
        <bf:code>aacr</bf:code>
      </bf:DescriptionConventions>
    </bf:descriptionConventions>
    <bf:identifiedBy>
      <bf:Local>
        <rdf:value>16614942</rdf:value>
      </bf:Local>
    </bf:identifiedBy>
```

Turtle is a RDF serialization format that supports namespace prefixes and a number of shorthand conventions, while not abandoning the hierarchical structure of RDFXML in favor of RDF statements, in syntax aligned to the SPARQL querying language. LC provides an alternative serialization of its BIBFRAME records into Turtle via its Web based comparison tool (Library of Congress, n.d.). Below is a snapshot of a BIBFRAME record serialized in the Turtle format:

```
@prefix bf: <http://id.loc.gov/ontologies/bibframe/> .
@prefix bflc: <http://id.loc.gov/ontologies/bflc/> .
@prefix madsrdf: <http://www.loc.gov/mads/rdf/v1#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
```

```

@prefix xml: <http://www.w3.org/XML/1998/namespace> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix zs: <http://docs.oasis-open.org/ns/search-ws/sruResponse> .

<http://bibframe.example.org/13714243#Item050-21> a bf:Item ;
  bf:itemOf <http://bibframe.example.org/13714243#Instance> ;
  bf:shelfMark [ a bf:ShelfMark ;
    rdfs:label "SDB 07008" ;
    bf:source <http://id.loc.gov/vocabulary/organizations/dlc> ] .

<http://bibframe.example.org/13714243#Work> a bf:Audio,
  bf:Work ;
  rdfs:label "Let it be-- naked" ;
  bf:adminMetadata [ a bf:AdminMetadata ;
    bflc:encodingLevel [ a bflc:EncodingLevel ;
      bf:code "1" ] ;
    bf:changeDate "2011-12-20T08:06:52"^^xsd:dateTime ;
    bf:creationDate "2004-09-09"^^xsd:date ;
    bf:descriptionAuthentication
<http://id.loc.gov/vocabulary/marcauthen/lcderive> ;
    bf:descriptionConventions [ a bf:DescriptionConventions ;
      bf:code "aacr" ] ;
    bf:descriptionModifier [ a bf:Agent ;
      rdfs:label "OCLCQ" ],
      [ a bf:Agent ;
        rdfs:label "JED" ],
      [ a bf:Agent ;
        rdfs:label "DLC" ] ;
    bf:generationProcess [ a bf:GenerationProcess ;
      rdfs:label "DLC marc2bibframe2 v1.3.1: 2018-06-
07T10:59:49-04:00" ] ;
    bf:identifiedBy [ a bf:Local ;
      bf:source
<http://id.loc.gov/vocabulary/organizations/dlc> ;
      rdf:value "13714243" ] ;
    bf:source [ a bf:Agent,
      bf:Source ;
      rdfs:label "GEC" ],
      [ a bf:Agent,
        bf:Source ;
        rdfs:label "GEC" ] ;
    bf:status [ a bf:Status ;
      bf:code "c" ] ] ;
  bf:capture [ a bf:Capture ;
    bf:date "1969-01-XX"^^<http://id.loc.gov/datatypes/edtf> ;
    bf:place [ a bf:Place ;
      bf:source [ a bf:Source ;
        rdfs:label "lcc-g" ] ;
      rdf:value "5754 L6" ] ] ;

```

N-Triples is a RDF serialization format that is noticeable for its lack of namespace prefixes and other shortcuts, resulting in larger file size than Turtle, but with the advantage of capacity of line-by-line parsing. An example of a BIBFRAME record serialized in N-Triple (which was run

through a conversion tool online that specializes in migrating between different RDF serialization formats) is below:

