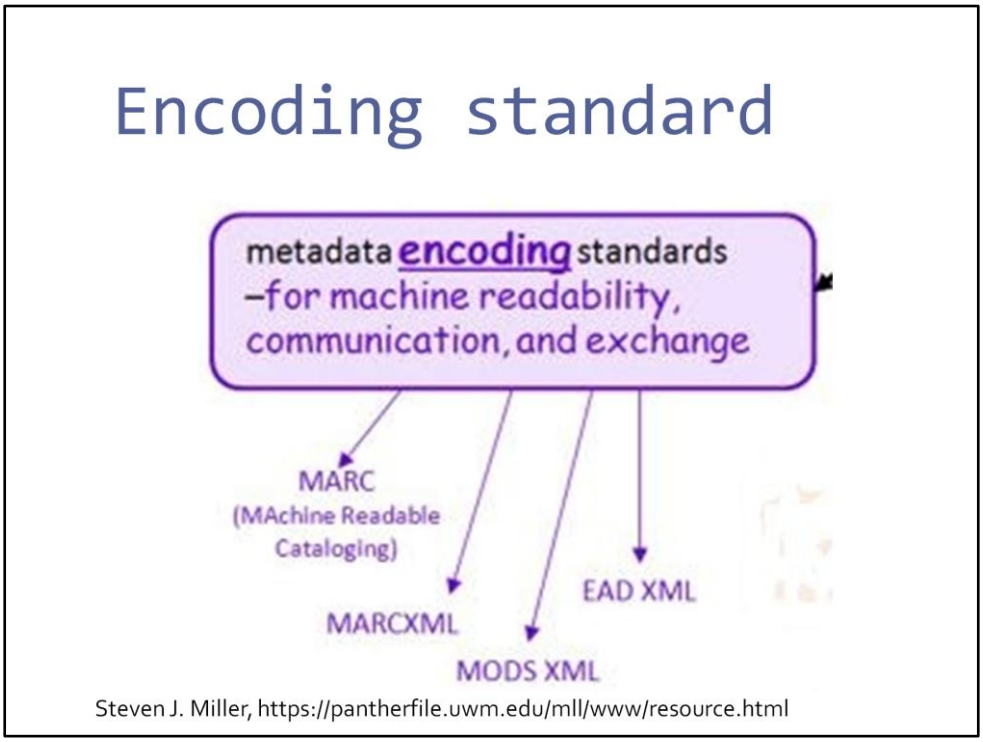
Kelley McGrath
University of Oregon
kelleym@uoregon.edu

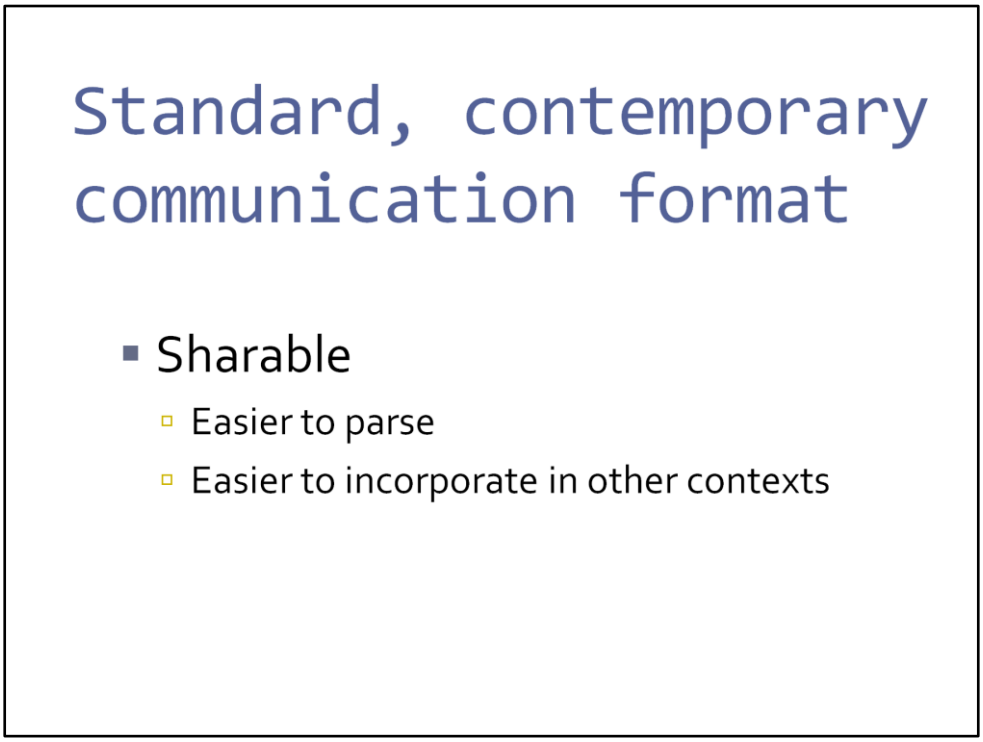
VISIONS OF A POST-MARC WORLD

Presentation for the MARC Format Interest Group at the ALA Midwinter Meeting
January 21, 2012
Kelley McGrath

I was asked to talk about what we need from a successor to MARC. These are some semi-random things I've been thinking about. It's not a comprehensive overview.

