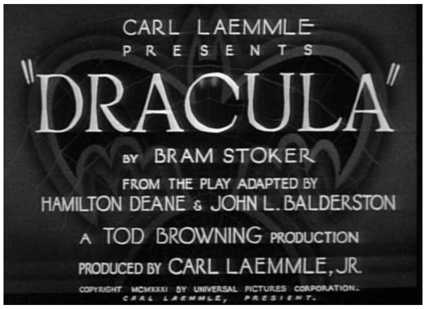
Finding Movies with FRBR & Facets

Kelley McGrath
University of Oregon
kelleym@uoregon.edu
OLAC CAPC meeting, June 22, 2012

Presentation at the meeting of the OLAC Cataloging Policy Committee (CAPC)
Anaheim, CA, June 22, 2012
Kelley McGrath

This presentation gives an overview of a project that I've been involved in that uses the Functional Requirements for Bibliographic Records (FRBR) model and faceted navigation to try to make it easier to find videos in libraries.

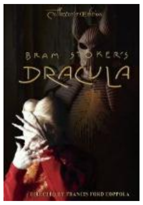
Users are looking for movies

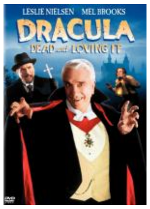

CARL LAEMMLE
PRESENTS
"DRACULA"
BY BRAM STOKER
FROM THE PLAY ADAPTED BY
HAMILTON DEANE & JOHN L. BALDERSTON
A TOD BROWNING PRODUCTION
PRODUCED BY CARL LAEMMLE, JR.
COPYRIGHT MCMXXXI BY UNIVERSAL PICTURES CORPORATION.
CARL LAEMMLE, PRESIDENT.

I'm going to use "movies" as shorthand for all the sorts of things that FRBR would call moving image works, including TV programs and other kinds of film and video.

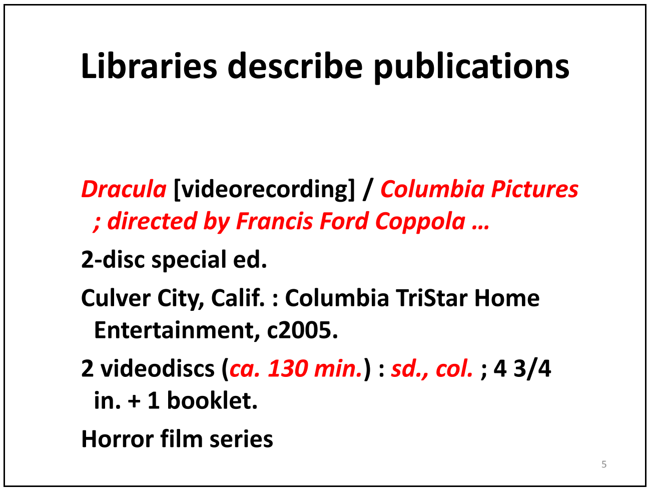
The typical person who comes to a library looking for a video is really looking for a movie, either a specific movie, such as the 1931 English-language Dracula with Bela Lugosi, or a category of movies, such as early German horror films, movies in the Friday the 13th series or documentaries on the history of horror films.

Libraries describe publications





Libraries, however, catalog publications (i.e., manifestations). For example, a catalog record might represent the set of DVDs issued by Warner Bros. Home Entertainment beginning in 2008 and with a particular ISBN.



Libraries describe publications

Dracula [videorecording] / *Columbia Pictures ; directed by Francis Ford Coppola …*

2-disc special ed.

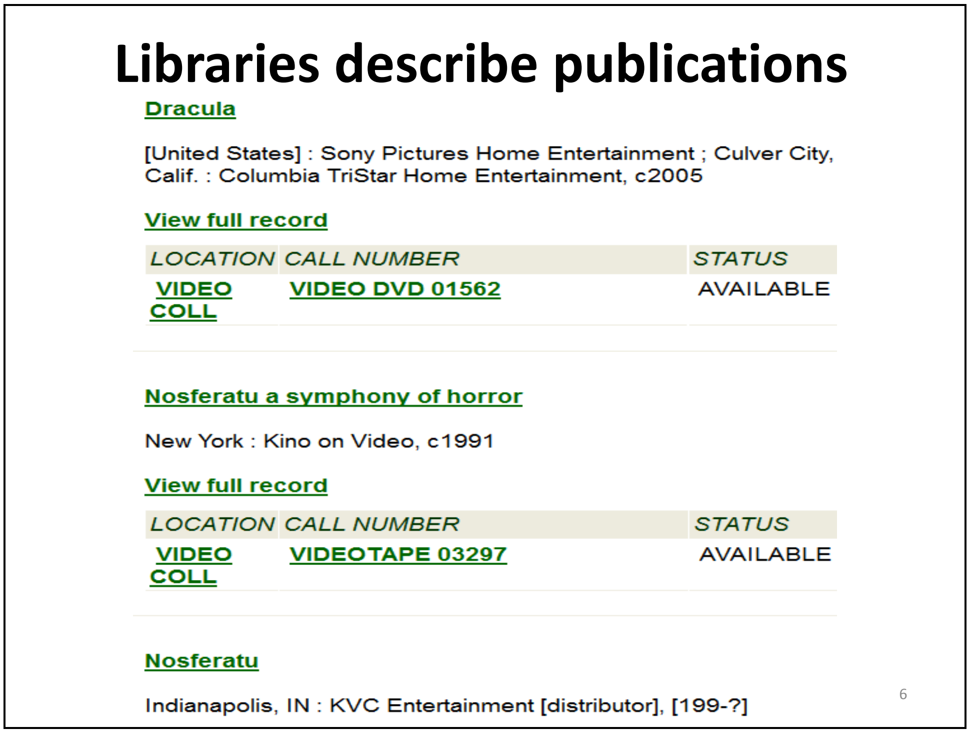
Culver City, Calif. : Columbia TriStar Home Entertainment, c2005.

2 videodiscs (*ca. 130 min.*) : *sd., col.* ; 4 3/4 in. + 1 booklet.

