OLAC CAPC
Moving Image Work-Level Records Task Force
Final Report and Recommendations
April 15, 2009

Part IV: Extracting Work-Level Information from Existing MARC Manifestation Records

Task Force Members:
Kelley McGrath (chair; subgroup 1 and 4 leader)
Susannah Benedetti (subgroup 2)
Karen Gorss Benko (subgroup 3)
Lynne Bisko (subgroup 4)
Greta de Groat (subgroup 1)
Scott M. Dutkiewicz (subgroup 4)
Ngoc-My Guidarelli (subgroup 2)
Jeannette Ho (subgroup 2 leader)
Nancy Lorimer (subgroup 1)
Scott Piepenburg (subgroup 3)
Thelma Ross (subgroup 3 leader)
Walt Walker (subgroup 3)

Advisors to the Task Force:
David Miller
Jay Weitz
Martha Yee

Introduction

This subgroup of the Moving Image Work-Level Records Task Force of Online Audiovisual Catalogers (OLAC) Cataloging Policy Committee (CAPC) was charged with identifying places in MARC manifestation-level bibliographic records where work-level information may be encoded and examining a sample of MARC records to see how reliably this information might be extrapolated from existing records. Currently we do not have work-level records for moving images, except for a relatively small number of uniform title authority records, which usually contain only title information. Moving image uniform title authority records usually represent works, but tend to include only enough information to uniquely identify the work or expression rather than a more complete description. However, information about moving image works is often embedded in our current manifestation-level bibliographic records. If we wish to move to an environment where we create and share work-level records for moving images, it would be helpful if we could use automated means to extract data from existing bibliographic records to populate provisional work-level records. These provisional records could later be enhanced, verified and corrected by human beings. Therefore, we are interested in determining the extent to

which it is possible to accurately extract work-level information from existing bibliographic records.

This subgroup of the OLAC task force was asked to conduct a pilot project to look at five characteristics:

- Original date (year)
- Original title
- Director
- Original language
- Original aspect ratio

We were interested in examining the following questions:

1. What data that might be used to construct provisional work-level records can we extract from existing MARC bibliographic records via automated methods that do not require human intervention or review?

2. How reliable is the data retrieved by these methods and what types of problems are encountered in this process?

3. Are there ways that we could change the way we code data in MARC bibliographic records in order to improve our ability to get this sort of data back out?

One Possible Scenario for Work-Level Records for Moving Images

Before discussing how we attempted to extract work-level information from manifestation-level bibliographic records, we would like to briefly discuss one possible scenario for using work-level records populated with extracted data.

It is possible that the most efficient approach to moving image cataloging is to record the reusable data in one record (what we refer to here as a work-level record and discussed in the task force's report, parts 1-2, as a work/primary expression record), the manifestation-specific data in machine-comprehensible form in another record, and to link the two (or for more traditional systems, to merge them in some form; if this data is machine-analyzable, the parts in the manifestation record that don't vary from the original could easily be suppressed).

Most of the time, it is unclear that explicit expression-level records offer any advantages for moving image cataloging. The exception is what might be called "named" expressions, e.g., director's cut or unrated versions, which cannot be reduced to exhaustive, controlled vocabularies and may require cross-references that cannot be anticipated prior to the creation of additional manifestations. It would be more practical to record most characteristics that may vary at the expression-level (e.g., color, duration, language access) in machine-readable form in the manifestation-level record and program the computer interface to offer this information as navigation options. In particular, for moving images in which given expressions tend to be multi-

faceted, it probably is not time-saving to try to locate or create an expression-level record that reflects a specific combination of options.

On the next page, we give an example of how this combination of work- and manifestation-level records could be presented to an end user. This is not intended to be a comprehensive example nor an ideal display, but merely to present a possible idea.

Limiters (from manifestation-level records)	Work
Available at: o Ball State University Libraries o Muncie Public Library Format: o DVD o Blu-ray o VHS Spoken language: o English o Spanish o French o Chinese Subtitle/caption language: o English o Spanish o Thai Accessibility options: o Audio-described o Captioned Aspect: o Fullscreen (1.33 : 1) o Widescreen (1.85 : 1) Publisher/Distributor: o Warner Home Video Special features: o Commentary track o The making of One flew over the cuckoo's nest (behind-the-scenes documentary) o Additional scenes o Cast/director career highlights o Theatrical trailer	Title: One flew over the cuckoo's nest Date: 1975 Director: Forman, Miloš Producer: Zaentz, Saul ; Douglas, Michael, 1944- Writers: Hauben, Lawrence ; Goldman, Bo Production company: Warner Bros. Pictures Cast: Nicholson, Jack.; Fletcher, Louise ; Redfield, William, 1927-1976 [additional creators and contributors could be included] Summary: Randall P. McMurphy, a free-spirited con, fakes insanity in order to get committed to the state mental hospital instead of going to prison. Once committed, his rebelliousness pits him against Nurse Ratched, the head nurse of the mental ward, and the full spectrum of institutional repression. Genre: Drama ; Adaptation Setting: Salem (Or.) ; Oregon ; Pacific Northwest ; United States Time period: Contemporary Language: English Country of production: United States Run time: 133 min. Color: Color Sound: Mono. Aspect ratio: 1.85 : 1 Awards: Academy Award (Best Picture ; Best Director ; Best Actor in a Leading Role ; Best Actress in a Leading Role ; Best Writing, Screenplay Adapted from Other Material) Based on: One flew over the cuckoo's nest (novel) Author of novel: Kesey, Ken

If the data in the work-level display on the right were recorded in a separate record, mechanisms currently exist to extract most of the data on the left from related MARC bibliographic records, assuming full and accurate records. The notable exceptions are that there is no reliable way to

extract aspect ratio or special features in the form given here. Missing or mistaken data will have some impact on implementation, but could be improved retrospectively.

Although it seems desirable to many to store data for bibliographic materials in a multi-record, FRBR-based structure, the transition by the diverse and under-funded library world to a new structure is likely to be difficult and to proceed at different paces in different institutions. Creation of work-based records that can be linked to and used both with existing manifestation records and future, leaner manifestation records created in a more robust model would provide one way of easing this transition.

