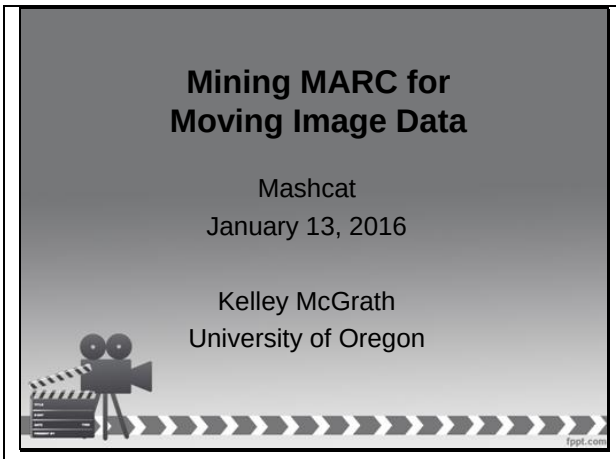
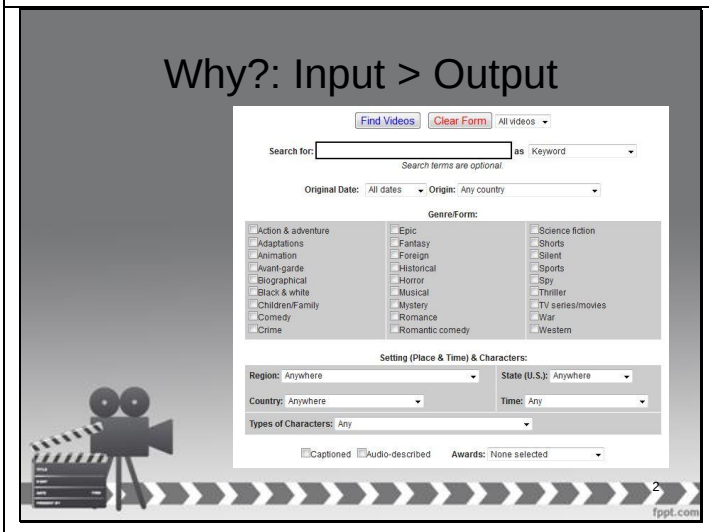


