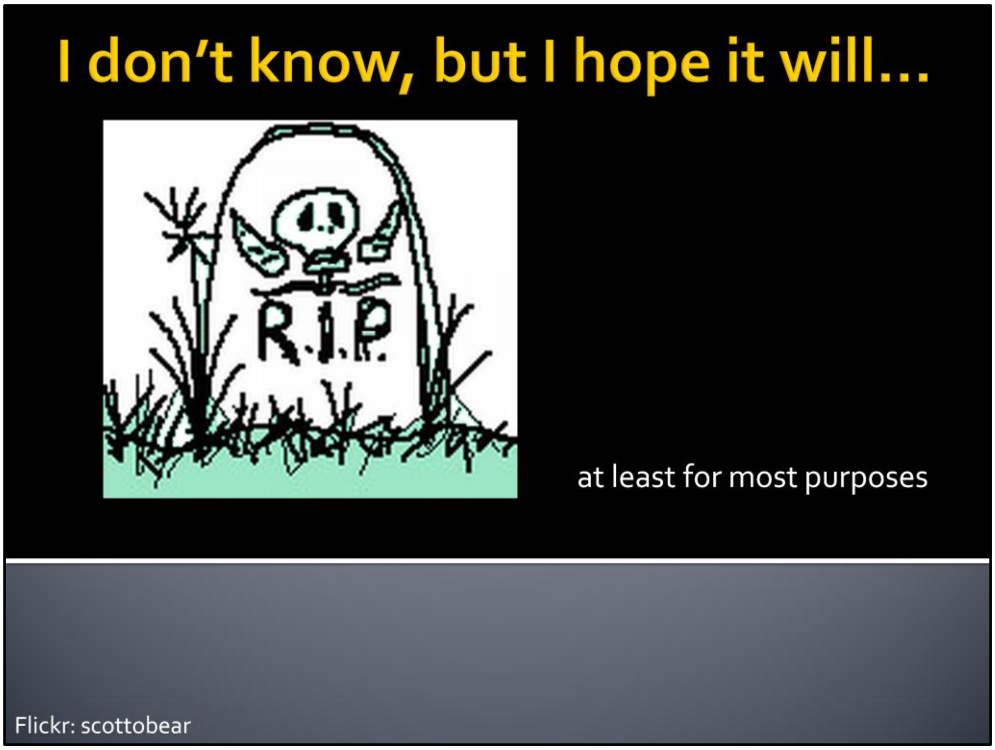
Will RDA Kill MARC?

Kelley McGrath
University of Oregon
kelleym@uoregon.edu

MARC Formats Interest Group
ALA Midwinter 2011

Flickr: Ashimjara

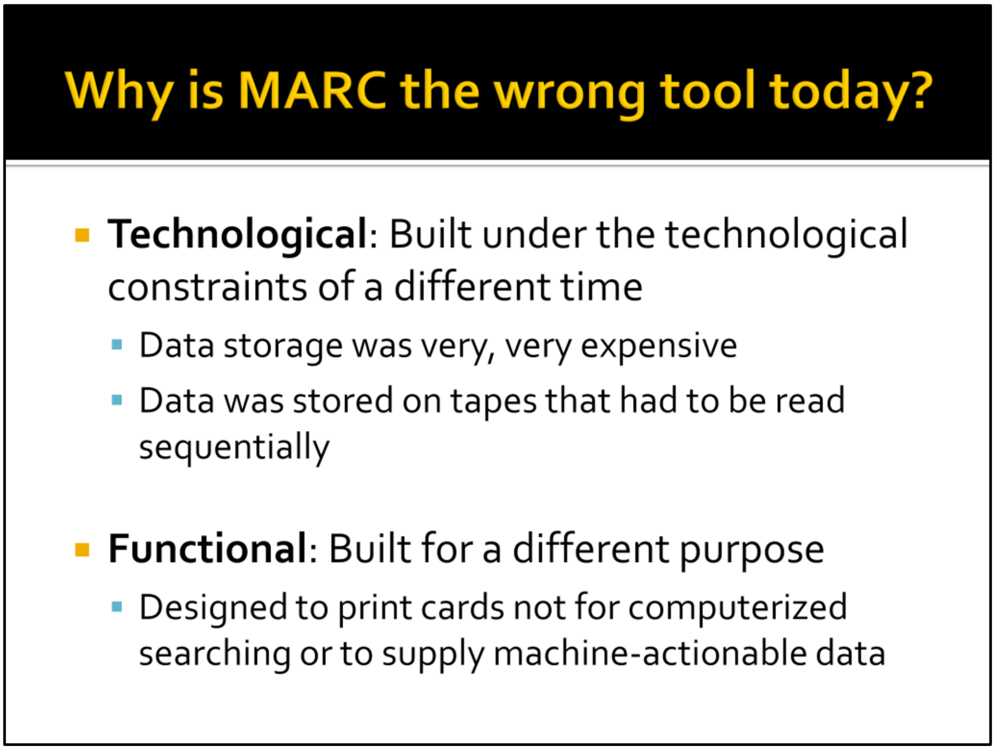


I don’t know, but I hope it will…

at least for most purposes

Flickr: scottobear

Why do I say this?
Not because I hate MARC, but because I think our needs have changed in ways that are difficult or impossible for MARC to fulfill. This was true even before RDA, but the conflict between the aspirations of RDA and the limitations of MARC may finally tip the balance in favor of changing data formats despite the inevitable pain and cost.



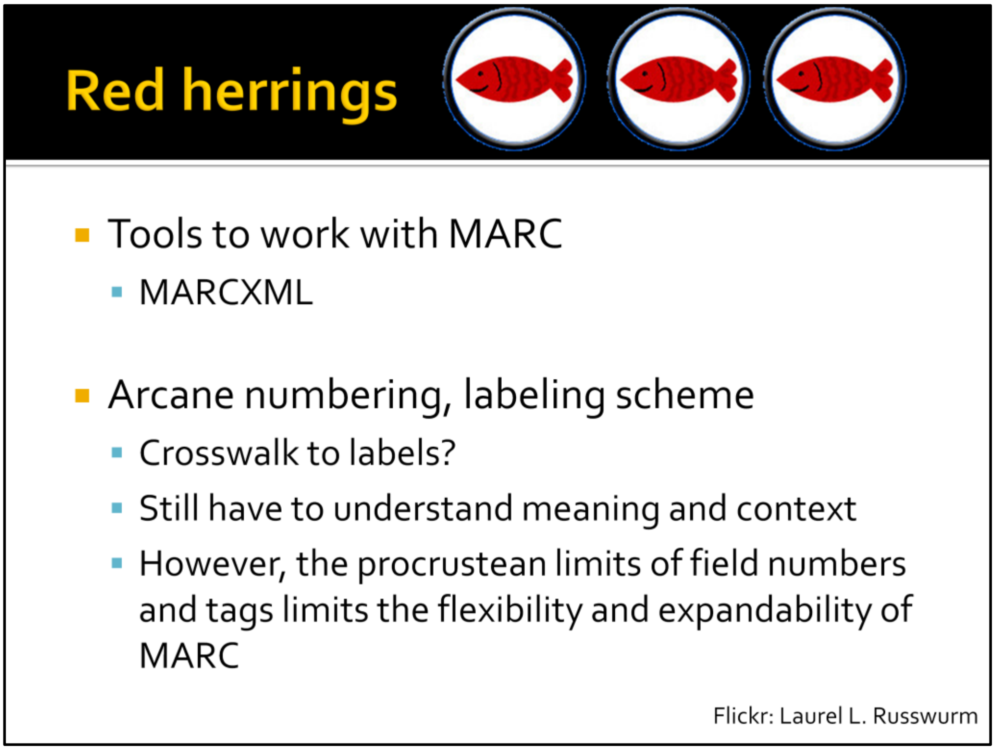
Why is MARC the wrong tool today?

- **Technological**: Built under the technological constraints of a different time
  - Data storage was very, very expensive
  - Data was stored on tapes that had to be read sequentially
- **Functional**: Built for a different purpose
  - Designed to print cards not for computerized searching or to supply machine-actionable data

MARC was a brilliant, visionary solution in its day, but it was conceived in different times when the limits of what technology could do were much more confining. MARC was designed for an environment where data storage was very, very expensive and data was read linearly from tapes.