Horror film series

5

The heart of the description in a catalog record reflects this emphasis on publications. The parts in red italics could possibly be interpreted as describing the movie, but everything else here clearly describes the publication—the edition statement, the publication statement, much of the physical description and the series statement. Even the information in red italics is mapped by RDA to transcribed information on the manifestation or to the expression. Most of the information about the movie is in notes and much of it is not even required by cataloging rules. The associated names are in 700 fields at the bottom of the record. The information about the movie and the publication is jumbled up so it's hard to pick out the data elements that relate to the movie.



Libraries describe publications

Dracula

[United States] : Sony Pictures Home Entertainment ; Culver City, Calif. : Columbia TriStar Home Entertainment, c2005

View full record

LOCATION	CALL NUMBER	STATUS
VIDEO COLL	VIDEO DVD 01562	AVAILABLE

Nosferatu a symphony of horror

New York : Kino on Video, c1991

View full record

LOCATION	CALL NUMBER	STATUS
VIDEO COLL	VIDEOTAPE 03297	AVAILABLE

Nosferatu

Indianapolis, IN : KVC Entertainment [distributor], [199-?]

6

Library hit lists also reflect this emphasis on publications. The information that is given to help identify a title usually describes the publication, as in this example. The date of publication, which is the date that displays here and is the one used for sorting and limiting in library catalogs, goes with the DVD and not the original movie. Because the hits are for publications, there can be more than one for the same movie. Can you tell which Dracula movie the first hit is describing or if the different hits represent the same or different movies?

Users care about versions

Rg1024: open clip art library

People do care about what I think of as versions. They have preferences or requirements as to how they want to access a particular movie. If I don't have a Blu-ray player, it does me no good to borrow a Blu-ray disc. If I just bought a Blu-ray player, maybe all I want to look at are Blu-rays so I can try one. If I don't speak Japanese, I don't want to borrow a video in Japanese with no English subtitles. Maybe I only want to see the director's cut or the unrated version of a movie.

Prototype: Movies & Versions

Funded: OLAC (Online Audiovisual Catalogers)

Developed by Chris Fitzpatrick

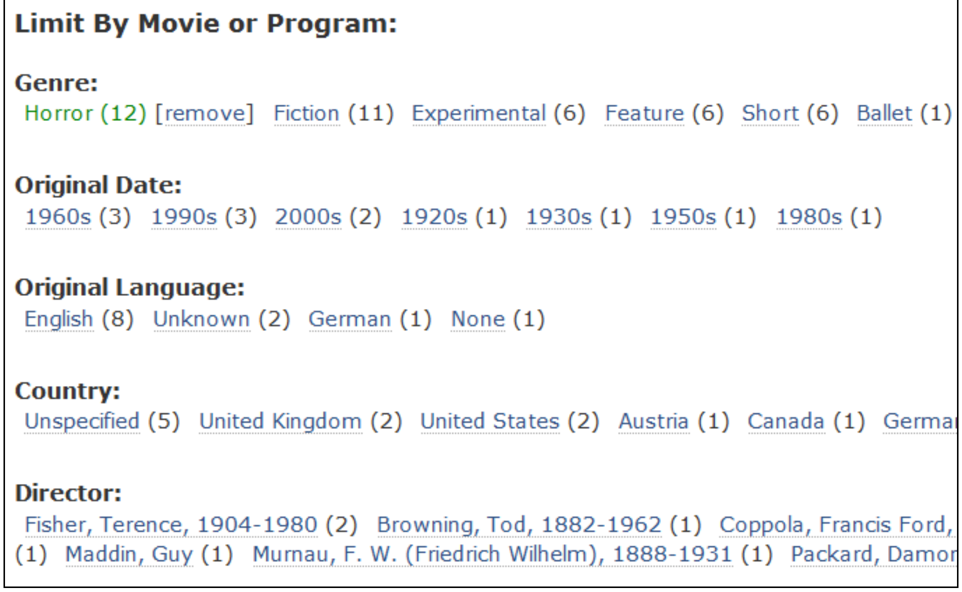
Small scale (limited data, few fields and records, simplified data model)

http://blazing-sunset-24.heroku.com

We built a prototype discovery interface that focuses on movies and versions rather than publications to experiment with what that might look like.

http://blazing-sunset-24.heroku.com

Movie (mostly work) facets



Limit By Movie or Program:

Genre:
Horror (12) [remove] Fiction (11) Experimental (6) Feature (6) Short (6) Ballet (1)

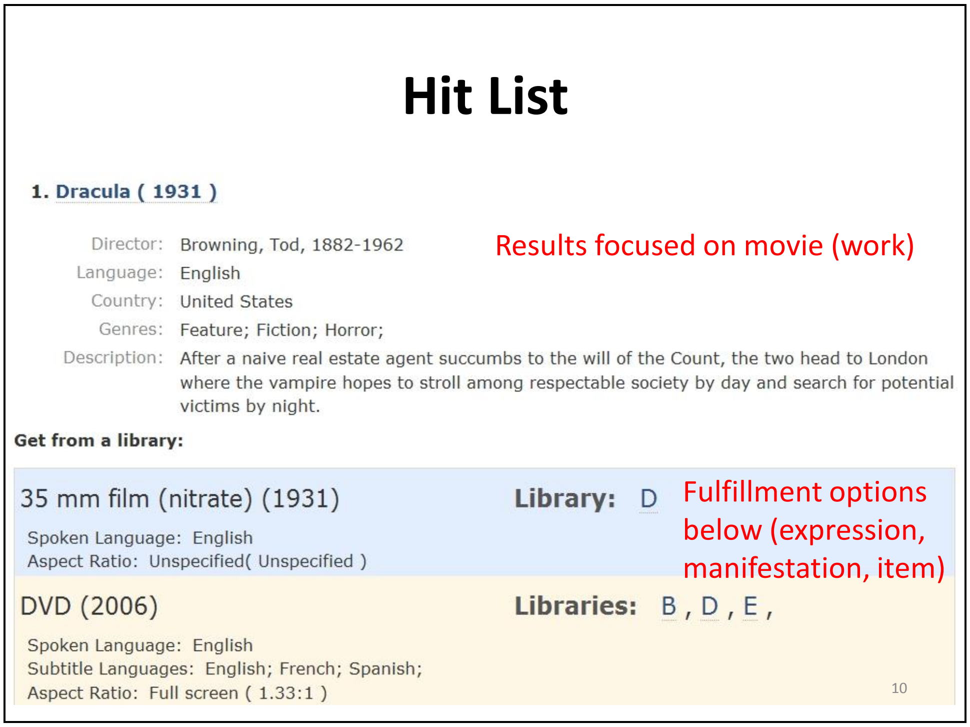
Original Date:
1960s (3) 1990s (3) 2000s (2) 1920s (1) 1930s (1) 1950s (1) 1980s (1)

Original Language:
English (8) Unknown (2) German (1) None (1)

Country:
Unspecified (5) United Kingdom (2) United States (2) Austria (1) Canada (1) Germa

Director:
Fisher, Terence, 1904-1980 (2) Browning, Tod, 1882-1962 (1) Coppola, Francis Ford, (1) Maddin, Guy (1) Murnau, F. W. (Friedrich Wilhelm), 1888-1931 (1) Packard, Damor

In addition to a search box, the prototype UI provides facets for important attributes of movies like genre and original date. This supports browsing of movies from many angles.



