Monday, March 14, 2022

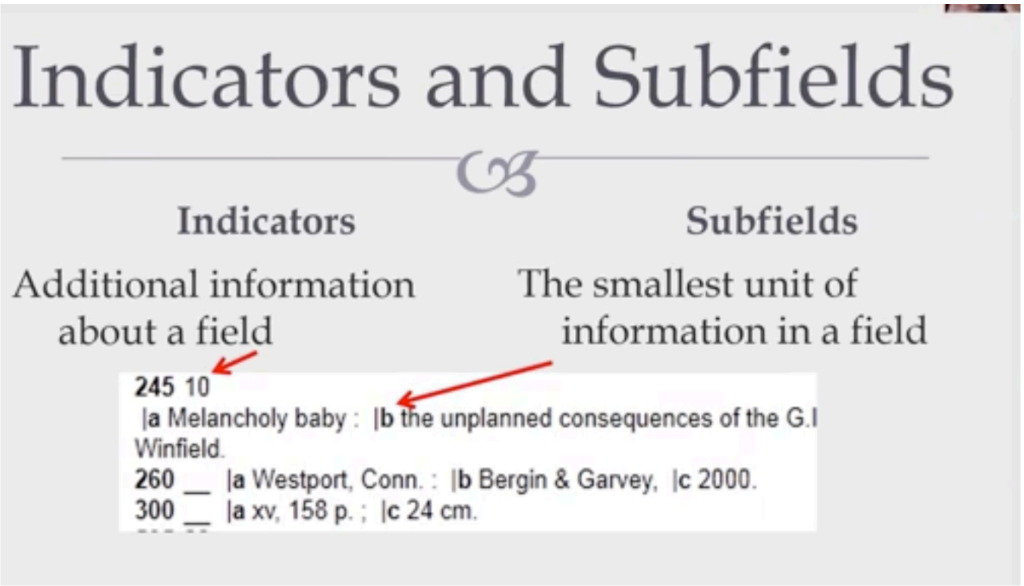
Class 9: Data Formats

LIBR 509: Foundations of Resource Description and Knowledge Organization

- Data Formats underlies everything we've been doing so far
- What are they?
- What do they determine? (What effects to they have)
- Exp, Dublin Core in Data Formats, CSV
    - How to identify what you're saying about a resources
- Exp. Dublin Core, in Data Formats, XML
    - Same information as CSV file - put data derived from one content standard into each format
- Exp. MARC (most common data in libraries)
    - Same information as CSV file and XML file but rather than DC title, you have row 245 (means title)
    - How well you can read the MARC records and what they say about values changes the familiarity
- Exp. An Index Card - also a data format!
    - Same information, including a classification number
    - Nothing that says "Title" even though title is included
    - What kind of information included is based on norms / standards on formatting (indenting, title, etc.)
- MARC21 records
    - A series of three-digit codes ; a space with numbers / underlines ; a series of pieces of data separated by letter and number codes
    - Relate to content standards

    - Uses RDA

- Indicators and Subfields
    - Indicators - two digits - Additional information about a field - tells you and the computer information about something that’s component
    - Subfields - The smallest unit of information in a field - breaks down information into smaller components
    - 245 field - 10 - |a (title) and |b (subtitle)



Indicators and Subfields

Indicators

Additional information about a field

Subfields

The smallest unit of information in a field

```
245 10
 |a Melancholy baby :  |b the unplanned consequences of the G.I
Winfield.
260 __  |a Westport, Conn. :  |b Bergin & Garvey,  |c 2000.
300 __  |a xv, 158 p. ;  |c 24 cm.
```