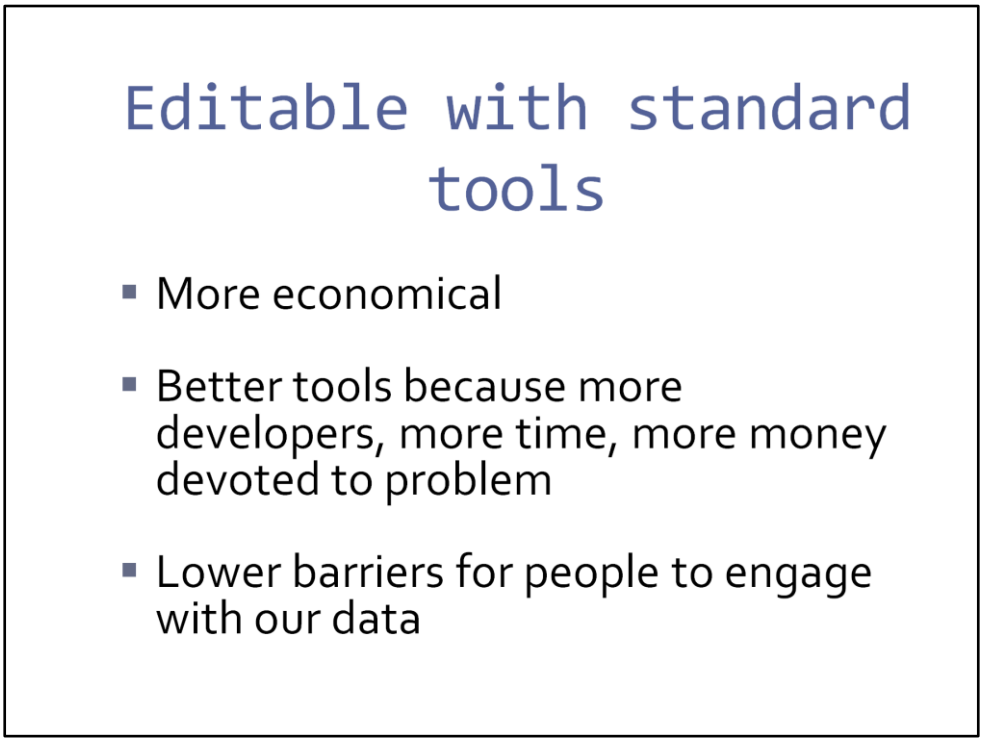


I'd like to start off by asking what is MARC and what is it doing now? Steve Miller at the University of Wisconsin has a nice diagram of different types of metadata standards (https://pantherfile.uwm.edu/mll/www/resource.html). He lists MARC as a metadata encoding standard for machine readability, communication and exchange. This is generally thought of as the basic function of MARC.



We need a new communication format to be a standard, contemporary one. Sure, you can convert MARC to MARCXML but you still have the same underlying structure and limitations. We need something designed to work with today's capabilities.

When MARC was developed, there were no widely-used, widely-accepted standards. MARC had to find its own way. We're in a different world now. If we use a more standard format for our data, it will be easier to share beyond the library world.

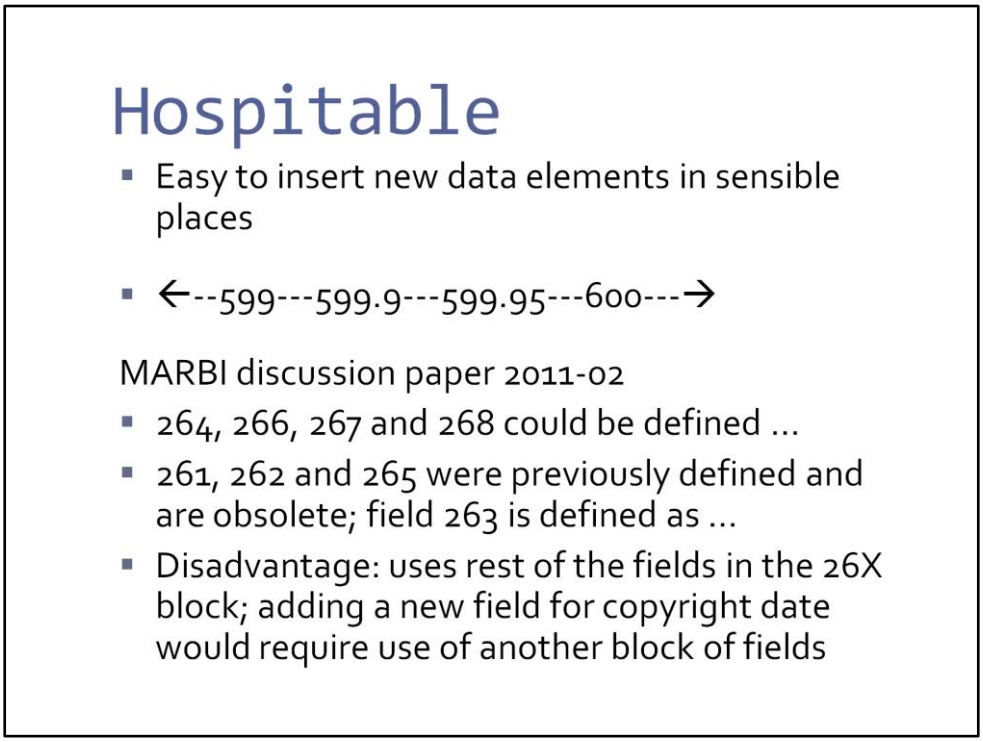


Another side benefit of changing to a widely-used contemporary standard would be a better variety of tools to work with our data. Working with our data would be more economical as we would not have to support so many specialized tools or programmers. We have some very good tools now and we would probably still need some geared toward our data, as well as programmers who understand bibliographic data. However, we would be able to take advantage of a range of existing tools. It would also lower the barriers for people new to library data to engage with our information.