Hit List

1. Dracula (1931)

Results focused on movie (work)

Director: Browning, Tod, 1882-1962
Language: English
Country: United States
Genres: Feature; Fiction; Horror;
Description: After a naive real estate agent succumbs to the will of the Count, the two head to London where the vampire hopes to stroll among respectable society by day and search for potential victims by night.

Get from a library:

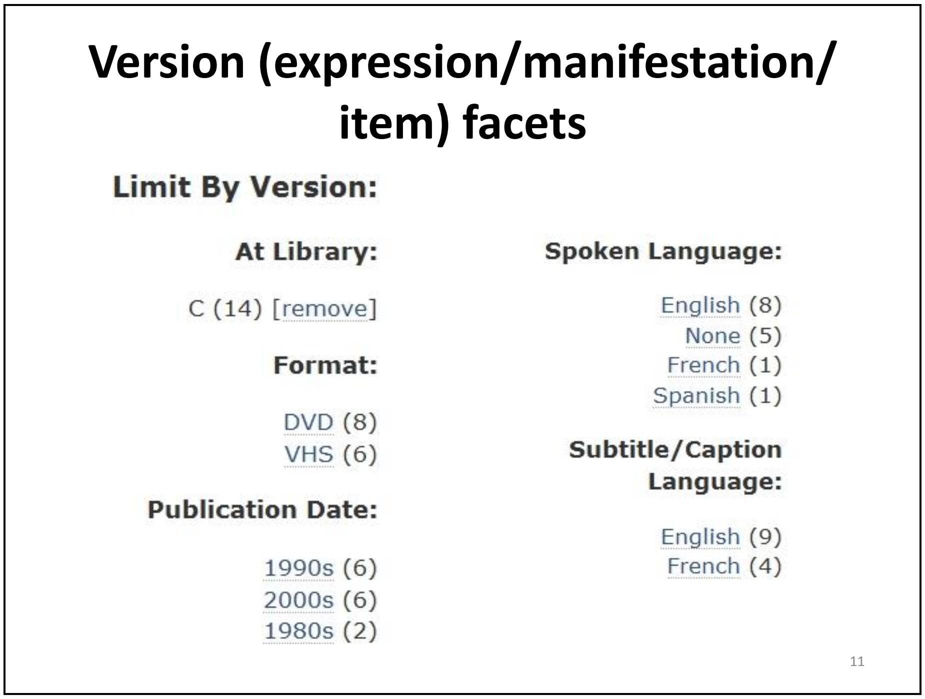
Fulfillment options below (expression, manifestation, item)

35 mm film (nitrate) (1931) — **Library:** D
Spoken Language: English
Aspect Ratio: Unspecified(Unspecified)

DVD (2006) — **Libraries:** B , D , E ,
Spoken Language: English
Subtitle Languages: English; French; Spanish;
Aspect Ratio: Full screen (1.33:1)

10

Our hit list features only one hit per movie and includes enough information to identify the movie. We also clearly present version information that is important for decision-making to enable easy selection.



Version (expression/manifestation/ item) facets

Limit By Version:

At Library:
C (14) [remove]

Format:
DVD (8)
VHS (6)

Publication Date:
1990s (6)
2000s (6)
1980s (2)

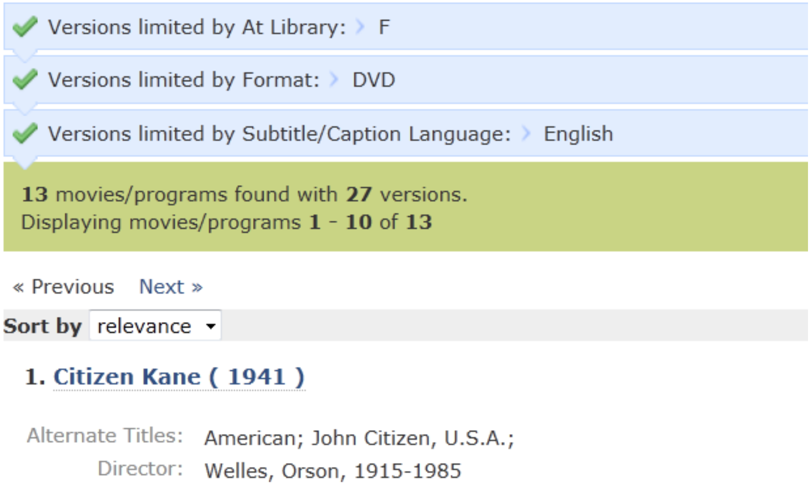
Spoken Language:
English (8)
None (5)
French (1)
Spanish (1)

Subtitle/Caption Language:
English (9)
French (4)

11

We include version-related facets, such as format and soundtrack and subtitle options.

Movie-version interaction



✓ Versions limited by At Library: › F

✓ Versions limited by Format: › DVD

✓ Versions limited by Subtitle/Caption Language: › English

13 movies/programs found with **27** versions.
Displaying movies/programs **1** - **10** of **13**

« Previous Next »

Sort by relevance ▾

1. Citizen Kane (1941)

Alternate Titles: American; John Citizen, U.S.A.;
Director: Welles, Orson, 1915-1985

The records and facets for movies and versions interact. This allows users to explore from the top down, say starting with horror films, or from the bottom up with Blu-rays at their local library. With each selection, the resulting movies and versions are appropriately narrowed. This was the part of the prototype that was most technically challenging to implement. Our approach contrasts with most FRBR implementations, which are designed around the scenario where users start at the top by picking a specific work first and then examining their options.

