Fishing for FRBR in MARC, Mining for Data in Free Text

Kelley McGrath
University of Oregon
kelleym@uoregon.edu

Cataloging & Classification Research Interest Group
June 24, 2012

What I'm going to talk about

- Overview of problem and proposed solution
- Discussion about what we're working on now:
  - trying to get the data we need out of MARC

Users are looking for movies

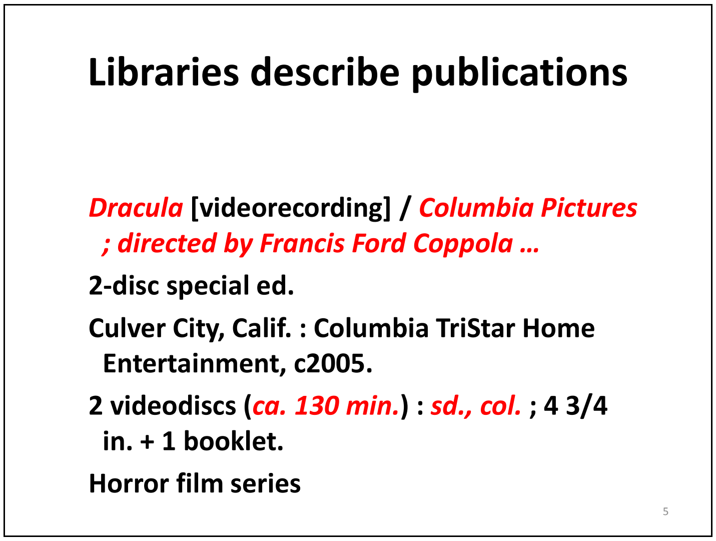
CARL LAEMMLE
PRESENTS
"DRACULA"
BY BRAM STOKER
FROM THE PLAY ADAPTED BY
HAMILTON DEANE & JOHN L. BALDERSTON
A TOD BROWNING PRODUCTION
PRODUCED BY CARL LAEMMLE, JR.
COPYRIGHT MCMXXXI BY UNIVERSAL PICTURES CORPORATION.
CARL LAEMMLE, PRESIDENT.

The typical person who comes to a library looking for something to watch is really looking for a movie (AKA a FRBR work), either a particular movie, such as the 1931 English-language Dracula, or a category of movies, such as recent horror releases, early German horror films, Friday the 13th films or documentaries on the history of horror films. In this presentation, I will use "movie" as shorthand for all types of moving image works, including TV, direct-to-video and born digital.

Libraries describe publications

CLASSIC MONSTER COLLECTION
DRACULA
BRAM STOKER'S DRACULA
LESLIE NIELSEN MEL BROOKS
DRACULA DEAD and LOVING IT

Libraries, however, catalog publications (AKA FRBR 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 …***

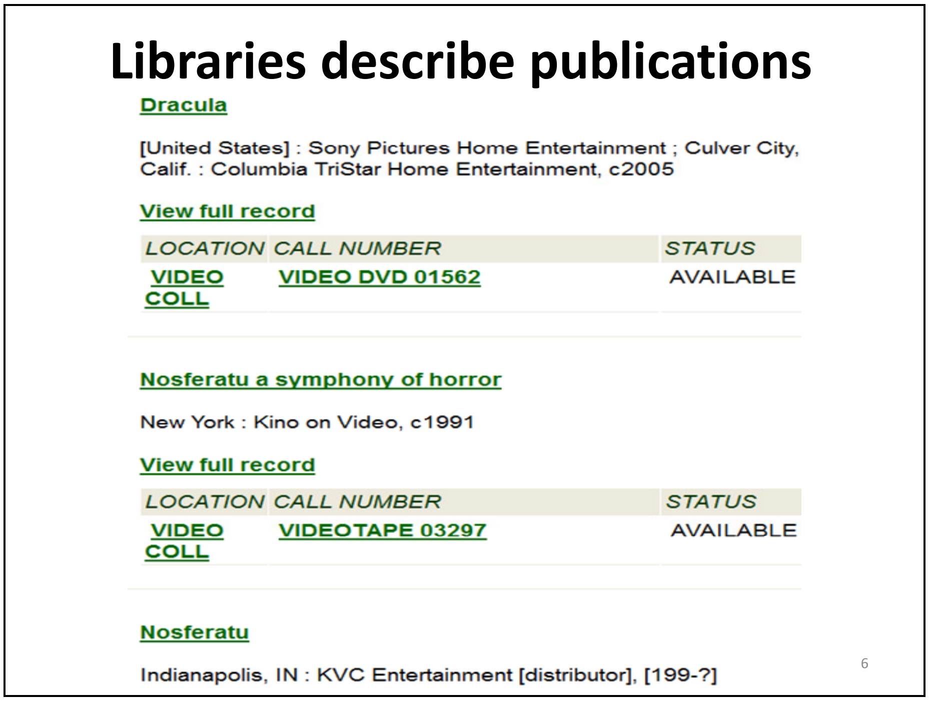
2-disc special ed.

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

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

Horror film series

The heart of the description in a catalog record reflects this emphasis on publications. The parts in italics could 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 title and statement of responsibility are transcribed from the publication and RDA maps these elements to the manifestation. The information about duration, sound and color is mapped to expression in RDA. Most of the information about the movie is in notes and much of it is not even required by cataloging rules. The information about the movie and the publication is all 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-?]

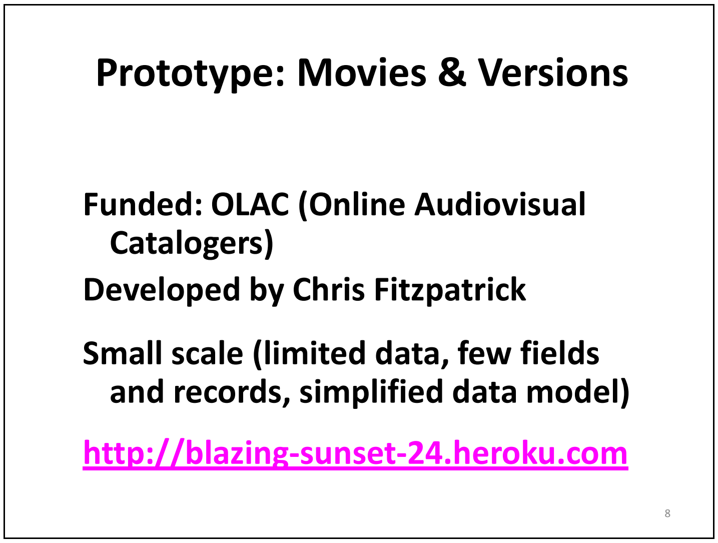
Library hit lists also reflect the emphasis on publications. The information that is given to help identify a title usually describes the publication, as in the example above. The date of publication, which is the date that displays here and is the one used for sorting and limiting in library catalogs, goes with the DVD and not the 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

Blu-ray Disc™
日本語
PRODUCTION
scene
take
camera
date
director

Rg1024 (open clip art library)

People do care about what I'm going to call 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