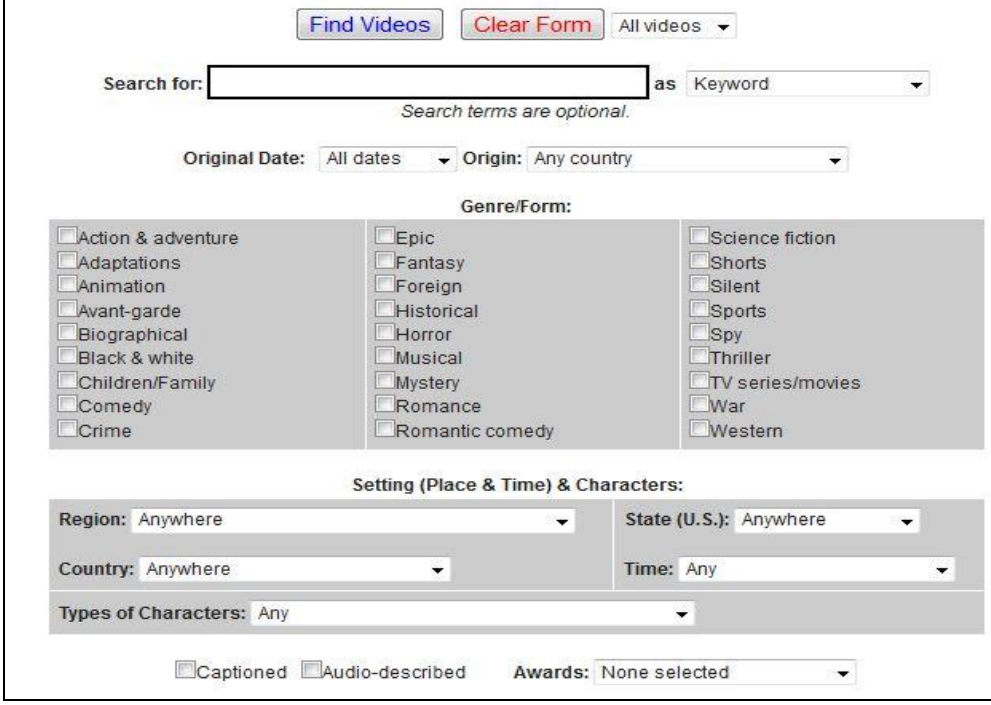
Mining MARC for Moving Image Data

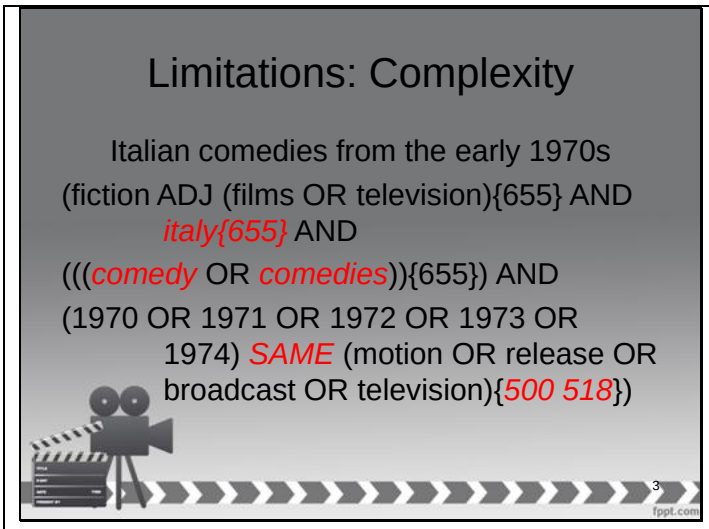
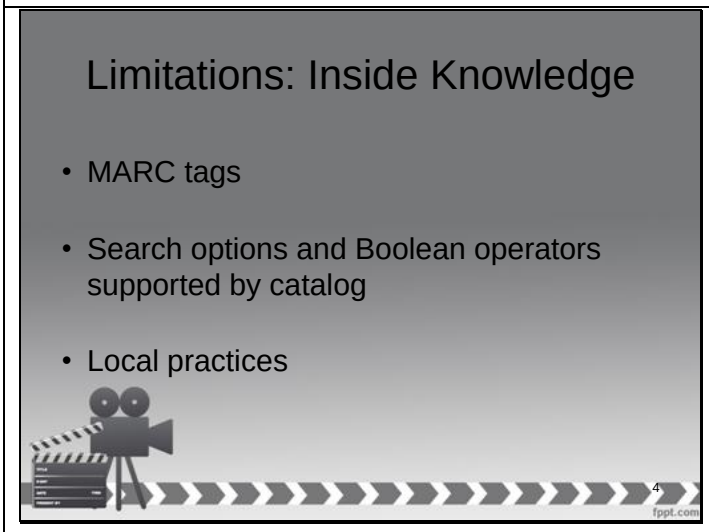
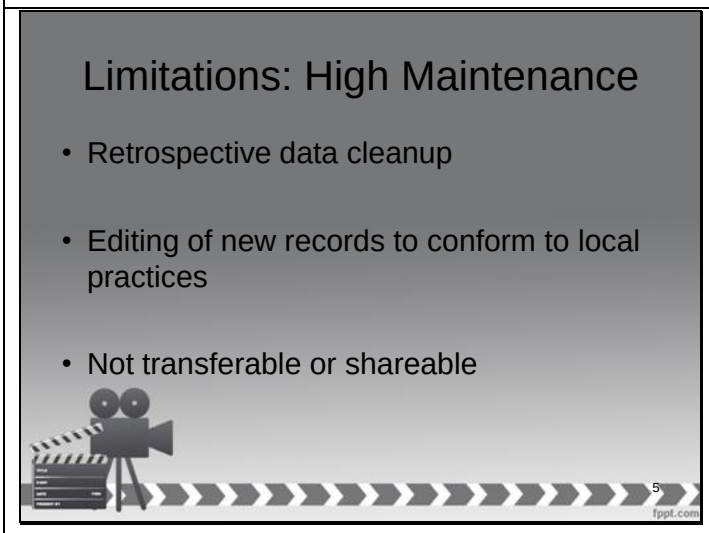
Mashcat
January 13, 2016