```
<http://bibframe.example.org/13714243#Work> <http://www.w3.org/1999/02/22-
rdf-syntax-ns#type> <http://id.loc.gov/ontologies/bibframe/Work> .
<http://bibframe.example.org/13714243#Work> <http://www.w3.org/1999/02/22-
rdf-syntax-ns#type> <http://id.loc.gov/ontologies/bibframe/Audio> .
<http://bibframe.example.org/13714243#Work>
<http://id.loc.gov/ontologies/bibframe/adminMetadata> _:genid1 .
<http://bibframe.example.org/13714243#Work>
<http://id.loc.gov/ontologies/bibframe/language>
<http://id.loc.gov/vocabulary/languages/eng> .
<http://bibframe.example.org/13714243#Work>
<http://id.loc.gov/ontologies/bibframe/genreForm> _:genid12 .
<http://bibframe.example.org/13714243#Work>
<http://id.loc.gov/ontologies/bibframe/genreForm> _:genid18 .
<http://bibframe.example.org/13714243#Work>
<http://id.loc.gov/ontologies/bibframe/genreForm>
<http://id.loc.gov/authorities/subjects/sh98004608> .
<http://bibframe.example.org/13714243#Work>
<http://id.loc.gov/ontologies/bibframe/capture> _:genid13 .
<http://bibframe.example.org/13714243#Work>
<http://id.loc.gov/ontologies/bibframe/contribution> _:genid16 .
<http://bibframe.example.org/13714243#Work>
<http://id.loc.gov/ontologies/bibframe/contribution> _:genid25 .
<http://bibframe.example.org/13714243#Work> <http://www.w3.org/2000/01/rdf-
schema#label> "Let it be-- naked" .
<http://bibframe.example.org/13714243#Work>
<http://id.loc.gov/ontologies/bibframe/title> _:genid17 .
<http://bibframe.example.org/13714243#Work>
<http://id.loc.gov/ontologies/bibframe/credits> "The Beatles ; with Billy
Preston on keyboards." .
<http://bibframe.example.org/13714243#Work>
<http://id.loc.gov/ontologies/bibframe/subject>
<http://id.loc.gov/authorities/subjects/sh87003327> .
<http://bibframe.example.org/13714243#Work>
<http://id.loc.gov/ontologies/bibframe/relatedTo>
<http://bibframe.example.org/13714243#Work710-33> .
<http://bibframe.example.org/13714243#Work>
<http://id.loc.gov/ontologies/bibframe/relatedTo>
<http://bibframe.example.org/13714243#Work740-34> .
<http://bibframe.example.org/13714243#Work>
<http://id.loc.gov/ontologies/bibframe/hasInstance>
<http://bibframe.example.org/13714243#Instance> .
```

RDFa is another RDF serialization format whose value for the NET Collection should be carefully considered since the format involves populating HTML documents with RDF triples, providing an opportunity for search crawlers and other Web applications to process the triples in the dataset published in this way. RDFa is an official W3C serialization format designed for simple and direct exposure of RDF triples to search crawlers and other Web applications, in

essence embedding URIs for subjects, objects, and predicates directly into HTML documents (both XHTML and HTML5). This supports both enriched content on the Web, such as enhanced search results, but also enables RDF parsers to process RDFa as structured Linked Data.

RDFa does this by using attributes in HTML documents to record URIs and literal values defining subjects, objects, predicates, and datatypes for literals. This allows institutions whose content management systems restrict Linked Data capabilities to be able to expose their content via HTML mediation. Below are the HTML attributes that are used for this purpose:

@rel – relationships (or predicates)

@rev – inverse relationships

@content – plain literal objects

@href – URI for resource object of a relationship

@source – URI for embedded resource object

RDFa also adds a number of new attributes to HTML, including:

@about – subject URI

@property – property for literal text

@resource – URI expressing a not clickable object

@datatype – datatype of a literal

@typeof – rdf types to associate with a resource

Below is an example of RDFa used in a WorldCat record (mapped to schema.org ontology, rather than BIBFRAME):

```
<div resource="http://www.worldcat.org/oclc/24251664" typeof="http://schema.org/CreativeWork
http://schema.org/Movie http://bibliograph.net/VHS"><a
href="http://www.worldcat.org/oclc/24251664">http://www.worldcat.org/oclc/24251664</a><span style="color:
orange"> # The Shining</span>

<br/>  a

<a href="http://schema.org/CreativeWork">schema:CreativeWork</a>, <a
href="http://schema.org/Movie">schema:Movie</a>, <a href="http://bibliograph.net/VHS">bgn:VHS</a> ;<br/>

<a style="text-decoration:none" href="http://purl.org/library/oclcnum">library:oclcnum</a> "<span style="color:
red" property="library:oclcnum">24251664</span>" ;<br/>
```

```

<a style="text-decoration:none" href="http://purl.org/library/placeOfPublication">library:placeOfPublication</a>
<<a href="http://id.loc.gov/vocabulary/countries/cau" property="library:placeOfPublication"
resource="http://id.loc.gov/vocabulary/countries/cau">http://id.loc.gov/vocabulary/countries/cau</a>> ;<br/>