Expansible

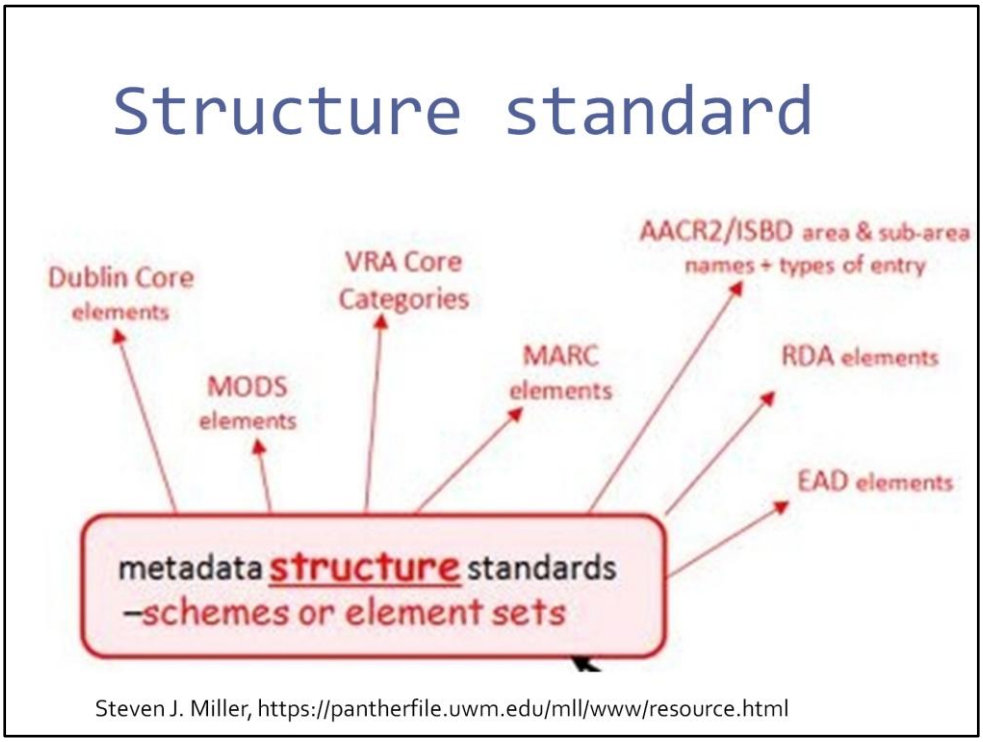
- Not arbitrarily limited by numbers or letters; able to add as many elements as we want
- 856 field (electronic location) only two letters (e, g) left for new subfields
- 246 (varying title) no indicators left for other types of titles (e.g., container)

We often think of expansibility in terms of classification schemes, but it applies here, too. There are many areas in MARC where we have trouble doing what we want to do because we've run out of numbers or letters. A new format should be designed not to have this limitation. A couple examples are the lack of lettered subfields left in 856 and the lack of indicators for specific types of varying titles in 246. I don't say this just because I'm tired of typing "Title on container."

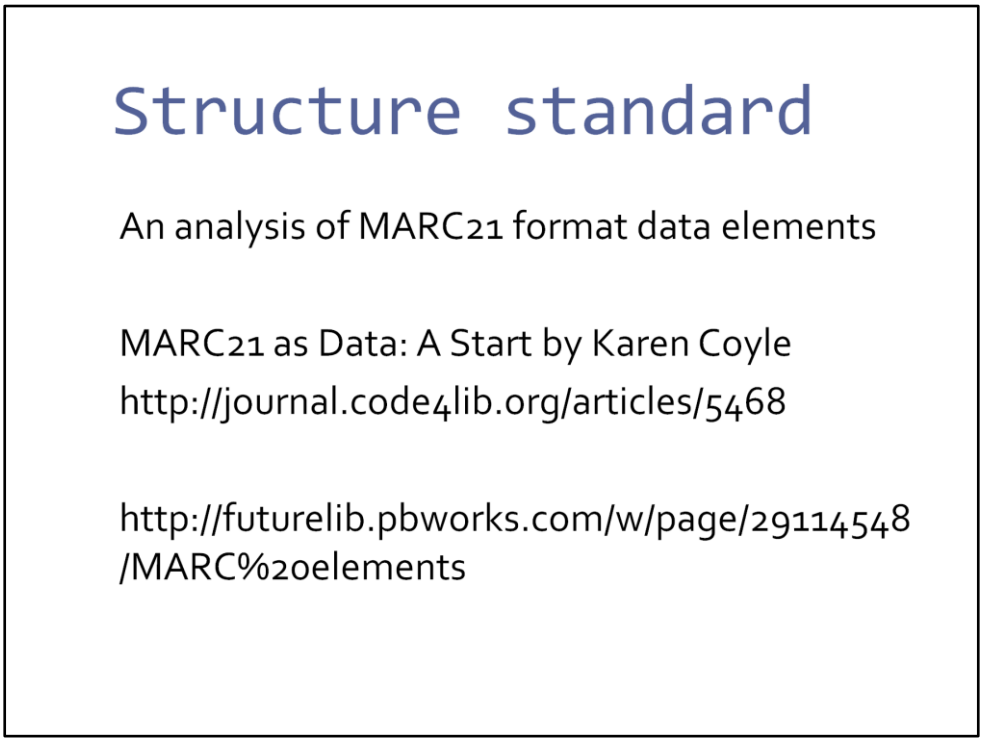


On a related point, the new format should also be hospitable and make it easy to insert new data elements in sensible places. I don't know that it's likely that the new format will be numbered, but we'll probably still want to group and maybe order data elements. If you imagine a more hospitable MARC, it would allow decimals and give us options as shown on the line here where there is always room to squeeze in another number.