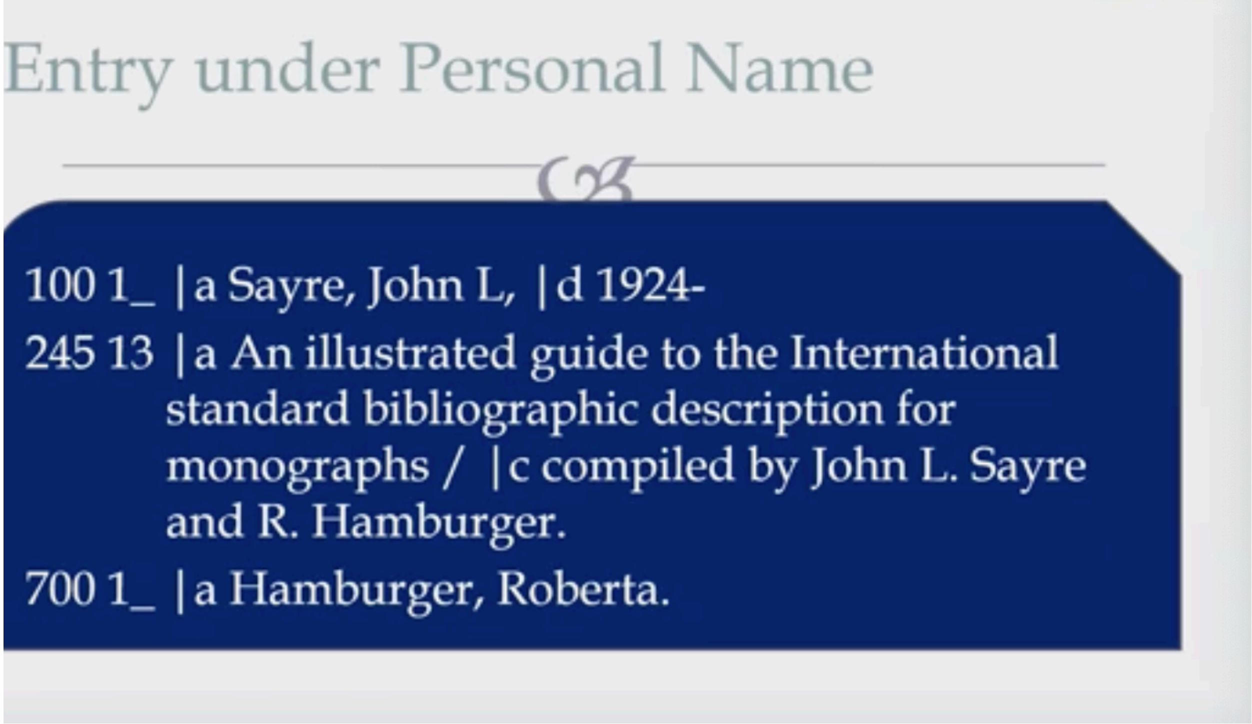
- MARC Record for “The Organization of Information” (textbook)
    - 245 12 |a The organization of information
    - Not all MARC fields have indicators - helpful for understanding source of information - check the quality of the data

- MARC for Main and Added Entries
    - Primary Access Point (Main Entry)
        - 100 (Personal Name)
        - 110 (Corporation)
        - 111 (Meeting)
    - Additional Access Point (Added entry)

        - 700 (Personal Name)
        - 710 (Corporation)
        - 711 (Meeting)

- **Personal names**
    - Birth to five / **Edward Short**
- Other Names
    - **Corporate Names**
        - Charter of the **United Nations**
    - **Meeting Names**
        - **International Conference on Continuing Professional Education for the Library and Information Professions**
- 100 & 700 Personal Name
    - Indicator 1: Type of personal name entry element
        - 0 - Forename (Liberace)
        - 1 - Surname (Carroll, Lewis)
        - 3 - Family name (Medici, House of)
    - Subfield used most often
        - | a - Personal name (Not repeatable)
        - | b - Numeration (Not repeatable)
        - | c - Titles and other words association with a name (repeatable)
        - | d - Dates associated with a name (generally, year of birth) (not repeatable)
        - | q - Fuller form of name (not repeatable)
- 245 Title Statement (Not repeatable)
    - Indicator 1: Title added entry
    - Indicator 2 : Nonfiling characters
    - Subfields used most often