In the 1960s our purposes for computerized catalog data were very different. MARC was designed to print cards not to be a searchable format for end users. Henriette Avram had the foresight to include a limited amount of machine-friendly data in the form of fixed fields, but most of the data in the MARC record is not designed for direct comprehension by a computer. Jason Thomale had a nice article in the Code4Lib Journal about the distinction between the kind of textual mark-up of catalog cards that he sees MARC doing and real computerized bibliographic data.

To stick with MARC and its constraints in 2011 is like trying to fight with both hands tied behind our backs.



Red herrings

- Tools to work with MARC
  - MARCXML
- Arcane numbering, labeling scheme
  - Crosswalk to labels?
  - Still have to understand meaning and context
  - However, the procrustean limits of field numbers and tags limits the flexibility and expandability of MARC

Flickr: Laurel L. Russwurm

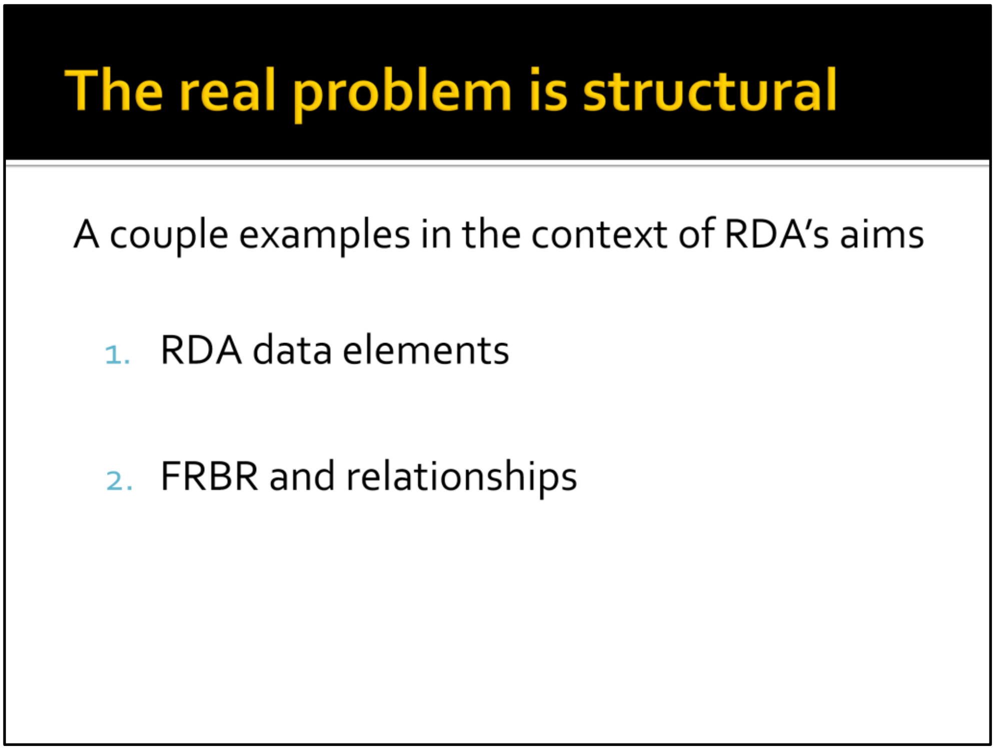
That said, I do think MARC is sometimes unfairly maligned.

People complain that MARC cannot be manipulated with general purpose editing tools. It is true that the MARC format is used only by a very narrow group of organizations and there are few tools for manipulating it directly. But we have MARCXML and it's easy enough to convert MARC to MARCXML and have access to all those XML tools, at least as long as you don't have bad MARC data. Since MARCXML is a round-trippable conversion, it's not hard to convert data back to MARC.

People also complain about the arcane numbering, indicators and subfields and say it's not easy to learn to speak MARC. As a cataloger, I find MARC to be a convenient shorthand, but there was a time when I, too, found it impenetrable.

However, I don't see that it would be that hard to make a crosswalk from MARCXML to something with English language labels that a programmer could work with and then transform back into MARCXML. This doesn't really solve the underlying problem of communicating the constraints and culture around the meaning of the MARC tags to outsiders, but it solves the simpler problem of having to remember what the field numbers and subfield tags label in a basic sense.

However, there is an important way in which MARC's field and subfield system does hold it back. In library school, I learned about Dewey and the procrustean ten where everything had to fit into ten categories even if more or less would have been better. MARC's system of fields and subfields has a similar limitation in that there are only so many available for use. In some places, such as the 245 field, we have maxed out the available subfields and can't fit anything else in. Or sometimes there aren't gaps left where it’s possible to fit a new field in in a logical order.

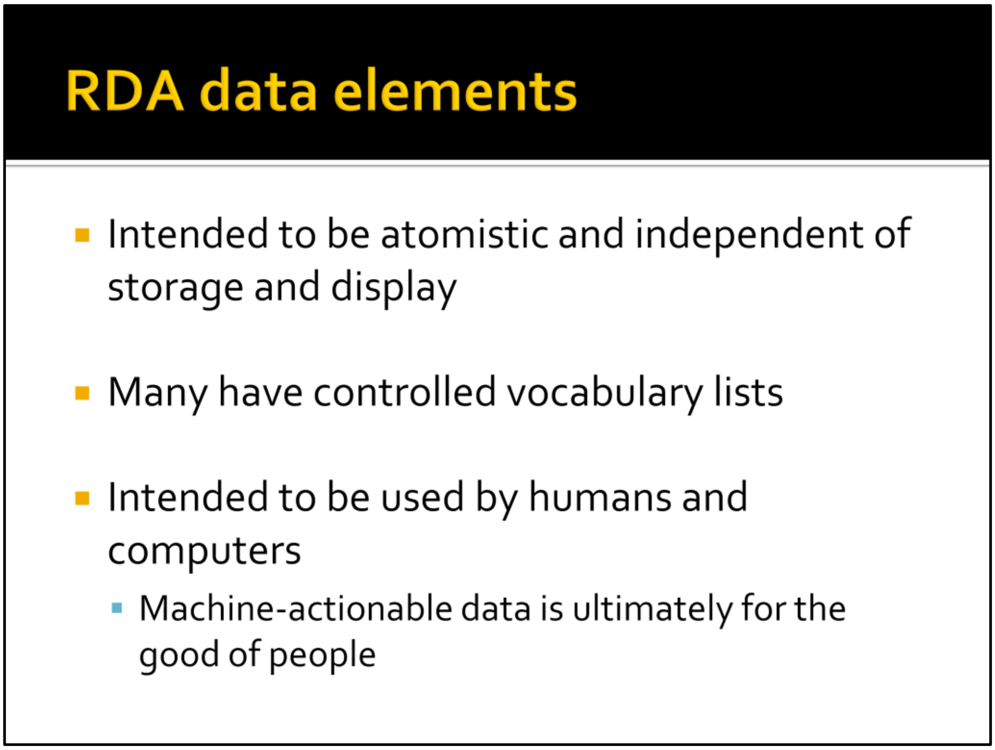


The real problem is structural

A couple examples in the context of RDA's aims

1. RDA data elements
2. FRBR and relationships

However, I think the XML and numbering complaints are red herrings and the most insurmountable trouble with MARC is structural. I'd like to talk about a couple of those limitations in the context of what RDA is trying to do.