There have been a lot of tortured discussions of potential changes to MARC at MARBI meetings that have focused on ways to get around the artificial constraints of MARC. For example, this recent discussion paper suggested an option for making separate fields for the RDA publication, distribution, manufacture and production elements. It suggested four possible fields and then pointed out that the main drawback of this option is that it would use up all the remaining fields in the 26x block. If we had a hospitable format, we could focus on more substantive issues of which there are plenty.



The second place that Steve Miller lists MARC is under structure standards, which he defines as schemes or element sets like Dublin Core or the RDA or EAD elements. This is the second way that we commonly think of MARC—as a list of elements or fields. When we talk about 245$a or 500 notes, we're thinking of data elements.

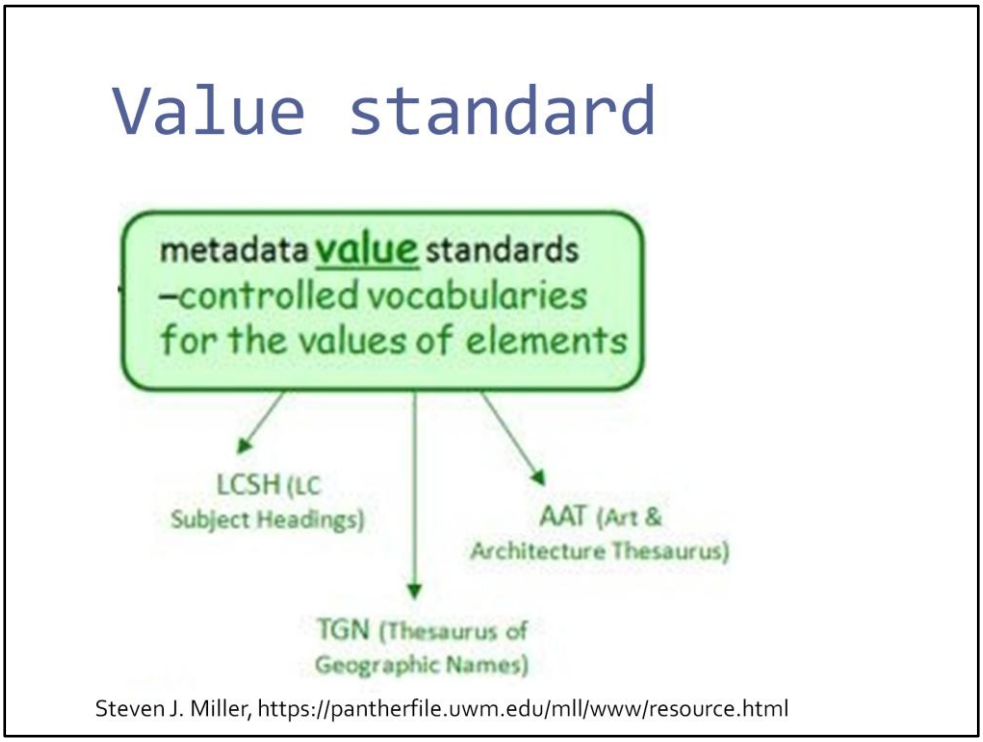


I'm not going to talk a lot about this aspect right now, but I want to point out an interesting and challenging project that Karen Coyle is working on where she's trying to analyze the meaning of the MARC21 format data elements. The idea is that if we don't have a good inventory of what we have now, it will be hard to effectively move forward. She has an article in the Code4Lib Journal, which I recommend, and also a wiki where she's keeping track of her progress. She's also looking for volunteers to help with this project.

One field (024) → 14 data elements

Element	MARC21 field/subfields
ISRC	024$a 024$c 024$d 024/02
UPC	024$a 024$c 024$d 024/02
ISMN	024$a 024$c 024$d 024/02
EAN	024$a 024$c 024$d 024/02
SICI	024$a 024$c 024$d 024/02
Other number	024$a 024$c 024$d 024$2 024/02
Unknown Number	024$a 024$c 024$d 024/02
ISRC Cancelled	024$z
UPC Cancelled	024$z
ISMN Cancelled	024$z
EAN Cancelled	024$z
SICI Cancelled	024$z
Other number cancelled	024$z 024$2
Unknown Number Cancelled	024$z

Just to highlight the complexity of this endeavor, here are the fourteen data elements that she found hidden in the 024 standard number field.



Now I'm going to talk about a couple types of metadata standards where Steve hasn't listed MARC, but where parts of MARC fit. The first is metadata value standards, which are controlled vocabularies for the values of elements, such as LCSH. MARC maintains a number of its own value lists, such as the language codes, geographic area codes and relator codes, as well as internal lists, such as the lists for type of material or illustrations in 008.

