Journal of e-Media Studies
Special Issue: Early Cinema Compendium, Mark Williams (editor)
Volume 7, Issue 1, 2024 Dartmouth College

Metadata and the Media Ecology Project

Shiyang Jiang
China Film Archive

Abstract

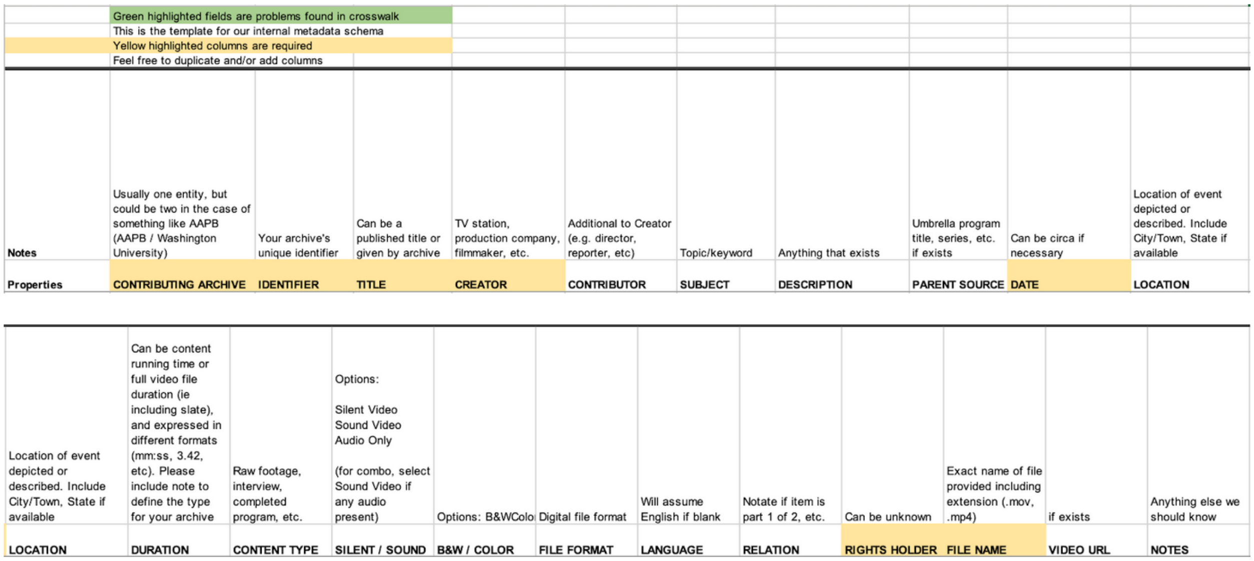
Over the years, the Media Ecology Project (MEP) has assembled a significant digital collection of moving images, books, and photographs of research value. From 2020 to 2022, MEP has taken several initiatives to improve the management and organization of the collection and its metadata, including creating a database and updating the existing schema. This article looks back at some of the challenges encountered during the process and forward to potential further developments.

The Media Ecology Project and Metadata

In its effort to provide access to valuable moving image resources, the Media Ecology Project (MEP) has worked closely with cultural heritage organizations to assemble a remarkable collection for scholarly research. Some highlights in the collection include extremely high-resolution digital scans of the paper prints at the Library of Congress and all volumes of American Mutoscope and Biograph Picture Catalogue. As the MEP collection grows impressively, more effort is required to improve the accessibility and usability of these digital assets. This article summarizes the progress made in MEP's metadata organization strategies from 2020 to 2022. While the work is still ongoing, this article summarizes what MEP has achieved thus far and reflects on some remaining limitations.

Metadata—the information we create, store, and share to describe things—allows researchers to interact with these collections to obtain the knowledge they seek. Institutions such as libraries, archives, and museums have long been creating, using, and sharing structured metadata. Our archive partners, including many such institutions, have been generous in their partnership with MEP by supplying the digital files and the metadata that describe them. For this purpose, MEP designed a spreadsheet template based on Dublin Core. It includes mandatory fields such as title, identifier, creator, and date, as well as optional fields, such as location, source, country of origin, languages, and more, that archivists could add as they see fit. MEP has been able to collect these spreadsheets and use them for further research.