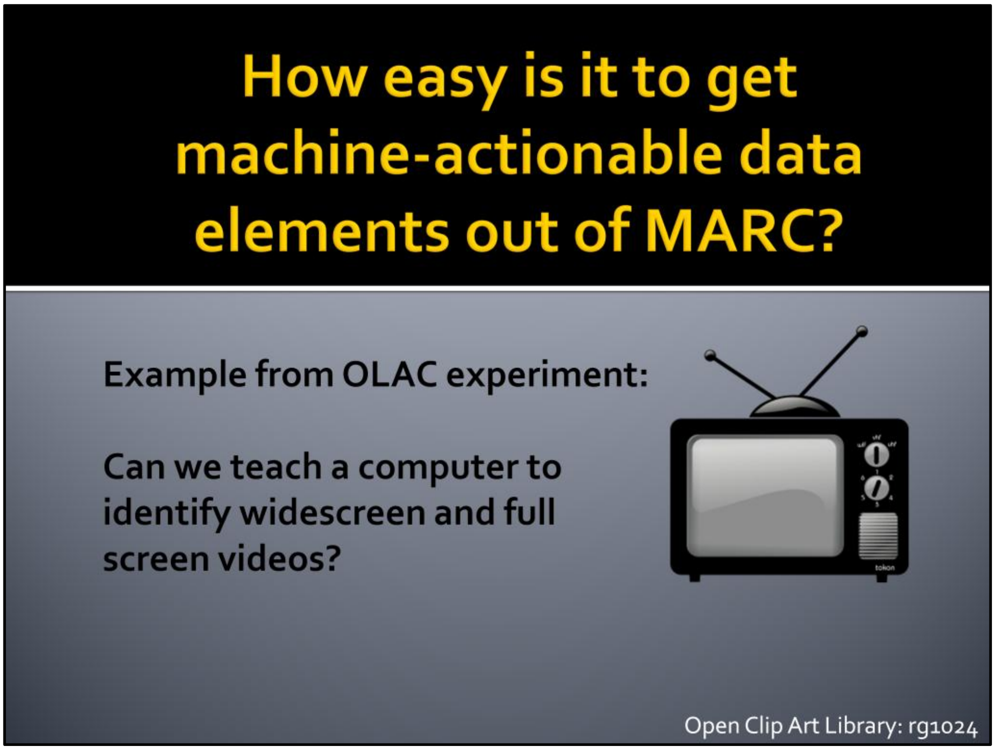


RDA data elements

- Intended to be atomistic and independent of storage and display
- Many have controlled vocabulary lists
- Intended to be used by humans and computers
  - Machine-actionable data is ultimately for the good of people

RDA defines individual data elements that are supposed to be independent of storage or display formats. RDA also tries to define atomistic elements, where each element contains only one type of information, rather than a conflation of different things, as is found in many MARC subfields. Just think about how many types of information can occur in 245 $c. Many of these elements have controlled vocabulary lists associated with them. It is hoped that these data elements will produce data that is in a form that computers can interpret and use to provide new and powerful services to help our users navigate the bibliographic universe.

Some people seem to have the impression that creating data that is machine-actionable is mutually exclusive with creating human-readable data. Nothing could be further from the truth. The whole point of creating data that is formatted so that a computer can use it is to be able to develop value-added ways of accessing and using bibliographic data for human consumption. The ultimate end is always helping human users, either directly or by making the behind-the-scenes work easier so we can do more with less time and money.



How easy is it to get machine-actionable data elements out of MARC?

Example from OLAC experiment:

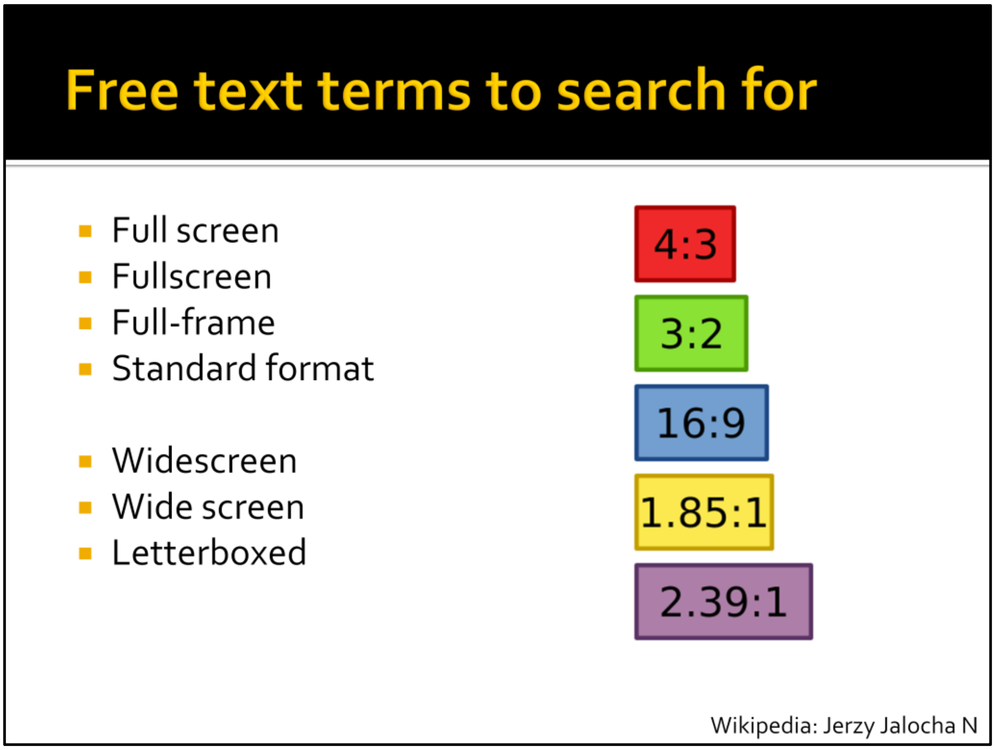
Can we teach a computer to identify widescreen and full screen videos?

Open Clip Art Library: rg1024

How easy is it to get discrete, machine-actionable data elements out of existing MARC records right now? In many cases, not all that easy.

I have been an active member of the A/V catalogers' group Online Audiovisual Catalogers, better known as OLAC, for many years. OLAC had a project related to improving access to moving images by taking advantage of the FRBR model. As part of this project, we investigated how easy it would be to automatically extract machine-comprehensible data for a sampling of moving image characteristics from existing MARC records. The results were written up in the Code4Lib Journal and in a somewhat fuller form on the OLAC website, but right now I just want to look at a single example from that project.

One of the things that we tried to identify is whether a video is widescreen or full screen.



Free text terms to search for

- Full screen
- Fullscreen
- Full-frame
- Standard format

- Widescreen
- Wide screen
- Letterboxed

4:3

3:2

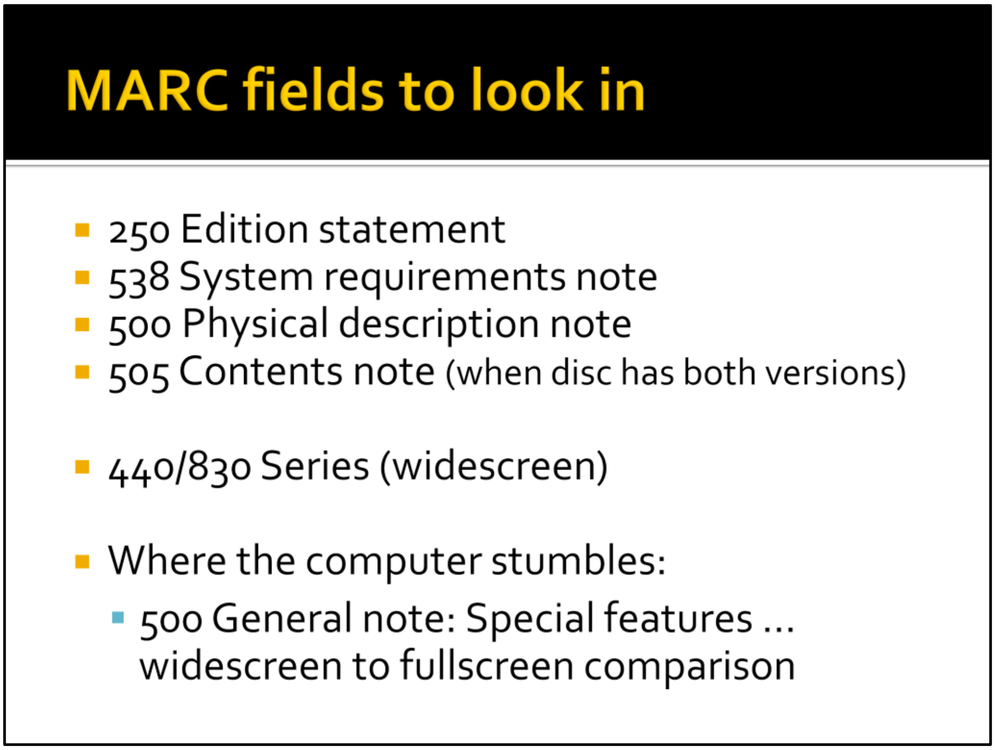
16:9

1.85:1

2.39:1

Wikipedia: Jerzy Jalocha N

Since this information is only recorded in free text fields in the MARC record, first we had to come up with a list of variant forms and spellings to look for, such as those listed on this slide.



