OLAC CAPC
Moving Image Work-Level Records Task Force
Final Report and Recommendations
August 21, 2009
Amended August, 28, 2009 to change Scarlet to Scarlett

Part IV: Extracting Work-Level Information from Existing MARC Manifestation Records
Appendix: Comparison of Selected Extracted MARC Data with External Sources

Task Force Members:
Kelley McGrath (chair; subgroup 1 and 4 leader)
Susannah Benedetti (subgroup 2)
Lynne Bisko (subgroup 4)
Greta de Groat (subgroup 1)
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)

Karen Gorss Benko (subgroup 3, 2008)
Scott M. Dutkiewicz (subgroup 4, 2008)

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

Titles Reviewed

We reviewed ten titles. These ten works were found in eighty-three records in our sample dataset. These represent primarily major feature films as those are the materials for which there are most likely to be reliable, independent secondary sources of information. The titles include several English language American film releases, one non-English Roman alphabet title, three non-English, non-Roman alphabet titles, and one English language television documentary. The dates of the works range from 1931 to 2001.

Five Online Sources Reviewed

We selected five reliable, comprehensive online sources to review. We did not use a print source as a sufficiently comprehensive one was not conveniently available to us at the time of the review.

1. allmovie (AMG) http://allmovie.com/

   Contains over 220,000 titles. Includes information from a variety of sources, including video packaging, promotional materials, press releases, watching the movies, etc.

2. Internet Movie Database (IMDb) http://www.imdb.com/

   Consists of a comprehensive database with over 2,250,000 filmographies covering some 180,000 movie titles and over 560,000 people. Includes primarily user-generated content; the main sources of staff-generated information are on-screen credits, press kits, official bios, autobiographies, and interviews.

3. Baseline (InBaseline) http://www.inbaseline.com/

   Provides credits and titles for film and television. Employs a full-time research staff to track and verify information. Covers features with theatrical distribution in the U.S., award-winning shorts, primetime television, and major cable network projects.

4. Turner Classic Movies Database (TCM) http://www.tcm.com/

   Focuses on pre-1990 films, particularly studio-era years from the late twenties to the early 1970s. The information that appears on TCM.com comes from several sources. It includes licensed material from the American Film Institute Catalogues (AFI), Baseline, Scarlett (Turner Entertainment's internal database), selected information from IMDB, TCM-produced content (articles and home video reviews, media and archival material), and user-generated content (user reviews and contributions).

5. Variety.com http://www.variety.com/

   Includes content from Variety, a premiere source of news and analysis about the entertainment industry since 1905, as well as additional content. Provides credits and reviews for many films and television programs.

Discussion of Data Sources

It quickly became obvious that all of our sources were not using the same rules to populate their data. This underlines the importance of identifying sources of work-level data in order to account for conflicting data. Most of the sites do not provide explicit explanations of their rules, but many of them can be inferred by looking at the data. The data we obtained are listed below:

Source	Title	Date	Director	Language	Aspect
allmovie	Le Fabuleux Destin d'Amélie Poulain	2001	Jean-Pierre Jeunet		widescreen
Baseline	Amelie	2001	Jean-Pierre Jeunet		
IMDB	Le Fabuleux destin d'Amélie Poulain	2001	Jean-Pierre Jeunet	French	2.35:1
TCM	Amelie	2001	Jean-Pierre Jeunet	French	widescreen
Variety.com	Amelie	2001	Jean-Pierre Jeunet		widescreen
allmovie	A Night at the Opera	1935	Sam Wood		
Baseline	A Night at the Opera	1935	Sam Wood		
IMDB	A night at the opera	1935	Sam Wood ; Edmund Goulding (uncredited)	English | Italian	1.37:1
TCM	A Night at the Opera	1935	Sam Wood		1.37:1
Variety.com	A Night at the Opera	1935	Sam Wood		
allmovie	Andrei Rublev	1966	Andrei Tarkovsky		
Baseline	Andrei Roublev	1966	Andrei Tarkovsky		
IMDB	Andrey Rublyov	1966	Andrei Tarkovsky	Russian | Italian | Tatar	2.35:1
TCM	Andrei Roublev	1966	Andrey Tarkovskiy	Russian	2.35:1
Variety.com	Andrei Roublev	1973	Andrei Tarkovsky		
allmovie	Citizen Kane	1941	Orson Welles		
Baseline	Citizen Kane	1941	Orson Welles		
IMDB	Citizen Kane	1941	Orson Welles	English	1.37:1
TCM	Citizen Kane	1941	Orson Welles		1.37:1
Variety.com	Citizen Kane	1941	Orson Welles		
allmovie	Dracula	1931	Tod Browning		
Baseline	Dracula	1931	Tod Browning		
IMDB	Dracula	1931	Tod Browning ; Karl Freund (uncredited)	English | Hungarian | Latin	1.37:1
TCM	Dracula	1931	Tod Browning		
Variety.com	Dracula	1931	Tod Browning		
allmovie	Invasion of the Body Snatchers	1956	Don Siegel		
Baseline	Invasion of the Body Snatchers	1956	Don Siegel		
IMDB	Invasion of the Body Snatchers	1956	Don Siegel	English	2.35:1
TCM	Invasion of the Body Snatchers	1956	Don Siegel		2.00:1
Variety.com	Invasion of the Body Snatchers	1956	Don Siegel		widescreen
allmovie	Invasion of the Body Snatchers	1978	Philip Kaufman		
Baseline	Invasion of the Body Snatchers	1978	Philip Kaufman		
IMDB	Invasion of the Body Snatchers	1978	Philip Kaufman	English	1.85:1
TCM	Invasion of the Body Snatchers	1978	Philip Kaufman		1.85:1
Variety.com	Invasion of the Body Snatchers	1978	Philip Kaufman		