Methodology

Overview

We identified a representative sample of work-level information for moving images and used our knowledge of cataloging rules and practices to identify all possible fields and subfields where this information might occur in MARC records. We then evaluated these fields and subfields, based on how commonly they are used and how amenable they are to reliable automatic extraction, and selected the most promising for processing.

In order to test the usefulness of our selected fields and subfields, we acquired from a variety of types of institutions a sample of MARC bibliographic records that describe a range of moving images, including features, television programs and nonfiction. We extracted from these MARC records the fields and subfields from which we wished to extract data, as well as those deemed useful for evaluating the accuracy of the extracted data. We wrote brief programs and queries to automatically extract the values of interest and then manually reviewed the results. The manual review was useful in that it allowed us to identify patterns of problems. This will enable us to improve future iterations of our program and also possibly to proactively identify records that are more likely to need manual intervention. The manual review also allowed us to make more accurate assessments of the relative usefulness and reliability of data from the various sources.

Our analysis has enabled us to suggest two types of improvements to enhance our ability to more effectively record and identify this type of data in the future. The first is to recommend the use of specific cataloging practices that are possible in the current infrastructure and that would support the machine-manipulable recording of data in which we are interested. The second is that, when we have identified areas where it is not possible to record useful data in machine-manipulable form, we can create proposals to expand the MARC format to support this type of data input.

Location of Data in MARC Records

We began by brainstorming about where in the MARC record these pieces of information might exist. The data sources we considered are listed below. For testing purposes, we then narrowed down the potential data sources to those that are shaded in gray. We selected those as the most promising based on the estimated accuracy of the data for our purposes and our perception of how often these fields are used. We limited our data sources to those that have a high probability of containing correct data in a form that can be extracted without manual review.

Category	Field	Subfield	Description	Notes
aspect ratio	250	a	Edition statement	Look for keywords such as widescreen, full screen or aspect
aspect ratio	500	a	General note	Look for keywords such as widescreen, full screen or aspect
aspect ratio	505	all	Formatted contents note	Look for keywords such as widescreen, full screen or aspect
aspect ratio	538	a	System details note	Look for keywords such as widescreen, full screen or aspect
date	008	07-10	Date 1	May be useful for archival cataloging
date	008	11-14	Date 2	
date	033	a	Formatted date/time of an event	
date	130	a	Main entry, uniform title	In form Title (Motion picture/Television program : [date]), e.g., King Kong (Motion picture : 1933)
date	260	c	Date of publication, distribution, etc.	May be useful for archival cataloging
date	261	d	Obsolete; date of production, release, etc. for films	May be useful for archival cataloging
date	500	a	General note	Look for year in combination with keyword
date	518	a	Date/time and place of an event note	Look for year in combination with keyword
director	130	a	Main entry, uniform title	In form Title (Motion picture/Television program : [date] : [director's last name]), e.g., Harlow (Motion picture : 1965 : Douglas)
director	245	c	Statement of responsibility	In combination with word for director/direction; use semi-colons to parse
director	505	ar	Formatted contents note	For multi-work items; not sure this will work in practice
director	508	a	Creation/production credits note	In combination with word for director/direction; use semi-colons to parse
director	700	4	Added entry, personal name with relator code	$4 = drt
director	700	e	Added entry, personal name with relator term	$e = direction
language	008	35-37	Language code	only useful if no 041 or no translation in 041
language	041	a	Language code of text/sound track or separate title	only if no translation involved
language	041	h	Language code of original and/or intermediate translations of text	

language	546	a	Language note	not sure how to get this information out automatically; not usually explicit
title	130	a	Main entry, uniform title	before first parenthesis only
title	245	ab	Title	Need to look out for parallel titles; items without collective titles
title	246	ab	Varying form of title	
title	505	t	Formatted contents note	probably hard to use
title	730	a	Added entry, uniform title	look out for TV series
title	740	a	Added entry, uncontrolled analytical title	2nd indicator 2 only

Selection of Records for Sample Testing

We obtained a sample consisting of 941 MARC records from six institutions, primarily via Z39.50. These included records from a public library, two medium-sized academic libraries, two large academic libraries and a film archive, all of whom do at least some local editing of their records.

We took several approaches to selecting records. We wanted to include some well-known movies that have been re-issued numerous times. To this end, we did title keyword searches for *Citizen Kane* and for Dracula. The Dracula search would enable us to pick up various different movies with the same or similar titles. We were also interested in examining some non-English language titles. We chose *Amélie* as a commonly-held Roman-alphabet title. We also searched for various spellings of Rublev to retrieve Tarkovsky's *Andrei Rublev* and the word samurai to retrieve, among others, Kurosawa's *Seven Samurai* whether it was listed under its English title or the original Japanese *Shichinin no Samurai*. We also used a general keyword search for a common word (sleep) to identify a more random sampling of titles that would include nonfiction and television shows, as well as features.

Searches	
Type	**Search**
Title	Amelie
Title	Citizen Kane
Title	Dracula
Title	Samurai
Title	Rublev OR Rubliev or Rublyov or Rubliov or Rublov
Keyword	Sleep

Processing and Review of Sample Records

Once we obtained the records, we used MarcEdit, a free Windows-based MARC editing tool, to export the relevant data to tab-delimited form and then imported the information into Microsoft Access. We normalized the data and then did some text processing to try to extract the relevant data. This process is described in more detail in the individual review sections.

Following this, we reviewed our results manually to determine if information that was present had been correctly extracted and to identify any patterns of problems. At this point, we have only been able to examine whether or not the data existing in the record was correctly extracted. We plan to assess at least a subset of our data against external sources for accuracy.

Other Issues

We do not believe that we can accurately extract data from multi-work records (e.g., records for a set of all the James Bond movies or a collection of animated shorts). The various pieces of information that pertain to the individual works in a multi-work MARC record are not linked in any way so it is impossible for a machine to identify, for example, which titles go with which dates or genres. It might be possible, once we have a set of provisional work-level records, to identify which works are contained in a given manifestation by matching information in the provisional work-level records to information in the manifestation records. This is an area that will require more manual intervention. We attempted to see how accurately we can identify the multi-work records in our dataset by looking for the presence of things like non-collective titles and analytical titles. We were able to identify almost all of the multi-work records through the presence of information such as contents notes in the record, but we did have a fairly high level of false drops (31%). Based on manual review, 79% of our records represent single works and an additional 6% are records for a main work that mention subsidiary work(s) not likely to interfere with extraction of data about the main work.