MARC fields to look in

- 250 Edition statement
- 538 System requirements note
- 500 Physical description note
- 505 Contents note (when disc has both versions)
- 440/830 Series (widescreen)
- Where the computer stumbles:
  - 500 General note: Special features … widescreen to fullscreen comparison

Then we had to come up with a list of fields in which these terms were likely to occur. We anticipated 250, 538, 500 and 505.

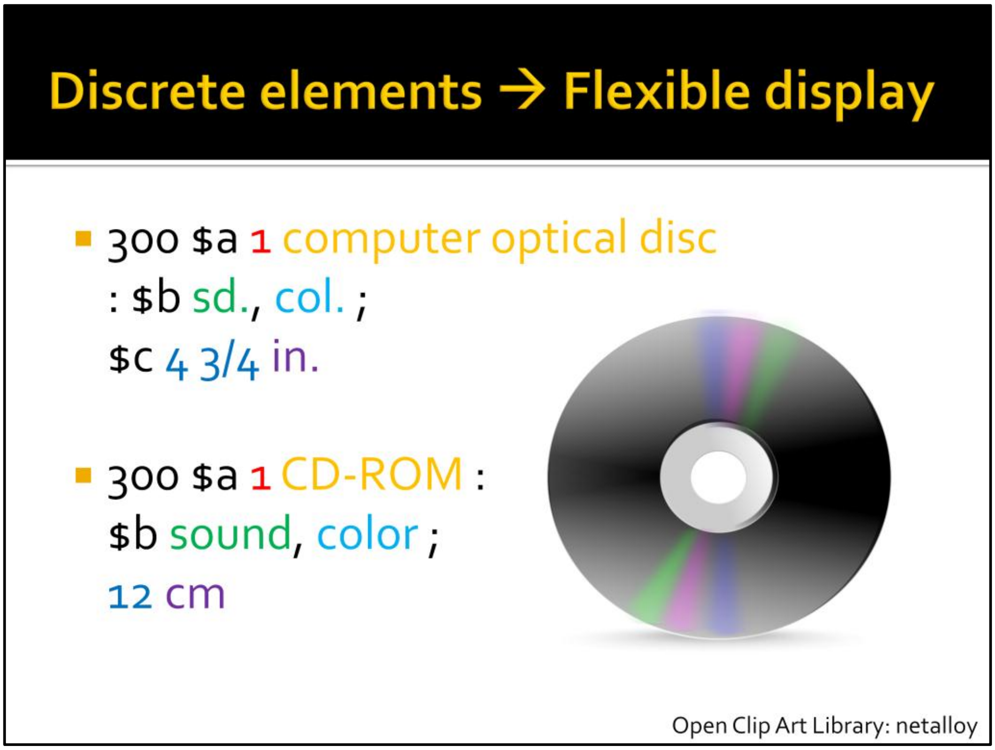
An analysis of the data after the fact revealed that widescreen also sometimes occurs in series statements in 440 or 830.

We also encountered some statements that, although clear to the human reader, were misinterpreted by the computer.

For example, one record had a statement in a 500 note field that it included a “widescreen to fullscreen comparison.” This led our program to conclude that the DVD included both versions, which does happen. However, in this case, the complete film was presented only in widescreen and this note refers to a special feature.

This process is not all that straightforward and this is just one minor element that occurs in a single format. Multiply that by all the types of information for which we'd like to have computer-comprehensible data and you begin to see the problem.

Sure, we could change the way we record data in MARC to more clearly define individual pieces of data. In fact, there's a MARBI discussion paper right now proposing to add new MARC fields and subfields to separately record carrier attributes (http://www.loc.gov/marc/marbi/2011/2011-dp04.html). Carrier attributes are a bunch of disparate things that currently go in the physical description in 300$b or in notes in undifferentiated form. For example, the base material on which pictures are mounted and DVD region codes are carrier attributes in RDA. However, to do this sort of thing comprehensively and effectively would be a big change from the way we do things now and if we're going to make that much effort, we might as well start over with another, more modern data format.



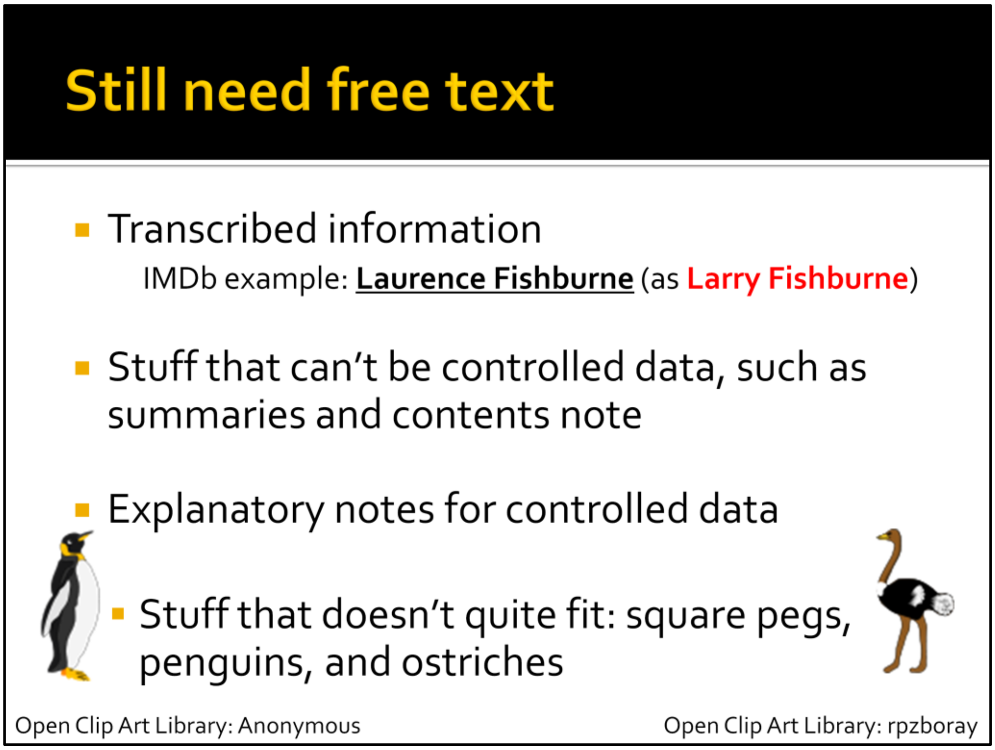
Discrete elements → Flexible display

- 300 $a 1 computer optical disc
  : $b sd., col. ;
  $c 4 3/4 in.
- 300 $a 1 CD-ROM :
  $b sound, color ;
  12 cm

Open Clip Art Library: netalloy

This doesn't mean that data couldn't be put back together again in a traditional display. Recording data as discrete elements gives us much more flexibility in designing displays of all sorts. In this example, if each of the differently-colored data types were a different element, it would be much easier to customize displays. We could then share cataloging data more easily in spite of our varying display preferences.

It’s easier to generate human-comprehensible data from machine-comprehensible data than the other way around.