Kelley McGrath
University of Oregon



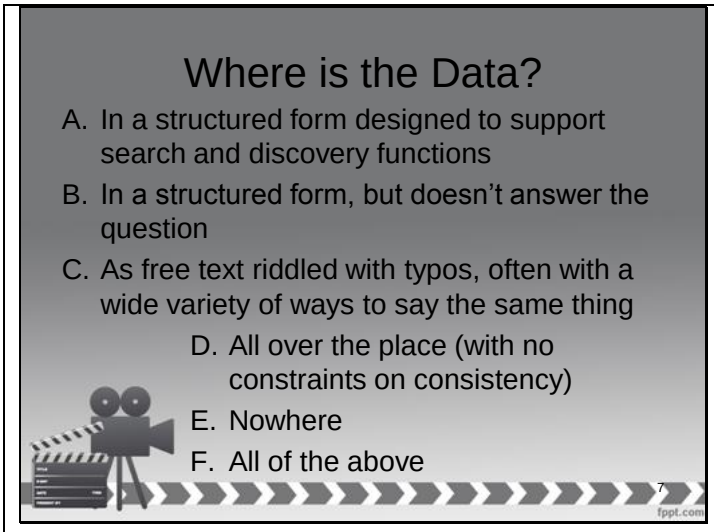
Why did I become interested in the problem of moving image data in MARC? My first job out of library school was as the a/v cataloger at a university with a large media collection. I was quickly dismayed by the disconnect between what I knew when I was cataloging something, what I could put in the record and what a user could get back out. This form was my first attempt to bridge that gap. The video collection was in closed stacks and the videos were only given accession numbers so there was no way to browse the collection. The form gives users a few ways to explore the collection.



	However, the form has some substantial drawbacks. First, the search strategies are very complex. You can see why a user would never be able to come up with this much less manage to type it in.
	The search strategies require inside knowledge of the way the catalog software works, of MARC tags and of local cataloging practices.
	The form is also high maintenance. Initially, we did substantial data cleanup to ensure that the data conformed to the expectations of the search strings. In order to make the form work with new records, we had to edit them according to our local practices. Because we relied on local practices, we couldn't effectively share our work.

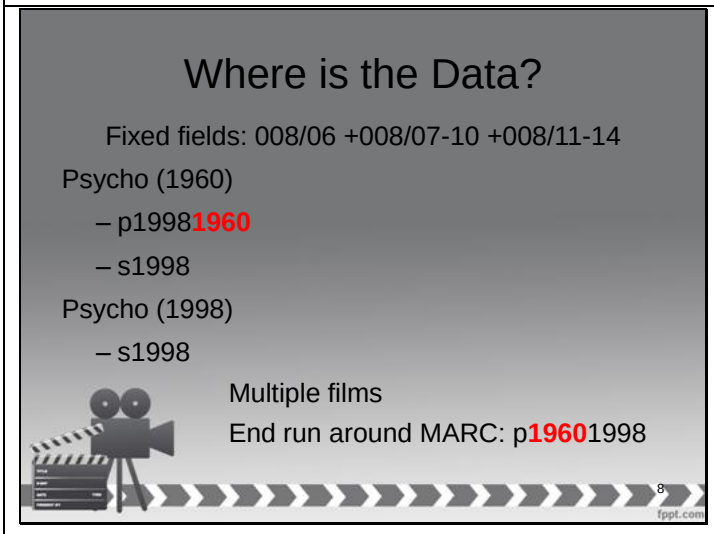
Facets: structured data, right data

Facets would be an even better way to enable users to explore the collection. However, you can't exactly facet on "Originally produced as a motion picture in 2006." We need machine-actionable data that is consistently-structured and that answers the questions we care about. The topsy-turvy, convoluted search strategies I used for that early form show that we often don't have that. It isn't practical to start all over again so I started trying to figure out how much we could extract from our existing records.