We are not sure what the threshold should be for reasonable reliability of this information. It is clear that information derived from manifestation-level bibliographic records will be incomplete and at times incorrect so we will eventually have to decide on an acceptable level of accuracy.

For works that have been issued in many versions, our results may be improved with clustering of manifestation-level records for the same work.

Analysis of Individual Characteristics

Original Date

Fields and Areas of the MARC Record Examined

We attempted to extract the original date from existing MARC bibliographic records for moving images via a number of methods.

1. **008 Date2 (Part of MARC 008 control field)**. When present in the record, this date is the most reliable method of determining the original date for moving image works. For many videos, “Type of date/Publication status” is coded “p” for “Date of distribution/release/issue and production/recording session when different,” the original motion picture date is given in Date2 and the publication date of the video is given in Date1. Date2 may be unreliable in the case of “m” for a range of dates. The only other “Type of date/Publication status” commonly used with Date2 for videos is “r” for

"Reprint/original date" where Date2 may be the original date or the date of a previous release. Note that works originally broadcast on television are generally not supposed to be coded "p."

2. **033: Date/Time and Place of an Event**. This field includes a formatted date/time of creation, capture or broadcast associated with an event. It seems to be more commonly used by archives.

3. **130: Uniform title (main entry)**. The original date is sometimes found here when needed to distinguish between two moving images with the same title.

4. **500: General note**. These notes were parsed to look for years in 18xx, 19xx or 20xx format in combination with a limited set of keywords that often indicate that the note refers to the original date of the work.

5. **518: Date/Time and Place of an Event Note**. Years were extracted from this field in the same manner as for General Note (500) fields above. Although most dates in Date/Time and Place of an Event Note (518) fields probably refer to the original date of recording, this note may also refer to the recording of the video in hand from some other source.

For dates in note fields (500 and 518) we looked for a year in combination with one of the following keywords:

Date Keywords
aired
broadcast
motion
produced
production
recorded live
release
telecast
television
copyright date

The original date may exist in other fields in the record, but we deemed the five listed above to be the most likely sources for reliable information about the original date.

The most common place the original date may be found, other than those described above, is in Date1 in the MARC 008 control field. However, we did not include Date1 in our project because there is no automated means to distinguish between the following scenarios:

1. The date of publication of the video and the date of the work are the same so there is only one date to put in the fixed fields and it is in Date1.

2. The date in Date1 is the date of publication of the video and there is no date in Date2 because:

    a. The cataloger forgot or chose not to do the research to determine the original date.

    b. The cataloger is following newer policies in which changes or additions (e.g., subtitle tracks, making-of featurettes) to the content of the original moving image work make the DVD a new publication with a single date.

We also considered dates in the Publication, Distribution, etc. (260) field, but again there is no reliable way to know when the date of publication is the same as the original date. It is possible that 008 Date1 and the Publication, Distribution, etc. (260) field dates might be useful when looking at archival cataloging where they are more likely to mirror the original production or release date, but we do not think they can be used to identify original dates in the case of general library cataloging.

Analysis

We examined 941 records from six sources. At this point we have only looked at whether we can extract dates that might potentially be the original date via the above methods. We have not assessed the extent to which these dates represent the correct original date.

We found that 72% of the records had some date that potentially could be identified as the original date, while 28% did not contain any information that we could leverage. Some adjustments to the program used to extract this information would improve our results slightly. However about one quarter of the records would still not contain information useful for automatic extrapolation of an original date, as these records include no identifiable dates in any of the fields we examined.

The two methods that worked best for extracting potential original dates were 008 Date2 (present in 41% of records) and the General Note (500) field (present in 39% of records). The other methods, Date/Time and Place of an Event (033), Main Entry-Uniform Title (130), and Date/Time and Place of an Event Note (518) fields, were each present in less than 10% of the records and 033 and 130 were disproportionately represented in records from the film archive, which may indicate a difference between archival and standard library cataloging.

Original Date Overview							
	008 Date2	General Note (500)	Date/Time and Place of an Event Note (518)	Date/Time and Place of an Event (033)	Main Entry-Uniform Title (130)	Overall	Any Date
Correctly-identified data	385	368	37	89	57	676	72%
Blank field or no identifiable date in field	556	407	891	829	846	265	28%
Multiple dates	0	137	13	23	17	0	
Missing keyword associated with presence of date (e.g., “produced”)	0	29	0	0	21	0	
Minimum presence of data**	30%	0%	0%	16%	0%	53%	
Maximum presence of data**	81%	26%	9%	70%	6%	91%	

** Minimum and maximum show variations in the availability of data by institution. That is, the number of records that contained useful data in 008Date2 ranged from 30% in the institution with the lowest use of this field to 81% in the institution with the highest use. These variations can reflect differences in the types of material collected, but also show the effects of local cataloging practices on the availability of data.

Some particular problems encountered in our data sample:

1. Many General Note (500) fields in our record set refer to the date associated with an external verification source, such as the publication year of the American Film Institute catalog or the date the cataloger checked the Internet Movie Database. Our program cannot distinguish between these dates and relevant dates and may incorrectly use the verification date as the original date. This could be resolved in many cases by having a hierarchy of date sources rather than just identifying the earliest date in the record as we are currently doing.

2. Records in which the General Note (500) field contains multiple dates, one of which is the release date, but the earliest date refers to an event other than the release.

3. Different or inconsistent dates in the Date/Time and Place of an Event (033) and Main Entry-Uniform Title (130) fields for the same video. For example, a record may contain a uniform title of “Simpsons (Television program : 1989),” qualified by the date the show began airing, as well as a Date/Time and Place of an Event (033) field of 19920507 that represents the date of a particular episode.

4. Incorrect cataloging practice for the 008 Date1 and Date2 fields, in which the dates are reversed so that the original date is in Date1 and the manifestation date is in Date2. Date1 is supposed to contain the publication date of the manifestation in hand and Date2 may contain the original release date under certain circumstances. Recording dates in reverse order is a non-standard use of MARC coding to achieve a desired end, i.e., sorting by original release rather than publication date in most OPACs, as OPACs generally sort on Date1.