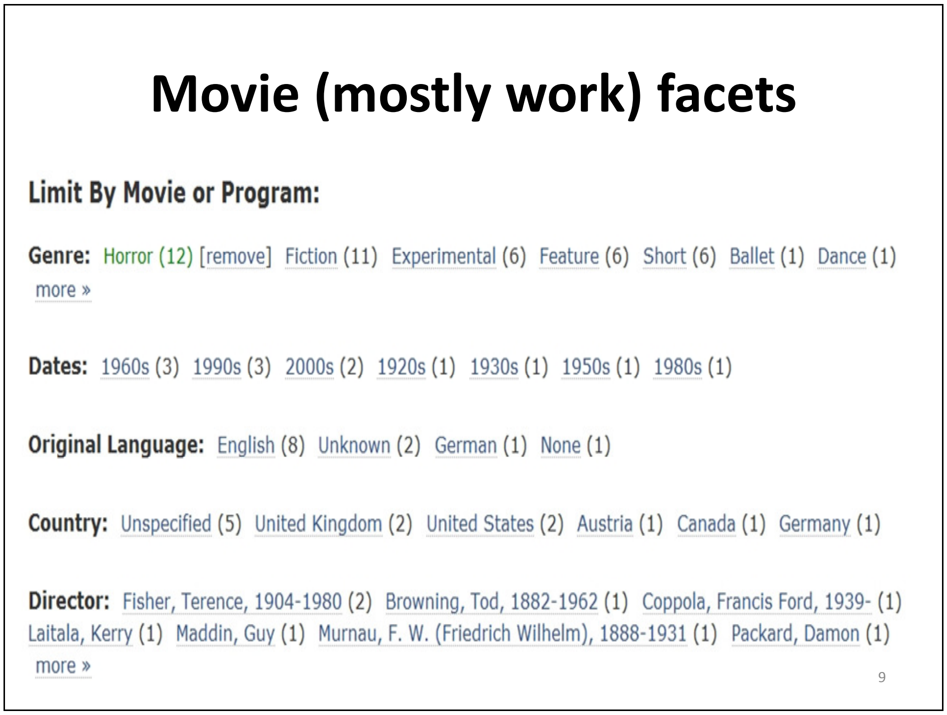
Funded: OLAC (Online Audiovisual Catalogers)

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) Dance (1) more »

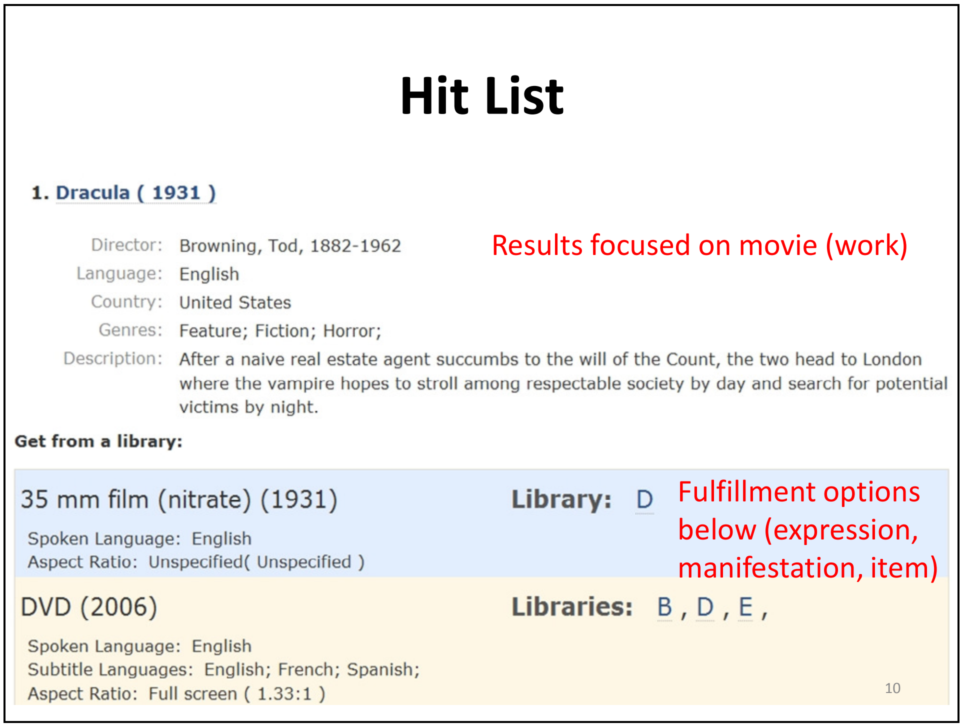
Dates: 1960s (3) 1990s (3) 2000s (2) 1920s (1) 1930s (1) 1950s (1) 1980s (1)

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 »

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

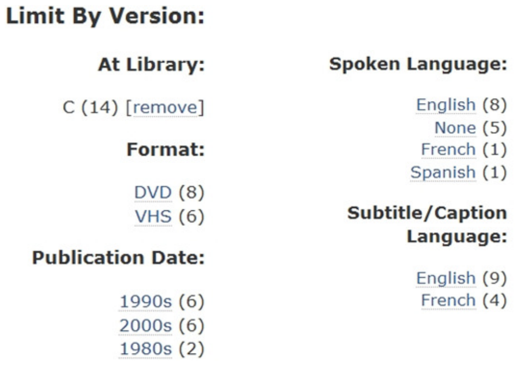
Spoken Language: English
Aspect Ratio: Unspecified(Unspecified)

DVD (2006) **Libraries:** B , D , E ,

Spoken Language: English
Subtitle Languages: English; French; Spanish;
Aspect Ratio: Full screen (1.33:1)

The 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)

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

How to get there from here?

We need…

- **Structured data**

  OriginalReleaseYear = 2011

  NOT

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

  OriginalReleaseYear = Date of the Work

How to get there from here?

- **How should we think about our data?**
- **How do we identify the data that's there?**
- **Can we teach a computer to identify that data in an automated fashion?**

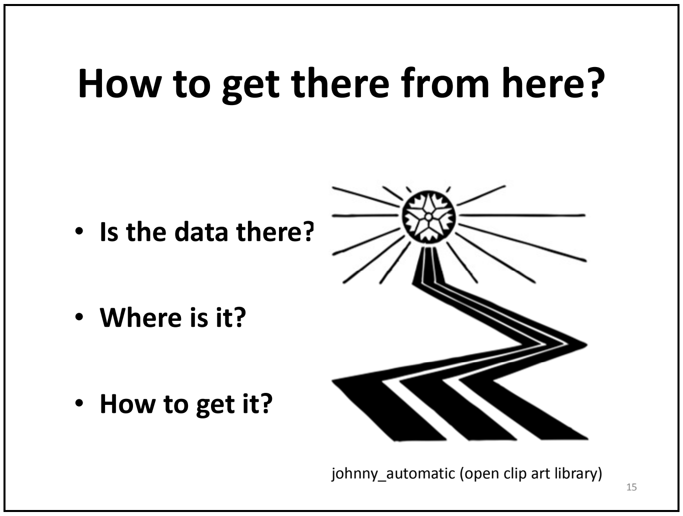
How to get there from here?

Original date of work

- **Identification: Ocean's Eleven (*2001*)**
- **Citation:**

  ***The Usual Suspects*. Dir. Bryan Singer. Perf. Kevin Spacey, Gabriel Byrne, Chazz Palminteri, Stephen Baldwin, and Benecio del Toro. Polygram, *1995*. Film.**
- **Browsing**
- **Limiting**

I'm going to talk about the problem in the context of an important characteristic of movies, original date. Movies are often identified and cited by title and date. This practice is followed by both the Internet Movie Database (IMDb) and standard reference works for moving images. It can also be seen in the MLA citation shown here. Users also want to use the original date for browsing and limiting. Unfortunately, libraries have not done a good job of providing consistent, machine-actionable access to this information.

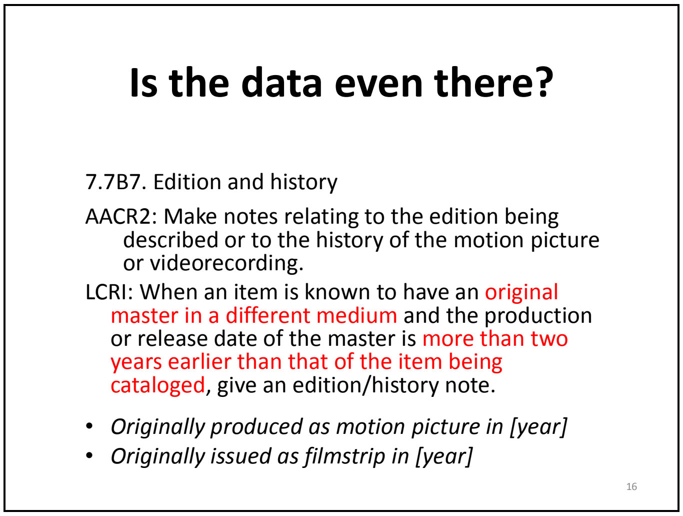


How to get there from here?