<a style="text-decoration:none" href="http://purl.org/library/placeOfPublication">library:placeOfPublication</a>
<<a href="http://dbpedia.org/resource/New_York_City" property="library:placeOfPublication"
resource="http://dbpedia.org/resource/New_York_City">http://dbpedia.org/resource/New_York_City</a>> ;<span
style="color: orange"> # New York</span>

<br/>

<a style="text-decoration:none" href="http://www.w3.org/2000/01/rdf-schema#seeAlso">rdfs:seeAlso</a> <<a
href="http://experiment.worldcat.org/entity/work/data/837159063#CreativeWork/shining" property="rdfs:seeAlso"
resource="http://experiment.worldcat.org/entity/work/data/837159063#CreativeWork/shining">http://experiment.w
orldcat.org/entity/work/data/837159063#CreativeWork/shining</a>> ;<span style="color: orange"> #
Shining.</span>

<br/>

<a style="text-decoration:none" href="http://schema.org/about">schema:about</a> <<a
href="http://id.worldcat.org/fast/960370" property="schema:about"
resource="http://id.worldcat.org/fast/960370">http://id.worldcat.org/fast/960370</a>> ;<span style="color:
orange"> # Horror films</span>

```

Using an online converter, the above code can be parsed as an unmediated RDF serialization. Below is the same code above parsed as N-Triples:

```

<http://www.worldcat.org/oclc/24251664> <http://www.w3.org/1999/02/22-rdf-
syntax-ns#type> <http://schema.org/CreativeWork> .
<http://www.worldcat.org/oclc/24251664> <http://www.w3.org/1999/02/22-rdf-
syntax-ns#type> <http://schema.org/Movie> .
<http://www.worldcat.org/oclc/24251664> <http://www.w3.org/1999/02/22-rdf-
syntax-ns#type> <http://bibliograph.net/VHS> .
<http://www.worldcat.org/oclc/24251664> <http://purl.org/library/oclcnum>
"24251664" .
<http://www.worldcat.org/oclc/24251664> <http://schema.org/datePublished>
"1981" .
<http://www.worldcat.org/oclc/24251664> <http://schema.org/datePublished>
"1980" .
<http://www.worldcat.org/oclc/24251664> <http://schema.org/description>
"Unforgettable images of terror envelope a family isolated and snowbound in a
huge resort hotel with a macabre history of violence. The caretaker's son's
psychic ability (his \"shining\") gives him visions of evil to come." .
<http://www.worldcat.org/oclc/24251664> <http://schema.org/inLanguage> "en" .
<http://www.worldcat.org/oclc/24251664> <http://schema.org/name> "The
Shining" .
<http://www.worldcat.org/oclc/24251664> <http://schema.org/productID>
"24251664" .

```