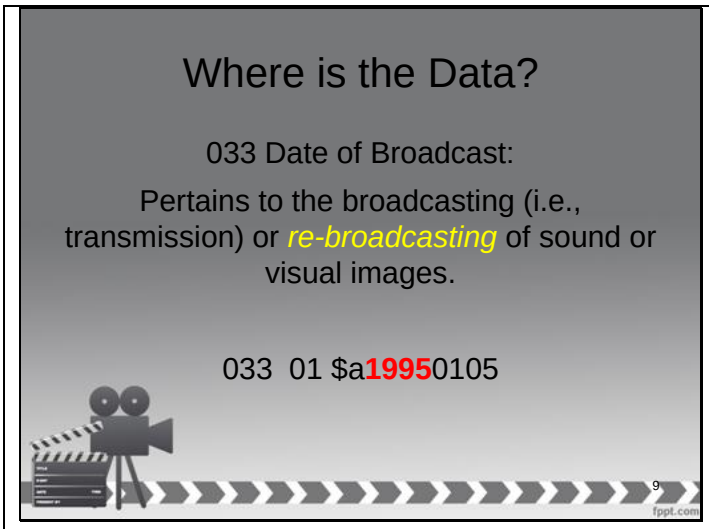
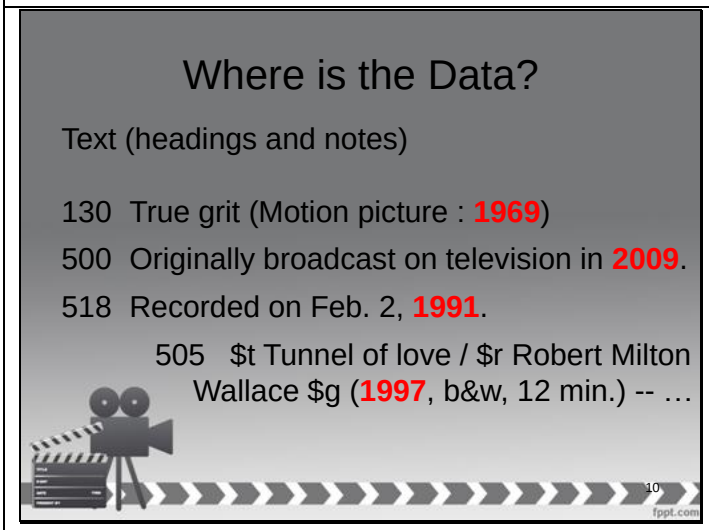
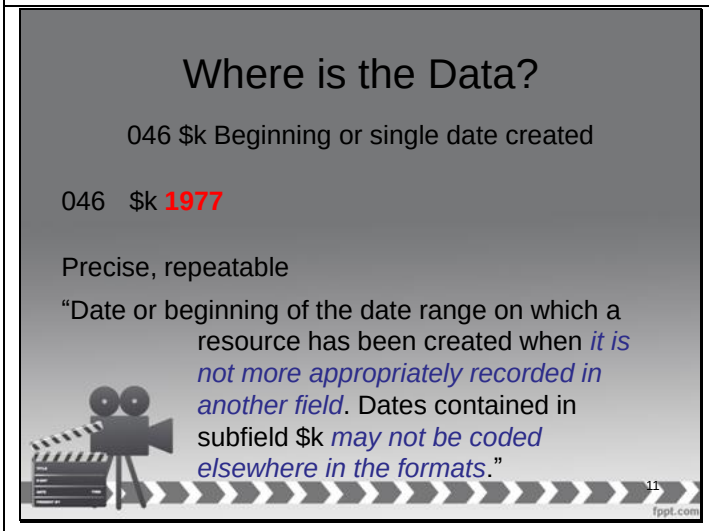


When you start looking for the data, it gets complicated very quickly. Sometimes, you have beautiful data, but more often you have data that isn't quite right or that is buried in text strings. It's also common to have too many choices when the same type of information occurs multiple times in a single record. For example, the format of a video can be recorded in several places and there's nothing that enforces consistency. During the transition from VHS to DVD, many catalogers copied a VHS record and edited it to describe the DVD version. Unfortunately, they often forgot to update one of the VHS-related values, which led to a lot of records with an identity crisis—they couldn't make up their minds if they were describing VHS or DVD.

Alternatively, sometimes that data just isn't there. The original theatrical release date is important for feature films, but the cataloging rules focus on the publication in hand and don't require this information.



Let's look at the various places that the original date of a moving image might appear in a MARC record. Sometimes the original date appears in date2 of the 008, as in the first example. However, this is only done if the video is a straight re-release. If new content of any sort, such as subtitles or special features, has been added, the record is coded with only the date of publication of the video, as in the second example. Sometimes the year of original release and the year the video was published are the same so there's only one date. Since date2 is not repeatable, this method doesn't work for videos that contain more than one film, such as a collection of animated shorts. In another twist, some libraries reverse date1 and date2 so that their public catalog will search and sort by the original date.

	There is a field for date of broadcast, but it isn't limited to the date of the original broadcast.
	The original date may appear in different text strings with all the parsing problems that come with that.
	Finally, there is a field that would seem to be a good fit except for constraints in the text of the MARC format. OLAC is in the process of trying to get these restrictions removed. No one at the Midwinter MARC Advisory Committee could think of a reason to keep them so I expect our proposal to remove these restrictions will be approved at Annual.

The kitchen sink that is 300$b

Other physical details

"Physical characteristics such as illustrative matter, coloration, playing speed, groove characteristics, presence and kind of sound, number of channels, motion picture presentation format, etc."