Source	Title	Date	Director	Language	Aspect
allmovie	original title not given (Samurai III)	1956	Hiroshi Inagaki		
Baseline	not found (Samurai III)				
IMDB	Miyamoto Musashi kanketsuhen: kettô Ganryûjima	1956	Hiroshi Inagaki	Japanese	??
TCM	Samurai (Part III)	1967	Hiroshi Inagaki		
Variety.com	not found (Samurai III)				
allmovie	Shichinin no samurai	1954	Akira Kurosawa		
Baseline	Seven Samurai	1954	Akira Kurosawa		
IMDB	Shichinin no samurai	1954	Akira Kurosawa	Japanese	1.37:1
TCM	The Seven Samurai	1954	Akira Kurosawa	Japanese	1.37:1
Variety.com	Shichinin no Samurai	1954	Akira Kurosawa		
allmovie	The Battle Over Citizen Kane	1996	Michael Epstein ; Thomas Lennon		
Baseline	The Battle Over Citizen Kane	1996	Thomas Lennon ; Michael Epstein		
IMDB	The Battle Over Citizen Kane	1996	Michael Epstein ; Thomas Lennon	English	1.85:1
TCM	The Battle Over Citizen Kane	1996	Thomas Lennon ; Michael Epstein		
Variety.com	The Battle Over Citizen Kane	1996	Thomas Lennon ; Michael Epstein		

English language titles were recorded consistently in our external data sources. As with AACR2 uniform titles, possessives from title frame titles (e.g., *Alfred Hitchcock's Notorious*) are omitted. In IMDb, the original title with the possessive is identified as the "complete title" in the "also known as" section. The other sources do not appear to record variants with possessives.

Significant variation was found in the practice of recording titles for non-English language films. IMDb prefers the original release title in the original language. AMG prefers an English title, but generally explicitly identifies the "original foreign title" in the AKA section. The other sources prefer the English release title and, although they generally give the original release title, do not explicitly identify it as such. Titles originally in non-Roman alphabets were transliterated inconsistently.

There are also variations in practices for date, many of which appear to be attributable to preferences for either the first public or the first U.S. release date. IMDb uses the first year of first public screening, which they define for television as "the broadcast year of the first episode" and for film as "either the year of general release or of a festival presentation if earlier" (http://www.imdb.com/help/show_leaf?titleformat). Some other resources, especially the more U.S.-centric ones, appear to use the date of the first U.S. screening. This occasionally leads to substantial discrepancies. IMDb lists the film *Cobra Verde* under 1987 whereas TCM lists it under 2007. TCM notes that the film was released in West Germany in 1987, but appears to base its primary date on the U.S. screening in New York in 2007 (http://www.tcm.com/tcmdb/title.jsp?stid=504438&category=Misc%20Notes). However, not all date variations can be explained in this manner. AMG and IMDb often vary by a year despite both claiming to use the year of first release.