Still need free text

- Transcribed information

  IMDb example: **Laurence Fishburne** (as **Larry Fishburne**)

- Stuff that can't be controlled data, such as summaries and contents note

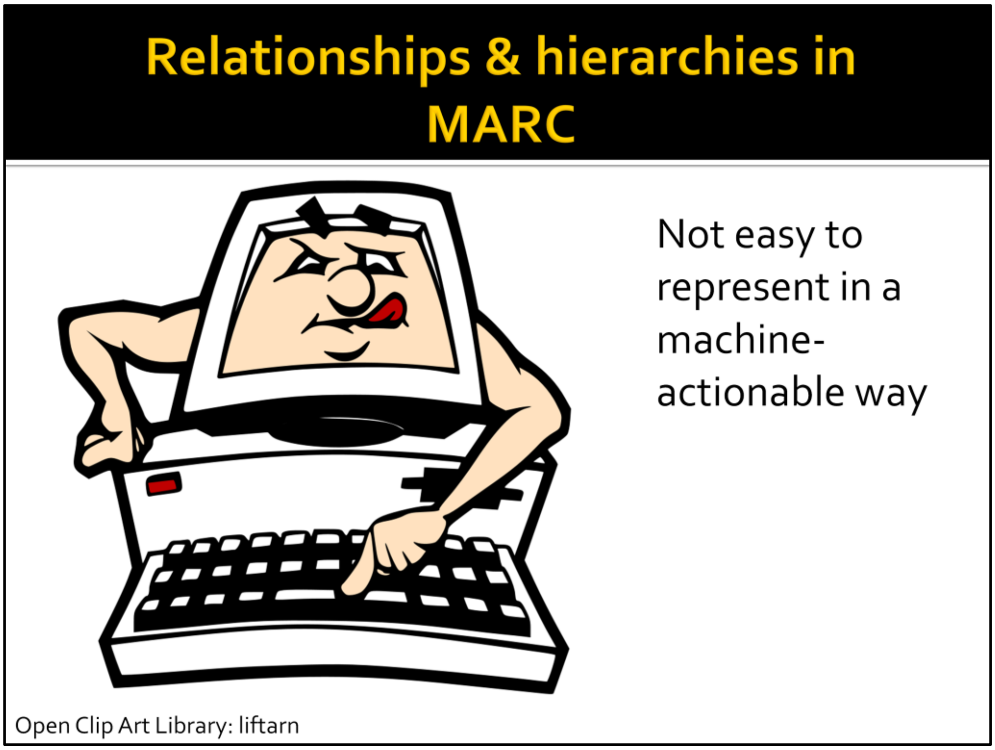
- Explanatory notes for controlled data

  - Stuff that doesn't quite fit: square pegs, penguins, and ostriches

Open Clip Art Library: Anonymous

Open Clip Art Library: rpzboray

This doesn't mean everything can be crammed into pull-down lists or that there is no place for free text. We still need transcribed data in some fields. This example shows how the Internet Movie Database displays a transcribed name from the title frames (Larry Fishburne in red) next to their authorized form of Laurence Fishburne. Some types of data, such as summaries and contents notes, do not reduce to controlled vocabulary. We will also still need notes to provide additional information about things that are captured by controlled vocabularies, as well as the inevitable things that don't fit into the existing category scheme very well.

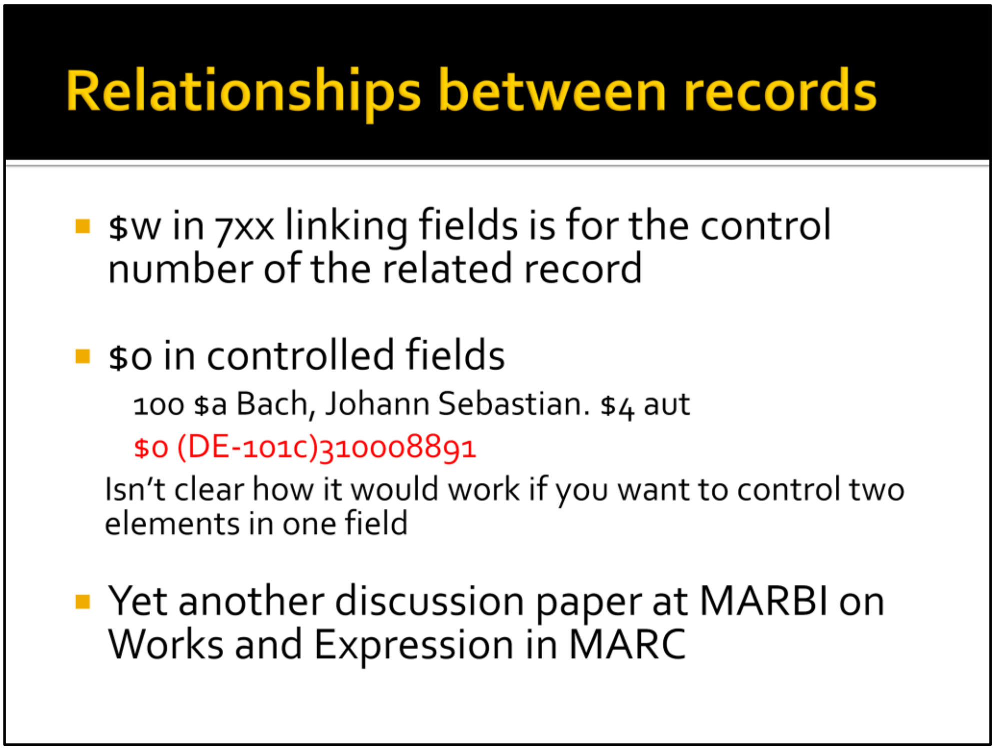


Relationships & hierarchies in MARC

Not easy to represent in a machine-actionable way

Open Clip Art Library: liftarn

RDA is also trying to follow the FRBR model and record relationships between different pieces of bibliographic information more effectively in a way that computers can use. Unfortunately, it's not easy to represent relationships and hierarchies in MARC.



Relationships between records

- $w in 7xx linking fields is for the control number of the related record
- $0 in controlled fields

  100 $a Bach, Johann Sebastian. $4 aut

  $0 (DE-101c)310008891

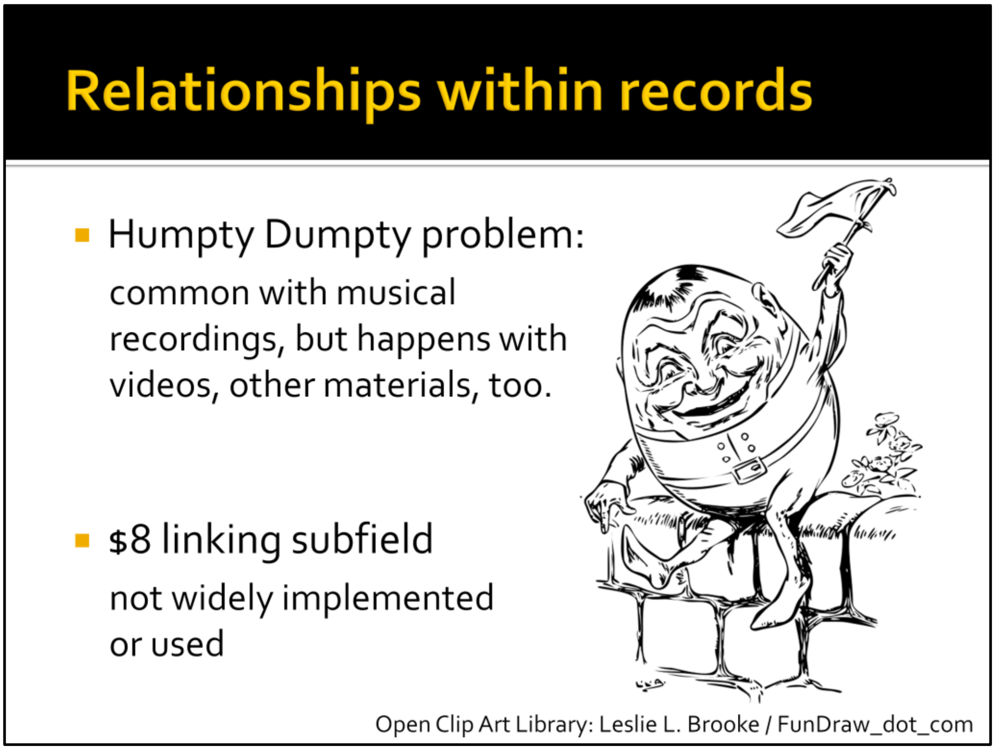
  Isn't clear how it would work if you want to control two elements in one field
- Yet another discussion paper at MARBI on Works and Expression in MARC

MARC wasn't designed to support machine-readable links between records, although there are a few isolated options for creating these links.

For example some of the 7xx linking fields include subfields that contain the control number of a related record, although few systems seem to take advantage of these links

The Germans have introduced $0 for identifiers for some controlled fields in MARC21. $0 links to authority records. This solution is not usable in all situations, such as if you needed to identify two different pieces of information in a single field.

There's yet another discussion paper at MARBI this conference about how to implement FRBR work and expression records in MARC, which suggests that this isn't a straightforward process.