Value standard

List: MARC21 Code List for Languages

041 1- $a eng $h spa

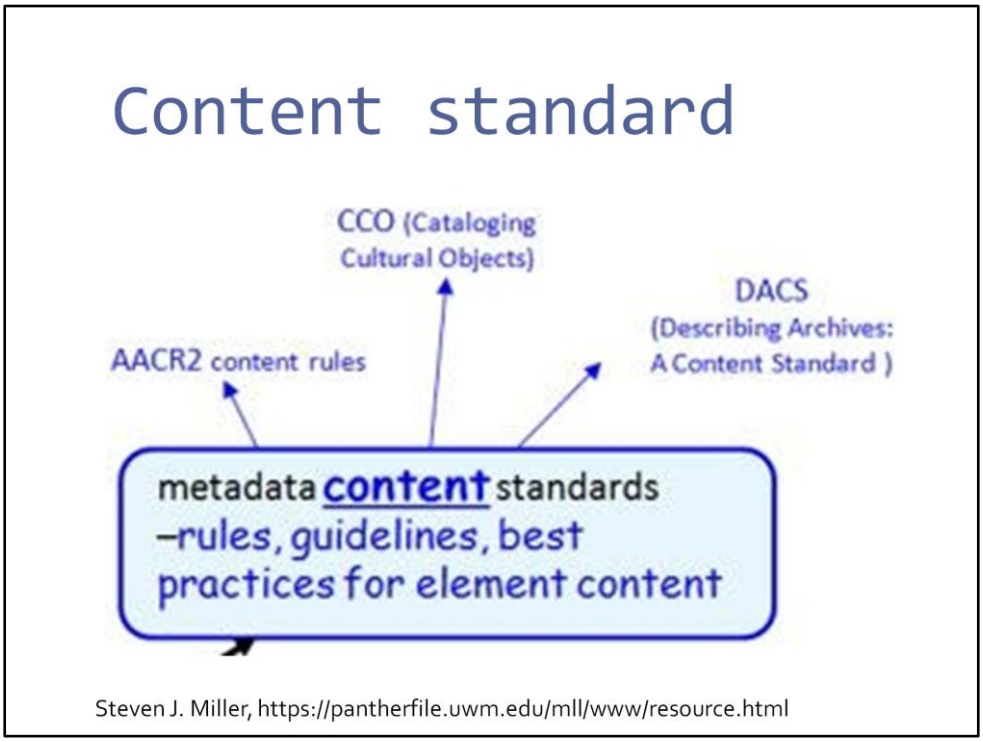
List of lists: Language Code and Term Source Codes

008/35-37 |||

041 07 $a en $a fr $a it $2 iso639-1

When MARC was developed there were no existing lists to draw from so MARC made its own. MARC can also use external lists. One limited way it does this is through indicators as in a 2nd indicator of 0 in 6xx subjects, which indicates that a value comes from LCSH. MARC has expanded its ability to incorporate external lists by creating lists of authorized lists and then noting the list used in a subfield 2 as shown in the bottom example. There are limits on this approach, too, as shown by the fact that the language information in the 008/fixed fields can't accommodate a $2.

Although many external lists have been shoehorned into MARC, we need a new format to be more flexible and to make it easier to add new lists. Do we still need our own lists or lists of lists? I think the answer here is maybe for some things, but we probably need far fewer since we can piggyback on other people's lists. For example, the ISO 639-3 list of languages is available as linked data. It may also not be necessary to have predetermined lists of lists.