5. Keywords that signal dates in General Note (500) and Date/Time and Place of Event (518) fields that were not included in our original program, e.g., "filmed," "copyright," "recorded." "Recorded" can be unreliable as it sometimes refers to the date a video copy was made.

6. In the Main Entry-Uniform Title (130) field, we also missed dates in titles that did not include the phrase "motion picture" or "television program," but our program could be revised to pick up those dates.

7. In addition, some dates are in notes in the form 28Feb36, which is harder to extract. We did remove "c" from in front of dates in the form c1999 so we were able to pick those up.

Recommendations

There should be a field in the MARC record where the original date of a moving image work can be unambiguously recorded. It is probably sufficient to record the year, but may be useful to include an option for recording exact dates, particularly for episodes of television programs. Perhaps the formatted Date/Time and Place of an Event (033) could be expanded to incorporate this use.

Original Title

Fields and Areas of the MARC Record Examined

We attempted to extract the original title from existing MARC bibliographic records for moving images via a number of methods.

1. **130: Uniform title (main entry)**. This is the only field that is likely to reliably contain the original title of a work. However, this field is not widely used for moving images, especially in older cataloging. Only 22% of the records in our sample contained Main Entry-Uniform Title (130) fields.

2. **245 $a: Title proper**. This is generally supposed to be the title on the title frames. However, not all videos have a title on the title frames. In addition, some catalog records are created from information on the container. Some distributors (e.g., Insight Media) often use a different title on the container and disc label from the title on the title frames. There are also inconsistencies in how titles are transcribed when more than one title appears on the title frames, particularly in the case of parallel titles and titles of works

that form a part of larger works (e.g., episodes of television programs). Sometimes the original title does not appear on the item at all and therefore may not appear in the record.

3. **245 $b: Other title information**. This subfield is unlikely to contain the original title except in instances where the original title has been transcribed as a parallel title and the translated title has been used in the Title Proper (245 $a) subfield. It may contain one or more of many original titles in the case of multi-work manifestations without a collective title.

4. **246: Varying form of title**. This title is not likely to be the original title, but occasionally an original title might be found here in the form of a note like "Originally released as…" or in the form of a parallel title where the English translation is given in the Title Proper (245 $a) subfield.

Analysis

The fundamental problem here is that although the original title is usually in the record somewhere, unless there is a Main Entry-Uniform Title (130), it is difficult to see how it would be possible to make an automated assessment as to whether a given title is the original title. It may be more realistic to create a cluster of titles associated with a work and then rely on later human intervention to identify one as the original title. Or perhaps some predictions could be made based on more complicated algorithms (e.g., if the original language can be identified and the language of the title in the Title Proper (245 $a) subfield is in the same language, assume that that is the original title).

We examined 941 records from six sources. We considered titles found in Main Entry-Uniform Title (130) fields to be correctly-derived and to mostly likely represent the original title or at least a title consciously chosen to represent the work. Unfortunately, only 22% of our sample had Main Entry-Uniform Title (130) fields and a disproportionate number of those (approximately half) came from the film archives in our example. Only 16% of the library records included a uniform title.

At this point we have not evaluated the titles found for accuracy against external sources. However, we manually reviewed the titles retrieved and made an assessment as to how likely the title in the Title Proper (245 $a) subfield, Remainder of Title (245 $b) subfield or Varying Form of Title (246) field is to be the original title. It seems probable that the Title Proper (245 $a) subfield title is the original title 92% of the time. Titles in the Remainder of Title (245 $b) subfield and the Varying Form of Title (246) field are far less likely to potentially be the original title.

Since in most cases there is no obvious reason to suspect that the Title Proper (245 $a) subfield title is not the original title, we examined the ones that seemed suspicious and found that 30 (44%) involved originally non-English language titles where an English language title had been given in the Title Proper (245 $a) subfield. The remainder consisted of variations between the Main Entry-Uniform Title (130) field and the Title Proper (245 $a) subfield. These include things like possessives at the beginning of a title and situations where a television uniform title is

given in a Main Entry-Uniform Title (130) field and episode titles in the Title Proper (245 $a) subfield. It is possible that in most cases, the Title Proper (245 $a) subfield title could be provisionally given as the original title.

Original Title Overview				
	Main Entry-Uniform Title (130)	Title Proper (245 $a)	Remainder of Title (245 $b)	Varying Form of Title (246)
Correctly-identified data	21.6%	0.0%	0.0%	0.0%
Blank field or no identifiable date in field	78.4%	0.0%	93.8%	58.6%
Possible/probable original title	0.0%	92.7%	0.5%	3.5%
Probably not original title	0.0%	7.3%	5.6%	37.9%

Reasons Why 245 $a is Probably Not Original Title	
English Title Proper (245 $a) not = Main Entry-Uniform Title (130)	38
Non-English Title Proper (245 $a) not = Main Entry-Uniform Title (130)	1
Non-English film but Title Proper (245 $a) subfield is English	29

Notes about the data:

1. If the Main Entry-Uniform Title (130) field or the Title Proper (245 $a) subfield contained a number in word format (e.g., *Magnificent Seven*) and the Varying Form of Title (246) field contained the number in numeral format, we selected “probably not original title” for the 246 assessment.

2. If the Main Entry-Uniform Title (130) field contained the words “television program,” “motion picture,” or “cartoon” after the title and the 245 or 246 title fields contained the same exact title, except it didn't include these words, we selected “possible/probable original title” for the 245 or 246 title. We also did this if the Main Entry-Uniform Title (130) included a date that wasn't included in the 245 or 246 title.

3. If we knew that the title wasn't the actual title (primarily for the *Samurai I*, *II* and *III* films where the original titles should be Japanese), but the Japanese title wasn't in the record, we still selected “probably not original title” even if there was enough information (usually subtitle information we found on the Internet Movie Database) in the record to convince us that it was that film.

Recommendations

Catalogers should include 130 (main entry) and/or 730 (added entry) uniform titles for works in moving image records.

Director

Fields and Areas of the MARC Record Examined

We attempted to extract the director's name from existing MARC bibliographic records for moving images via a number of methods. We took as the desired endpoint correctly identifying the 700 field (Added Entry–Personal Name) containing the authorized, standardized form of the director's name. It is possible that the director's name might occur in a 100 field, but this is relatively rare and we did not account for this possibility in our sample. Director can also be traced in the Added Entry-Corporate Name (710) field. During out post-processing analysis, we found this type of added entry in the case of the director team The Brothers Quay in our sample.