However, as the collections at MEP grew more and more diverse, some issues with this workflow arose. The original spreadsheet template was set up only for newsfilm collections, so it features a small set of controlled vocabulary like raw footage and the completed program that may not apply to other collections (see Table 1). Furthermore, each metadata spreadsheet delivered by MEP archive partners was in a separate file instead of consolidating information in a central resource. As a result, team members spent excessive time looking for the correct file. Finally, different collections have different cataloging rules, which results in varying degrees of detail. The lack of consistency hindered MEP's goal of using machine-learning analysis to unlock the collection's potential.



	Green highlighted fields are problems found in crosswalk										
	This is the template for our internal metadata schema										
	Yellow highlighted columns are required										
	Feel free to duplicate and/or add columns										
Notes	Usually one entity, but could be two in the case of something like AAPB (AAPB / Washington University)	Your archive's unique identifier	Can be a published title or given by archive	TV station, production company, filmmaker, etc.	Additional to Creator (e.g. director, reporter, etc)	Topic/keyword	Anything that exists	Umbrella program title, series, etc. if exists	Can be circa if necessary	Location of event depicted or described. Include City/Town, State if available
Properties	CONTRIBUTING ARCHIVE	IDENTIFIER	TITLE	CREATOR	CONTRIBUTOR	SUBJECT	DESCRIPTION	PARENT SOURCE	DATE	LOCATION

Location of event depicted or described. Include City/Town, State if available	Can be content running time or full video file duration (ie including slate), and expressed in different formats (mm:ss, 3.42, etc). Please include note to define the type for your archive	Raw footage, interview, completed program, etc.	Options: Silent Video Sound Video Audio Only (for combo, select Sound Video if any audio present)	Options: B&WColo	Digital file format	Will assume English if blank	Notate if item is part 1 of 2, etc.	Can be unknown	Exact name of file provided including extension (.mov, .mp4)	if exists	Anything else we should know
LOCATION	DURATION	CONTENT TYPE	SILENT / SOUND	B&W / COLOR	FILE FORMAT	LANGUAGE	RELATION	RIGHTS HOLDER	FILE NAME	VIDEO URL	NOTES

Table 1. MEP Original Metadata Template.

By improving metadata organization, researchers can better access and connect MEP's collections to digital humanities platforms like the Minimum Viable Annotator and Scalar. With consistent metadata, users can easily reference a film while creating annotations in the Minimum Viable Annotator or articles on Scalar. To address these concerns and facilitate better collection management, MEP needs new strategies for organizing, cleaning, and consolidating the metadata.

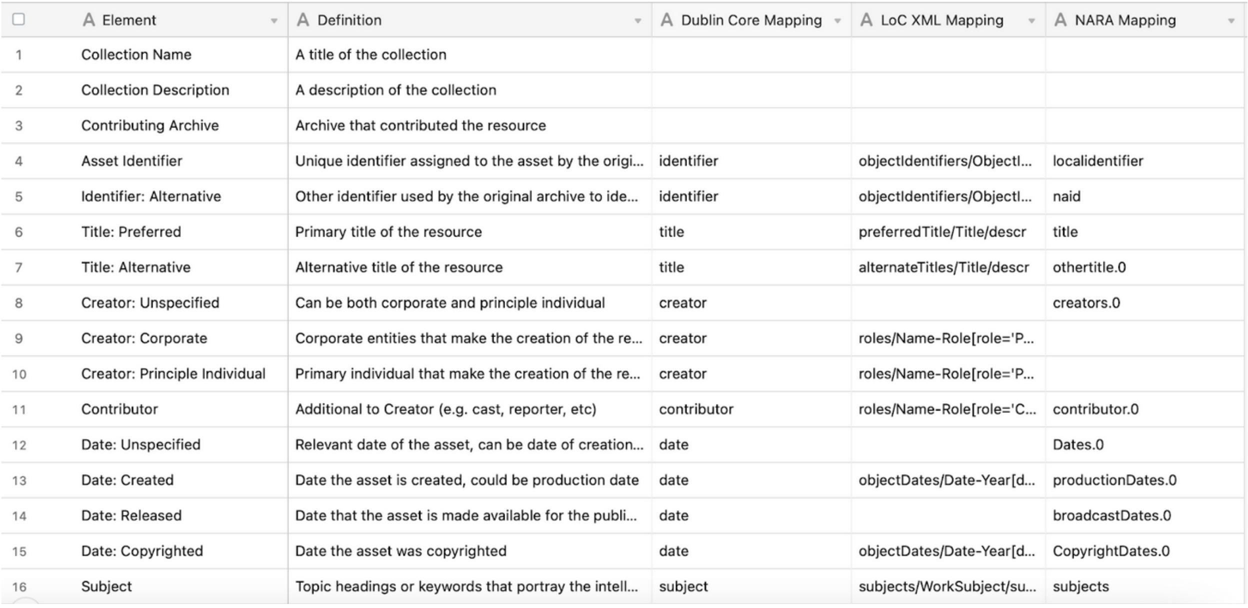
Schema

> **"We recognize that preserving the original records is the best way to capture the stories of how these materials were created, distributed, and archived. As such, we chose to retain the original descriptions as faithfully as possible."**