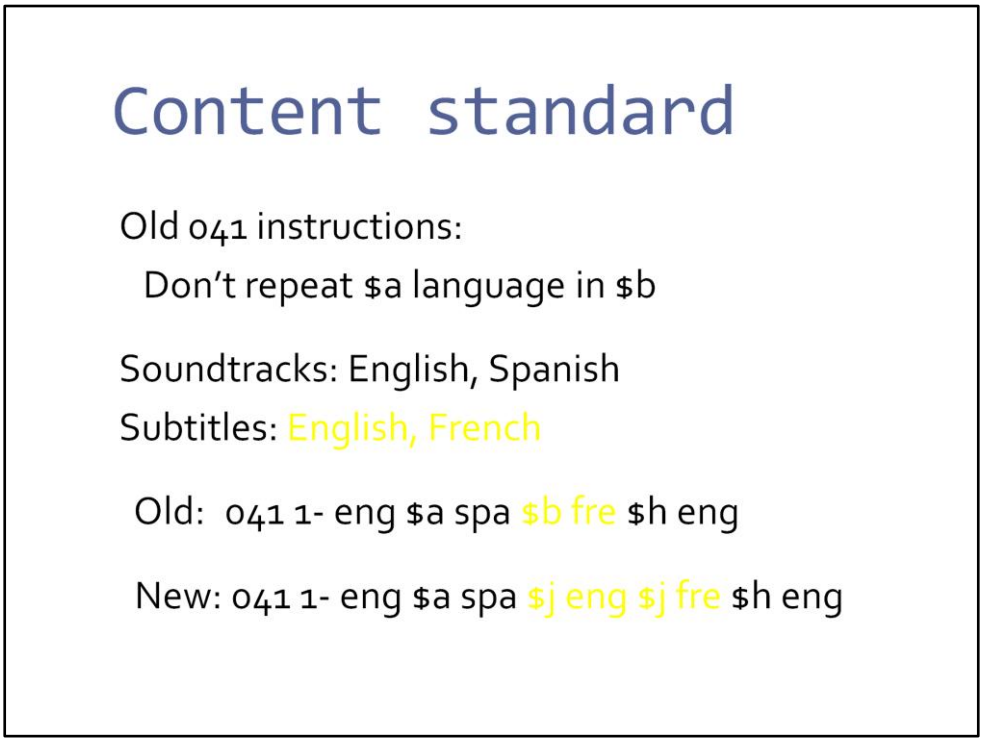


Finally, I would like to talk about metadata content standards or rules, guidelines and best practices for element content. For most of us, the content of most MARC fields is determined by external guidelines, like AACR2 or RDA or the Subject Headings Manual. There are a number of elements not included in those rules and for which the guidance comes from MARC, such as date information in the 008/fixed fields or language or place of publication coding. Even for fields that we think of as coming from AACR2, MARC describes how to handle the indicators and gives some other instructions.

Content standard

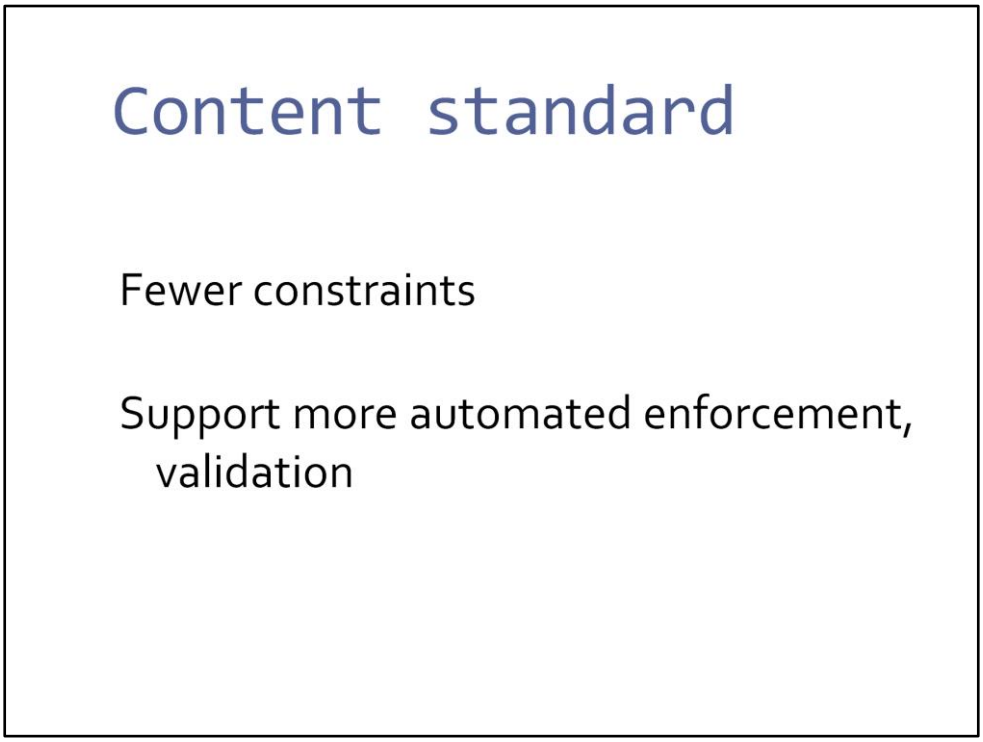
008/35-37: Three-character alphabetic code that indicates the language of the item. Code from: *MARC Code List for Languages*. Choice of a MARC code is based on the predominant language of the item.

Here's an example of the beginning of some content instructions in MARC for the language information in the 008/fixed fields.



The MARC instructions are in textual form. These instructions are easy to overlook and sometimes hard to keep integrated when MARC is updated.

One example of a problem with the instructions that I have been involved in concerns the coding of video subtitles. These were formerly coded in 041$b along with summaries, as a result of format integration. The instructions said not to repeat a language given in $a in $b, which makes sense for summaries, but not so much for subtitles. It was only an occasional problem until the advent of DVDs, which often have subtitles and soundtracks in the same language as in this example. If you ever wanted to limit or facet by subtitle languages, you would get incomplete results. Now we have a new separate $j and better instructions.



I think we need fewer constraints and, for the ones that are important, we need to make them less textual and more incorporated into the format so they're easier to validate. No system enforced the former instructions not to repeat languages from 041$a in 041$b and lots of catalogers ignored it. In this case, it didn't matter, but for many things more effective validation would be helpful.

Data-driven

What can be made into computer-comprehensible data, should be

$a A short /

$c directed by John J. Smith. Mr. X makes a short / directed by J. Johnson, Jr. and Anna Allen.

$a A short /

$c directed by John J. Smith. $t Mr. X makes a short / $c directed by J. Johnson, Jr. and Anna Allen.