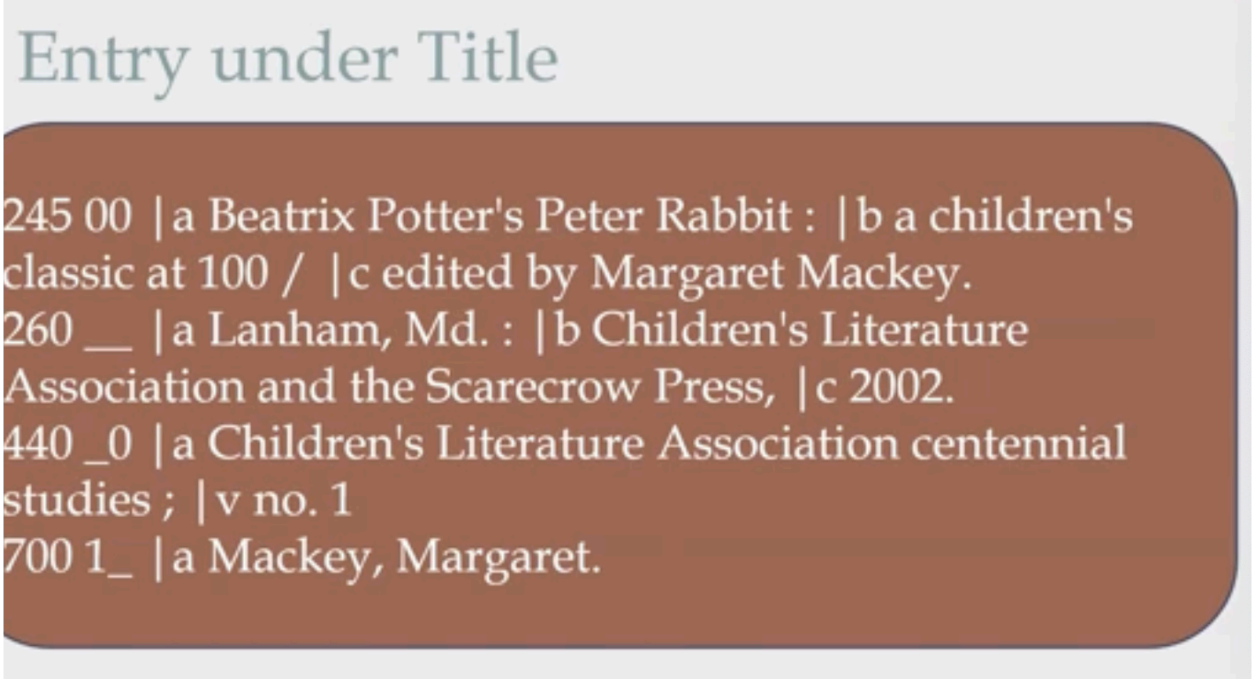
  - Title (Not repeatable)
  - Remainder of title (subtitles) (not repeatable)
  - Statement of responsibility (Not repeatable)
- Entry Under Personal Name



Entry under Personal Name

100 1_ |a Sayre, John L, |d 1924-
245 13 |a An illustrated guide to the International standard bibliographic description for monographs / |c compiled by John L. Sayre and R. Hamburger.
700 1_ |a Hamburger, Roberta.