12

Going forward, it would be best if there is only one place to record a given type of data. Conversely, only one type of data should be recordable in a given field. Unfortunately, this isn't true of MARC. 300$b, other physical details, is a particularly egregious example. Nobody in their right mind would define a field like this for data processing.

Bibframe and 300$b

bf:colorContent
bf:illustrationNote "sd., col. and b&w ;"

http://bibfra.me/vocab/marc/color
otherPhysicalDetails "sd., col. and b&w ;"

13

Both the LC and the Zepheira versions of Bibframe have specific fields for at least some of the data, such as color content, that are currently recorded in 300$b. Unfortunately, these all appear only to take literal values. This is a step back from MARC, which at least had coded values for color content in 007. In addition, the conversion algorithms that I could find are not very sophisticated and keep everything glommed together in a single textual field.

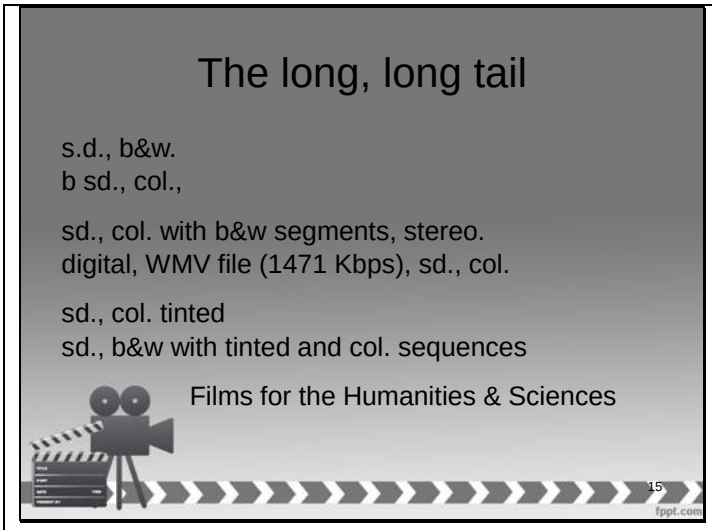
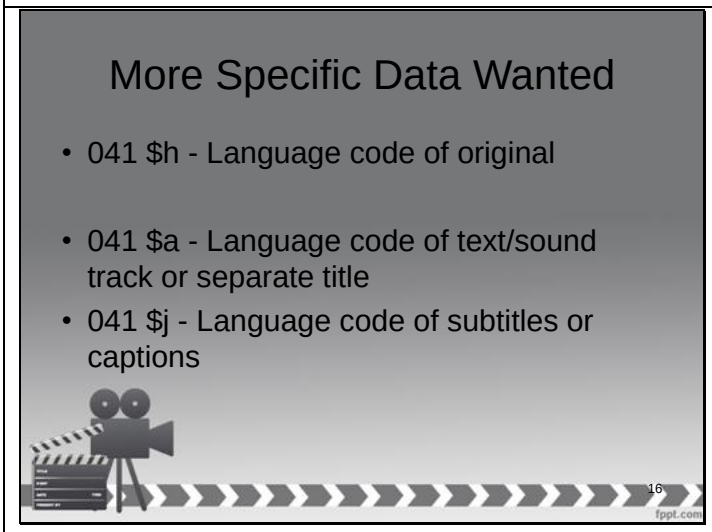
The long, long tail

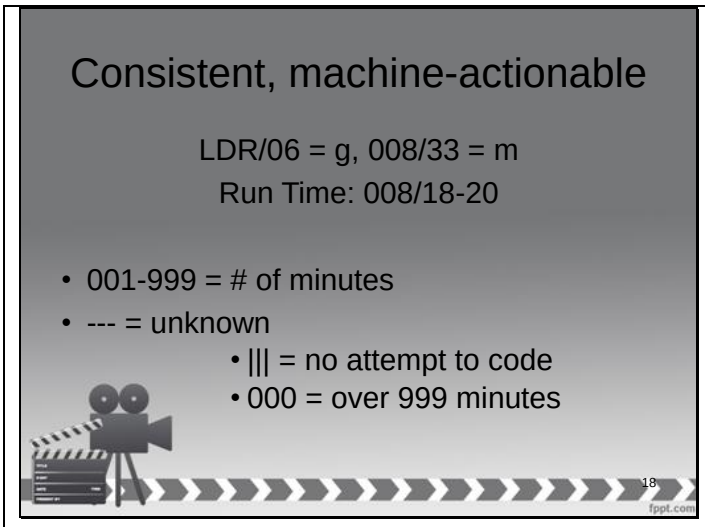
950 variants
523 appear only once

sd., col.	63.42%
	12.22%
sd., b&w	8.74%
sd., col. with b&w sequences	4.78%
sd., col. and b&w	1.41%