Prototype

Prototype **http://blazing-sunset-24.heroku.com**

Sample searches and use cases **http://blazing-sunset-24.heroku.com/page/samples**

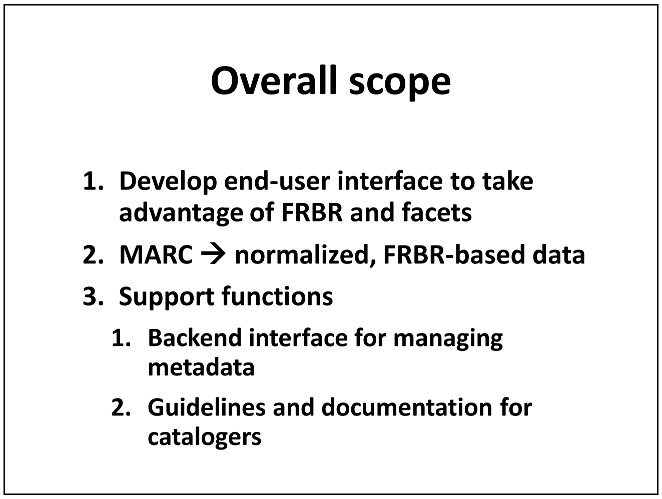
Code **http://github.com/cfitz/olac**

Because there isn't a lot of data in the prototype, I recommend checking out the sample searches to get a fuller sense of the possibilities. We're hoping to build on the prototype and eventually create a functional system.

Prototype: http://blazing-sunset-24.heroku.com

Sample searches: http://blazing-sunset-24.heroku.com/page/samples

Code: http://github.com/cfitz/olac



Overall scope

1. **Develop end-user interface to take advantage of FRBR and facets**
2. **MARC → normalized, FRBR-based data**
3. **Support functions**
   1. **Backend interface for managing metadata**
   2. **Guidelines and documentation for catalogers**

The overall project has three main elements.

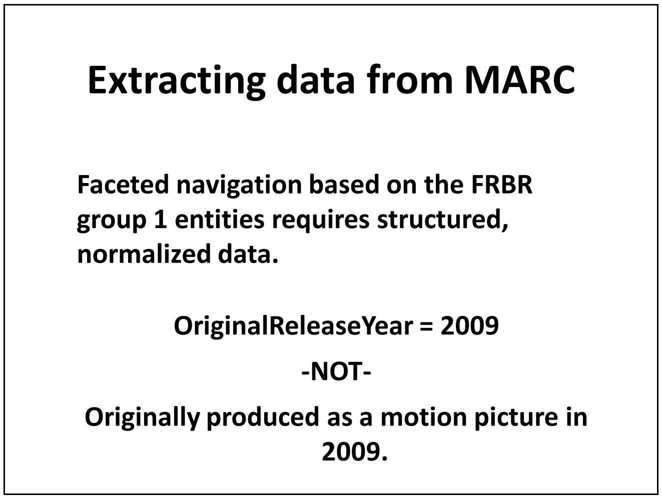
1. Above, I talked about the development of a new discovery interface that would work better for users. To achieve this end, we propose to

- Separate information about the movie from information about the versions and publications
- Present movie-centric results with the ability to easily select a specific version or publication
- Use structured data for types of information that are important to users to enable faceted navigation and support browsing

2. However, for numerous reasons, our existing MARC records can't support this type of interface. We want to maximize the use of existing data and get to a place where we have data that will work as quickly as possible. To do this, we plan to extract structured values from existing MARC data where possible and map those values to the appropriate FRBR entities. We will then perform selective quality control and data enhancement. We will also be able to clearly identify gaps in our data.

3. Finally, since transforming data from MARC is inherently inefficient and error-prone, going forward we plan to develop a backend interface for inputting or importing metadata and then managing it. In order to achieve reliable results, we will need documentation and guidelines for those working with the metadata.

I'm now going to talk a little bit about these aspects of the project, starting with the data extraction.



Extracting data from MARC

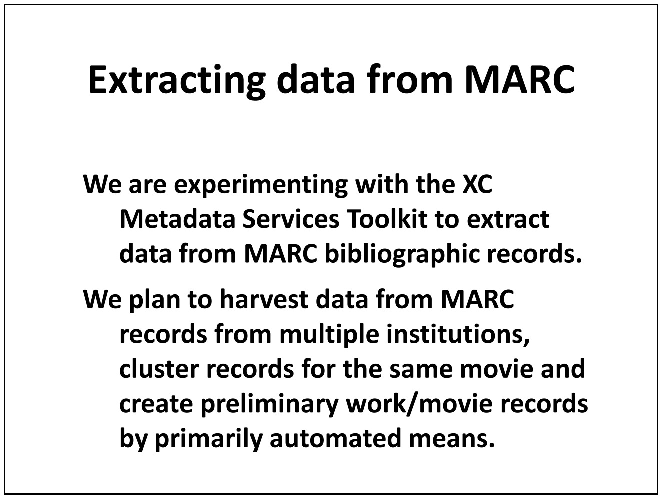
Faceted navigation based on the FRBR group 1 entities requires structured, normalized data.

OriginalReleaseYear = 2009

-NOT-

Originally produced as a motion picture in 2009.

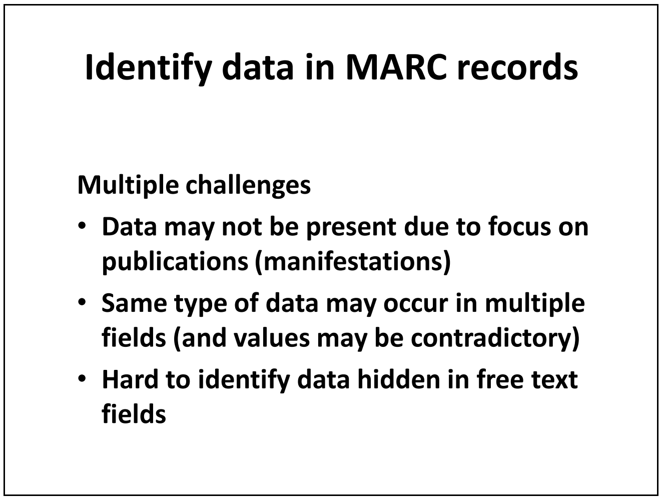
The data in our existing MARC records usually looks more like the note on the bottom: "Originally released as a motion picture in 2011." You can't use that as a facet. For one thing, it's too long. Plus all the notes about 2009 movies won't use the same wording so they won't collocate into the same facet. We're trying to map that kind of note into the simpler statement shown above. In the example shown, the pattern is obvious, but there are many, many patterns and outliers.