- Entry under Title (no primary creator but has a contributor



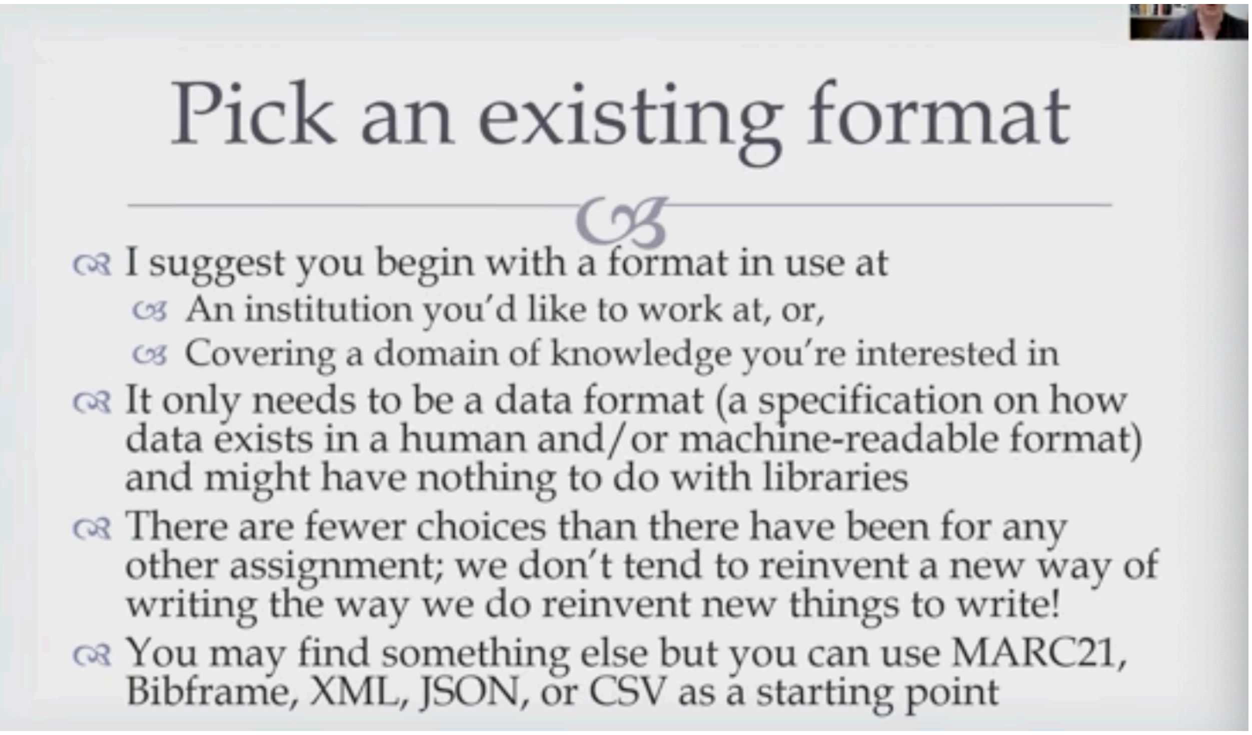
Entry under Title

245 00 |a Beatrix Potter's Peter Rabbit : |b a children's classic at 100 / |c edited by Margaret Mackey.
260 __ |a Lanham, Md. : |b Children's Literature Association and the Scarecrow Press, |c 2002.
440 _0 |a Children's Literature Association centennial studies ; |v no. 1
700 1_ |a Mackey, Margaret.

- 650 - Subject Added Entry-Topical term (R)
  - Indicator 2: Thesaurus
  - Controlled subfield used most often
- Further MARC
  - MARC 100 field:
  - MARCH 245 field:
  - More subfield codes and examples
- ***Don't have to memorize, just know what kind of info gets embedded in MARC as a data format so that you know what's worth checking against if you want to decipher something odd going on with the records
- **RECAP**
  - **Content Schema and Data Formats are separate things**
    - Content Schema / Standards has:
      - A set of values
      - Some instructions on which elements are necessary
      - Some instructions on how to modify elements
      - Some instructions on how to fill out the values
        - Gets to practical stuff but not intractable until it has a data format - still conceptual realm and what is worth describing about this thing
    - A data format determines:
      - Where all this data will be (all the things you've said about this resource)
      - How to express connections between attributes and values (CSV value - they occur in the same row)
      - How to express connections between attributes themselves (exp. Contributor's name and a contributor's role - how to say this person is associated as a creator because they are an illustrator)
      - What characters you can use, how many characters you can use (key words)

        - How records can relate to, overlap, add to each other (whether or not, they relate to each other - side by side ; in a row)
        - How you can use (sort, search, filter, combine) the records
        - If you are in danger of paper cuts