- Is the data there?
- Where is it?
- How to get it?

johnny_automatic (open clip art library)

I am going to discuss these three questions using the original date in MARC bibliographic records as an example.



Is the data even there?

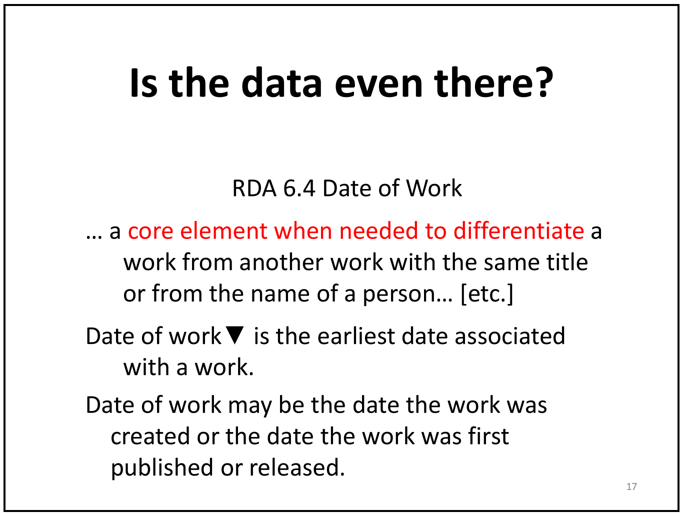
7.7B7. Edition and history

AACR2: Make notes relating to the edition being described or to the history of the motion picture or videorecording.

LCRI: When an item is known to have an original master in a different medium and the production or release date of the master is more than two years earlier than that of the item being cataloged, give an edition/history note.

- *Originally produced as motion picture in [year]*
- *Originally issued as filmstrip in [year]*

The focus of current cataloging rules is on the publication or manifestation (the “item-in-hand”). Original date is included in uniform titles to distinguish movies with the same name, but is otherwise not required. In AACR2 it is only mentioned obliquely in the notes section of the chapter on motion pictures and videorecordings. The LCRI is more restrictive, limiting the use of the note to situations involving a change of medium (e.g., film to DVD) and a gap of more than two years between the original release and the publication of the item being cataloged. Those restrictions are often ignored, but it is clear that neither AACR2 nor the current LCRI tell catalogers to consistently give the original date when it is known.



Is the data even there?

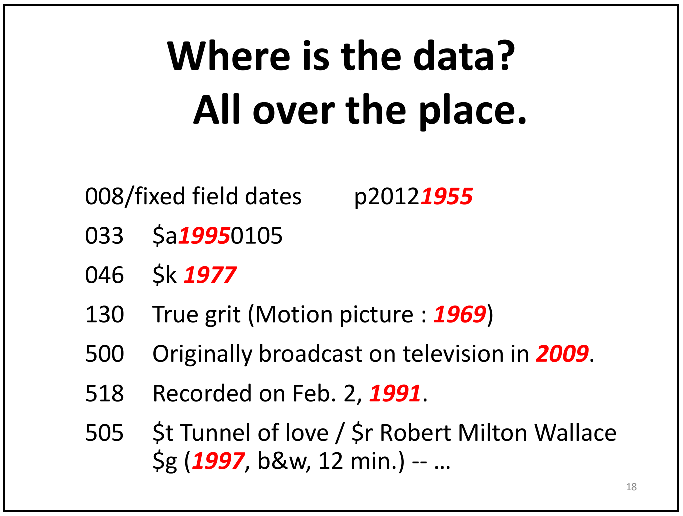
RDA 6.4 Date of Work

… a core element when needed to differentiate a work from another work with the same title or from the name of a person… [etc.]

Date of work▼ is the earliest date associated with a work.

Date of work may be the date the work was created or the date the work was first published or released.

RDA does better. It has an explicit element for date of work. However, date of work is only required when it is needed to differentiate between two movies or other works with the same title. It would be useful for community best practices to encourage catalogers to always record the date of work for a moving image when it’s known.



Where is the data? All over the place.

008/fixed field dates p2012***1955***

033 $a***1995***0105

046 $k ***1977***

130 True grit (Motion picture : ***1969***)

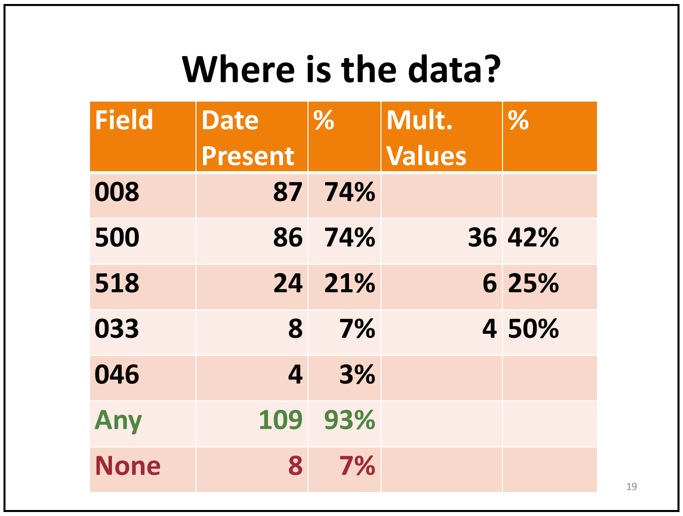
500 Originally broadcast on television in ***2009***.

518 Recorded on Feb. 2, ***1991***.

505 $t Tunnel of love / $r Robert Milton Wallace $g (***1997***, b&w, 12 min.) -- …

18

When we started to look for original dates in MARC bibs, we found many possibilities. A given record could have any of these, none of these, or some combination. If there is more than one date in a record, they may not agree with each other.

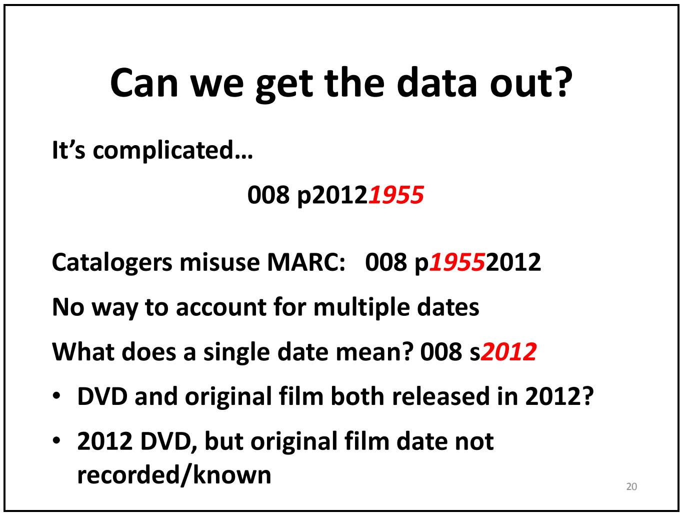


Where is the data?

Field	Date Present	%	Mult. Values	%
008	87	74%		
500	86	74%	36	42%
518	24	21%	6	25%
033	8	7%	4	50%
046	4	3%		
Any	109	93%		
None	8	7%		

19

In our small initial sample (117 titles), almost three quarters had a date in 008 or 500. The multiple values column tells how many records with at least one date in a given field had more than one date in that field. The percentage column shows how common it is to have multiple values. So of the 86 records that had at least one date in 500, 36 (or 42%) had more than one date. The “any” row shows that we were able to find *some* date in at least one field for 93% of the records in our sample. It doesn’t say anything about the correctness of those dates. There were eight records for which our current process did not find any dates. For many of those, a date could be identified by a human looking at the record.