Extracting data from MARC

We are experimenting with the XC Metadata Services Toolkit to extract data from MARC bibliographic records.

We plan to harvest data from MARC records from multiple institutions, cluster records for the same movie and create preliminary work/movie records by primarily automated means.



Identify data in MARC records

Multiple challenges

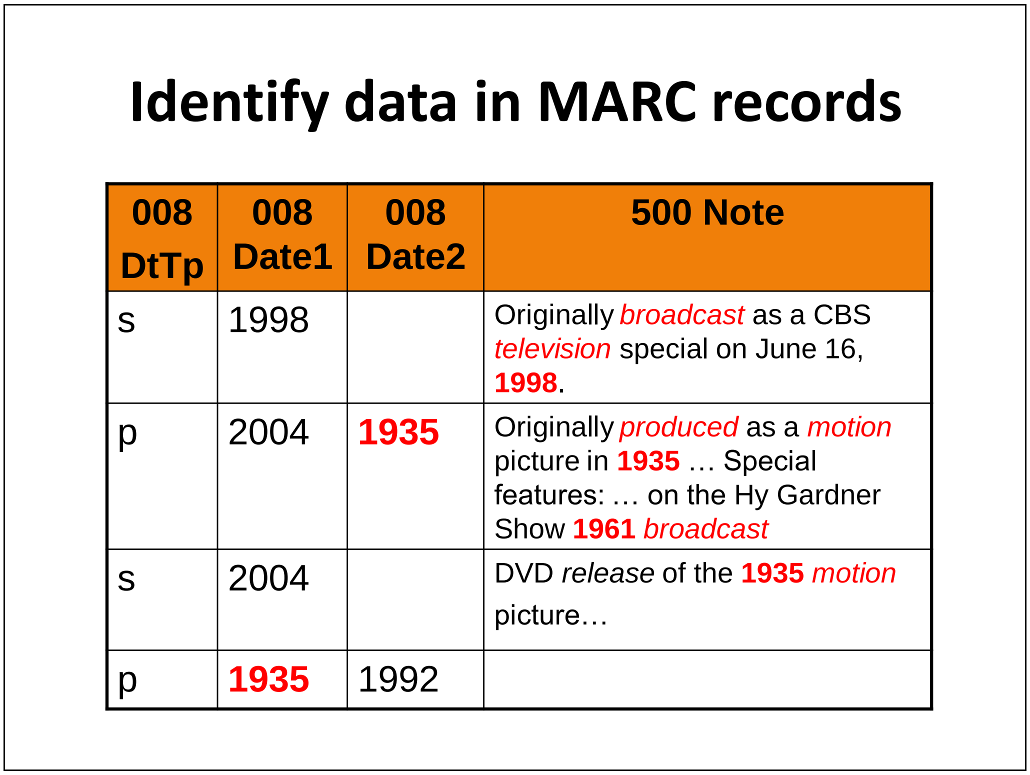
- **Data may not be present due to focus on publications (manifestations)**
- **Same type of data may occur in multiple fields (and values may be contradictory)**
- **Hard to identify data hidden in free text fields**

There are a number of reasons our attempts to identify an original release date might fail.

Cataloging rules focus on manifestations and don't require the original date so it might not even be in the record. Of course, it might also be the case that the original date is unknown.

Information about the original date can occur in more than one place in a record and those sources might contradict each other.

There are also many free text notes where it is difficult or impossible to teach a computer how to spot the original date.



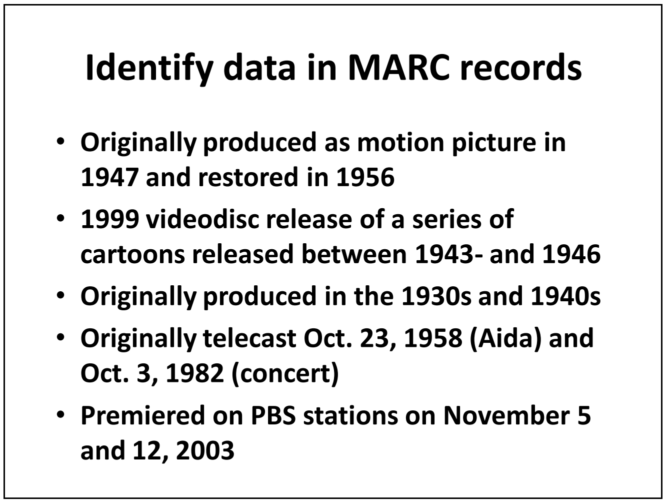
Identify data in MARC records

008 DtTp	008 Date1	008 Date2	500 Note
s	1998		Originally *broadcast* as a CBS *television* special on June 16, **1998**.
p	2004	**1935**	Originally *produced* as a *motion* picture in **1935** … Special features: … on the Hy Gardner Show **1961** *broadcast*
s	2004		DVD *release* of the **1935** *motion* picture…
p	**1935**	1992	

The two primary challenges for identifying and extracting data about movies from MARC bibliographic records can be seen in this table of selected strategies for identifying the original date of a movie.

- There are often multiple potential sources for the same data element in MARC records. The original date of a movie is often found in 008 Date2. Although not correct MARC, this date may also appear in 008 Date1. The most common place to find the original date is in a free text note, usually tagged 500. There are other, less common, locations where original date can be found. More than one of these options may occur in the same record and the different sources may contain contradictory data.