1. **245 $c: Statement of responsibility, etc.** Many records contain a transcribed statement of responsibility including the director's name and the function, usually as they appear on the title frames. Moving images often list multiple functions in the statement of responsibility, with each distinct function separated by specific punctuation, i.e., space-semicolon-space. We used this prescribed punctuation to parse each statement of function and attempt to match it with its associated authority-controlled name entry.

   We identified each statement of function that included the letter sequence "direct" to pick up variations such as "director," "directed, "direction," etc. We did not attempt to account for non-English terms for director or directing in our test run.

   Since we had no way to automatically identify names as opposed to other types of information, we went through all the words occurring in a given directing function statement and attempted to match at least (1) two consecutive words or (2) two words separated by a single word with words occurring in a 700 field. The latter helped with names that had middle initials in the Statement of Responsibility, etc. (245 $c) subfield, but not in the matching Added Entry-Personal Name (700) field. On the whole, this method worked well, but did lead to a few false hits (erroneously matched headings), generally involving names with initials, which more sophisticated programming could probably eliminate.

2. **508: Creation/Production Credits.** The type of credits included in this field on moving image records varies. Creation/Production Credits Note (508) fields often include only credits considered to be more minor than director, producer and screenwriter, particularly for feature films. On the other hand, some institutions, at least under some circumstances, use this field for the main or all credits for a moving image. Like the Statement of Responsibility, etc. (245 $c) subfield, this field consists of statements of function and related names, with each function separated by space-semicolon-space or possibly just

semicolon-space. We processed the data in this field using the same procedure described for the Statement of Responsibility, etc. (245 $c) subfield above.

One additional difficulty with this field is that it often includes various types of directors other than the primary director, e.g., statements such as “director of photography” or “art direction.” Our program was not sophisticated enough to identify those by methods such as prospectively accounting for variations or attempting to limit occurrences of “director” to those occurring at the very beginning of a statement of function. Since data in moving image Creation/Production Credits Note (508) fields is usually given in the form of function followed by name, the easiest shortcut to eliminating most false drops would be to require “direct” to appear at the beginning of the statement. It would, however, still be necessary to explicitly exclude “director(s) of photography” and many less commonly-occurring phrases, e.g., “directing animators.” It is unlikely to be practical to achieve 100% accuracy in discriminating between main directors and other types of directors and directing functions. This problem can also occur in the Statement of Responsibility, etc. (245 $c) subfield, but is less frequent.

Many libraries do not usually trace these other types of directors so there often is not a matching Added Entry-Personal Name (700) field in the record, which cuts down on the number of false drops. On the other hand, since the Creation/Production Credits Note (508) field is a note field and not a transcribed field, it is unusual to find non-English language data in a Creation/Production Credits Note (508) field in an English language bibliographic record. Therefore, in the majority of cases, it is only necessary to match on variations of “direct,” unlike with Statement of Responsibility, etc. (245 $c) subfield information, which is more likely to include non-English terms for director or directing.

3. **700: Added Entry–Personal Name with $e direction**. Some 700 personal name fields include a relator term of “direction” in 700 $e identifying that person as the director.

4. **700: Added Entry–Personal Name with $4 drt**. Some 700 personal name fields include a MARC relator code of “drt” in 700 $4 identifying that person as the director.

The director’s name may exist in other places in the record, such as in Formatted Contents Note (505) fields in multi-work records, but we deemed the four listed above to be the most commonly-occurring.

Analysis

We examined 941 records from six sources. We found that we could identify at least one Added Entry-Personal Name (700) field representing a director in 62% of the records. The vast majority of these (84%) were derived from matching statements of responsibility from the Statement of Responsibility, etc. (245 $c) subfield with Added Entry-Personal Name (700) fields. 700 $e (relator term “direction”) and 700 $4 (MARC relator code “drt”) each identified directors in about 15% of the records. Relator Terms ($e) were used almost exclusively by the film archive, which included a relator term for director in 98% of its records. The remaining works likely did not have directors or did not have named directors. Use of the Relator Code ($4) identified

directors in about 15% of the records. The use of Relator Code ($4) subfields varied widely among institutions and ranged between 0-83% for a given institution. This reflects the impact of local cataloging practices on the usability of data for our purposes. Most of the directors identified by Relator Term ($e) and Relator Code ($4) subfields were also identified by matching Added Entry-Personal Name (700) fields with the Statement of Responsibility, etc. (245 $c) subfield and the Creation/Production Credits Note (508) field, but the use of relator terms ($e) and relator codes ($4) has the advantage of eliminating all of the hard matching problems (e.g., accounting for foreign language terms for director and variations in spelling, transliteration and form of name). The Creation/Production Credits Note (508) field was the least successful method and was useful in identifying a director in only 5% of our records.

On the other hand, a quarter of the records did not include identifiable director information in the fields we examined and a further 9.6% did not include a matching Added Entry-Personal Name (700) field with a controlled name for the director(s) identified in the Statement of Responsibility, etc. (245 $c) subfield or Creation/Production Credits Note (508) field. Less than 10% of the records with no director information included director in a Formatted Contents Note (505) field. The rest either had no director information, used a different form (e.g., “a film by…”) or the cataloger omitted that information.

Some of the names in the Statement of Responsibility, etc. (245 $c) subfield and Creation/Production Credits Note (508) fields that our program was unable to match correctly could be resolved with more sophisticated programming. For example, thirty names (3%) in the Statement of Responsibility, etc. (245 $c) subfield failed to match because we did not look for non-English director functions such as “Regie” or “kantoku”. However, accounting for all variations, would be time-consuming vis-à-vis the number of records affected. This problem is somewhat mitigated by the fact that not all libraries transcribe original language credits; many prefer to use English language credits from another source.

Some names failed to match because of variations in spelling or transliteration between the transcribed and authorized forms (e.g., “Pierre Schoendorffer” vs. “Schoendoerffer, Pierre” and “Andrei Tarkovsky” vs. “Tarkovskii, Andrei Arsenevich”). In some cases the name was traced under a different form entirely (e.g., “T. C. Frank” vs. “Laughlin, Tom”). Some match failures could be resolved by using both the official Added Entry-Personal Name (700) field form of name and the forms of name in the cross-references in the relevant authority record.