Can we get the data out?

It’s complicated…

008 p2012*1955*

Catalogers misuse MARC: 008 p*1955*2012

No way to account for multiple dates

What does a single date mean? 008 s*2012*

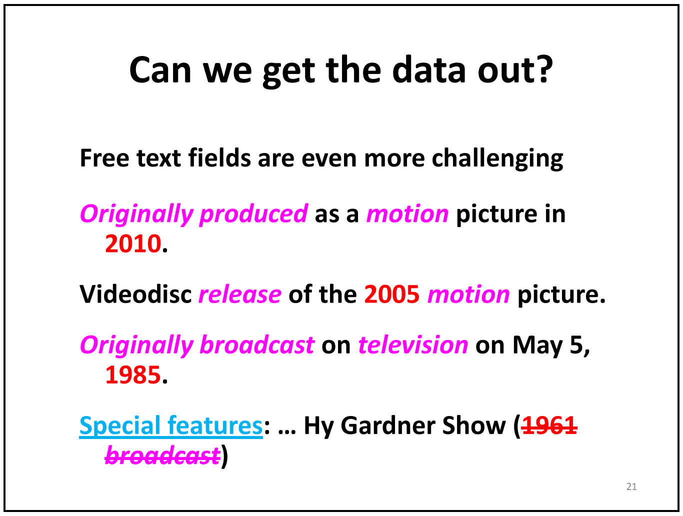
- **DVD and original film both released in 2012?**
- **2012 DVD, but original film date not recorded/known**

20

Even what appears to be straightforward, structured data, like dates in the 008, have pitfalls for the unwary. The top example is correctly-formatted MARC with the date of the publication in Date1 and the original date in Date2. In the wild, we found some data with the two dates reversed. There is actually a practical reason why catalogers might do that. Most catalogs sort and limit using 008 Date1 so if catalogers put the original date there, their users can search by that date.

008 also has some structural limitations. Date2 can only express a single date so it provides incomplete information when a DVD contains more than one

movie. It is also impossible to interpret a single date. A single date might mean that both the DVD and original film were released in the same year. Or it might not. It might mean that the cataloger didn’t know or chose not to record the original date. The latter is more common now since many have interpreted the release of a movie on DVD with multiple language options and/or special features as a new work which should be coded with a single “s” date for the DVD.



Can we get the data out?

Free text fields are even more challenging

Originally produced **as a** ***motion*** **picture in 2010.**

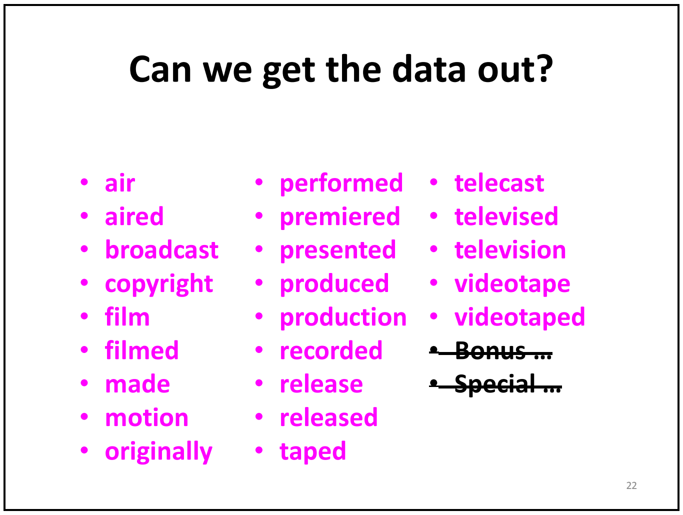
Videodisc ***release*** **of the 2005** ***motion*** **picture.**

Originally broadcast **on** ***television*** **on May 5, 1985.**

Special features: **… Hy Gardner Show (~~1961 *broadcast*~~)**

21

Free text fields are more challenging, but you can see that it’s not too hard to teach a computer to recognize some common patterns. Our approach was to look for 500 general notes with a year (18xx, 19xx or 20xx) in association with some keywords that suggest that the note is about the original date. We took out notes starting with the words “special” or “bonus” since these usually apply to content other than the main movie.

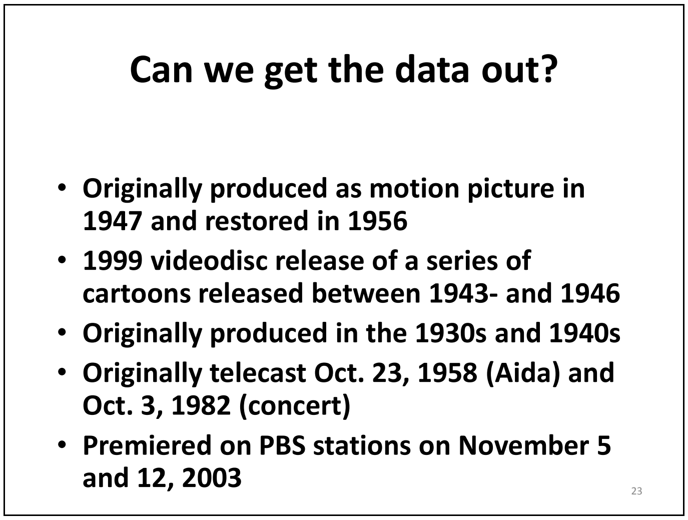


Can we get the data out?

- **air**
- **aired**
- **broadcast**
- **copyright**
- **film**
- **filmed**
- **made**
- **motion**
- **originally**
- **performed**
- **premiered**
- **presented**
- **produced**
- **production**
- **recorded**
- **release**
- **released**
- **taped**
- **telecast**
- **televised**
- **television**
- **videotape**
- **videotaped**
- ~~**Bonus …**~~
- ~~**Special …**~~

22

Our current list of keywords

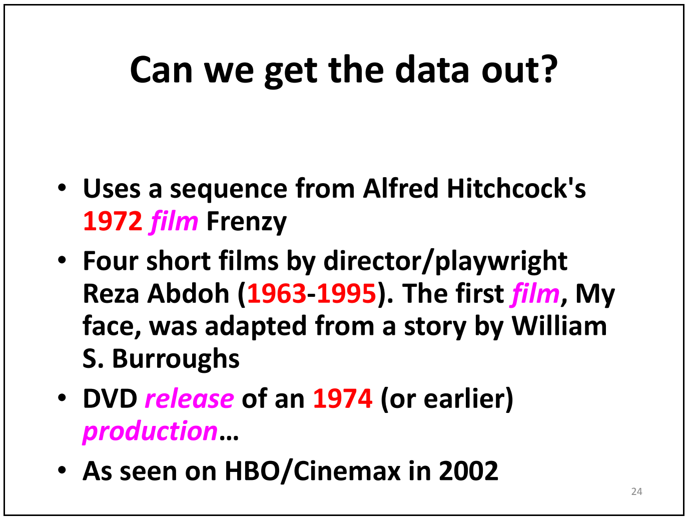


Can we get the data out?

- **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**

23

Although the examples given two slides ago are largely straightforward to process, a great many more complicated variations exist in the wild. These are all real examples of 500 notes from UO's catalog.

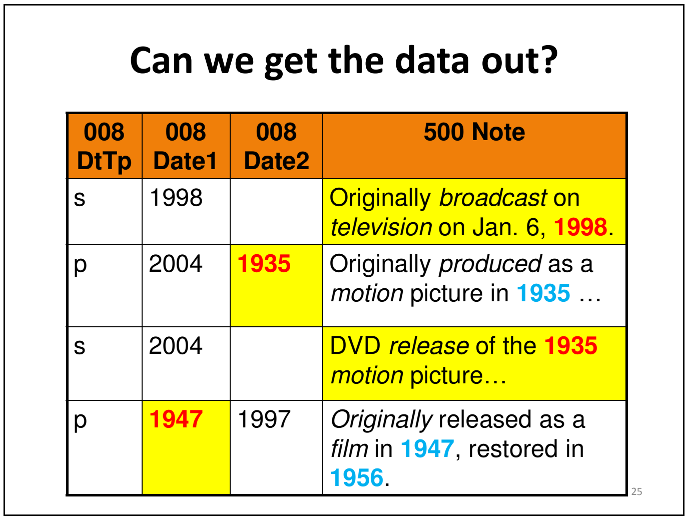


Can we get the data out?

