Data Formats in Libraries
March 15th, 2022

Data Formats in Libraries

Agenda

- Introduction to what data format is
- Closer look at MARC
    - Acknowledging the novelty of this topic, due to no "Creation" assignment for Data Formats earlier in the course

Data Formats

- Underlying a lot of the standards/processes in libraries
    - Where data formats come in: "pen to paper"
- Two main questions:
    - What are they?
    - What do they determine?

E.g., Dublin Core, in Data Formats, CSV

- Common content schema
- One column labels what is coming and another column labels what the value is

E.g., Dublin Core, in Data Formats, XML

- Same work of Dublin Core description could happen in another data format too: such as XML
    - Does not change what data is; just changes interaction with the data

In MARC

- A little less obvious looking at it to know what's going on
- 245 = title
- 110 = 110 corporate name of resource
    - However, data is roughly the same

On an Index Card

- Pen on paper is its own Data Format
- No labels
    - On an index card, what kind of information each piece is, is determined by norms around formatting and spacing

MARC21

- Machine readable cataloguing of the 21st century
- MARC formats are already laying out what data can do in each space; MARC must change if you want to bring in Content Standards for other materials
    - E.g., 260 - change to RDA

Breaking MARC down…

- 245 field: tells you what kind of information this row contains
- Indicators: two digits that tell you and computer something about the information that's coming

- Subfields: breaks down information into its components

MARC record for "The Organization of Information"

- 245 field - information on title
    - 1 field - telling computer something about placement of record
    - 4 field - telling computer how to alphabetize record; saying to skip 4 characters when alphabetizing
        - A and C - indicating different parts of title statement
            - Title
            - Statement of responsibility
- Not every one of MARC fields have these indicators
- Important for understanding source of the data
    - Important for librarians to check quality of data
    - Instructions for computer

MARC for Main and Added Entries

- 100 fields - can only be used once per record; indicates primary creator of work
    - Could be case where more than one person is responsible for work; co-authors, illustrators --> put into 700, 710, 711 fields
    - Some works will not have 100 fields at all, with no primary creator; in that case, some will be listed in 700 fields as additional creators
- 700 fields
- 600 fields
    - The "aboutness" of the record
    - 650 - subject added entry - Topical Term (R)

Content and Format

- A content schema or standard has:
    - A set of values or attributes
    - Some instructions on which elements are necessary (required fields)
    - Some instructions of how to modify elements
    - Some instructions for how to fill out the values (what to do in case of misspellings, inferred rather than transcribed data)
    - Still not interactable without a data format
- A data format determines:
    - How to express connections between attributes and values
    - How to express connections between attributes (e.g., contributor's name and contributor's role)
    - What characters you can use, how many characters can you use
    - How records can relate to, overlap, add to each other
    - How you can use (sort, search, filter, combine) the records
    - If you are in danger of papercuts

In-Class Session

MARC21

- 21 = Machine readable cataloguing for the 21st century

Analyze Assignment

- Only about 5 data formats to analyze this week:
    - MARC21
        - Makes sense to pick something specific to focus on in a MARC21 record
            - Picking oddities
                - What it means for you as a user, looking up items
    - BIBFRAME
        - Trying to get libraries to switch to BIBFRAME
            - Showcasing specific collections that have adopted BIBFRAME, will be pilot studies or test cases
    - XML
        - HTML editing
    - JSON
        - Has many logical features similar to XML
        - Will look funny unless you download JSON viewer in browser
    - CSV
        - Excel spreadsheets
- Looking at:
    - Standard features
    - To what extent are text tags readable
    - Scholarly and practitioner resources
    - Why it matters to a computer, to a user

Dealing with Data Format

- Eligible to be a file extension

Semantic networks

- Best practices available
- Much more limited data formats that work with it
    - RDF/XML
        - Combo content and data format
    - 3 part statements:
        - X is a Y of Z

Schema vs Standard

- Schema is just a list of attributes and some indication of how to fill those out
- Standard includes an editorial board, committee meetings, etc. Widely shared across institutions
    - "a content schema with cops"

Controlled Vocabulary vs. Thesaurus

- Controlled Vocabularies are just thick lists of terms
- A particular type of CV is a Thesaurus - once you have programmed in term relationships to CV, it is a Thesaurus

★ Controlled Vocabulary
NT Thesaurus

Data Formats, Linked Data, Wikidata (Oh My) - By Bri Watson

Linked Open Data

- Linked data is the practice of creating formal sentences called triples
- Virginia Wrote a Room of One's Own
    - Subject predicate and Object (3)

Why Link?

- Libraries are investing in these links
- Building out into semantic web

How to Link?

- Wikidata
    - Structure
    - 1 item = 1 page
    - Items have properties and values
        - Values can have qualifiers
    - A claim is a wrapper that includes: property, value,