Director Overview						
	Statement of Responsibility, etc. (245 $c)	Creation/ Production Credits Note (508)	Added Entry-Personal Name with Relator Term (700 $e)	Added Entry-Personal Name with Relator Code (700 $4)	Overall	Overall %
Correctly-identified data	310	53	142	144	584	62.1%
Blank field or no identifiable relevant information	492	576	799	797	237	25.2%
Problem with matching algorithm and initials; fixable with better programming	4	4	0	0	3	0.3%
Director is corporate body (710)	1	1	0	0	2	0.2%
No matching authorized name (700) for transcribed name	84	6	0	0	90	9.6%
Non-English term for director	30	0	0	0	9	1.0%
Difference in spelling or transliteration between transcribed and authorized forms of name	16	1	0	0	11	1.2%
Stage director	0	1	0	0	1	0.1%
Other difference between transcribed and authorized form of name (e.g., use of variant names or pseudonyms)	4	3	0	0	4	0.4%
Wrong director type (e.g., director of photography)	0	296	0	0	0	0.0%
Minimum presence of data**	44%	0%	0%	0%		
Maximum presence of data**	69%	12%	43%	83%		

** Minimum and maximum show variations in the availability of data by institution. That is, the number of records that contained useful data in Added Entry-Personal Name fields with relator codes (700 $4) ranged from 0% in the institution with the lowest use of this field to 84% in the institution with the highest use. These variations can reflect differences in the types of material collected, but also show the effects of local cataloging practices on the availability of data.

Recommendations

Although the matching algorithm found corresponding authorized names in Added Entry-Personal Name (700) fields for most directors transcribed in the corresponding Statement of Responsibility, etc. (245 $c) subfield, a certain number of matches will inevitably be missed due to variations in form of name or non-English terms for director.

Accuracy is still unlikely to reach 100%, even if we take into account authority record cross-references and include additional non-English director keywords. The process of matching transcribed and authorized forms after the fact is inherently more complex than indicating during cataloging that this particular authorized form accurately identifies the director. The use of $4 (MARC relator code) or $e (relator term) is more reliable and more amenable to machine-based processing than even the most sophisticated matching algorithm and it is recommended that one of these options be used whenever possible. This is particularly useful for moving image records, which usually record a variety of functions.

Original Language

Fields and Areas of the MARC Record Examined

We attempted to extract language data from existing MARC bibliographic records for moving images via two methods in order to determine whether we could identify the original language(s) of those moving images.

1. **008 Language Code (Part of MARC 008 control field)**. The MARC code for the main, first or only language associated with an item is given in the language positions of the 008 field. If there is no additional language information given in the record, it is likely that the language in 008 is both the language of the item in hand and the original language of that moving image. However, some records which should have additional language information don't, either because the cataloger didn't have the information (e.g., some dubbed nonfiction videos are difficult to identify as such) or for whatever reason did not include the information in the record. The percentage of records with missing language information is unknown.

2. **041 $h: Language code of original and/or intermediate translations of text**. If additional language information is supplied and an item includes a translation, the original language of an item can be coded in the Language Code of Original and/or Intermediate Translations of Text (041 $h) subfield. Although the definition of this subfield includes languages of intermediate translations, these are unlikely to happen with moving images and if they should occur, are even less likely to be known to catalogers. So if data exists in Language Code of Original and/or Intermediate Translations of Text (041 $h) subfield, it is likely to be a reliable source of information about original language.

Analysis

Original language has a fairly high percentage of correctly-derived data. 78% of records examined include a language or languages that can be inferred to be the original language. However, the impact of missing data on the accuracy of these results is unknown. Some omissions could probably be identified and resolved by clustering of records for various manifestations of a given work.

The majority of records examined (66%) have only a single language in 008. Of the remaining records, 115 (12%) include an original language coded explicitly in Language Code of Original and/or Intermediate Translations of Text (041 $h) subfield. 198 records (21%) include a Language Code (041) field without a $h. For various reasons, including inconsistency in the practicing of coding the Language Code (041) field indicator for whether or not an 041 includes a translation, it is impossible to accurately infer original language in this situation. For example, two languages in the Language Code of Text/Sound Track or Separate Title (041 $a) subfield could be parallel soundtracks or a single mixed soundtrack. The likely conclusions to be drawn about these two situations would be different. In the first, one of the languages is probably the original language. In the second, both are probably original languages.

Original Language Overview

	008 Language Code	Language Code of Original and/or Intermediate Translations of Text (041 $h)	Overall	Overall %
Correctly-identified data	618	115	733	78%
Blank field or no identifiable relevant information	0	825		
Invalid code	0	1	1	0.1%
Fill character	9		9	1%
Original language in 041$h	116			
Includes 041 without $h	198		198	21%

Notes about the data:

1. Nine records had fill characters in the 008 language code and no other language data. It is not clear if this is an omission, an attempt to represent silent film or an error.

2. One record had an invalid two-letter code in Language Code of Original and/or Intermediate Translations of Text (041 $h) subfield so we counted this separately.

Recommendations

Catalogers should include a Language Code (041) field as well as a Language Code of Original and/or Intermediate Translations of Text (041 $h) subfield in moving image records when applicable. Practice in recording Language Code of Original and/or Intermediate Translations of Text (041 $h) subfield should be standardized so that both parallel soundtracks and subtitles are coded with a first indicator of one for including a translation. Language Code of Original and/or Intermediate Translations of Text (041 $h) subfield should be used consistently after both spoken and written (e.g., subtitled) translations of the moving image's dialogue or original intertitles.

OLAC has recommended to MARBI that a subfield be included in the Language Code (041) field where the original language can be explicitly coded in all cases. If this subfield is implemented, it should be used to bring out the original language explicitly whenever it is known.

Original Aspect Ratio

Fields and Areas of the MARC Record Examined

We attempted to extract aspect ratio data from existing MARC bibliographic records for moving images via a number of methods in order to support inferences about the original aspect ratio of those moving images.

1. **250: Edition statement**. Statements such as widescreen or fullscreen are often found in the edition statement area. Publishers issue many popular films in both formats. In addition, many libraries include this information in the Edition Statement (250) field so that it displays more prominently to users even when only one version exists.