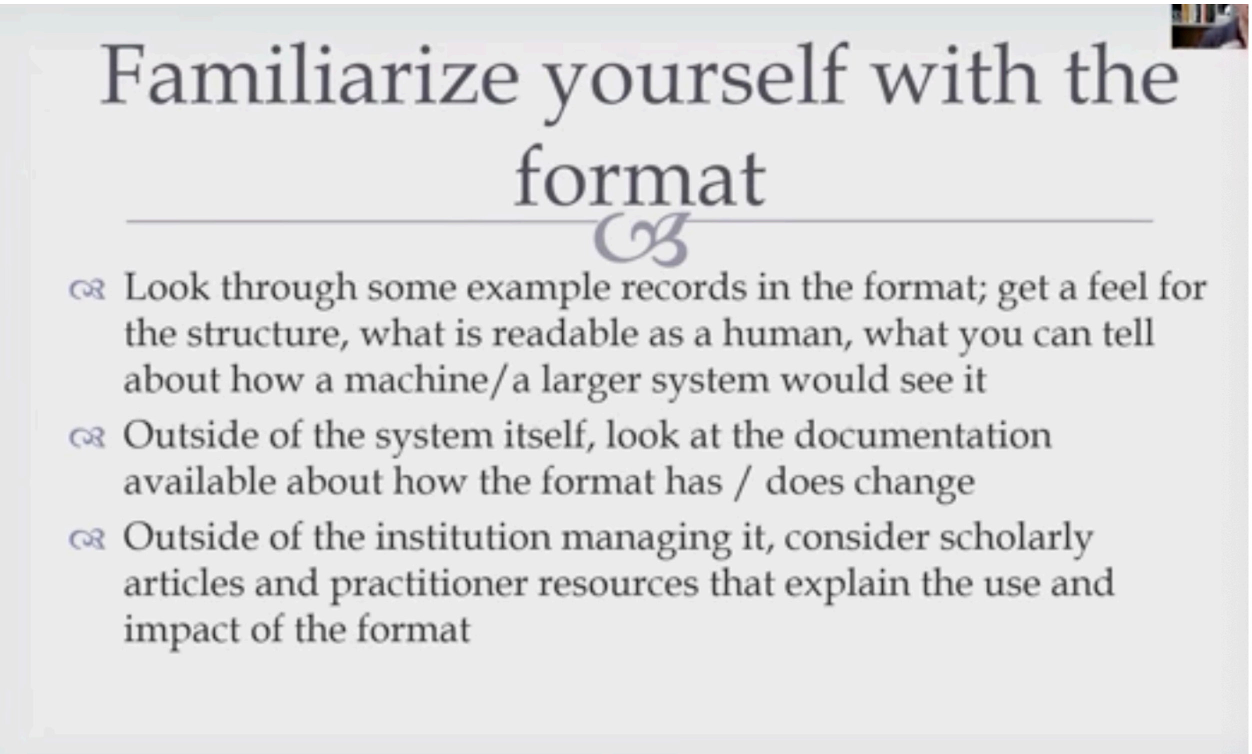
- **PART 2: Analysis Assignment Instructions - Analyze Data Format Assignment**
    - Pick an existing format - format used for the collection of records you're looking in
        - Exp. JSON, MARC21 (library), CSV, Bibframe (library), XML
            - How data is going to be read by a machine



Pick an existing format

- I suggest you begin with a format in use at
    - An institution you'd like to work at, or,
    - Covering a domain of knowledge you're interested in
- It only needs to be a data format (a specification on how data exists in a human and/or machine-readable format) and might have nothing to do with libraries
- There are fewer choices than there have been for any other assignment; we don't tend to reinvent a new way of writing the way we do reinvent new things to write!
- You may find something else but you can use MARC21, Bibframe, XML, JSON, or CSV as a starting point

    - Familiarize yourself with the format
        - Look at the structure - how do you know what a piece of information in the record is. (Exp. MARC 245 field = author)