```

<http://dbpedia.org/resource/New_York_City> <http://www.w3.org/1999/02/22-
rdf-syntax-ns#type> <http://schema.org/Place> .
<http://dbpedia.org/resource/New_York_City> <http://schema.org/name> "New
York" .
<http://experiment.worldcat.org/entity/work/data/837159063#Agent/warner_home_
video> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://bibliograph.net/Agent> .
<http://experiment.worldcat.org/entity/work/data/837159063#Agent/warner_home_
video> <http://schema.org/name> "Warner Home Video" .
<http://experiment.worldcat.org/entity/work/data/837159063#CreativeWork/shini
ng> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://schema.org/CreativeWork> .
<http://experiment.worldcat.org/entity/work/data/837159063#CreativeWork/shini
ng> <http://schema.org/name> "Shining." .
<http://id.loc.gov/vocabulary/countries/cau> <http://www.w3.org/1999/02/22-
rdf-syntax-ns#type> <http://schema.org/Place> .
<http://id.loc.gov/vocabulary/countries/cau>
<http://purl.org/dc/terms/identifier> "cau" .
<http://id.worldcat.org/fast/922105> <http://www.w3.org/1999/02/22-rdf-
syntax-ns#type> <http://schema.org/Intangible> .
<http://id.worldcat.org/fast/922105> <http://schema.org/name> "Feature
films" .
<http://id.worldcat.org/fast/960370> <http://www.w3.org/1999/02/22-rdf-
syntax-ns#type> <http://schema.org/Intangible> .
<http://id.worldcat.org/fast/960370> <http://schema.org/name> "Horror
films" .

```

The fact that RDFa can be parsed into unmediated serializations like N-Triples or Turtle is an important feature of this semantic annotation approach to encoding front-facing discovery systems on the Web with Linked Data. This can be compared to the idea of inserting URIs into XML or MARC code without using semantic relationships defined through properties or using semantic vocabularies defining classes. For instance, PBCore includes the common attributes for all elements *source* and *ref*, but without exploiting relationships between URIs inserted in these attributes, then the most that's been accomplished at this point is a reference to a Linked Data resource and not a semantically defined actionable link to navigate from one node to another. This is not to say that there is no merit to insertion of URI's as an aid to conversion to RDF, but that insertion of URIs is not sufficient to facilitate RDF statements that can be serialized in unmediated formats and that can be represented as graphs.

The above examples of RDFa from a WorldCat record represent the importance in Linked Data publishing of establishing a plan for access and discovery of Linked Data for users through establishing relationships between data but also for designing Linked Data relationships to be actionable for machine processing as well. This dual aspect of publishing, aiding human use and re-use of Linked Data and machine use and re-use of Linked Data through semantic querying, search crawlers, data extraction and parsing, etc. is integral to the best publication strategies. Whether one is providing SPARQL endpoints to our relational databases, using a semantic annotation format to aid in HTML mediation, providing third party Linked Data interfaces, or

making static RDF files accessible through server hosting, this dual aspect of Linked Data publishing will likely vary according to the needs, resources, and capabilities of publishers.

Below, specific challenges to publishing the NET Collection dataset as Linked Data will be discussed, specifically issues with conversion tools to aid in the conversion from MAVIS to BIBFRAME and issues with authorities.

NET Collection dataset conversion process

The NET Collection project has created a catalog of records describing content produced and distributed by the National Educational Television public broadcasting service during the tumultuous period between the mid-1950s and the early 1970s. MBRS has produced these records in its moving image and record media asset management system, MAVIS, and shared them with WGBH for inclusion in the American Archive of Public Broadcasting web site as PBCore exports. One of the goals of this project has been to test an alignment of PBCore with BIBFRAME and EBUCore, so our data modeling and crosswalking has been focused on an alignment of a sample NET Collection dataset with BIBFRAME.

Publishing such a dataset within the timeframe of the project is challenging due to the lack of conversion tools, particularly ones aiding transformation from PBCore to MARC and aiding bridging gaps between BIBFRAME and PBCore with EBUCore. It is intended that the crosswalk and data modeling examined for the report will aid in future efforts in this area.

A preliminary dataset based on the stylesheet designed for the NET Collection has been created. This dataset and stylesheet (converting PBCore standardized XML to MARC-XML) is designed for the NET Collection, but it is hoped to aid future attempts to align PBCore with BIBFRAME and other Linked Data models and to aid Linked Data implementation more broadly for the AAPB. The dataset will also be limited to programs in the NET journal umbrella series and the NET playhouse umbrella series and published to the AAPB's Github account, alongside documentation of evaluation, crosswalking, and other data that supported the efforts documented in this report of examining implementation of Linked Data for the NET Collection.