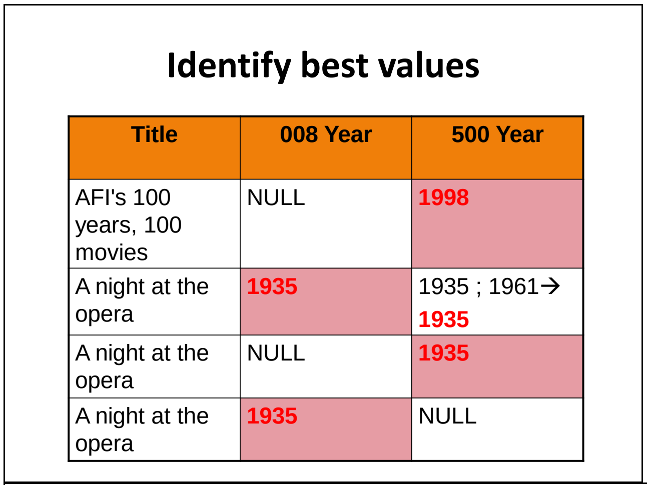
- Much data is in free text fields where it may be possible to develop heuristics to extract normalized values, but accuracy will inevitably be less that 100%. Since 500 fields are used for other purposes, in this example we have looked for years (18xx, 19xx or 20xx) in combination with a keyword that suggests that it is a note about the original date, such as broadcast, motion or produced.



Identify data in MARC records

- **Originally produced as motion picture in 1947 and restored in 1956**
- **1999 videodisc release of a series of cartoons released between 1943- and 1946**
- **Originally produced in the 1930s and 1940s**
- **Originally telecast Oct. 23, 1958 (Aida) and Oct. 3, 1982 (concert)**
- **Premiered on PBS stations on November 5 and 12, 2003**

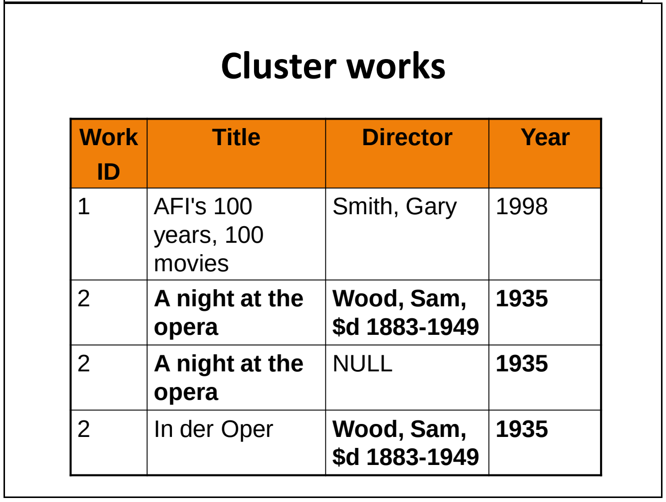
Although the data given in the table on the previous slide is largely straightforward to process, a great many more complicated variations exist in the wild. These are all real examples.



Identify best values

Title	008 Year	500 Year
AFI's 100 years, 100 movies	NULL	**1998**
A night at the opera	**1935**	1935 ; 1961→ **1935**
A night at the opera	NULL	**1935**
A night at the opera	**1935**	NULL

Once we have extracted various potential values for a specific data element, we propose to rank the possibilities in order to obtain a single likeliest value from each record. Here we have considered the 008 date to be more reliable than the 500 note and when there are multiple dates in 500, we have chosen the earliest one. Both of these rules can lead to incorrect conclusions, but they work more often than not.

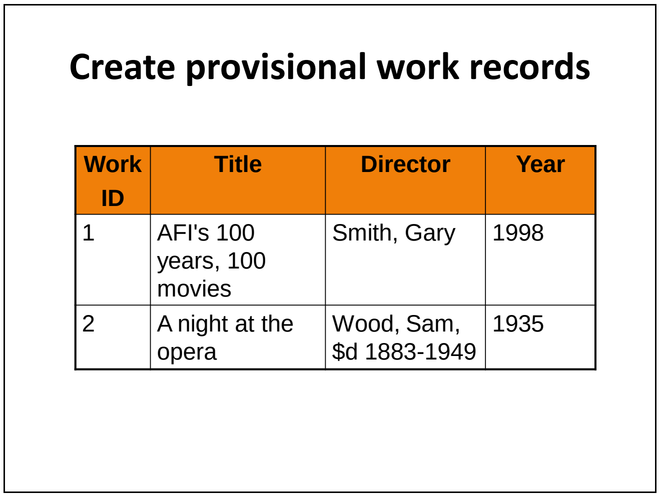


Cluster works

Work ID	Title	Director	Year
1	AFI's 100 years, 100 movies	Smith, Gary	1998
2	**A night at the opera**	**Wood, Sam, $d 1883-1949**	**1935**
2	**A night at the opera**	NULL	**1935**
2	In der Oper	**Wood, Sam, $d 1883-1949**	**1935**

We then propose to cluster records describing the same work/movie based on matching data. Title, director and original year are examples of good candidate data elements for clustering.

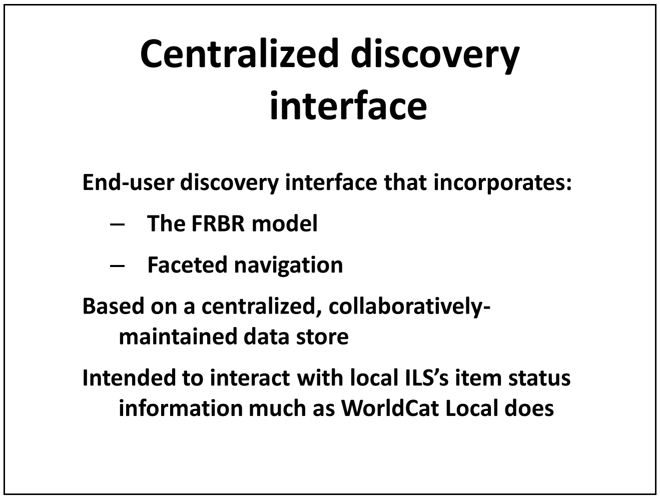
An alternative approach would be to take these core data elements and match them against an external service such as Freebase. We could then cluster on the IDs provided by the external service. This might be more effective if done with all the possible variations of extracted data elements prior to picking a best value as was done above. A drawback of this approach would be the incomplete overlap between works/movies held by libraries and those described by an external service.



Create provisional work records

Work ID	Title	Director	Year
1	AFI's 100 years, 100 movies	Smith, Gary	1998
2	A night at the opera	Wood, Sam, $d 1883-1949	1935

The clustered data would be used to populate provisional movie/work records. Provisional movie/work records could be selectively targeted for manual clean up.