Relationships within records

- Humpty Dumpty problem:

  common with musical recordings, but happens with videos, other materials, too.

- $8 linking subfield

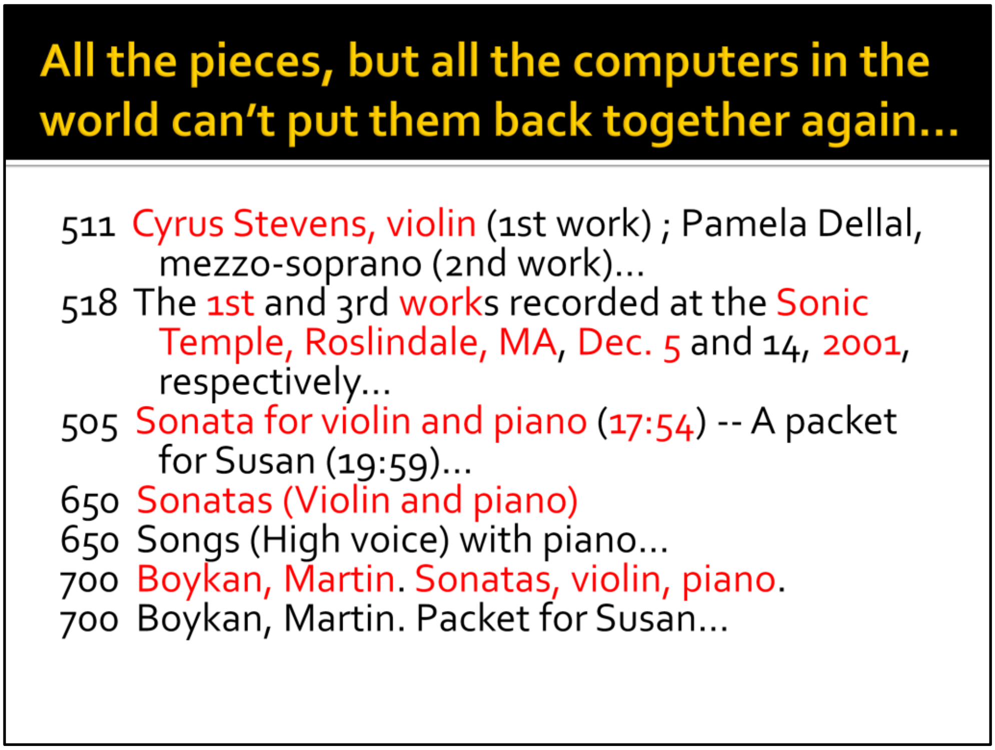
  not widely implemented or used

Open Clip Art Library: Leslie L. Brooke / FunDraw_dot_com

There is also no easy way to represent relationships or hierarchies within a single MARC record. It is not uncommon to have more than one work on a single bibliographic record. In this situation, there may be pieces of information about a given work scattered throughout various fields with no explicit, machine-comprehensible connections among them.

I think of this as the Humpty Dumpty problem. All the pieces might be there, but all the computers in the world can't put them back together again. This problem occurs with DVDs and many types of records, but is particularly acute for musical recordings.

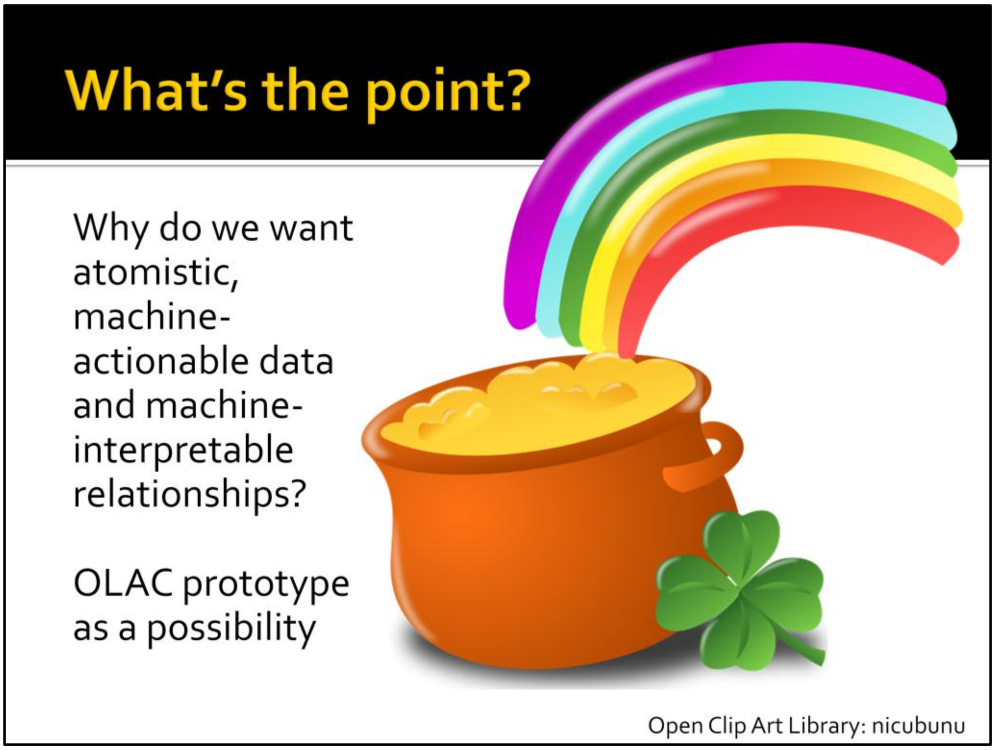
There is one option currently in MARC for linking fields within a single record. Many of you might not know this, but the MARC bibliographic record has a $8, which can be used to link whole fields. However, $8 is not supported by any systems that I know of nor does it seem to be widely used.



All the pieces, but all the computers in the world can't put them back together again…

511 Cyrus Stevens, violin (1st work) ; Pamela Dellal, mezzo-soprano (2nd work)…
518 The 1st and 3rd works recorded at the Sonic Temple, Roslindale, MA, Dec. 5 and 14, 2001, respectively…
505 Sonata for violin and piano (17:54) -- A packet for Susan (19:59)…
650 Sonatas (Violin and piano)
650 Songs (High voice) with piano…
700 Boykan, Martin. Sonatas, violin, piano.
700 Boykan, Martin. Packet for Susan…

Here's a somewhat random sample that shows in red the information related to the first musical piece. If you assume that everything is in the same sequential order, you might be able to parse some of it. However, I doubt you can count on that and I think it would be pretty challenging to get the computer to understand what respectively means in this 518 field. In many cases, even if everything is spelled right and punctuated correctly, there's still no way for the computer to connect all the dots



What's the point?

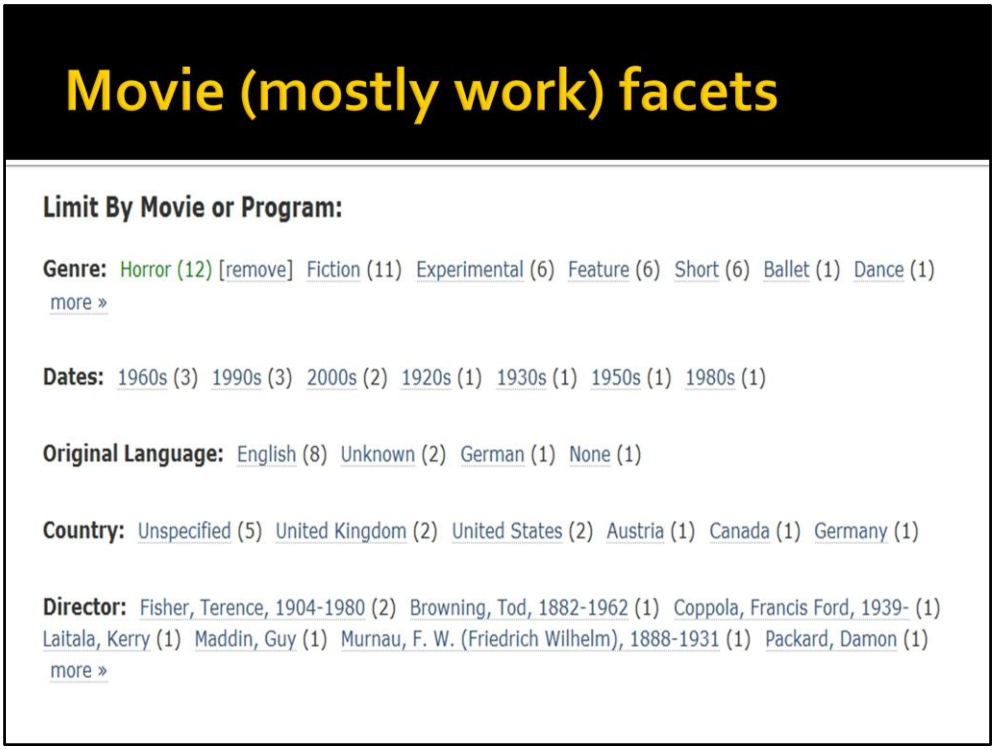
Why do we want atomistic, machine-actionable data and machine-interpretable relationships?