2. **538: System requirements**. Physical description notes that contain words or ratios designating the aspect ratio of the item are often combined with System Details Note (538) fields describing playback requirements.

3. **500: General note**. Physical description notes that are recorded in General Note (500) field may contain words or ratios designating the aspect ratio of the item.

4. **505: Contents note**. Information about aspect ratio is occasionally found here when a DVD contains both full screen and widescreen versions.

In order to identify when the listed fields actually included aspect ratio information, we looked for some key phrases in our selected fields as follows:

Aspect Ratio Keywords
aspect (in combination with a ratio)
fullframe, full frame, full-frame
fullscreen, full screen, full-screen
letterbox, letterboxed
ratio (in combination with a ratio)
standard format
widescreen, wide screen, wide-screen

Analysis

The primary difficulty with trying to extract original aspect ratios from current bibliographic records is that if an aspect ratio is given, it is the aspect ratio of the item in hand and it is difficult to say whether that is the same as the original or not. However, it may be possible to make some reasonable inferences based on

1. Other information in the record. For example, it might be possible to conclude that television shows produced prior to a certain date would all be in the 4:3 aspect ratio.

2. Clustering of various manifestations of a given work. If only widescreen or both widescreen and full screen versions exist, it is probably reasonable to infer that the original was widescreen, although we may not know the exact ratio.

Looking at our sample of data for aspect ratios of items in hand, another problem is that this data seems to be given in any form in only about a quarter of the records that we examined. The existing data was fairly evenly split between the Edition Statement (250), System Details Note (538) and General Note (500) fields (9%, 8% and 9% correctly derived respectively). However, since the data usually occurs in only one of these fields, the aggregate percentage of records with a correctly-identified aspect ratio in at least one field is 23%. The field preferred for recording this data seems to vary by library.

Aspect Ratio Overview						
	Edition Statement (250)	System Details Note (538)	General Note (500)	Formatted Contents Note (505)	Overall	Overall %
Correctly-identified data	82	75	81	3	216	23%
Blank field or no identifiable date in field	843	863	854	937	702	75%
Missing aspect keyword	14	2	5	1	20	2%
Aspect keyword, wrong context	0	0	1	0	0	0%
"Theatrical format"	2	0	0	0	2	0%
Unclear	0	1	0	0	1	0%

Notes about the data:

1. Two records were identified in the Edition Statement (250) field as “Original theatrical format,” which probably means widescreen.

2. One record stated in a System Details Note (538) field “Technirama not letterboxed.” This probably means that the video is in full screen format, but it is not absolutely clear so this one is marked “unclear.” This also demonstrates a significant pitfall that impacts the accuracy of using keyword searches on free-text note fields. Sometimes keywords are used in the context of stating negatives, e.g., not letterboxed or not closed-captioned.

3. One record had a statement in a General Note (500) field that it included a “widescreen to fullscreen comparison.” This led our program to conclude that the DVD included both versions when, in fact, the complete film was presented only in widescreen. This one is marked “aspect keyword, wrong context.” Again, this demonstrates a potential shortcoming of using notes rather than controlled data fields for information that we want to be able to retrieve consistently.

4. In one case, we accidentally noted that a video was part of a “Widescreen Collection” series. This suggests that series fields may be an additional place to look for aspect ratio information.

5. We used the presence of a colon between two numbers to identify aspect ratios. In some cases in notes, the program misidentified times as aspect ratios (e.g., 20:34 for 20 minutes and thirty-four seconds).

Recommendations

There currently does not seem to be anywhere in the MARC bibliographic record that aspect ratios can be recorded unambiguously. It is desirable to create such a field so that this data can be encoded in a form that can be consistently used for retrieval when it is known.

Multiple Works on One Bibliographic Record

Background

The types of inferences we are attempting to make are only possible with bibliographic records that represent one work or one main work. In the case of multi-work items, it does not seem possible to automatically answer such questions as which title goes with which director goes with which date. In order to estimate the potential impact of this difficulty, we tried various strategies to automatically identify multi-work bibliographic records and then matched this against a manual review.

1. **245 $b: Non-collective title**. When a physical item does not have a title that refers to the whole item, but it does have titled parts, a non-collective title is recorded in the 245 field. The first part title is recorded in the Title Proper (245 $a) subfield and the second often in the Remainder of Title (245 $b) subfield preceded by a semi-colon. We have identified these based on the punctuation. In some cases, the second part title is recorded in the Statement of Responsibility, etc. (245 $c) subfield, but we were unable to come up with a method to systematically identify these, as these semi-colons cannot be distinguished from semi-colons used to separate different statements of responsibility.

2. **505: Enhanced contents note**. These are contents notes where individual titles and authors are contained in separate subfields. It seems more likely that these usually represent separate works than the unenhanced contents note described below.

3. **505: Unenhanced contents note**. These are contents notes where titles and authors have not been explicitly identified, either because the record predates the ability to enhance a Formatted Contents Note (505) field, the cataloger chose not to make an enhanced contents note or because it makes no sense to make an enhanced contents note (e.g., a 505 field noting chapter titles for keyword searching).

4. **740 02: Analytical title**. This field can be used to make added entries for titles of parts of an item if they are deemed important.

Analysis

Based on our manual review, 740 (79%) of the records represent single works. An additional 60 (6%) records include substantial supplemental work(s) which are mentioned in notes and which may warrant a link to a separate moving image work record, but which probably do not contain additional non-title data that would become confused with the data we might be able to extract

about the main work. This suggests that our method of extracting data would not be compromised by the presence of data about multiple works in most cases.

We also attempted to automatically identify records that might contain multiple works in order to assess how accurately we could identify potentially problematic records. It appears that the majority of records that include multiple works include some clue as to their presence. Of the 37 records that include multiple works, but were not automatically identified as potential multi-work records, 35 were for records that include supplementary works not likely to interfere with data extraction and two incorrectly included the entire non-collective title in the Title Proper (245 $a) subfield.