- **Uses a sequence from Alfred Hitchcock's 1972 *film* Frenzy**
- **Four short films by director/playwright Reza Abdoh (1963-1995). The first *film*, My face, was adapted from a story by William S. Burroughs**
- **DVD *release* of an 1974 (or earlier) *production*…**
- **As seen on HBO/Cinemax in 2002**

24

Inevitably, rules don't always work. We get some false drops, as in the first two examples here. There can also be a problem with unclear data, as in the third note. Finally, our approach misses some data that really is there. The fourth note is an example of data with no obvious hook to identify it as a note about the original date.

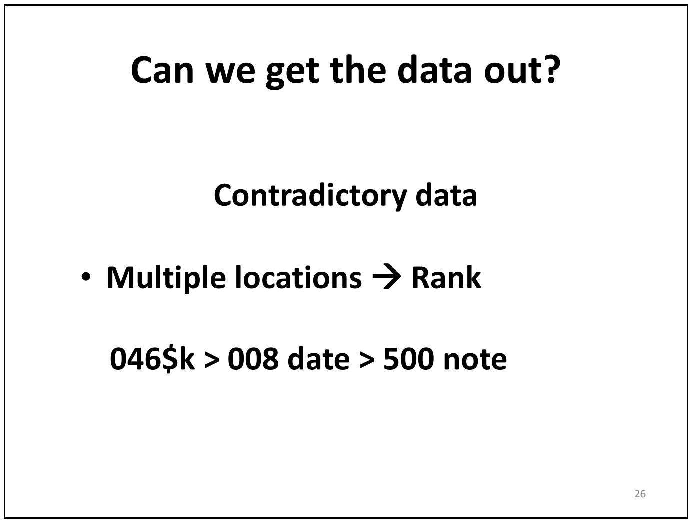


Can we get the data out?

008 DtTp	008 Date1	008 Date2	500 Note
s	1998		Originally *broadcast* on *television* on Jan. 6, **1998**.
p	2004	**1935**	Originally *produced* as a *motion* picture in **1935** …
s	2004		DVD *release* of the **1935** *motion* picture…
p	**1947**	1997	*Originally* released as a *film* in **1947**, restored in **1956**.

25

I have described the problem of sometimes having no data and the challenges of trying to get what data is there out in a machine actionable form. We also face the problem of too much data that may be contradictory. This table shows two examples (lines 2 and 4) where there are dates in both 008 and 500. Line 4 also shows and example where there are two dates in a 500 note.



Can we get the data out?

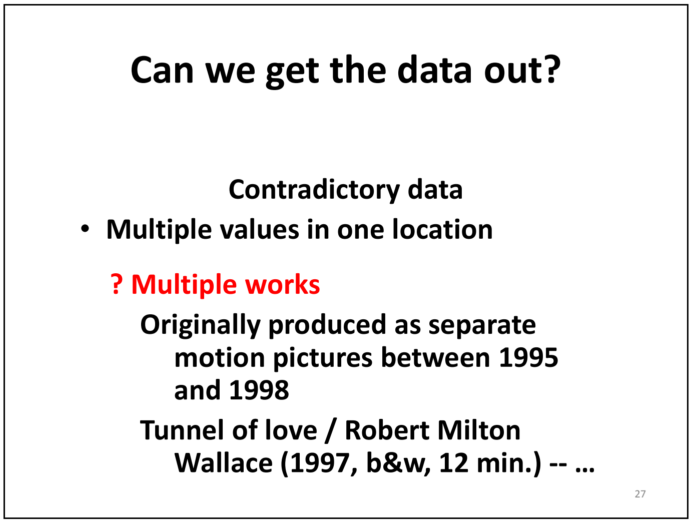
Contradictory data

- **Multiple locations → Rank**

046$k > 008 date > 500 note

26

When there is more than one potential location (field or subfield) where the original date could potentially be found, the most promising approach seems to be to rank them in terms of likely accuracy. Here we consider an 046 field more reliable than an 008, which is more reliable that a 500 note. There have to be some caveats, though. If there are signals elsewhere in the record that the bib is describing multiple movies, we would need to remove the 008 from the hierarchy completely since it can’t express multiple dates.



Can we get the data out?

Contradictory data

- **Multiple values in one location**

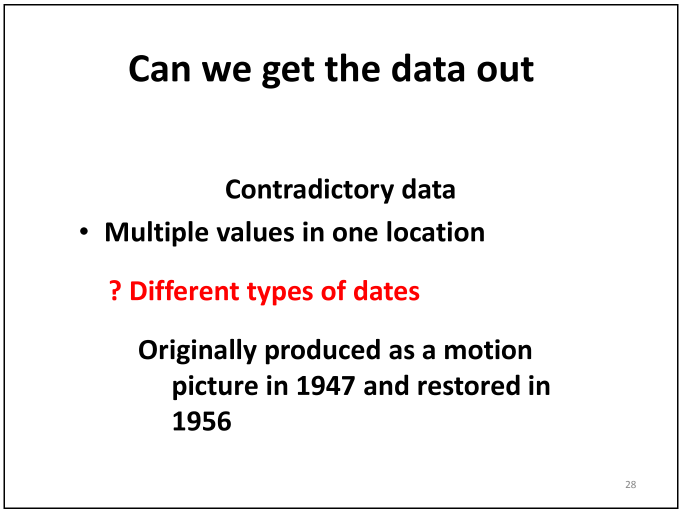
? Multiple works

Originally produced as separate motion pictures between 1995 and 1998

Tunnel of love / Robert Milton Wallace (1997, b&w, 12 min.) -- …

27

Multiple dates in one field or types of field are more complicated to deal with since there are several possible explanations. One is that there are multiple movies on the DVD. These can often be identified by the presence of a 500 note or a non-collective title in the 245 field.



Can we get the data out

Contradictory data

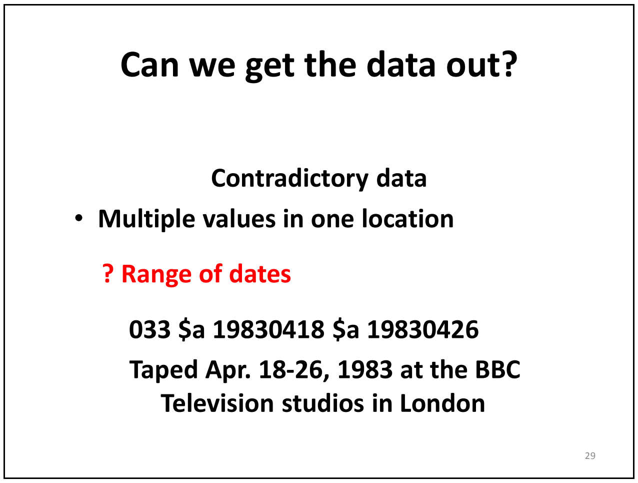
- **Multiple values in one location**

? Different types of dates

Originally produced as a motion picture in 1947 and restored in 1956

28

A second common situation is that the dates represent different types of dates. In some cases, as in this example, it is possible to pick out the date that is embedded in a common pattern (“originally produced as a motion picture in”) indicating the original date. We also experimented with the heuristic of taking the lowest date. This works most of the time, but is still wrong a significant percentage of the time.



Can we get the data out?