OLAC prototype as a possibility

Open Clip Art Library: nicubunu

Why do we want to be able to do this? What's the point of showing more relationships and providing discrete data elements, and doing both in a machine-actionable form?

I'd like to talk a little bit about a project that is close to my heart and that I think demonstrates the potential of having machine-actionable data and using the FRBR model as a framework. OLAC recently sponsored the development of a prototype discovery interface for moving images that uses both of these concepts.



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)

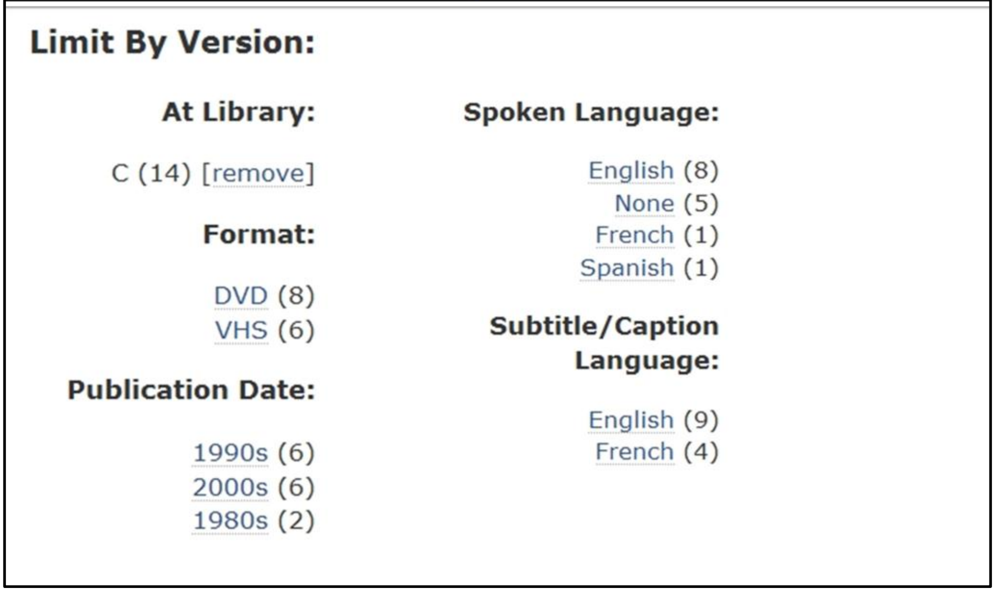
Country: Unspecified (5) United Kingdom (2) United States (2) Austria (1) Canada (1) Germany (1)

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 »

Keep in mind that this is a prototype designed to demonstrate a concept and includes a limited number of records and fields. The interface uses FRBR as an organizing principle and facets based on machine-actionable data to enable users to explore and navigate.

Here are some facets for movies or works with horror selected.

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)

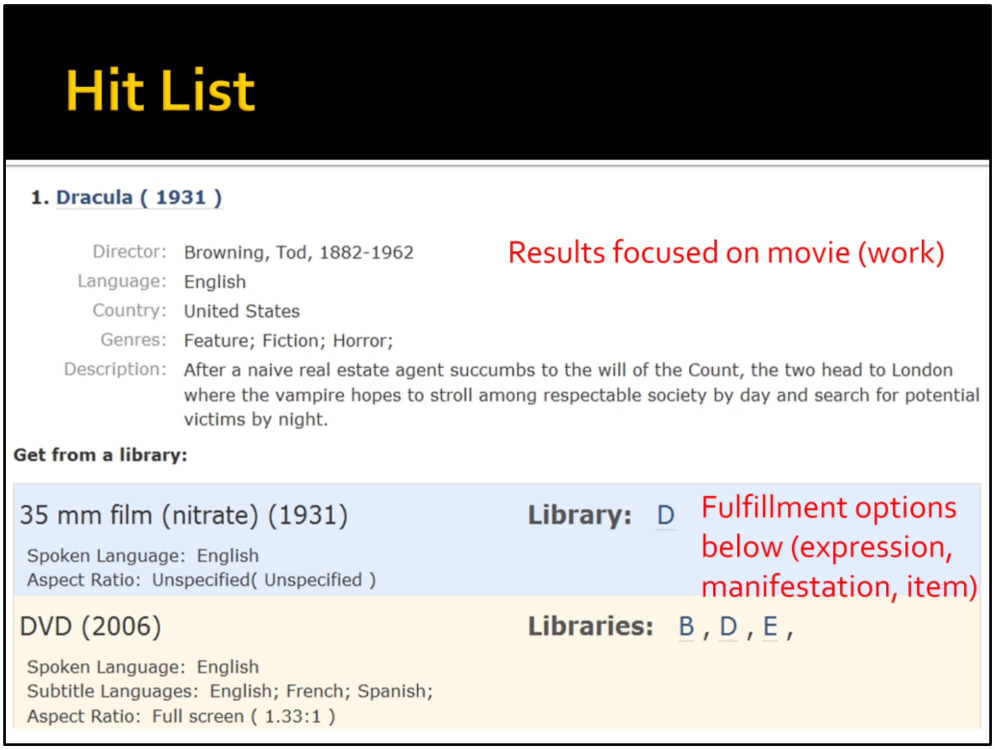
Spoken Language:

English (8)
None (5)
French (1)
Spanish (1)

Subtitle/Caption Language:

English (9)
French (4)

Here are some facets for what we're calling versions that include information from FRBR expressions, manifestations and items. Users can limit to items available at their local library in their preferred format with language options that make sense for them.



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:

35 mm film (nitrate) (1931) — **Library:** D

Spoken Language: English
Aspect Ratio: Unspecified(Unspecified)

Fulfillment options below (expression, manifestation, item)

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

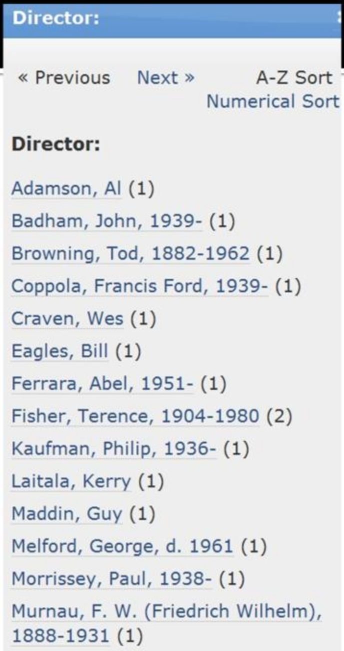
Spoken Language: English
Subtitle Languages: English; French; Spanish;
Aspect Ratio: Full screen (1.33:1)

Here is an example result set. The display is focused on the movie or work and offers a number of fulfillment options below.

Slicing and dicing data:

horror film directors

Open Clip Art Library: Oscar S.R. / miutopia



Director:

« Previous Next » A-Z Sort
Numerical Sort

Director:

Adamson, Al (1)
Badham, John, 1939- (1)
Browning, Tod, 1882-1962 (1)
Coppola, Francis Ford, 1939- (1)
Craven, Wes (1)
Eagles, Bill (1)
Ferrara, Abel, 1951- (1)
Fisher, Terence, 1904-1980 (2)
Kaufman, Philip, 1936- (1)
Laitala, Kerry (1)
Maddin, Guy (1)
Melford, George, d. 1961 (1)
Morrissey, Paul, 1938- (1)
Murnau, F. W. (Friedrich Wilhelm), 1888-1931 (1)

The interface allows users to do all kinds of slicing and dicing of the data that aren't possible with MARC records. For example, once the user has picked horror, he or she can get an alphabetical list of horror directors in the database. You could create many different views of the data, such as a list of directors of horror comedies or 1930s horror films. Unlike a typical library catalog browse list, there are no cross-references in the demo, but that is fixable.