Centralized discovery interface

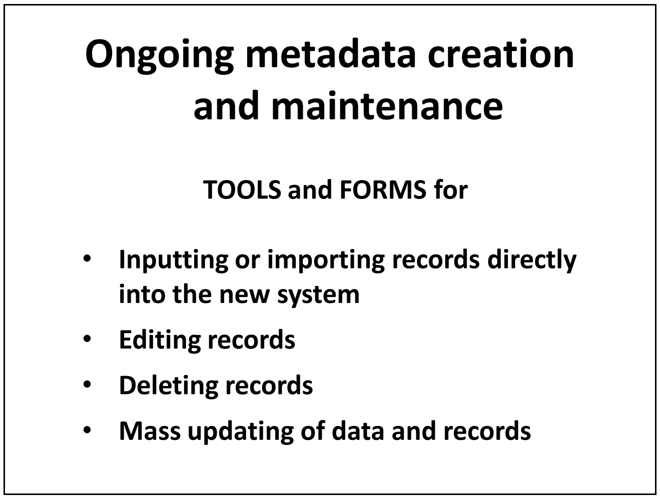
End-user discovery interface that incorporates:

- The FRBR model
- Faceted navigation

Based on a centralized, collaboratively-maintained data store

Intended to interact with local ILS's item status information much as WorldCat Local does

As described earlier, we plan to build an end-user discovery interface that uses the FRBR model and faceted navigation to support the needs of users looking for moving image materials in libraries. Our discovery interface will be based on a centrally-maintained data store initially derived from records supplied by participating institutions. We intend for the discovery interface to interact with local ILS's item status information much as WorldCat Local does. We hope to be able to support local views of this interface for institutions and consortiums and to be able to make the underlying data available for download or via API.

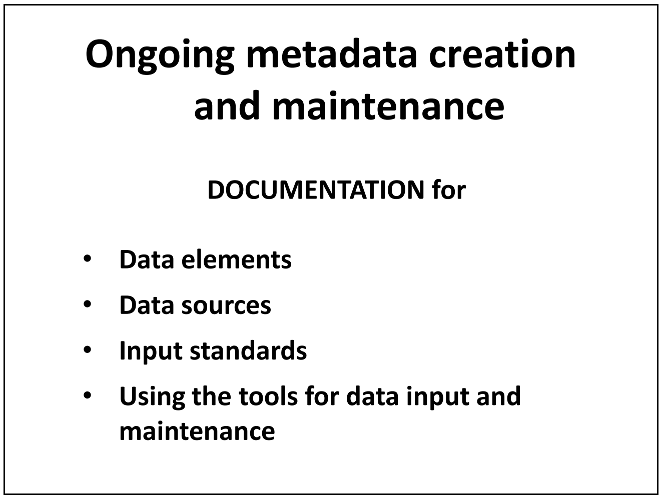


Ongoing metadata creation and maintenance

TOOLS and FORMS for

- Inputting or importing records directly into the new system
- Editing records
- Deleting records
- Mass updating of data and records

We need tools for performing the functions described on this slide. Although in the short run it will be necessary to extract data from existing MARC bibliographic records, this is an inefficient and error-prone process. Since it is easier to derive human-comprehensible text from controlled values than vice versa, a more viable approach would be to enter and maintain data about movies and versions directly in a new system and derive MARC records from that.



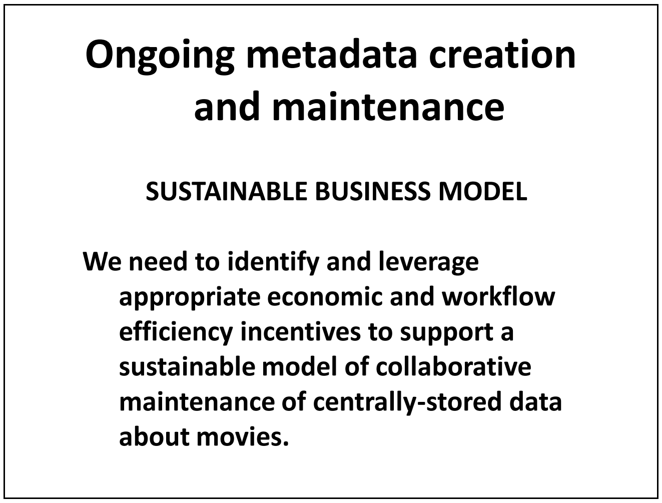
Ongoing metadata creation and maintenance

DOCUMENTATION for

- Data elements
- Data sources
- Input standards
- Using the tools for data input and maintenance

In order to remove redundancy and improve quality, movie records must be centrally stored and maintained, as the Library of Congress/NACO Authority File is today. These movie records must be editable by a broad community so it will be essential to have easily-understood definitions of the data elements and other content editing guidelines, as well as input forms that support accurate, efficient metadata entry and import.

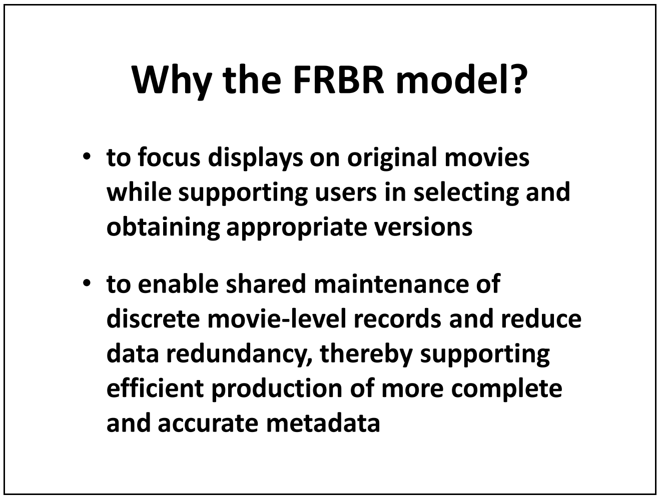
Some initial work on the first two topics has been done. See parts 1-3 at http://www.olacinc.org/drupal/?q=node/27



Ongoing metadata creation and maintenance

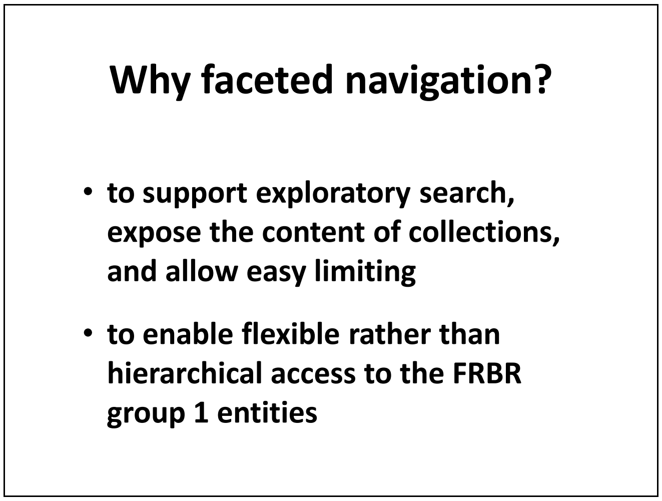
SUSTAINABLE BUSINESS MODEL

We need to identify and leverage appropriate economic and workflow efficiency incentives to support a sustainable model of collaborative maintenance of centrally-stored data about movies.



