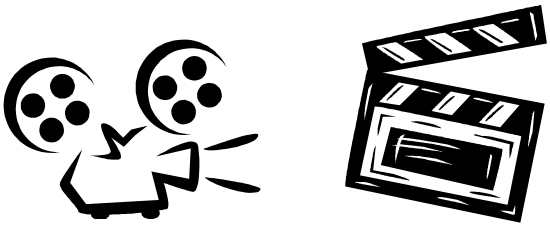
FRBR for Movies and Finding FRBR in MARC

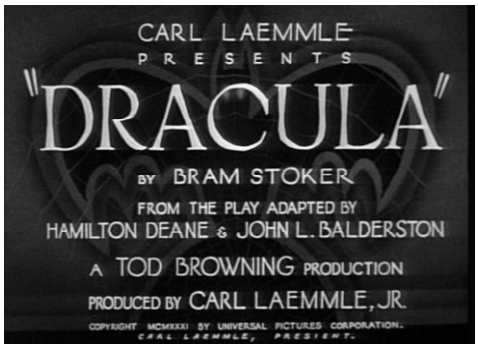
OLAC meeting

Kelley McGrath
University of Oregon
January 27, 2013

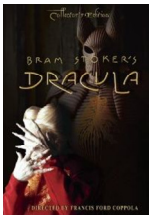
FRBR for Movies

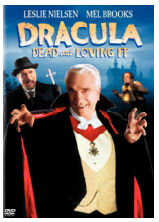


Users Are Looking for Movies



Libraries Describe Publications





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

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-?]

Users Care About Versions

日本語

Rg1024: open clip art library

7

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

8

Movie (Mostly Work) Facets

Limit By Movie or Program:

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

Dates: 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) Germany (1)

Director: Fisher, Terence, 1904-1980 (2) Browning, Tod, 1882-1962 (1) Coppola, Francis Ford, 1939- (1) Laitala, Kerry (1) Maddin, Guy (1) Murnau, F. W. (Friedrich Wilhelm), 1888-1931 (1) Packard, Damon (1) more »

9

Results List

1. Dracula (1931)

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.

Results focused on movie (work)

Get from a library:

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)

Fulfillment options below (expression, manifestation, item)

10

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

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

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

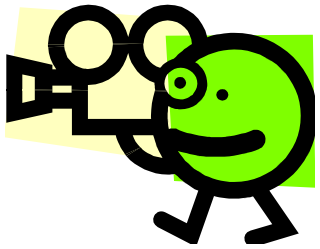
Code http://github.com/cfitz/olac

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

13

Finding FRBR in MARC



14

Machine-Actionable Data

- Structured data

  OriginalReleaseYear = 2011

  NOT

  Originally released as a motion picture in 2011.
- Mapped to FRBR entities and attributes

  OriginalReleaseYear = Date of the Work

15

Machine-Actionable Data

- Supports *faceted access* and the creation of more *readable, grid-like displays*
- Enables *targeted search* and *flexible display*

Dracula (1992 : Francis Ford Coppola)
Format: DVD, NTSC
Languages: English, German
Subtitles: English, French, Spanish
Accessibility: Closed-captioned
Region: Region 1 (U.S. and Canada only)

16

ALA Annual 2012 presentation on extracting work data from MARC records:

http://goo.gl/BrpuJ

Names and Functions

Want to link authorized names with controlled vocabulary for functions

Director = Clint Eastwood

directed by Clint Eastwood

700 $a Eastwood, Clint, $d 1930- $4 drt

18

Identify Individual Statements Based on Punctuation

245 $c Metro-Goldwyn-Mayer picture ; screenplay by George S. Kaufman and Morrie Ryskind ; directed by Sam Wood

→

1. Metro-Goldwyn-Mayer picture
2. screenplay by George S. Kaufman and Morrie Ryskind
3. directed by Sam Wood

19

Make a List of Terms for Each Function (Synonyms)

aus =

- screenplay
- screen play
- screenwriter
- script
- scriptwriter
- writer
- written by

20

Make a List of Terms for Each Function (Translations; Unwanted Variations)

drt =

- directed by
- direction
- director
- 監督
- Regie
- режиссер-постановщик
- ~~director of photography~~
- ~~animation director~~

21

Map Transcribed Names

1. screenplay by *George S. Kaufman* and *Morrie Ryskind*
2. directed by *Sam Wood*

1. 700 $a *Kaufman, George S.* …
1. 700 $a *Ryskind, Morrie*, …
2. 700 $a *Wood, Sam*, …

22

Map Transcribed Roles

1. screenplay by *George S. Kaufman* and *Morrie Ryskind*
2. directed by *Sam Wood*

1. 700 $a *Kaufman, George S.* … $4 aus
1. 700 $a *Ryskind, Morrie*, … $4 aus
2. 700 $a *Wood, Sam*, … $4 drt

23

Many Possibilities…

700 $a *Wood, Sam, 1883-1949* $4 drt

→ → → → →

http://id.loc.gov/authorities/names/n85151535.html
= *http://www.imdb.com/name/nm0939992/*
http://id.loc.gov/vocabulary/relators/drt.html

→ → → → →

Director: *Wood, Sam, 1883-1949*
Regie: *Sam Wood*

24

Natural Language Processing (NLP)

"deals with analyzing, understanding and generating the languages that humans use naturally"—Webopedia

- Artificial intelligence
- Automatic summarization
- Machine translation
- Named entity recognition (NER)

25

What we are doing with trying to parse statements of responsibility and notes about names and roles in MARC records falls within the area of computer science known as natural language processing. Natural language processing involves getting a computer to analyze and work with naturally-occurring human language as opposed to the kind of structured input I talked about earlier. It has many applications, including those listed here.