Schema is an essential metadata component, describing the overall structure used to capture information about a resource. According to ISO 23081's definition, "A schema is a logical plan showing the relationships between metadata elements, normally through establishing rules for the use and management of metadata specifically as regards the semantics, the syntax and the optionality (obligation level) of values."[1]

We looked closely at the metadata MEP received to lay the groundwork for the new schema. Some items are described in standard archival language, while others have more unique descriptions like shot lists, probably from the hand of local television crews or stock footage archivists. We recognize that preserving the original records is the best way to capture the stories of how these materials were created, distributed, and archived. As such, we chose to retain the original descriptions as faithfully as possible.

Following this approach, we modified the existing schema and created a new application profile (see Table 2). The new schema contains most elements of the old one, as well as additional fields that reflect the research needs of film historians. The new schema has multiple "title" and “date” elements to capture the nuances better. Allowing multiple titles is especially important for early cinema. Due to the low survival rate of early cinema, titles are often the only thing that has withstood the ravages of time. In addition, many early films were referred to by different titles in different publications, making every title valuable to historical research. Furthermore, each item can also have metadata about its production date, release date, and copyright date, which are vital for understanding film production and distribution history. The new schema also defines more comprehensive sets of controlled vocabularies, such as the “color” fields. It covers more than “B&W” and “color” but also includes “tinted” and “toned”—a detail especially relevant for early cinema collections.



□	Element	Definition	Dublin Core Mapping	LoC XML Mapping	NARA Mapping
1	Collection Name	A title of the collection			
2	Collection Description	A description of the collection			
3	Contributing Archive	Archive that contributed the resource			
4	Asset Identifier	Unique identifier assigned to the asset by the origi...	identifier	objectIdentifiers/ObjectI...	localidentifier
5	Identifier: Alternative	Other identifier used by the original archive to ide...	identifier	objectIdentifiers/ObjectI...	naid
6	Title: Preferred	Primary title of the resource	title	preferredTitle/Title/descr	title
7	Title: Alternative	Alternative title of the resource	title	alternateTitles/Title/descr	othertitle.0
8	Creator: Unspecified	Can be both corporate and principle individual	creator		creators.0
9	Creator: Corporate	Corporate entities that make the creation of the re...	creator	roles/Name-Role[role='P...	
10	Creator: Principle Individual	Primary individual that make the creation of the re...	creator	roles/Name-Role[role='P...	
11	Contributor	Additional to Creator (e.g. cast, reporter, etc)	contributor	roles/Name-Role[role='C...	contributor.0
12	Date: Unspecified	Relevant date of the asset, can be date of creation...	date		Dates.0
13	Date: Created	Date the asset is created, could be production date	date	objectDates/Date-Year[d...	productionDates.0
14	Date: Released	Date that the asset is made available for the publi...	date		broadcastDates.0
15	Date: Copyrighted	Date the asset was copyrighted	date	objectDates/Date-Year[d...	CopyrightDates.0
16	Subject	Topic headings or keywords that portray the intell...	subject	subjects/WorkSubject/su...	subjects

Table 2. New Application Profile (Preliminary).

An important decision is whether to have one set of metadata elements (single entity) or establish groups of metadata elements (multiple entities). Multiple-entity models have the advantage of grouping elements around what we need to describe.[2] We chose the multiple-entity model for MEP as we can cover more details about collections, assets (films), files (local video files), streaming (online URLs), and annotation records. Additionally, we established rules about the relationships between entities, which would provide linkages between records and agents. For example, we define the relationship between assets and files as “one to many.” That way, even when a video has multiple parts, there will still be only one comprehensive record that describes it. These changes make digital assets much more manageable.