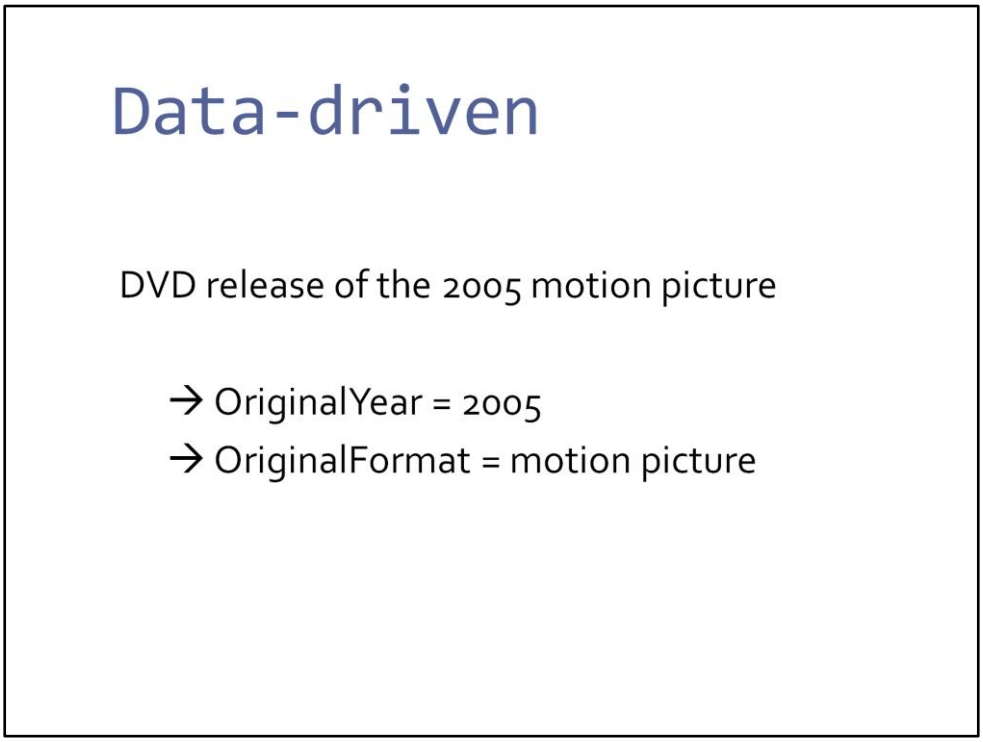
Now I'm going to talk about a couple other points that I think will be important moving forward. The first is that we need more data and less text. What can be made usable by a computer, should be made so. Note that that's not everything. We also need to clearly mark our data and get rid of things like the heterogeneous data in 245$c.

Data-driven

1. DVD release of the 2005 motion picture
2. OriginalYear = 2005

- Search
- Limit
- Sort
- Display (flexible):
  - Charlie and the Chocolate Factory (2005)
  - Original Release Year: 2005

Here are two ways to say that a movie on a DVD was originally released in 2005. Which is better for searching? For limiting? For sorting? For display? Display is the only one where you can make a case for the first option, the kind of note we use today. But even here, we can create a lot of displays from the data, even one that looks exactly like that note, but it's much harder to get data out of display text.



I am involved in a project where we are trying to automate this kind of mapping using the XC Metadata Services Toolkit. For this common example, it's easy to train a computer to do this.

Data-driven

- 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

How about for these examples? How many rules and exceptions would you need to teach a computer? Can you even teach a computer to parse these?

Data-driven

Not everything fits neatly

- IMDb: Ivan the Terrible, Part II (1958)
- AllRovi: Ivan the Terrible: Part 2 (1946)

"Although filmed shortly after Part One in 1946, the film was suppressed and was not released until 1958"—AllRovi.com

Data is nice, but it's true that in the real world not everything fits nicely into a little box. For example, the Internet Movie Database gives a date of 1958 for the movie Ivan the Terrible Part 2. AllRovi says it was 1946, but goes on to explain, "Although filmed shortly after Part One in 1946, the film was suppressed and was not released until 1958."

Data-driven

1. Use more specific data elements
   - OriginalProductionYear = 1946
   - OriginalReleaseYear = 1958
2. Pick one and make a note to explain

A couple potential ways to deal with this come to mind. One is to use more specific data elements and the other is to just pick one, but make a note to explain the situation.

Data-driven

Balance between

- Desirability and power of controlled data
- Cost of controlling the infinite variety that exists in the real world

I think we want a balance between…

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

As I pointed out in my presentation at this group last year, we still need free text for many purposes.

Relationships

WITHIN A FIELD

$0 or $2 can't refer to specific subfields

600 10 $a Abbey, Edward, $d 1927-1989 $x Criticism and interpretation

$0 http://id.loc.gov/authorities/names/n78093802.html

$0 http://id.loc.gov/authorities/subjects/sh99005576.html

Finally, I would like to talk a little about the potential of relationships between pieces of data and some of the problems we have recording those now. Some of our current problems are related to structural limitations of MARC, such as not being able to connect two subfields in a field except loosely via order.

Relationships

BETWEEN FIELDS: TWO MOVIES

Roy Rogers collection. 1 (DVD)

500 DVD release of the 1939 (Days of Jesse James) and 1943 (King of the cowboys) motion pictures.

505 Days of Jesse James / director, Joseph Kane ... (54 min.) -- King of the cowboys / directed by ... (71 min.)

508 Photography, Reggie Lanning, William Bradford ; art directors, Russell Kimball, Paul Youngblood ...

This is one of my pet peeves; the inability to keep information about individual works together when you have more than one in a record for an item. This is more obvious when you look at works where we traditionally record more data, like music or moving images. Here is what a MARC record for a DVD with two films on it might look like. This is based on a real record. I have given the notes in numerical order because that is how many systems present them. Notice also that there are a number of ways of connecting the data about the films with the correct film. Or, as in the credits note at the bottom of this page, not connecting them. Another approach I've not shown is to say things like "first work."