Natural language processing toolkits

Named entity recognition (NER)

- "April Stevens"
- "Twentieth Century Fox"
- "an Austrian-French co-production, Wega Film, MK2 Productions and Les Films Alain Sarde, Arte France Cinéma"

26

Named entity recognition is where a computer goes through and identifies the proper names in a text. There are existing toolkits for this whose functionality we can use for our project. However, name recognition can be tricky for many reasons. For example, our processor had trouble with the name April Stevens because it wanted to make her first name into a month. With Twentieth Century Fox, it wanted to identify only Fox as the name. The program we have been using generally works better on personal names than corporate names, especially when those names are embedded in long and complex statements such as the one shown here.

Named entity recognition

Approaches to matching

- Start with authorized names and match to statements
- Start with statements and match to authorized names

1. screenplay by *George S. Kaufman* and *Morrie Ryskind*
2. directed by *Sam Wood*

1. 700 $a *Kaufman, George S.* …
1. 700 $a *Ryskind, Morrie*, …
2. 700 $a *Wood, Sam*, …

27

There are a couple possible approaches to matching the names in free-text statements to the authorized versions of the names. In OLAC's early experimentation, we started with the authorized names and tried to match them to the free-text statements. That is, we started with Kaufman, George S and found all the statements with that name; then we moved on to Morrie Ryskind. We are currently using a program that tries to identify the names in the statements and then match them to the authorized names. This is a more complex task, but has the advantage of including names that don't match an access point.

Hard-Coded Rules vs. Machine Learning

- Rules:
  - Manually-compiled lists and decision trees
- Machine learning:
  - Usually based on statistical models
  - Supervised vs. semi-supervised vs. unsupervised learning

28

In the early days, NLP largely relied on manually-compiled lists and decision trees. These have largely been replaced by machine learning models, which are usually based on statistical models. There are various types of learning, but I am only going to talk about supervised learning, which is based on a curated set of training examples.

Supervised Learning

- Training data
  - Set of hand-annotated inputs and desired outputs

  directed by Sam Wood →
  drt = *$a Wood, Sam, $d 1883-1949*
- Computer then generalizes from training data when working on novel data

29

Basically, what happens is that humans develop a set of training data that consists of the inputs, in our case the statements and the authorized names from MARC records, and the correct answers. This is done in a form that the computer can digest and the computer then uses the patterns in these examples to inform its processing of new data.

What You Can Do Soon

- Help us create a hand-annotated set of correct answers for
  - Training data
  - Evaluation
- Online web form coming soon…

30

We are hoping to develop a corpus of correct answers for a pool of sample MARC bibliographic records for moving images. This can be used as training data or alternatively, it can be used to automate and quantify the assessment of various tweaks to the program's approach. We are in the process of creating a web form that will allow anyone to parse these statements and identify a standardized form of the name and function where possible. When the form is ready, it will be announced on various lists or feel free to email me if you want to be sure to get the announcement.

What You Can Do Soon

directed by Sam Wood

directed by *Sam Wood*

- English
- Sam Wood
- Wood, Sam, $d 1883-1949
- Person
- Directed by
- Director

31

We're currently planning to ask people to identify the language of the role, the transcribed name, the authorized name if it appears in the record, whether the name represents a person or organization, the transcribed role and possibly a standardized form of the role.

What You Can Do Now

Use

- 130 uniform titles
- 257 country of producing entity
- 046 $k for original date
- 041 $h for original language
- 1xx/7xx $4/$e relator codes or terms

32

There are some things that we need that can't be squeezed into the existing MARC format, but there are many things you can do now. For example, we wouldn't have to go through all these contortions if the data was already in the record in a way that is easy for a computer to use. It would be better if all of us catalogers could just translate the things we know into computer-speak up front. This slide shows some examples of ways to do that that are important for videos.

What You Can Do Now

130 0- $a ***Lawrence of Arabia (Motion picture)***
257 -- $a ***Great Britain*** $a ***United States*** $2 naf
046 -- $k ***1962***
041 0- $a eng $h ***eng***
700 1- $a Lean, David, $d 1908-1991. $4 ***drt***
700 1- $a O'Toole, Peter, $d 1932- $4 ***act***

33

What You Can Do Now

130 0- $a ***My neighbor Totoro (Motion picture)***
257 -- $a ***Japan*** $2 naf
046 -- $k ***1988***
041 1- $a jpn $a eng $j eng $h ***jpn***
700 1- $a Miyazaki, Hayao, $d 1941- $4 ***drt*** $4 ***aus***

34

Overview of Project

1. Develop end-user interface to take advantage of FRBR and facets
2. Extract and transform existing data
   - MARC → normalized, FRBR-based data
   - Cluster records for FRBR entities
   - Create provisional work (movie) records
   - Assess and correct errors where possible
3. Create backend interface for ongoing input and management of metadata
4. Develop guidelines and documentation for catalogers

35

Interested in Participating?

Contact me at

Kelley McGrath
Metadata Management Librarian
University of Oregon Libraries
kelleym@uoregon.edu
(541) 346-8232