Contradictory data

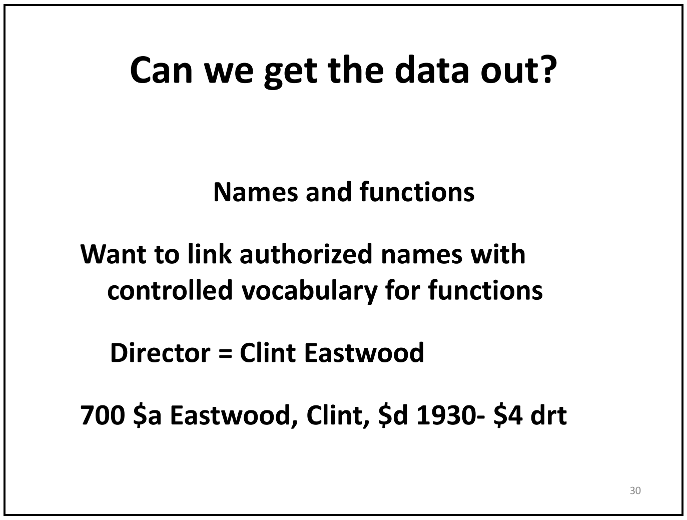
- **Multiple values in one location**

? Range of dates

033 $a 19830418 $a 19830426

Taped Apr. 18-26, 1983 at the BBC Television studios in London

A third, less common, possibility is that multiple dates represent a range of dates. It’s not clear right now how we might easily identify these.



Can we get the data out?

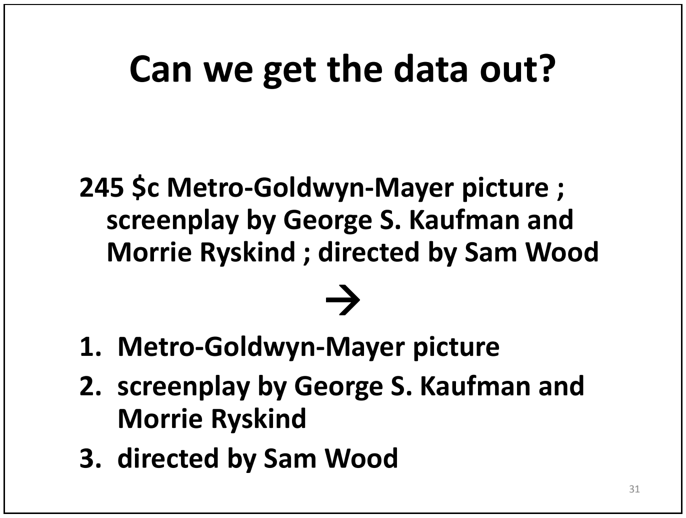
Names and functions

Want to link authorized names with controlled vocabulary for functions

Director = Clint Eastwood

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

Many movies are created by large numbers of people performing many different functions. Another thing we’re trying to do is link authorized forms of names with the functions they performed. We want to come up with statements like “director = Clint Eastwood” only using controlled data (ideally identifiers) rather than free text. In some cases, as in the 700 shown, this work is already done for us, but in most cases, the connection is not there in machine-actionable form.



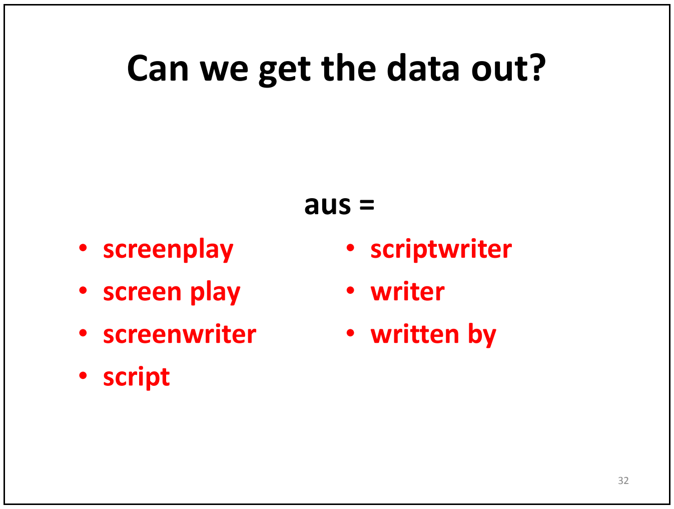
Can we get the data out?

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**

Generally in the statement of responsibility (245$c), participant or performer note (511) or creation/production credits note (508), there is a transcribed or quasi-transcribed set of statements connecting names and functions. The type of statements in the 245$c shown here are typical for movies. According to ISBD punctuation, individual statements are separated by space-semicolon-space (or in the note fields officially by semicolon-space, although few catalogers do it that way). So theoretically each field or subfield can be broken apart based on this punctuation into separate statements. Each statement would then include one or more functions that are related to one or more names while eliminating unrelated information. Of course, no one has ever forgotten to include a space-semicolon-space, right?



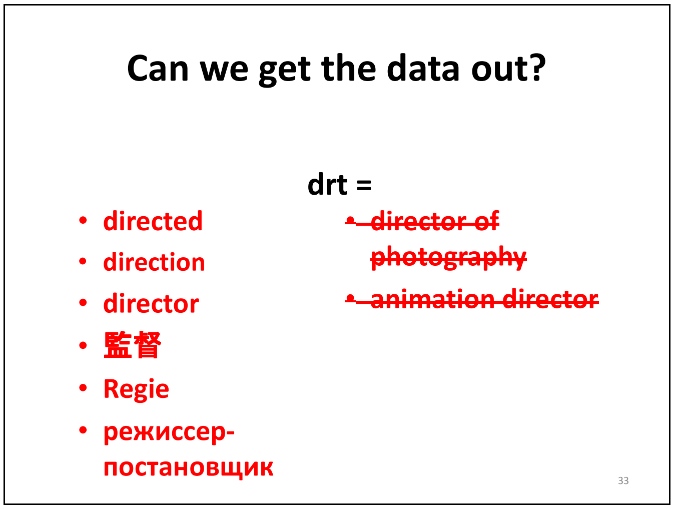
Can we get the data out?

aus =

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

32

We then have to try to identify the names and functions in the statements. For the functions, the starting point is to come up with a list of free-text usage that maps to a given standard term. Here the MARC relator code aus (author of screenplay) is mapped to a number of common variations that occur in movie credits.



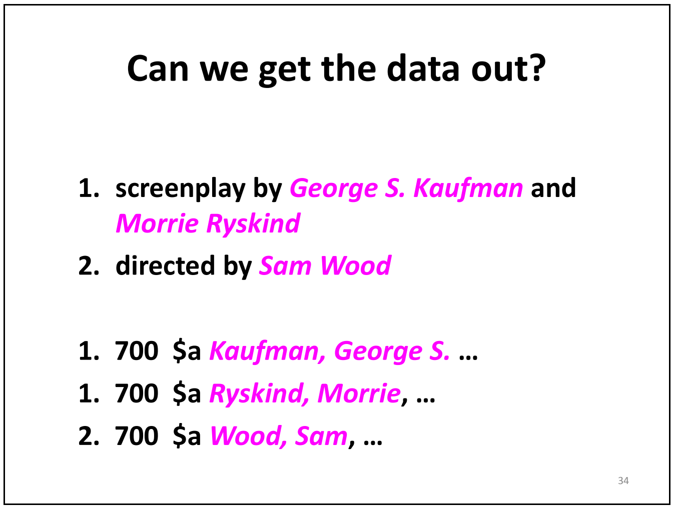
Can we get the data out?

drt =

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

33

These are some terms for the MARC relator code drt (director). Since we are taking data from transcribed fields, we have to account for non-English variations. We also have to watch out for similar terms that contain our base term. For example, “director” is contained within the functions “director of photography” and “animation director.”



Can we get the data out?