Authorities

Throughout this report, there are references to challenges handling authorities. Below is a summation of these issues:

- MAVIS identifies authorities by use of an automatically assigned MAVIS id number; LCCNs for authorities that were manually copied from LCSH, LCNAF, and LCGFT are referenced as local MAVIS ids.
- No identifiers are converted to PBCore
- Conversion of MAVIS title records limits labels to the preferred label

- Identifying local authorities not already established in authority lists and developing a strategy for what to do about them
- Reconciling authorities with LC's Linked Data Service (id.loc.gov)— or when and how to insert URIs

It is clear that some strategy for dealing with these issues needs to be established for publishing an RDF dataset since authorities are connected to critical access points to leverage Linked Data relationships. In the following discussion, a few critical conclusions about the state of authorities in MAVIS in relation to conversion to RDF will be shared as a basis to examine a workflow aimed at the last two bullet points above: 1) identifying local authorities and establishing a strategy for what to do about them, and 2) reconciling with LC authorities – or when and how to insert LC URIs (or other URIs, such as VIAF).

To address the issues raised by the first three bullet points – 1) the fact that MAVIS title records identify MAVIS authorities by MAVIS ids (with LCCN numbers referenced in the MAVIS record and not appearing on a title record), 2) the fact that identifiers are not transferred in conversion to PBCore, and 3) the fact that conversion from MAVIS title records limits labels to the preferred labels – it is important to consider the extent to which these issues affect publishing these access points as URIs.

The fact that MAVIS title records identify authorities by MAVIS ids instead of LCCN numbers silo-ed in MAVIS authority records is not as much of a problem when it is considered that name matching should allow reconciliation between these authorized access points and URIs assigned to these identities and concepts in external Linked Datasets. Preferably there would be URIs present in the original dataset to aid matching, but the matching can occur without them. This holds true for the fact that identifiers are not transferred in the conversion to PBCore, although PBCore is a candidate for hosting URIs related to authorities due to the attributes *source* and *ref* (which were not used by the XSLT, likely reflecting lingering design from earlier versions of PBCore). The main effect the third issue has is on local authorities whose authority metadata has not been recorded in a centralized authority file or Linked Data graph.

As long as there is a method of reconciliation to a centralized authority file or to a Linked Dataset of identities, then the problem is establishing a procedure for dealing with local authorities and for reconciling authorities against a Linked Data service (through an API, SPARQL endpoint, etc.) to be inserted into the NET catalog dataset. Below, the procedure the MBRS followed regarding these issues is discussed.

First, it was decided to let local authorities not authorized through LCNAF remain literal values rather than minting new URIs. A separate dataset of identities was created where name authorities not established with the LCNAF are marked for future work in a Linked Data environment and the URI's inserted into the NET collection dataset.

Second, OpenRefine's reconciliation capabilities were used to match the separate dataset of authorities pulled from this dataset mentioned above against an API (Harlow, 2015/2018) that handles reconciliation to LC Linked Data services (for names and subjects). This was a process that required creating four separate datasets of authorities pulled from MAVIS exports: credits (creator and contributor), subjects, names as subjects, and genres. In the Evaluation section, the possibility of future named entity recognition technology to pull identities from these notes where necessary and the different types of note elements and types of notes themselves where this would be valuable was discussed, but for the current task of managing the transition and publication of NET catalog metadata as RDF, MBRS decided to focus on identities used in the access points listed above.

For each of these datasets – credits, subjects, names as subjects, and genres – the process was roughly the same as outlined below. Significant challenges will be discussed after:

1. Use XSLT to pull from MAVIS NET catalog dataset (of NET journal and NET playhouse titles).
2. Upload to OpenRefine where the dataset is reconciled to LCNAF using an API (Harlow, 2015/2018). Credits and names as subject were reconciled also against an API that handles reconciliation for Wikidata.
3. Use the templating export function in OpenRefine to export the reconciled identities with URIs into structured XML data.
4. Use XSLT to reconcile the new XML datasets with URIs with the NET catalog dataset (of NET journal and NET playhouse titles) transformed from MAVIS to PBCore (with the MAVIS to PBCore XSLT). Attributes *ref* (for URIs) and *source* (to name the source of the element value and URI).