Relationships

WITHIN RECORDS: TWO MOVIES

Roy Rogers collection. 1 (DVD)

511 King of the cowboys: Roy Rogers, Smiley Burnette.

511 Days of Jesse James: Roy Rogers, George "Gabby" Hayes.

520 In Days of Jesse James, Roy proves that Jesse James ... In King of the cowboys, Roy infiltrates ...

Here is the record continued with cast notes and summaries. Think about how confusing this record is for the ordinary person looking at it. It's also a disaster for the kind of data mining I was showing earlier for trying to get original dates out of notes because it's much more difficult or impossible to teach a computer how to match dates with titles.

Relationships

WITHIN RECORDS: TWO MOVIES

Roy Rogers collection. 1 (DVD)

- Western films.
- Feature films.
- Fiction films.
- Rogers, Roy, 1911-1998. $4 act
- Hayes, George, 1885-1969. $4 act
- Burnette, Smiley, 1911-1967. $4 act
- Kane, Joseph, 1894-1975. $4 drt

We also have genres and names not connected to titles.

Relationships

WITHIN RECORDS: TWO MOVIES

Roy Rogers collection. 1 (DVD)

1. **Days of Jesse James (1939, 54 min.)**
   Director: Kane, Joseph, 1894-1975
   Cast: Rogers, Roy, 1911-1998 Hayes, George, 1885-1969
   Summary: Roy proves that Jesse James … (more info)
2. **King of the cowboys (1943, 67 min.)**
   Director: Kane, Joseph, 1894-1975
   Cast: Rogers, Roy, 1911-1998 Burnette, Smiley, 1911-1967
   Summary: Roy infiltrates … (more info)

Wouldn't it be so much easier to follow, if we could group the data related to each film? We need a format that makes this sort of thing easier to do.

Relationships

BETWEEN RECORDS

Anna Karenina (Motion picture : 1967)

Based on the novel by Leo Tolstoy

Tolstoy, Leo, graf, 1828-1910. Anna Karenina

http://id.loc.gov/authorities/names/no2009034643

http://rdvocab.info/RDARelationshipsWEMI/motionPictureAdaptationOfWork

http://id.loc.gov/authorities/names/no2009049512.html

Relationships between records are also important and need to be much more data-driven. We need something more like the second example than the first. The second example can be self-explanatory for computers while simultaneously being linked to a variety of human-readable displays.

Relationships

WITH OTHER DATASETS

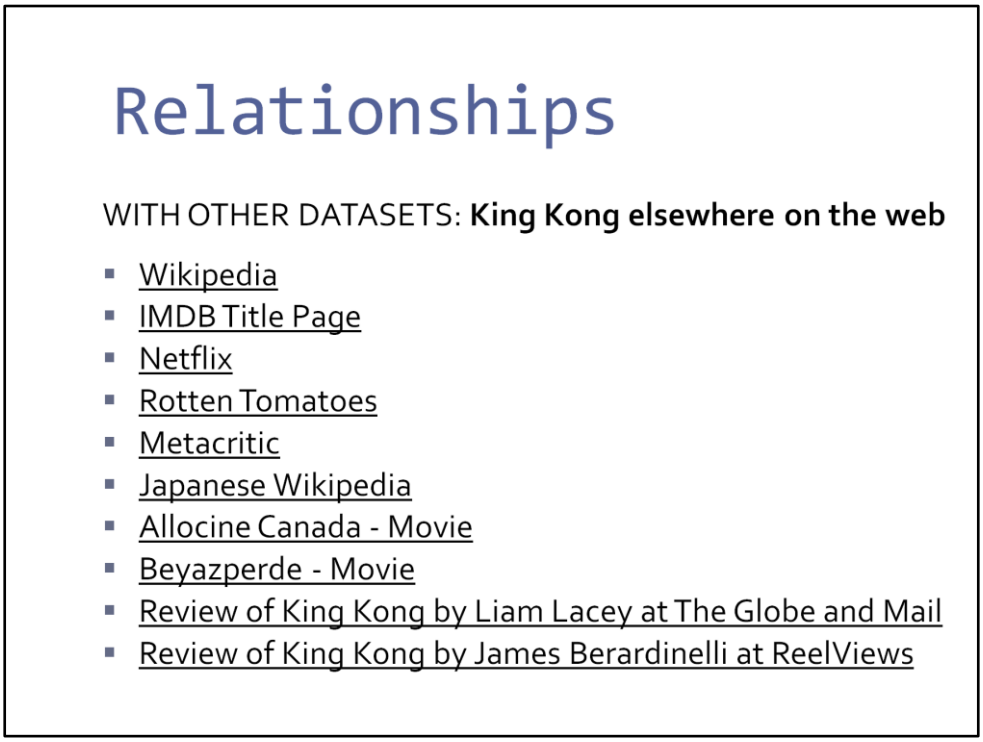
King Kong (Motion picture : 2005)

http://id.loc.gov/authorities/names/no2005102675.html

http://www.freebase.com/view/en/king_kong_2005

http://www.imdb.com/title/tt0360717/

Finally, we need to support relationship with other datasets. So we need to be able to say things like the movie King Kong represented by this authority record with this number is the same as this Freebase entity. Freebase can then link to all sorts of other things, like the Internet Movie Database.



Or a whole bunch of other stuff. The other point here, which isn't about the format, is that we need more authority records for works and other entities so that we have something to link from.

Those are a few thoughts on where we might want to go post-MARC. Thank you for listening.