Crosswalk and Cleanup

The metadata crosswalk is an important procedure for maintaining good metadata quality. Crosswalk is the process of mapping the elements and values from one schema to those of another, but mistakes can occur regardless of how careful you are. We keep all original spreadsheets as a last resort for this issue. For example, identifiers for the United States Information Service films have leading zeros, which get automatically removed when the spreadsheets are in comma-delimited formats (CSV). It took us some time to discover the root of the problem. Thankfully, we kept all the original files and developed a workaround to preserve these zeros.

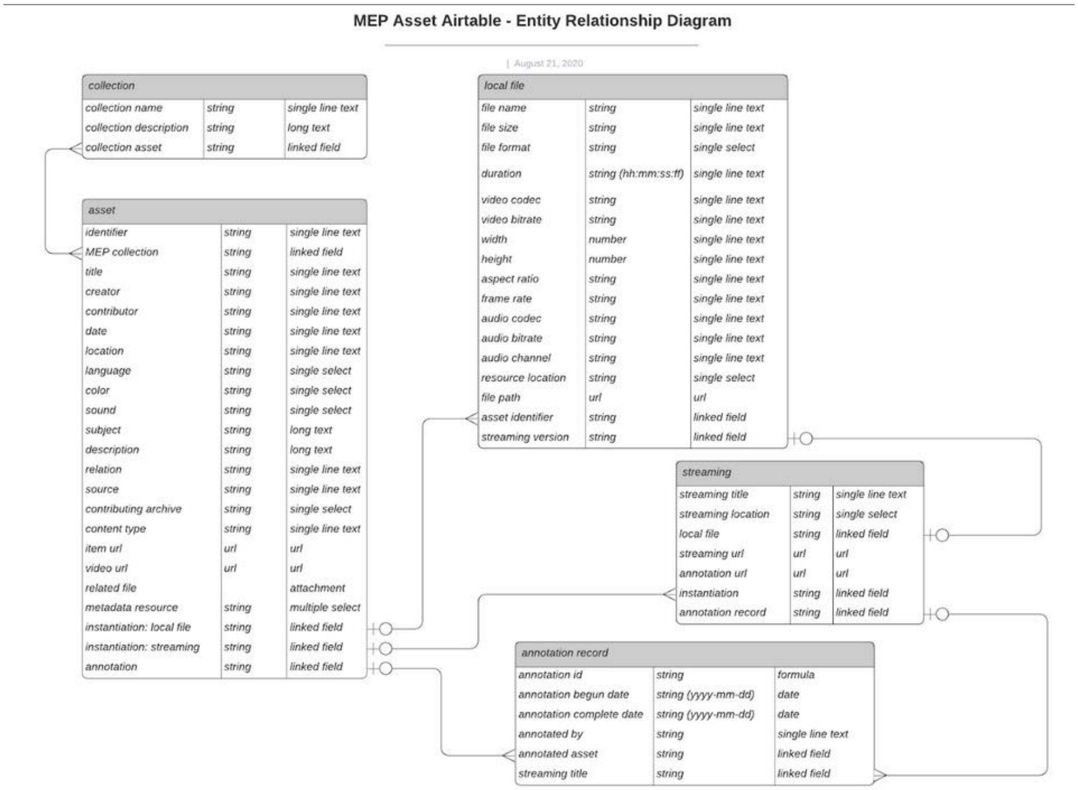
Consistency is vital to high-quality metadata, so we ensured the new application profile was applied consistently. We used Open Refine for many procedures, including formatting all date fields to YYYY-MM-DD, implementing controlled vocabularies, and consolidating duplicated records.

So far, the work has primarily focused on structured metadata. However, certain collections present unique challenges involving valuable data buried in plain text. The issue results from using an obsolete system that hosts the metadata or relying on optical character recognition (OCR) texts from books. To transform the plain text into a spreadsheet, we employed Python Regular Expressions to identify recurring headings in the texts and retrieve the necessary data. Additionally, Tesseract, an open-source program, was utilized to improve the recognizability and accuracy of the OCR output. After running numerous experiments, we were able to achieve a good result.

Digital Assets

Unique identifiers link the metadata records and thousands of video files. Any discrepancy between the two would cause massive confusion and problems for researchers. Derivative files add a further layer of complexity. We created many derivative videos from the originals to put on streaming platforms like Minimum Viable Annotators to reduce loading time. It is important to note that these files are not for close analysis.