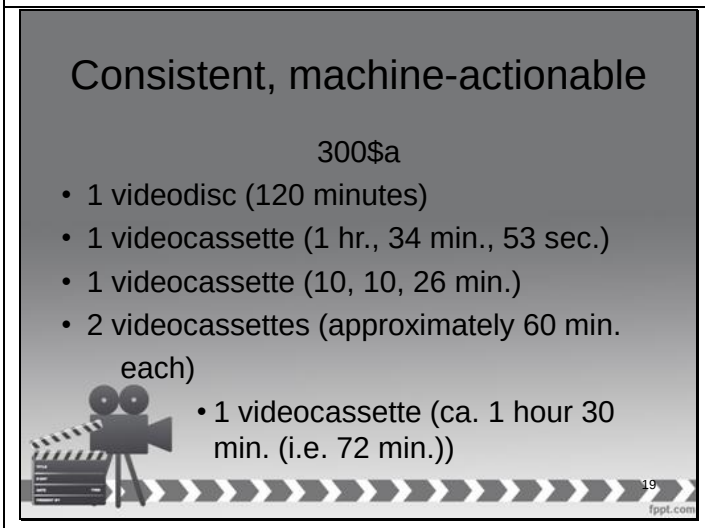
14

I looked at a sample of around 90,000 moving image records and almost all instances of 300$b fall into a narrow range of patterns. I found over 900 variations, although this could be reduced by normalizing the punctuation. Over half of these occur only once. 90% of the 300$b fields contain one of these strings or are blank so it would be pretty straightforward to write something that would do a better job with conversion to Bibframe.

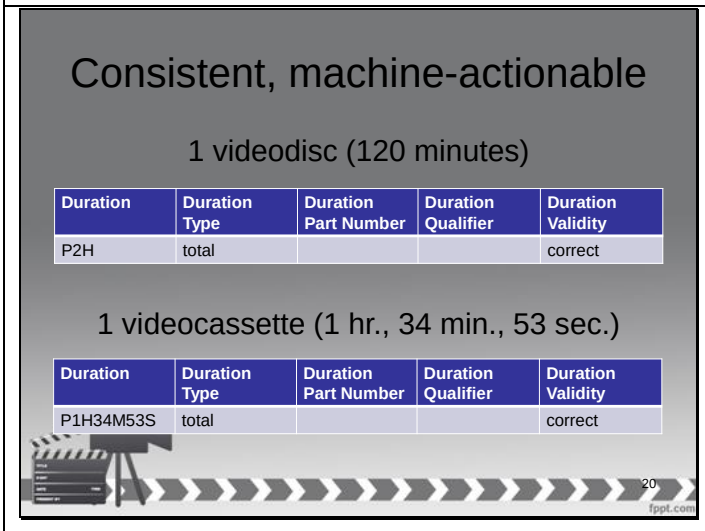
<table>
<tr>
<td> 
</td>
<td>There is a long, long tail. There are typos and weird punctuation. There are variant phrasings, such as “segments” instead of the more common “sequences.” Although color and sound are the most common types of information in 300$b for moving images, other information does appear. The cataloging rules do not provide guidance on recording information about tinting and toning, which increases the number of variant forms. And then there are the outliers that are mis-tagged or just make no sense.</td>
</tr>
<tr>
<td> 
</td>
<td>With the transition to a new data carrier, we have the opportunity to think about where we might want different or more detailed data. For example, most moving image language data is currently coded in these three subfields.</td>
</tr>
<tr>
<td> 
</td>
<td>If we really wanted to give users a clear picture of what they’re getting, it would be better to have more specific categories.</td>
</tr>
</table>



We currently don't have consistent, machine-actionable data in many places that we need it. Take duration or run time. It can be recorded in the 008, but only three characters are allotted so only times under 1000 minutes can be recorded. Times can only be recorded in one-minute increments, which is not helpful for very short clips.



More complex and exact information about duration can be recorded in 300$a, but with all the drawbacks of text strings.



As part of a project I'm working on, I'm trying to convert information from MARC into standardized forms that can be used for analysis and comparison. These five pieces of information enable almost all the variations on duration in my dataset to be normalized in a consistent manner. The time is recorded using a standardized method.