1. screenplay by *George S. Kaufman* and *Morrie Ryskind*

2. directed by *Sam Wood*

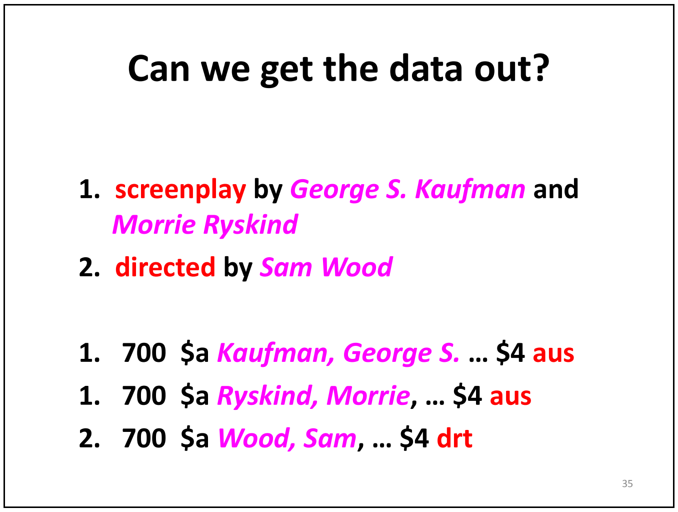
1. 700 $a *Kaufman, George S.* …

1. 700 $a *Ryskind, Morrie*, …

2. 700 $a *Wood, Sam*, …

34

For the names, we try to match the transcribed versions with the authorized forms in 1xx or 7xx fields. In many cases, as in the examples shown here, this is reasonably straightforward. It is not so hard to teach a computer to test names in inverted order. Obviously, this doesn’t work in all cases. Some more complicated strategies are discussed in my Code4Lib Journal article with Lynne Bisko at http://journal.code4lib.org/articles/775.

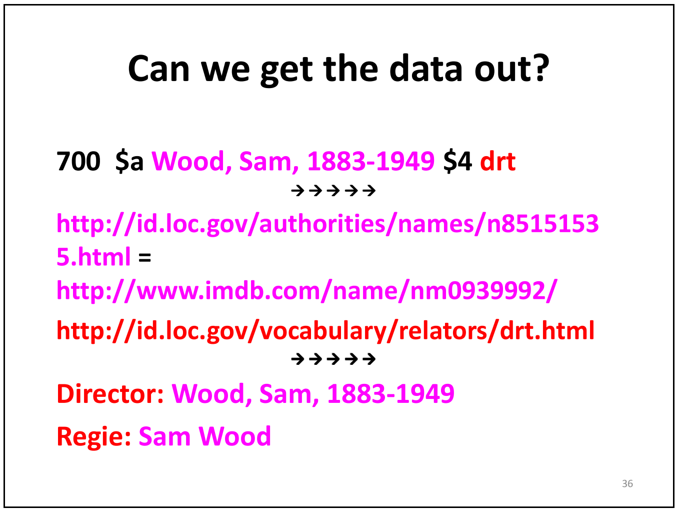


Can we get the data out?

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

We can then combine the names we've identified with the functions we've identified to add relator codes for functions to the appropriate names.



Can we get the data out?

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

→→→→→

http://id.loc.gov/authorities/names/n85151535.html =

http://www.imdb.com/name/nm0939992/

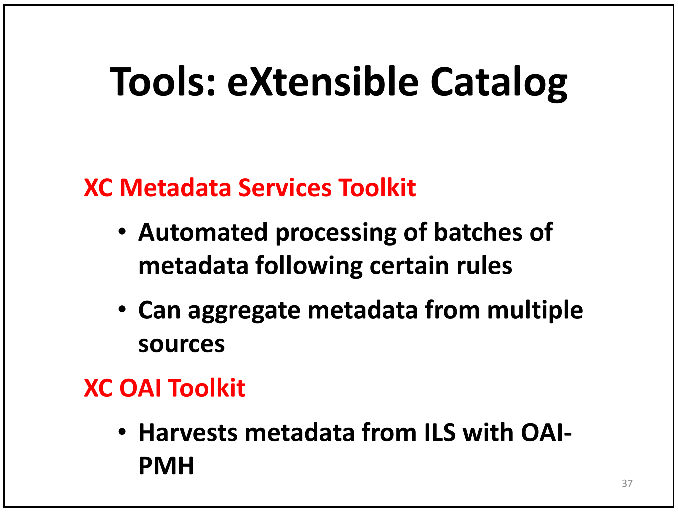
http://id.loc.gov/vocabulary/relators/drt.html

→→→→→

Director: Wood, Sam, 1883-1949

Regie: Sam Wood

Now that we have a controlled form for both the name and the function and they're connected, we can map both to some sort of identifier. Here I show the LC NAF (national authority file) and the MARC relator code identifiers from id.loc.gov. The LC NAF identifier could potentially then be linked to the IMDb identifier for the same person. This would allow us not only to create the first display strictly from existing library data, but also to create many variations. In the second display, the director function has been translated into German and the display form is taken from IMDb. Very many different displays could potentially be generated.



Tools: eXtensible Catalog

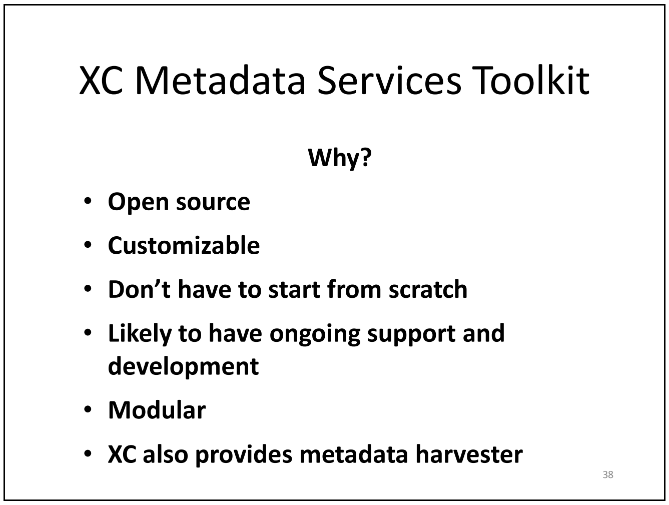
XC Metadata Services Toolkit

- Automated processing of batches of metadata following certain rules
- Can aggregate metadata from multiple sources

XC OAI Toolkit

- Harvests metadata from ILS with OAI-PMH

For our current experimentation, we are using the eXtensible Catalog's XC Metadata Services Toolkit.

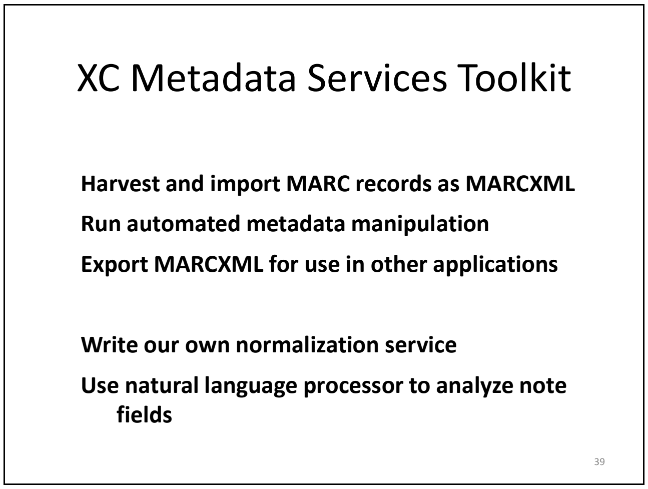


XC Metadata Services Toolkit

Why?