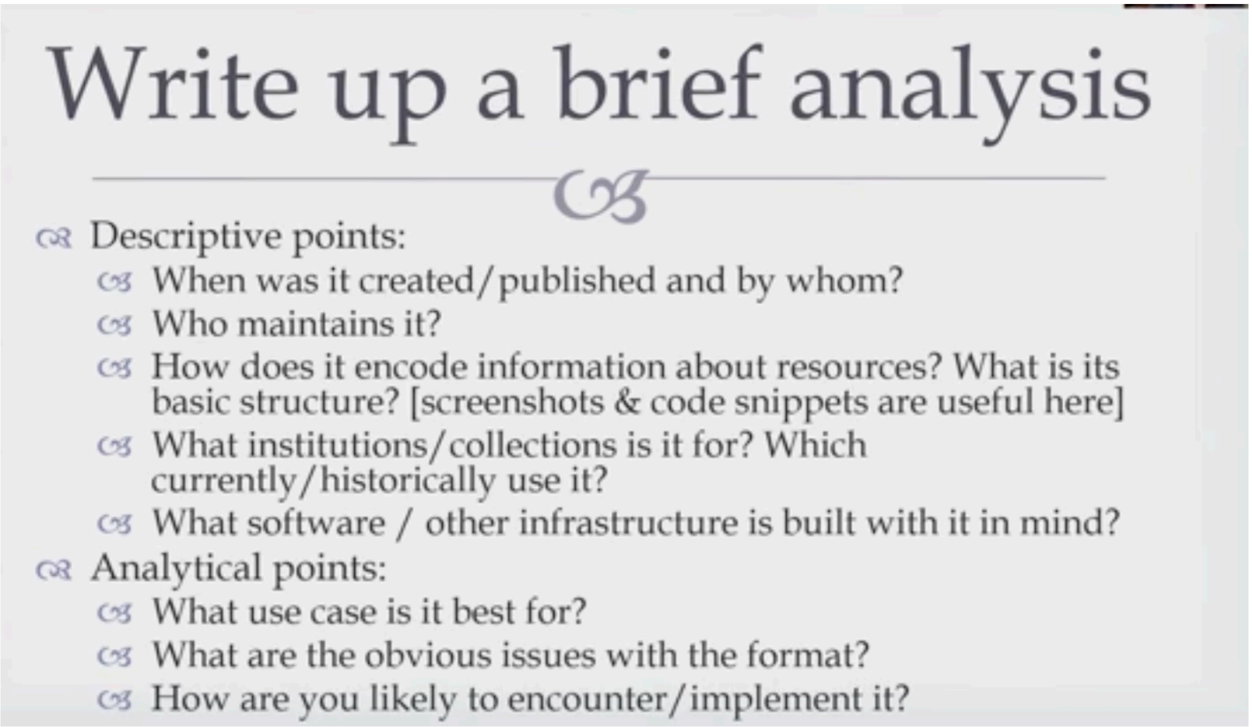


Familiarize yourself with the format

- Look through some example records in the format; get a feel for the structure, what is readable as a human, what you can tell about how a machine/a larger system would see it
- Outside of the system itself, look at the documentation available about how the format has / does change
- Outside of the institution managing it, consider scholarly articles and practitioner resources that explain the use and impact of the format

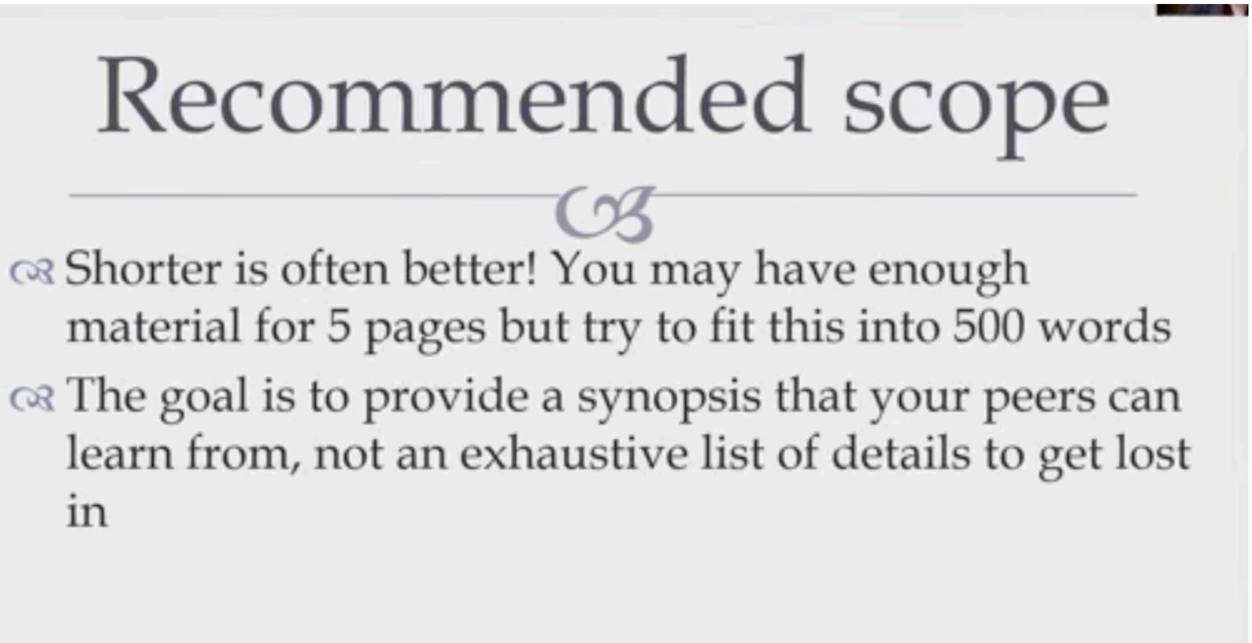
- Strong and week points of using the data format