The sites also vary in the types of information that they provide. All sources provide titles (but not necessarily original-release titles identified as such), dates, and directors. Original aspect ratio and language (particularly for English language works) are given less frequently. IMDb is the source that most consistently provides aspect ratio and language data. However, the original language data needs to be examined carefully as IMDb provides major and minor languages without distinguishing between the two. For the primarily English language film *The Godfather*, IMDb shows English, Italian, and Latin. For the film *Joyeux Noël*, which has a mixed soundtrack in English, French, and German, IMDB lists French, German, English, and Latin without revealing the relative importance of those languages. Major languages are more consistently identified on video packaging.

Director was the element with the least variation. It was given consistently in all sources except for the fact that IMDb includes uncredited directors. Uncredited roles are not included in the other sources and not generally included in library cataloging. However, as these are explicitly marked as "uncredited," this is unlikely to cause difficulties. Some variation was found in the way director's names were transliterated from non-Roman alphabets and none of the sources use the LC-ALA Romanization system.

Interestingly, all five sources gave Michael Epstein and Thomas Lennon as the directors of the documentary *The Battle Over Citizen Kane*. However, the title frames of this documentary list no director and credit Epstein and Lennon as producers (producers are not infrequently the primary creative force behind television productions). It is unclear why this was done unless the sources had access to uncredited attributions or chose to shoehorn Epstein and Lennon into the director area in order to convey their primacy.

Source	**Title**	**Director**	**Date**	**Aspect**	**Language**
AMG	Generally gives U.S. release title for non-English language films; usually includes original title as a cross-reference and identifies it as "original foreign title."		Uses first release year whether limited or wide release.	Usually not given; numerical ratios not given.	Not given.
IMDb	Give original release title for all works, including non-English language films; Romanized when necessary.	Includes uncredited directors marked as such. Performs authority control on credits.	Gives year of first public screening.	Usually provides numerical aspect ratio information.	Provides language information for non-silent films; gives both major and minor languages without distinction
InBaseline	Generally gives U.S. release title for non-English language films; usually includes original title as a cross-reference, but not explicitly identified as such.		Appears to give first U.S. release date.	Not given.	Not given.

Source	Title	Director	Date	Aspect	Language
TCM	Generally gives U.S. release title for non-English language films; usually includes original title as a cross-reference, but not explicitly identified as such.		Appears to give first U.S. release date.	Sometimes includes aspect ratio information under Misc Notes or Theatrical Aspect Ratio.	Often provided under Misc Notes" for films not originally in English.
Variety	Generally gives U.S. release title for non-English language films; usually includes original title as a cross-reference, but not explicitly identified as such.		Date taken from review data for review-derived data; project data appears to give first U.S. release date.	Usually not given; numerical ratios not given.	Not given.

Summary of Consistency of Data Extracted from MARC Records and External Sources

In the sections below, we compare the results of the data we extracted from MARC bibliographic records with the data from external sources. Total number of records for each title is given in the second column. The answer we used for evaluation is listed in the third column. The following columns contain the number of records with data deemed to be correct in the relevant MARC field. Records from which our algorithm was not able to derive the correct data from any of the examined MARC fields are counted in the "not found" column. Some records may be counted under more than one MARC field if correct data was identified in more than one field. We also briefly discuss the results for each element.

Title:

It is unclear what the correct original title is for *Andrei Rublev*. All the sources checked give some variant of *Andrei Rublev*, which could be attributed to variations in Romanization practice. It appears likely to have been released in the U.S. under the title Andrei Roublev. However, some videos have been issued with the title *Stastri po Andreiu*.on their title frames (search Andrei Rublev at http://www.shillpages.com/movies/aa.shtml). We have examined its occurrence in both forms below.

Also, one record provided the U.S. release title Amelie in the 130 field of the bib record despite the existence of a NAF record for Fabuleux destin d'Amélie Poulain (Motion picture) so even 130 fields are not always reliable sources of the original title.

In general, for originally English language titles, the correct title (or possibly the title with an initial possessive) appears in 245 $a (title proper). For originally non-English language titles, the original title usually appears somewhere, but its location is difficult to predict. The most reliable way to identify original titles, 130 uniform title fields, is not commonly used, particularly outside of the archive in our sample, even when a LC NAF uniform title exists. Due to the lack of ability to algorithmically identify the original title when 130 uniform

titles are not used, the numbers listed below, which only indicate where the data is present, are greater than the numbers that could be automatically extracted from the MARC records.