Can we get there from here?

Challenges for development

Who?

With what $?

Wikimedia Commons; Open Clip Art Library: SRD

Open Clip Art Library: johnny_automatic

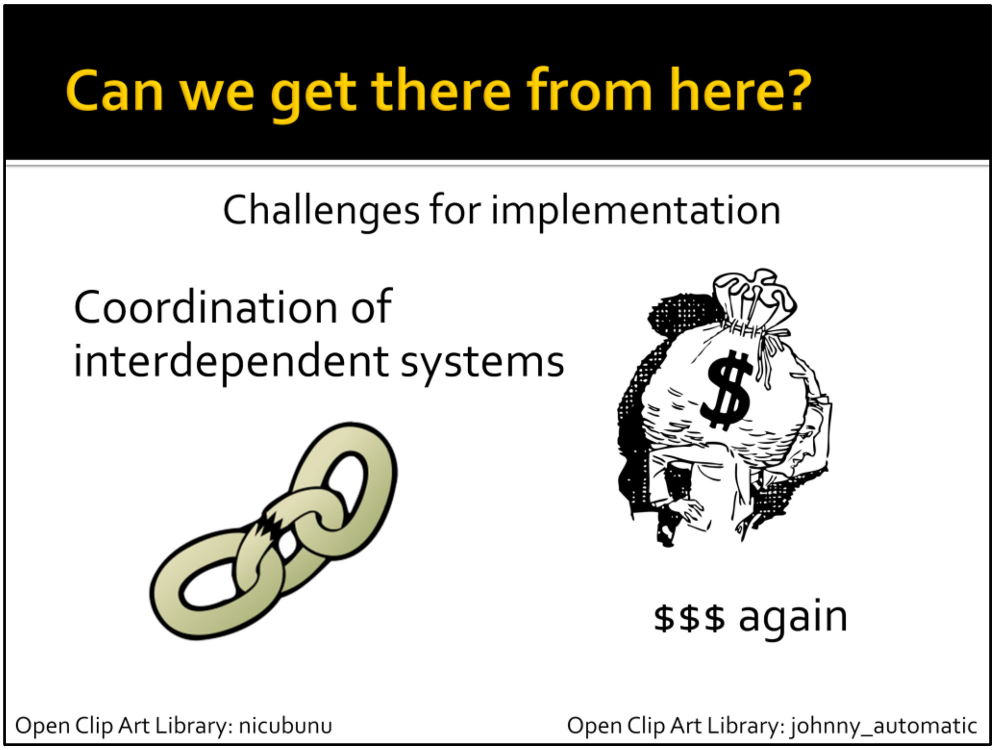
However, even if we agree that we need to change our data format and agree on the general direction that we need to go, many challenges remain.

Development is a challenge.

Let's start with money. Who is going to pay for this? The library world always seems to be underfunded and the situation is not improving. RDA, whatever you think of it, cost a lot of money. RDA was financed by the major Anglo-American national libraries and library organizations. They want a return on their investment so RDA is locked down behind a pricey subscription pay wall. It’s out of reach of many of its potential users. MARC21 has always been a freely-available standard and it’s hard to imagine that the vision that we have of a brave new data format will work if that new data standard, too, is not out there for all interested parties to use. The Library of Congress has bankrolled MARC21, but do they have the resources and political will to pay for what comes next?

Who is going to do all this work, especially with little or no pay? RDA was developed with a great deal more transparency and stakeholder input than any previous cataloging standard. This was wonderful in many ways, but I wonder if there is a cost to losing the focused vision you get when a single person or a few individuals create something. In the end, many things in RDA seem to be the result of compromises among competing interests, which has left a standard that satisfies nobody.

Some have suggested that more people should be involved in this conversation than just the traditional library and MARBI stakeholders. Who needs or wants input into this discussion?



Can we get there from here?

Challenges for implementation

Coordination of interdependent systems

$$$ again

Open Clip Art Library: nicubunu

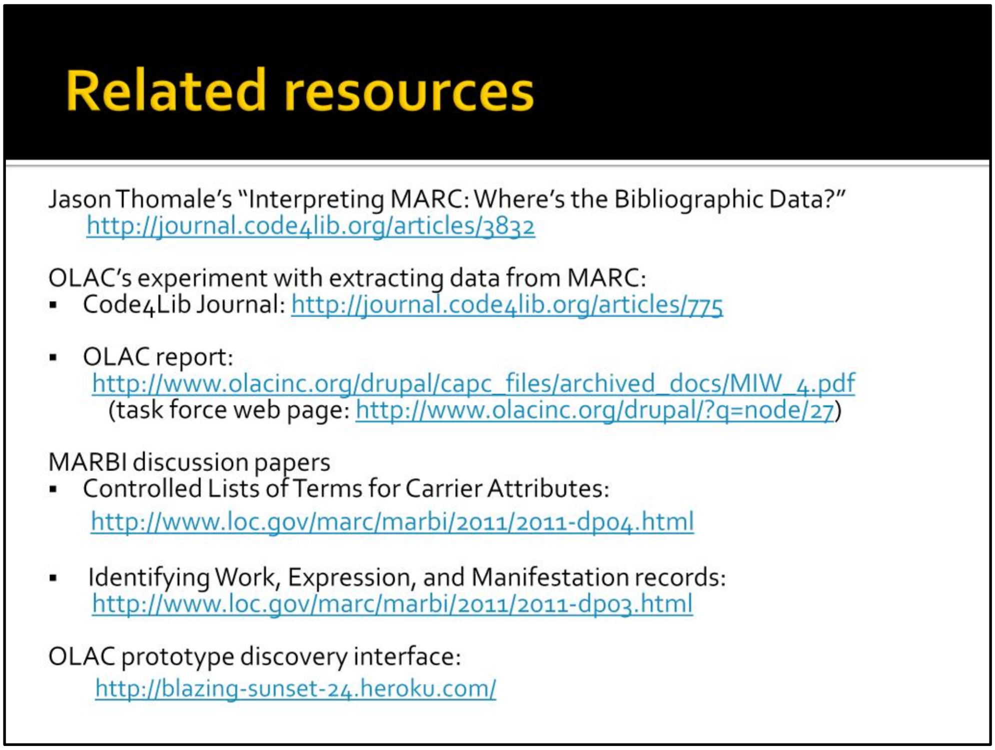
Open Clip Art Library: johnny_automatic

There are also challenges for implementation

I sometimes think that the reasons we're still using MARC in 2011 have much in common with why the banking world was still using software that was jerry-rigged on top of code written in the 1960s in the late nineties. The banking system was complicated, interconnected, and relied on by a diverse group of users who would have to transition in unison so it took a powerful outside force like Y2K for the benefits of getting rid of that old code to outweigh the costs and inertia. The library world needs its own Y2K

There are a lot of interlocking pieces of the library world that rely on MARC. How do we shift all those pieces to a new format, especially since it is unlikely that all organizations will transition at the same time?

If development of a new standard will cost money, implementation will cost far more. Just think of all the applications that will have to be updated and all the people who will have to be trained.



Related resources

Jason Thomale's "Interpreting MARC: Where's the Bibliographic Data?"
http://journal.code4lib.org/articles/3832

OLAC's experiment with extracting data from MARC:
- Code4Lib Journal: http://journal.code4lib.org/articles/775
- OLAC report:
  http://www.olacinc.org/drupal/capc_files/archived_docs/MIW_4.pdf
  (task force web page: http://www.olacinc.org/drupal/?q=node/27)

MARBI discussion papers
- Controlled Lists of Terms for Carrier Attributes:
  http://www.loc.gov/marc/marbi/2011/2011-dp04.html
- Identifying Work, Expression, and Manifestation records:
  http://www.loc.gov/marc/marbi/2011/2011-dp03.html

OLAC prototype discovery interface:
http://blazing-sunset-24.heroku.com/