Consistent, machine-actionable

1 videocassette (10, 10, 26 min.)

Duration	Duration Type	Duration Part Number	Duration Qualifier	Duration Validity
P46M	total			correct
P10M	part	1		correct
P10M	part	2		correct
P26M	part	3		correct

21

Oftentimes, the duration of both the whole and the parts is recorded and this can be accounted for.

Consistent, machine-actionable

2 videocassettes
(approximately 60 min. each)

Duration	Duration Type	Duration Part Number	Duration Qualifier	Duration Validity
P2H	total		approximately	correct
P1H	part	1	approximately	correct
P1H	part	2	approximately	correct

22

Modifiers such as approximately and over can be added where appropriate.

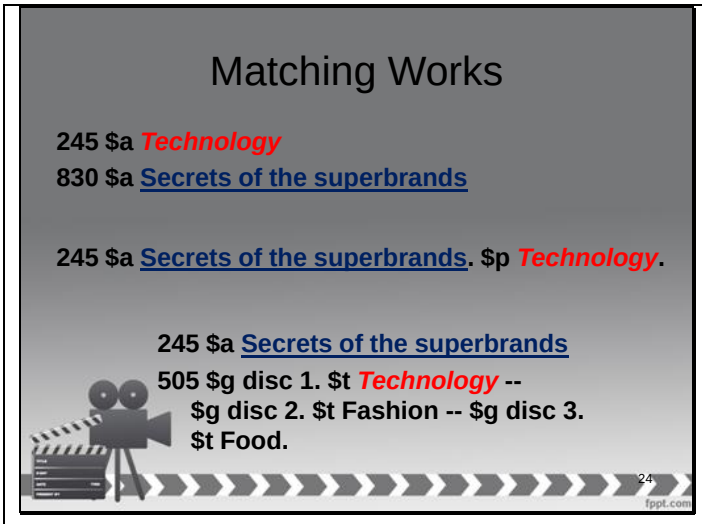
Consistent, machine-actionable

1 videocassette (ca. 1 hour 30 min.
(i.e. 72 min.))

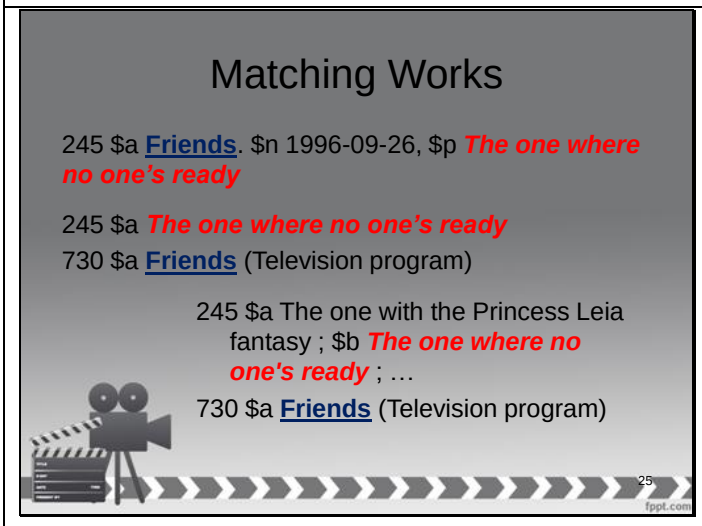
Duration	Duration Type	Duration Part Number	Duration Qualifier	Duration Validity
P1H30M	total		approximately	incorrect
P1H12M	total			correct

23

In some cases, the cataloger corrects the time stated on the piece and this can also be accounted for. RDA is at least interested in supporting more machine-actionable data and has been investigating how to do this. I am not sure that Bibframe is considering this angle as much as it should.