	Recs	**LC NAF uniform title**	**130 uniform title**	**245 title proper**	**245 $b parallel title**	**246 variant title**	**Not found**
Andrei Rublev (variations in spelling ignored)	8	Andreĭ Rublev (Motion picture)	0	4	1	4	0
Andrei Rublev (as Strasti po Andreiu)	8	Andreĭ Rublev (Motion picture)	0	3	0	0	5
The Battle Over Citizen Kane	2	N/A [Uniform titles are not made for individual television program episodes.]	0	2	0	0	0
Citizen Kane	22	Citizen Kane (Motion picture)	10	22	0	1	0
Dracula	16	Dracula (Motion picture : 1931)	11 (plus 2 without motion picture qualifier)	16	0	2 records include “Carl Laemmle presents Dracula”	0
Le Fabuleux destin d’Amélie Poulain	8	Fabuleux destin d’Amélie Poulain (Motion picture)	2	5	0	2	1
Invasion of the Body Snatchers (1956)	4	Invasion of the body snatchers (Motion picture : 1956)	2	4	0	0	0
Invasion of the Body Snatchers (1978)	3	N/A	2	3	0	0	0
Miyamoto Musashi kanketsuhen: kettô Ganryûjima (Samurai III)	5	N/A	0	2	0	0	3
A Night at the Opera	5	Night at the opera (Motion picture)	0	5	0	0	0
Shichinin no samurai	10	Shichinin no samurai (Motion picture)	0	2	0	0	8

Director:

Foreign language directors sometimes not found due to the fact that our program was not designed to deal with transcribed credits in languages other than English. There were some inexplicable non-matches. For *Andrei Rublev,* T. Ogorodnikova was listed in English as the director in the credits of two records. Subsidiary "director" functions (such as "director of photography") were picked up in some records.

The situation where the people listed as directors of *The Battle Over Citizen Kane* in secondary sources, but listed as producer on the title frames was discussed in the introduction.

	Recs	**Director**	**245 $c statement of responsibility**	**508 creation/ production credits**	**700 $4 relator code**	**700 $e relator term**	**Not found**
Andrei Rublev	8	Tarkovskiĭ, Andreĭ Arsen'evich, 1932-1986	0	0	3	3	2
The Battle Over Citizen Kane	2	Epstein, Michael, film director + Lennon, Thomas	0	0	0	0	2
Citizen Kane	22	Welles, Orson, 1915-1985	18	2	4	4	2
Dracula	16	Browning, Tod, 1882-1962	12	4	2	2	
Le Fabuleux destin d'Amélie Poulain	8	Jeunet, Jean-Pierre, 1955-	7	0	1	1	1
Invasion of the Body Snatchers (1956)	4	Siegel, Don, 1912-1991	4	0	2	0	0
Invasion of the Body Snatchers (1978)	3	Kaufman, Philip, 1936-	3	0	1	0	0
Miyamoto Musashi kanketsuhen: kettô Ganryûjima (Samurai III)	5	Inagaki, Hiroshi, 1905-1980	4	0	1	1	0
A Night at the Opera	5	Wood, Sam, 1883-1949	5	0	0	1	0
Shichinin no samurai	10	Kurosawa, Akira, 1910-1998	9	0	1	6	0

Date:

Some records attributed dates to the work that were close, but not the same as the dates listed in our secondary source. This may be due to different dates appearing on the packaging of the items in hand.

Some records include multiple dates in notes that we examined as potentially referring to the original date of the work. If the original date is not the earliest date mentioned, it is impossible for us to programmatically identify the correct date.

	Recs	**Original date**	**008 Date2**	**033 $a date of event**	**130 $a uniform title**	**500 general note**	**Not found**
Andrei Rublev	8	1966 (Variety.com gives 1973)	2 (+ 3 records with incorrect dates in Date2)	0	0	5 (3 records include 1965, 1966, 1969, 1971, 1979, etc. in 500 note and are not counted as correct due the earliest date not being the correct one)	3 that include the correct date, but not algorithmically identifiable due to earlier (1965) date also appearing in a 500 note and do not contain the correct date in another field.
The Battle Over Citizen Kane	2	1996	0	1	0	1	0
Citizen Kane	22	1941	15 (+1 record with incorrect date in Date2)	0	0	18 (5 records include1996 in 500 note; + 4 records with only date of 1942 in a 500 note)	1
Dracula	16	1931	8	0	14	14 (2 records additionally include 1940 in a 500 note)	0