72 (31%) of the 236 records automatically identified as potential multi-work records turned out to be single works. All of the records with non-collective titles were multi-work records, but this was also the least frequent situation. Most of the records with enhanced Formatted Contents Note (505) fields that were manually identified as single works were compilations of musical works so they do contain multiple works, but not necessarily multiple moving image works in our context. Unenhanced Formatted Contents Note (505) fields were the biggest source of false drops. 57 (41%) of 140 identified as multi-work records turned out to be single works. Many of these are for chapter titles or for non-title information such as widescreen and full screen versions. Occasionally, contents notes are given for accompanying materials (e.g., music CDs), which can also cause false drops in this area. In the case of analytic titles in Added Entry-Uncontrolled Related/Analytical Title (740) fields, the four that turned out to be single works include two errors and two analytic titles for things that were deemed not to be separate moving image works. One incorrect Added Entry-Uncontrolled Related/Analytical Title (740) field was for a television show title that should have been in a non-analytical Added Entry-Uniform Title (730) field and one was for an English translation of the title of the novel that the film is based on. One record included the title of a DVD-ROM feature that might be considered a supplemental work, but insufficient information was available to make this judgment or to tell whether it was a moving image or some other content type. The final record included a Added Entry-Uncontrolled Related/Analytical Title (740) field for a single song that was part of a live performance.

It seems that barring cataloging errors, the majority of records potentially containing information about multiple works can automatically be identified in advance. The records which contain substantial supplemental works would ideally be linked both to their main work and to their supplemental work(s) when they can be identified, but the supplemental works generally should not interfere with our goal of data extraction. The records identified as potential multi-work records would probably have to be routed for some sort of manual review.

Multiple Works on One Bibliographic Record						
	Manual review	Auto review	Remainder of Title (245 $b) for non-collective titles	Enhanced Formatted Contents Note (505)	Non-enhanced Formatted Contents Note (505)	Added Entry-Uncontrolled Analytical Title (740, 2nd indicator 2)
Single work	740	702				
Multi-work record (all)	201	236	10	75	140	51
Multi-work subcategories						
Multi-work record (general)	58					
Multiple TV episodes	83					
Includes substantial supplementary work	60					
Misidentified as multi-work		72	0	12	57	4
Not automatically identified as multi-work	37					

Recommendations

Ideally, all moving image manifestation records would contain uniform title(s) for the main work(s), as well as uniform titles for significant supplemental works so that the number of works represented on a given record could be easily ascertained. However, this does not solve the underlying problem of how to connect data related to different works. Although the MARC format includes linking subfields, these are rarely used and systems do not seem to be able to make use of them.

In the future, it is hoped that manifestations that include multiple works can be linked to work or expression records with more detailed information that would eliminate the current confusion. Many current records for multiple moving image works, if they attempt to give many details at all, are not only incomprehensible for machines, but are confusing and jumbled from the point of view of human users.

Summary of Recommendations for Improving Machine-Based Access to Work-Level Information in MARC Bibliographic Records for Moving Images

It would be desirable to be able to easily extract work-level information from existing MARC manifestation-level bibliographic records. It would also be useful, so long as current MARC bibliographic records are used, to be able to automatically insert previously verified work-level information into a new MARC bibliographic record for a manifestation or to update existing MARC manifestation records with corrected or expanded work-level information.

In order to do this, it is necessary to be able to easily and accurately identify the location of this data in the bibliographic record. Our exercise has shown that this is not always straightforward.

However, there are some things that catalogers can do in the existing context that will ensure that data is available for machine processing.

1. Use 130 and, when applicable, 730 uniform titles for all moving image works.

2. Use relator codes or terms after 1xx and 7xx fields for responsible entities wherever possible.

3. Use Language Code of Original and/or Intermediate Translations of Text (041 $h) subfield to bring out the original language(s) whenever possible.

There are also some ways in which the MARC bibliographic format could be modified to enable machine-readable encoding of data for data elements that currently do not have such fields. Some possible changes that could improve automatic identification of the data elements we examined are listed below

1. OLAC has submitted a proposal to add a subfield to the Language Code (041) field that would allow the original language to be explicitly coded.

2. A field should be created in the MARC bibliographic record where the original date of a moving image work can be unambiguously recorded. This could be an expanded use of 033 (Formatted date/time of an event) or a new field.

3. Although original aspect ratio is not currently recorded in MARC bibliographic records nor is it likely to be, it would be beneficial for later analysis and FRBR-based implementations to encode the aspect ratio of the item in hand in a machine-readable format as this is often important to users in selecting appropriate expressions.

Conclusion

Our preliminary assessment of a sample of records suggests that varying amounts of work-level data can be extracted from MARC bibliographic records. About 20% of our sample consisted of multi-work records which are not likely to prove amenable to automated extraction of work-level data. We extracted potential work-level data for original language in 78% of the records in our dataset, for original date in 72% and for director in 62%. Although about 20% of the records we examined had Main Entry-Uniform Title (130) fields, original title and original aspect ratio are difficult to directly derive from most single bibliographic records, but potentially could be identified by looking at patterns in clusters of records for the same work.

We have not examined the data extracted for accuracy by verifying against external sources, but the percentage of correct data that is extracted will be lower than the percentage of possibly correct data that we have currently identified.

We have identified a number of areas in which cataloging practices or the MARC bibliographic format could be changed to improve our ability to automatically identify work-level data. We have provided recommendations for accomplishing these aims. These include encouraging catalogers to consistently add this information when known and to add information to machine-

parsible fields if possible, as well as suggestions for several new fields or subfields for the MARC bibliographic format.

In particular, we would like to raise awareness among catalogers about the type of information that is likely to be useful in creating work-level records and what methods are most effective for recording it in a machine-readable manner. We hope that this will increase recording of this information, as well as standardization. Local practices and individual catalogers clearly have an influence on the prevalence and retrievability of the data elements we examined. For example, the percentage of records from a given institution for which we could extract an authorized form of a director's name ranged from 47-84%. However, the percentage from which we could use the most reliable method of extracting director name, i.e., a MARC relator term or code, ranged from 2% to 83%. Certainly factors other than cataloging practices, such as the availability of the director's name at the time of cataloging and the relative importance and applicability of the director function to a given resource, affect this percentage, but it seems clear that catalogers have an opportunity to increase the usefulness of this data for later use at the time of input.

It is unlikely that complete, accurate work-level records could reliably be derived from existing MARC bibliographic records in most cases, but it is possible that "good-enough" provisional records could be created and then revised and upgraded by human beings. We think this approach bears further investigation and testing.