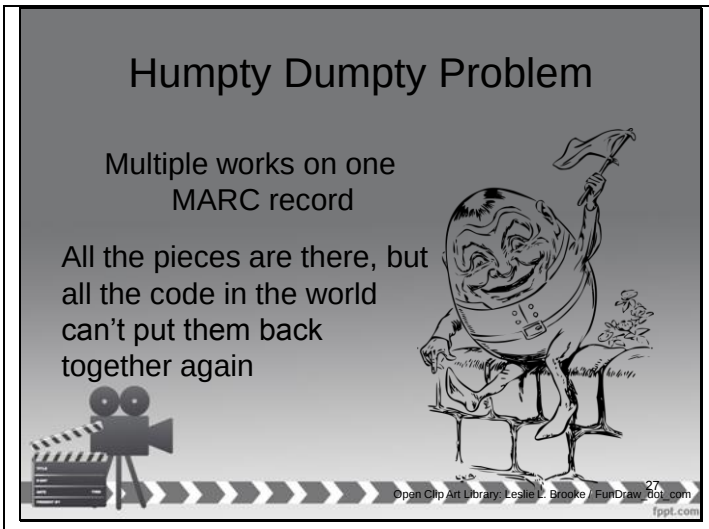
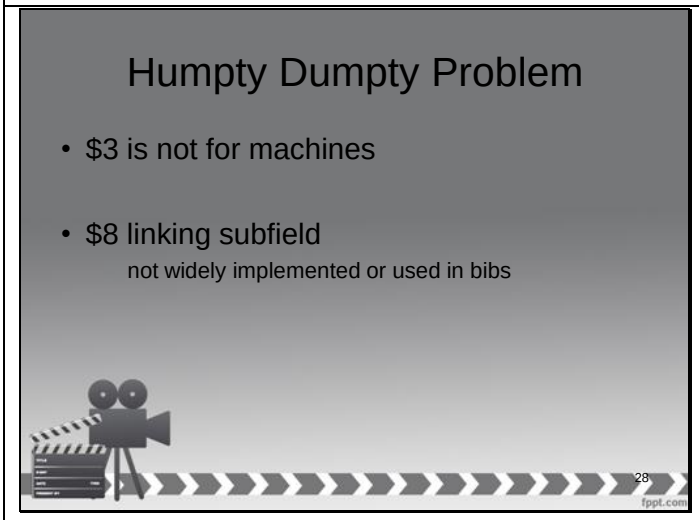
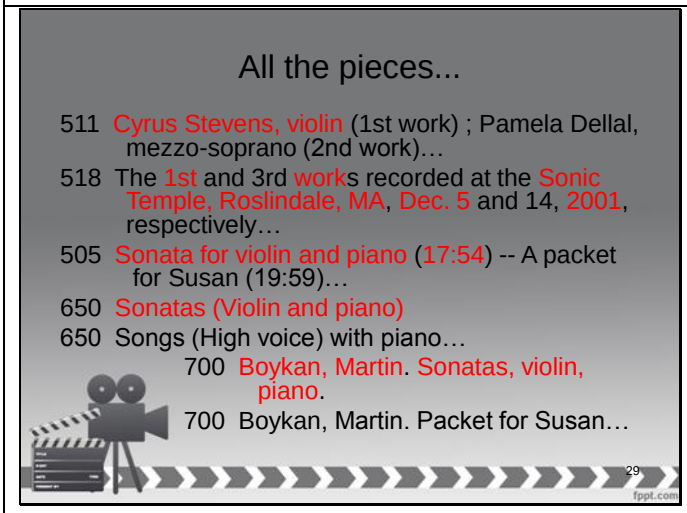


In MARC, the same film or program can be described in more than one way. It would be helpful if all the variations could be identified so that they could be presented to the user as equivalent. OCLC tries to group "works," but, for various reasons, their works are not the same as FRBR works. It is much more difficult to identify all the instances of a true FRBR work. These are three different ways that the Technology episode of Secrets of the Super Brands might appear. The need to manipulate and compare different forms of titles is a strong reason not to reduce titles to flat strings.



Here are several ways that the title of a particular episode of Friends might appear.

And here are even more variations.

<table>
<tr>
<td> 
</td>
<td>Finally, I want to talk about one of my major frustrations with MARC, which I'm not sure that Bibframe is resolving. I think of this as the Humpty Dumpty problem. Many videos contain multiple works, such as a collection of animated shorts. Various things can be said about each short, such as who the director is, when it was made, what language it's in, and so on. These things are usually in different MARC fields. Combining the director of one film with the creation date of another in search results or facets misleads the user.</td>
</tr>
<tr>
<td> 
</td>
<td>Sometimes subfield 3 is used to try to connect these pieces of information, but the free text used in $3 is not very machine friendly. Subfield 8 is designed for this purpose, but I don't know of any systems that implement it for bibliographic records. Nor does $8 help sort out situations where information about more than one work is combined in a single field.</td>
</tr>
<tr>
<td> 
</td>
<td>This is a music example, but the same principle applies to videos. A person looking at a record can usually untangle which pieces of information go with what, but it's hard to imagine how an algorithm could do so.</td>
</tr>
</table>