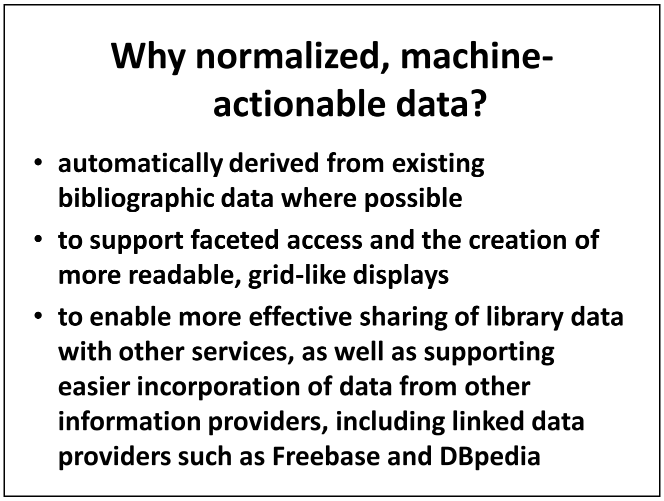
Why the FRBR model?

- to focus displays on original movies while supporting users in selecting and obtaining appropriate versions
- to enable shared maintenance of discrete movie-level records and reduce data redundancy, thereby supporting efficient production of more complete and accurate metadata



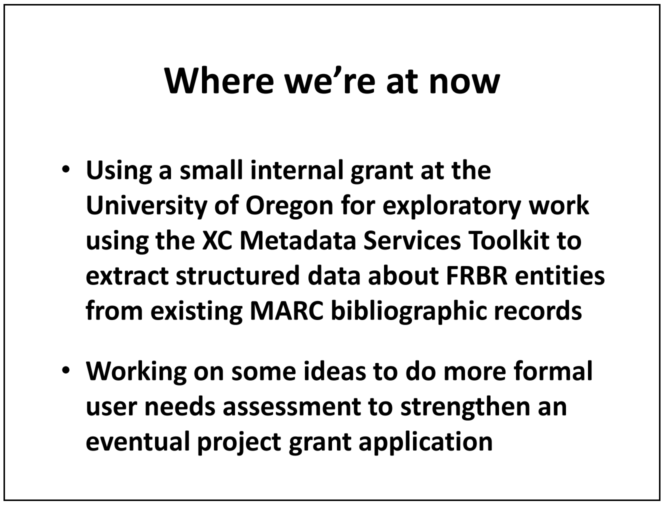
Why faceted navigation?

- to support exploratory search, expose the content of collections, and allow easy limiting
- to enable flexible rather than hierarchical access to the FRBR group 1 entities



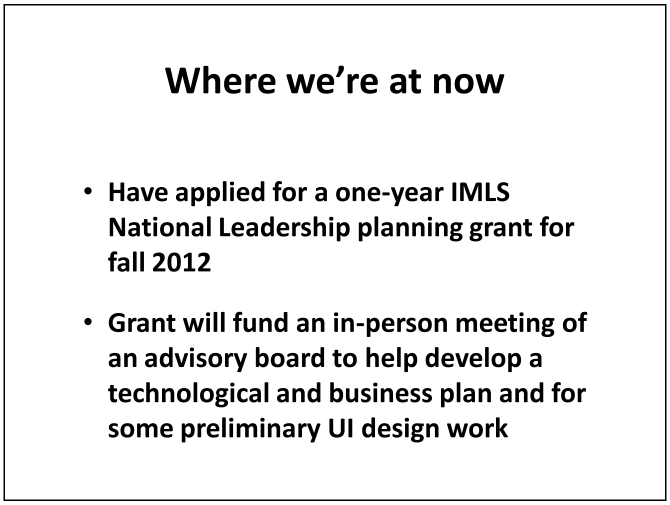
Why normalized, machine-actionable data?

- automatically derived from existing bibliographic data where possible
- to support faceted access and the creation of more readable, grid-like displays
- to enable more effective sharing of library data with other services, as well as supporting easier incorporation of data from other information providers, including linked data providers such as Freebase and DBpedia



Where we're at now

- Using a small internal grant at the University of Oregon for exploratory work using the XC Metadata Services Toolkit to extract structured data about FRBR entities from existing MARC bibliographic records
- Working on some ideas to do more formal user needs assessment to strengthen an eventual project grant application



Where we're at now

- Have applied for a one-year IMLS National Leadership planning grant for fall 2012
- Grant will fund an in-person meeting of an advisory board to help develop a technological and business plan and for some preliminary UI design work

More info

OLAC Moving Image Work-Level Records Task Force Reports
http://www.olacinc.org/drupal/?q=node/27

OLAC discussion group (includes lit review)
http://www.olacinc.org/drupal/?q=node/434

McGrath & Bisko. "Identifying FRBR Work-Level Data in MARC Bibliographic Records for Manifestations of Moving Images"
http://journal.code4lib.org/articles/775

McGrath, Kules, and Fitzpatrick "FRBR and Facets Provide Flexible, Work-Centric Access to Items in Library Collections"
http://pages.uoregon.edu/kelleym/publications/JCDL_OLAC_FRBR_prototype.pdf

Upcoming presentation

Workshop at OLAC conference in Albuquerque in October:
http://olac2012.weebly.com

Want to get involved?

Contact:
Kelley McGrath
Metadata Management Librarian
University of Oregon Libraries
kelleym@uoregon.edu