	Recs	Original date	008 Date2	033 $a date of event	130 $a uniform title	500 general note	Not found
Le Fabuleux destin d'Amélie Poulain	8	2001	8	0	1	7 (+1 record with only date of 2005 in a 500 note)	0
Invasion of the Body Snatchers (1956)	4	1956	0 (+1 record with incorrect date in Date2)	0	2	2 (+1 record with only date of 1955 in a 500 note)	1
Invasion of the Body Snatchers (1978)	3	1978	3	0	2	3	0
Miyamoto Musashi kanketsuhen: kettô Ganryûjima (Samurai III)	5	1956 (TCM gives 1957)	2	0	0	4 (+1 record with only date of 2005 in a 500 note)	0
A Night at the Opera	5	1935	3	0	0	4 (all records additionally include 1961 in a 500 note)	0
Shichinin no samurai	10	1954	5	0	0	4 (+2 records with only 1991, 2002, 2007 in a 500 note)	4

Language:

The original language is often, but not always, in the MARC record. However, in some situations it is not safe for a computer to infer the original language. For example, a copy of Andrei Rublev might have its language information coded as "$a rus $a eng $b eng" where $a gives the soundtracks and $b the subtitles. Although it is more likely that the first language listed (which is also the language in 008) is the original language, it is not always true.

A substantial number of records did not include unambiguous data about the original language. This could be fixed with more consistent, explicit coding practice.

	Recs	Language	008 language code	041 $h original/ intermediate language	Not unambiguously identifiable
Andrei Rublev	8	Russian (IMDB also notes Italian and Tatar sequences)	0	0	8
The Battle Over Citizen Kane	2	English	1	0	1
Citizen Kane	22	English	15	1	6
Dracula	16	English (IMDB also notes Hungarian and Latin sequences)	3	3	10
Le Fabuleux destin d'Amélie Poulain	8	French	0	2	6
Invasion of the Body Snatchers (1956)	4	English	1	1	2
Invasion of the Body Snatchers (1978)	3	English	1	1	1
Miyamoto Musashi kanketsuhen: kettô Ganryûjima (Samurai III)	5	N/A	0	0	5
A Night at the Opera	5	English (IMDB also notes Italian sequences)	1	1	3
Shichinin no samurai	10	Japanese	0	3	7

Aspect ratio:

We listed the total with the right term (some variation on widescreen or full screen) in each MARC field column and then gave the number with the right numerical aspect ratio in parentheses. There is a tendency not to note full screen presentations. Some ratios that IMDb showed as 1.37:1 were listed in the MARC records as 1.33:1.

	Recs	**IMDB aspect ratio**	**250 edition statement**	**500 general note**	**505 contents note**	**538 system details**	**Not appearing**
Andrei Rublev	8	Widescreen (2.35:1)	4 (0)	1 (1)	0	0	3
The Battle Over Citizen Kane	2	Widescreen (1.85:1)	0	0	0	0	2
Citizen Kane	22	Full screen (1.37:1)	0	0	0	0	22
Dracula	16	Full screen (1.37:1)	1 (0)	1 (says 1.33:1)	0	3 (all 3 say 1.33:1)	12
Le Fabuleux destin d'Amélie Poulain	8	Widescreen (2.35:1)	5 (1)	3 (2)	0	0	0
Invasion of the Body Snatchers (1956)	4	Widescreen (2.35:1) (TCM gives 2.00:1)	1 (1)	1 (1)	2 (0)	1	0
Invasion of the Body Snatchers (1978)	3	Widescreen (1.85:1)	1 (1)	0	1 (0)	0	1
Miyamoto Musashi kanketsuhen: kettô Ganryûjima (Samurai III)	5	Unknown (no data was found in the MARC records, either)	0	0	0	0	5
A Night at the Opera	5	Full screen (1.37:1)	0	1 (0)	0	1(0)	3
Shichinin no samurai	10	Full screen (1.37:1)	0	0	0	1 (says 1.33:1)	9

Summary

Most of the data extracted from MARC records was accurate when compared to data in widely available and reasonably reliable online secondary sources. The more common problem was the lack of extractable data in the MARC records. Not all the data in secondary sources was consistent, but most of it was, and the inconsistencies can largely be explained by varying rules for data determination. We did not examine any multi-work manifestations, but data extracted from these is likely to be less reliable as the data about each work cannot be algorithmically disentangled. There are also likely to be some problems matching title and names originally in

non-Roman alphabets due to varying practices in Romanizing and recording these. For some elements, particularly title, it remains impossible for a computer to pick out the original title in most cases even when the original title is present in the MARC record. It would be helpful for future re-use of this type of data if more of it were recorded explicitly and unambiguously in the MARC record in a way that is amenable to machine extraction.