It is recommended to be prepared for occasionally lengthy human review of the reconciliation to LC Linked Data services, particularly when reconciling against LCNAF. Since the reconciliation relies on label matching rather than URI matching, the LCNAF's extensive list of names provides much more opportunity for mistakes in matches in the automated process. When compared to the Wikidata API (which had its own separate issues of aligning to types of names, such as television stations or countries), this process was lengthy. One method of shortening this process was de-duplicating to narrow the number of names to double check. Initially, de-duplication was not done since there was initial concern about needing to disambiguate amongst similar names within the NET dataset, but in hindsight, especially considering how few instances of disambiguation were encountered in this specific dataset, de-duplicating names would have been ideal early in the process since it would have meant less time de-duplicating. MBRS de-duplicated *after* matching credits to LC Linked Data services, but for the other datasets were de-duplicated *before* reconciliation. MBRS did not find an API for reconciling to the Genre/Form

Linked Data service, so it manually matched these after de-duplicating the genres (leaving 65 total authorized access points).

As mentioned above, the MBRS will share the XML datasets for credits, subjects, names as subject, and genres, which will be useful for tracing local name authorities that had not yet been established in LCNAF for future work in creating URIs for these identities.

Conclusion

For this report, four phases of Linked Data implementation were presented: Evaluation, Data Modeling, Crosswalking, and Publishing. Ideally, the work done here would also support metadata creation using Linked Data by building the foundation for the metadata environment necessary for metadata creation. Presently, this report instead uses these phases to model the conversion of a dataset of NET content (consisting of two anthology series, one fiction and one non-fiction – NET playhouse and NET journal) from XML to RDF in an alignment between MAVIS/PBCore and BIBFRAME, with some recommended extensions to EBUCore.

For the Data Modeling phase, different standards (BIBFRAME, PBCore, MAVIS) were compared and contrasted in a process of examining the impact of each standard on the conversion process of transforming a dataset exported from MAVIS to BIBFRAME standardized RDF. Data modeling “works” was also examined in relation to EIDR ids.

For the Evaluation phase, a series of studies on different types of controlled fields in MAVIS was conducted in order to understand potentials in access points and text fields for Linked Data exposure. The MAVIS to PBCore stylesheet was also examined for idiosyncrasies and for the reliability of its preservation of potential Linked Data exploits in unstructured text.

For the crosswalking phase, a crosswalk from MAVIS to PBCore to MARCXML to BIBFRAME was developed with recommended extensions to EBUCore. The crosswalk also represents the baked in mappings of the MAVIS to PBCore XSLT, while also providing more ideal mappings where necessary (such as for pbcoreCoverage).

For the publishing phase, publishing methodologies were examined, and the method of data dumping RDF file on a server for this static dataset was chosen as a preliminary step to publishing NET collection data as Linked Data. However, MBRS also positioned its choice against the backdrop of other choices, including choices between different serialization formats and different forms of mediation from HTML to SPARQL endpoints to third party APIs. MBRS also explored publishing authorities such as credits and subjects through reconciliation with Linked Data services in OpenRefine and inserting URIs into PBCore attributes *source* and *ref*. Authorities that did not reconcile with the LCNAF were isolated, so consequently MBRS will be

publishing separate datasets, where some are marked as not matching to the LCNAF for future catalogers to add to the LCNAF.

Publishing the NET collection catalog as an alignment of PBCore/MAVIS metadata with BIBFRAME (and via extensions, EBUCore) faces the significant challenge of the lack of a tool to convert from PBCore to MARCXML or to other standards, including BIBFRAME. Consequently, three separate datasets will be published: a PBCore dataset with URIs inserted into PBCore attributes *source* and *ref* in XML, a MARC-XML dataset transformed based on the crosswalk work discussed during the crosswalking phase, and a BIBFRAME dataset with converted with LC's official stylesheet (modified a bit to manage the needs of the NET collection. EBUCore mappings have not yet been tested and added to the dataset, only proposed. Additionally, the original MAVIS-XML exports will be published as well.