- Write up a brief analysis



Write up a brief analysis

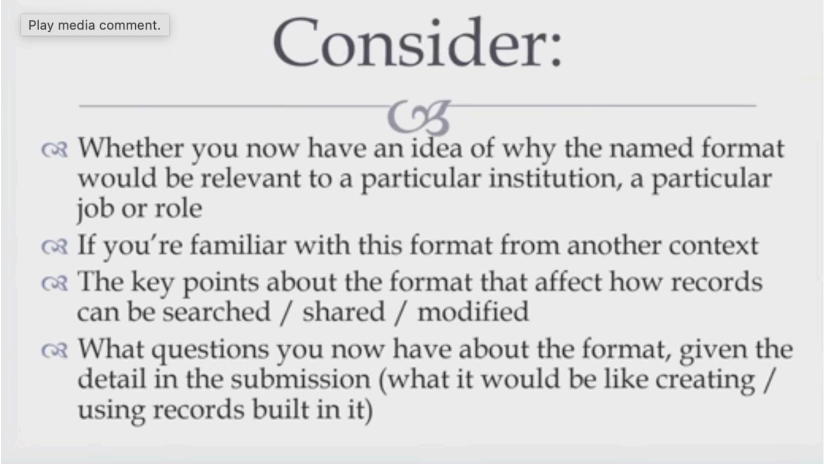
- Descriptive points:
  - When was it created/published and by whom?
  - Who maintains it?
  - How does it encode information about resources? What is its basic structure? [screenshots & code snippets are useful here]
  - What institutions/collections is it for? Which currently/historically use it?
  - What software / other infrastructure is built with it in mind?
- Analytical points:
  - What use case is it best for?
  - What are the obvious issues with the format?
  - How are you likely to encounter/implement it?



Recommended scope

- Shorter is often better! You may have enough material for 5 pages but try to fit this into 500 words
- The goal is to provide a synopsis that your peers can learn from, not an exhaustive list of details to get lost in

- can rely on screenshots
- For Peer Review, consider:


Play media comment.

Consider:

- Whether you now have an idea of why the named format would be relevant to a particular institution, a particular job or role
- If you're familiar with this format from another context
- The key points about the format that affect how records can be searched / shared / modified
- What questions you now have about the format, given the detail in the submission (what it would be like creating / using records built in it)

<u>IN-CLASS LECTURE NOTES</u>

- MARC21 - MARC for the 21st century
- Bibframe - trying to get libraries to switch to - introduction to Bibframe on Canvas
- Not specific to libs: XML (html editing); JSON (same logical features of XML); CSV (excel spreadsheets)
- Focus on what you think the standout features are
- You know you're dealing with a data format when you're dealing with a file extension (like .mrc)
- Controlled vocab vs thesaurus
    - Controlled vocals - fixed list of terms - exerted some form of control
    - A TYPE of controlled vocabulary is a thesaurus

~

<u>IN-CLASS LECTURE NOTES</u>

- **Into, or: Why Data?**
    - Us figuring out things in context
    - Linked Data: Is the practice of creating formal sentences called triples
    - Virginia Woolf (VW) example:
        - **VW (subject) wrote (predicate) a Room's of One's Own (object)**
    - Why do we care?
        - Cools things about linked data!
            - We cab bring together a lot of related linked data to build more context
            - Libraries / museums are using linked data now
            - Semantic Web - a painting that depicts the Irish City
            - Translation - can translate instantly!
    - Okay, how do we link data?
        - Structure - Wikidata

    - One item, one page, items have properties, values, claims, and statement
        - Claim = property, value, qualifier - a wrapper that includes property, value, qualifier
        - Statement = claim and reference - a wrapper that includes claim and reference
- Title, qualifier number - Subject, predicate, and objects