The multiple-entity metadata model is essential to solving this challenge (see Table 3). By categorizing each video as an entity, one film can have multiple versions listed under the “related file” field. We used MediaInfo—an open-source metadata reading application—to gather each file’s name, size, resolution, color space, and duration. A Python script collected all this information and turned it into CSV files. This comprehensive metadata not only helps researchers compare different versions and decide which one they need but also makes it simpler for project managers to calculate the size of the entire collection.



MEP Asset Airtable - Entity Relationship Diagram

| August 21, 2020

collection		
collection name	string	single line text
collection description	string	long text
collection asset	string	linked field

asset		
identifier	string	single line text
MEP collection	string	linked field
title	string	single line text
creator	string	single line text
contributor	string	single line text
date	string	single line text
location	string	single line text
language	string	single select
color	string	single select
sound	string	single select
subject	string	long text
description	string	long text
relation	string	single line text
source	string	single line text
contributing archive	string	single select
content type	string	single line text
item url	url	url
video url	url	url
related file		attachment
metadata resource	string	multiple select
instantiation: local file	string	linked field
instantiation: streaming	string	linked field
annotation	string	linked field

local file		
file name	string	single line text
file size	string	single line text
file format	string	single select
duration	string (hh:mm:ss:ff)	single line text
video codec	string	single line text
video bitrate	string	single line text
width	number	single line text
height	number	single line text
aspect ratio	string	single line text
frame rate	string	single line text
audio codec	string	single line text
audio bitrate	string	single line text
audio channel	string	single line text
resource location	string	single select
file path	url	url
asset identifier	string	linked field
streaming version	string	linked field

streaming		
streaming title	string	single line text
streaming location	string	single select
local file	string	linked field
streaming url	url	url
annotation url	url	url
instantiation	string	linked field
annotation record	string	linked field

annotation record		
annotation id	string	formula
annotation begun date	string (yyyy-mm-dd)	date
annotation complete date	string (yyyy-mm-dd)	date
annotated by	string	single line text
annotated asset	string	linked field
streaming title	string	linked field

Table 3. Multiple-Entity Data Model (Preliminary).

Database

A new information management system was needed to accommodate the new schema and thousands of records. After extensive research, we chose Airtable as our online database; an offline copy was also saved and regularly updated for security. Airtable combines the user-friendly features of a spreadsheet with the functionality of a traditional relational database. Furthermore, it is not static—any changes made to the records will be immediately visible on the user end.

MEP uses Airtable to provide researchers with multiple points of access. With password-protected links, researchers can view, sort, and search selected metadata of one collection quickly and easily. These views can be embedded on any website to enable multitasking. Moreover, researchers with access to the entire database can search across all the records in one go using Airtable's advanced search tool. Additionally, Airtable also has an application programming interface (API) to facilitate integration with other tools as a backend, and integration with Scalar is currently in progress.

To make Airtable the central hub for metadata organization, we developed an ingest-and-export workflow. Excel and Google spreadsheets are still the default format for delivering new metadata to MEP, but adding and verifying new input are done through Airtable tools. The tool is handy for updating existing records. The import tool can take any spreadsheets or XML files and add them to the database as long as the headers are consistent with the schema. We also created tutorials that guide users on routinely maintaining the system.

Conclusion

MEP is continuously striving to make metadata from archives more accessible for researchers. Looking forward, MEP will continue making improvements to the current workflow. For example, an automated process could provide more reliable upkeep of an offline database copy. The creative use of automation tools with Google spreadsheets and Airtable could also save considerable effort while helping the ever-growing collection stay updated without constant monitoring. With these ongoing improvements, we anticipate that MEP's metadata organization will become more efficient so that more researchers can benefit from its offerings.

About the Author

Shiyang Jiang is an archivist at China Film Archive. She was previously a media annotator for the Media Ecology Project at Dartmouth College, where she worked to improve the interoperability and standardization of metadata. Shiyang holds a master's degree in moving image archiving and preservation from New York University and a BFA in theatre, film, television, and literature from Xiamen University.

[1] ISO 23081-1:2017(en), "Information and documentation—Records management processes—Metadata for records—Part 1: Principles." https://www.iso.org/obp/ui/#iso:std:iso:23081:-1:ed-2:v1:en.

[2] National Information Standards Organization, "ISO/TC 46/SC11N800R1 Building a Metadata Schema—Where to Start." www.niso.org/internation/tc46.

Published by the Dartmouth College Library.

http://journals.dartmouth.edu/joems/
Article DOI: 10.1349/PS1.1938-6060.A.494