- **Open source**
- **Customizable**
- **Don’t have to start from scratch**
- **Likely to have ongoing support and development**
- **Modular**
- **XC also provides metadata harvester**

38



XC Metadata Services Toolkit

Harvest and import MARC records as MARCXML

Run automated metadata manipulation

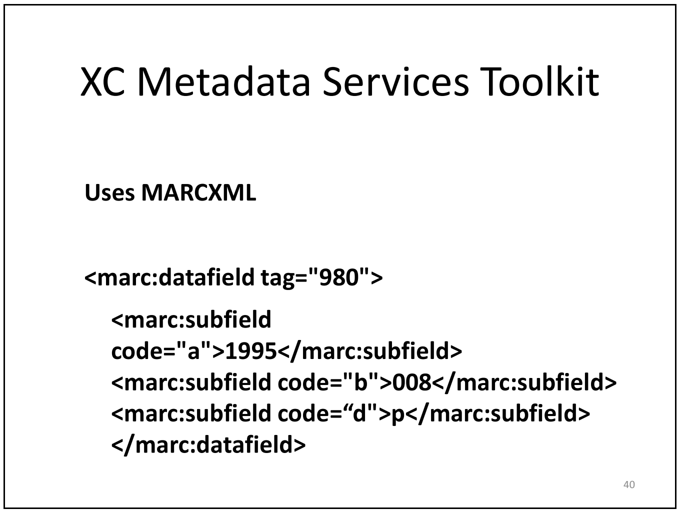
Export MARCXML for use in other applications

Write our own normalization service

Use natural language processor to analyze note fields

39

At the top are the general steps of the process. We are currently working on our own normalization service, which is the program that will do the automated metadata manipulation.



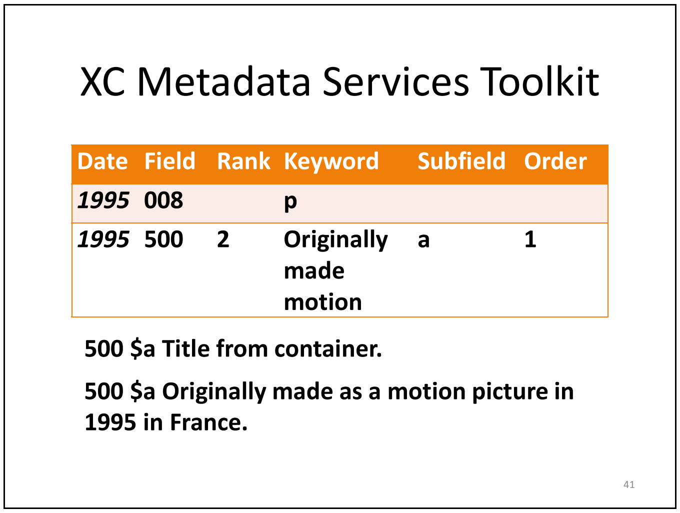
XC Metadata Services Toolkit

Uses MARCXML

```
<marc:datafield tag="980">
    <marc:subfield
    code="a">1995</marc:subfield>
    <marc:subfield code="b">008</marc:subfield>
    <marc:subfield code=“d">p</marc:subfield>
    </marc:datafield>
```

40

The XC Metadata Services Toolkit uses MARCXML and we put the data we extract into 9xx (local) fields.



XC Metadata Services Toolkit

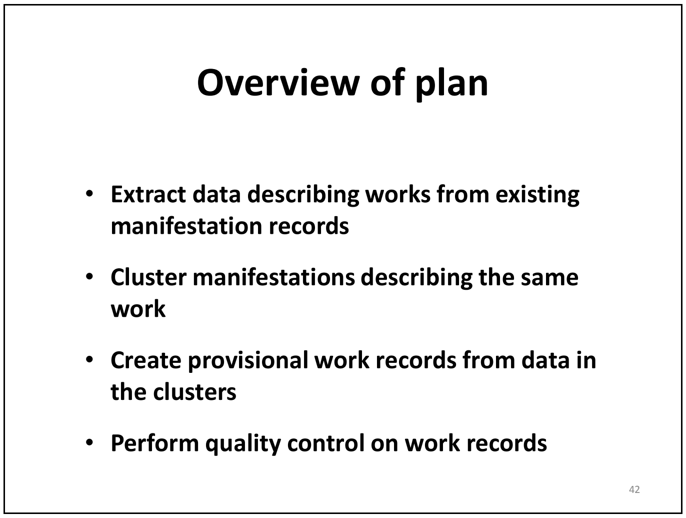
Date	Field	Rank	Keyword	Subfield	Order
1995	**008**		**p**		
1995	**500**	**2**	**Originally made motion**	**a**	**1**

500 $a Title from container.

500 $a Originally made as a motion picture in 1995 in France.

41

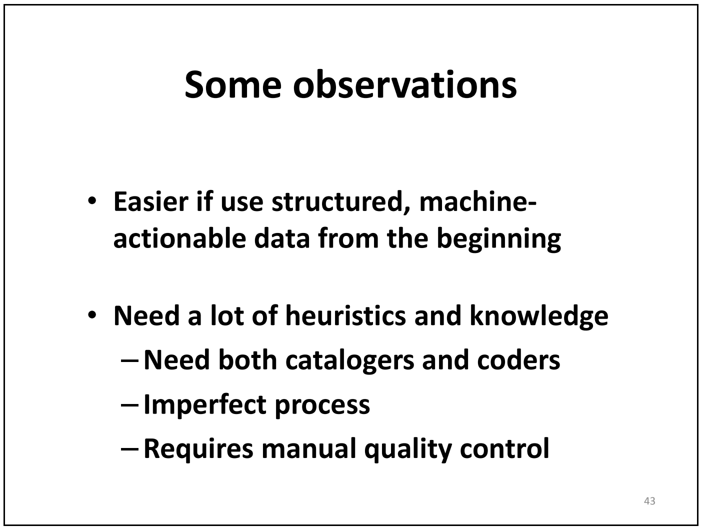
It’s hard to read MARCXML so here’s the same thing in a table. One thing to note is that we are keeping all the data that we have harvested, along with some contextual information about where it came from. For example, in the second row, our program found the date 1995 in a 500 field. It was the 2nd 500 field in the record, it had three keywords (originally, made, motion), the date was in subfield a and it was the first date found in that 500 field.



Overview of plan

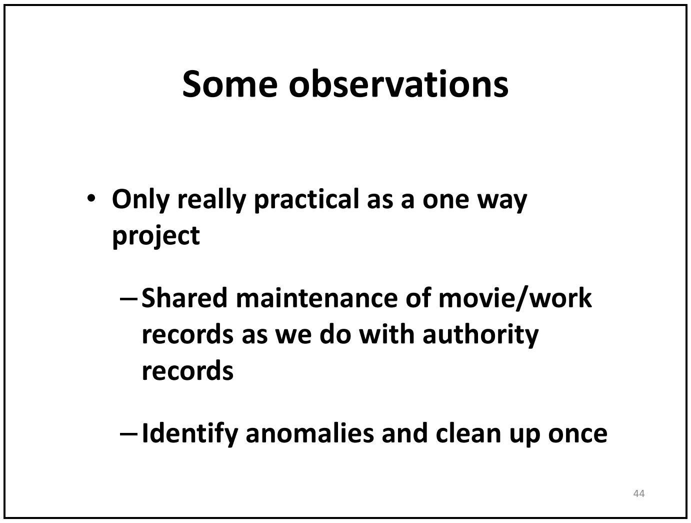
- Extract data describing works from existing manifestation records
- Cluster manifestations describing the same work
- Create provisional work records from data in the clusters
- Perform quality control on work records

This is a general outline of our long-term plan. There are some potential alternative approaches for widely-distributed or popular movies. For example, we could take the identifying elements we found in the MARC records and match them against an external service such as Freebase. We could then cluster movies on the IDs provided by the external service.



Some observations

- Easier if use structured, machine-actionable data from the beginning
- Need a lot of heuristics and knowledge
  - Need both catalogers and coders
  - Imperfect process
  - Requires manual quality control



Some observations

- Only really practical as a one way project
  - Shared maintenance of movie/work records as we do with authority records
  - Identify anomalies and clean up once

Prototype:

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

Sample searches and use cases (good to check out because of small size of prototype):
http://blazing-sunset-24.heroku.com/page/samples

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

More info:

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

OLAC discussion group (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

Want to get involved?

Contact:

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