Transitioning to a Linked Data environment is a difficult and frequently costly process, particularly if an institution is already managing complex metadata environments either internally and/or externally through collaboration. In this preliminary report on Linked Data implementation for the NET Collection Catalog, the goal has been to provide a model for different phases of the process, either during an early analysis of the current metadata environment, during the often bewildering choice of data models and vocabulary terms that confront attempts to transition to RDF and still represent collections with expected granularity and expressivity, or during publishing where simply deciding how to share Linked Datasets can be an opaque process. This implementation study has only "gotten our feet wet," but also is an essential step to form the basis for future work to enable more deeply embedded Linked Data that improves discovery and access, bring more attention to collections, and help provide authoritative information services to the Web overall.

Bibliography

- Allemang, D., & Hendler, J. A. (2012). *Semantic web for the working ontologist: Modeling in RDF, RDFS and OWL*. Amsterdam: Morgan Kaufmann Publishers/Elsevier
- Alemu, G., Stevens, B., Ross, P., & Chandler, J. (2012). Linked Data for libraries: Benefits of a conceptual shift from library-specific record structures to RDF-based data models. Retrieved from <https://www.ifla.org/past-wlic/2012/92-alemu-en.pdf>
- S. Bechhofer, K. Page, & D. De Roure. (2013). Hello Cleveland! Linked Data publication of live music archives. *2013 14th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, 1–4. <https://doi.org/10.1109/WIAMIS.2013.6616155>
- Berners-Lee, T. (2006, July 7). Linked Data - Design Issues. Retrieved April 20, 2017, from <https://www.w3.org/DesignIssues/LinkedData.html>
- EBU. (2016, September). Tech 3293: EBU CORE metadata set (EBUCore). Retrieved from <https://tech.ebu.ch/docs/tech/tech3293.pdf>
- Gracy, K. F. (2015). Archival description and Linked Data: a preliminary study of opportunities and implementation challenges. *Archival Science*, 15(3), 239–294.
- Gracy, K. F., Zeng, M. L., & Skirvin, L. (2013). Exploring methods to improve access to Music resources by aligning library Data with Linked Data: A report of methodologies and preliminary findings. *Journal of the American Society for Information Science and Technology*, 64(10), 2078–2099. <https://doi.org/10.1002/asi.22914>
- Hardesty, J., & Young, J. B. (2017). The semantics of metadata: Avalon Media System and the move to RDF. *Code4Lib*, (37). Retrieved from <http://journal.code4lib.org/articles/12668>

- Hardesty, J. L. (2016). Transitioning from XML to RDF: Considerations for an effective move towards Linked Data and the Semantic Web. *Information Technology and Libraries*, 35(1), 51. <https://doi.org/10.6017/ital.v35i1.9182>
- Kroon, R. W., Drewry, R., Leigh, A., & McConnachie, S. (2015). Content identification for audiovisual archives. *IASA Journal*, (45). Retrieved from <http://search.ebscohost.com/login.aspx?direct=true&profile=ehost&scope=site&authtype=crawler&jrnl=1021562X&AN=118493716&h=dVULeBSQyfvRPBFtmY7fu0i%2Bh2toVBFM%2B1oXCOwwL6x%2BFett1qQZpGlzFpSxDNbrwViVa6yV4sjvskh9bMKJjg%3D%3D&crl=c>
- Library of Congress, WGBH. (n.d.). National Educational Television (NET) Collection Catalog Project. Retrieved August 9, 2017, from <http://americanarchive.org/about-the-american-archive/projects/net-catalog>
- Library of Congress. (n.d.). Compare MARC/XML BIBFRAME/RDF. Retrieved June 7, 2018, from <http://id.loc.gov/tools/bibframe/compare-ccn/full-ttl?find=2004597860>
- Lyons, B., & Malssen, K. V. (2015). BIBFRAME AV Assessment: Technical, Structural, and Preservation Metadata, 49.
- Madsen, B. N., & Erdman Thomsen, H. (2009). Ontologies vs. classification systems. Retrieved from <http://dspace.ut.ee/handle/10062/9840>
- Mixer, J. (2014). Using a Common Model: Mapping VRA Core 4.0 Into an RDF Ontology. *Journal of Library Metadata*, 14(1), 1–23. <https://doi.org/10.1080/19386389.2014.891890>
- Morgan, E. L. (2014, April 23). Linked archival metadata: a guidebook. Retrieved from <http://infomotions.com/sandbox/liam/tmp/guidebook.pdf>

- Pattueli, M. C., Provo, A., & Thorsen, H. (2015). Ontology Building for Linked Open Data: A Pragmatic Perspective. *Journal of Library Metadata*, 15(3–4), 265–294.
<https://doi.org/10.1080/19386389.2015.1099979>
- Schreur, P. E. (2012). The academy unbound. *Library Resources & Technical Services*, 56(4), 227–237.
- Smith, M., Stahmer, C. G., Li, X., & Gonzalez, G. (2017, March 14). BIBFLOW: A roadmap for library Linked Data transition. University Library, University of California, Davis; Zepheira Inc.
Retrieved from <http://roytennant.com/BIBFLOWRoadmap.pdf>
- Tennant, R. (2002, October 15). MARC Must Die. Retrieved March 9, 2018, from
<https://lj.libraryjournal.com/2002/10/ljarchives/marc-must-die/>
- W3C. (2002). Resource Description Framework (RDF): Concepts and Abstract Data Model.
Retrieved August 9, 2017, from <https://www.w3.org/TR/2002/WD-rdf-concepts-20020829/#xtocid48014>
- W3C. (2014). RDF Schema 1.1. Retrieved August 14, 2017, from <https://www.w3.org/TR/rdf-schema/>
- Wiggins, B., McCallum, S., Frank, P., & Hess, K. (2016, September 9). Bibframe on the move.
Retrieved April 20, 2017, from <https://stream-media.loc.gov/webcasts/captions/2016/160906dfy1000.txt>
- Zeng, M. L., Gracy, K. F., & Skirvin, L. (2013). Navigating the Intersection of Library Bibliographic Data and Linked Music Information Sources: A Study of the Identification of Useful Metadata Elements for Interlinking. *Journal of Library Metadata*, 13(2–3), 254–278.
<https://doi.org/10.1080/19386389.2013.827513>