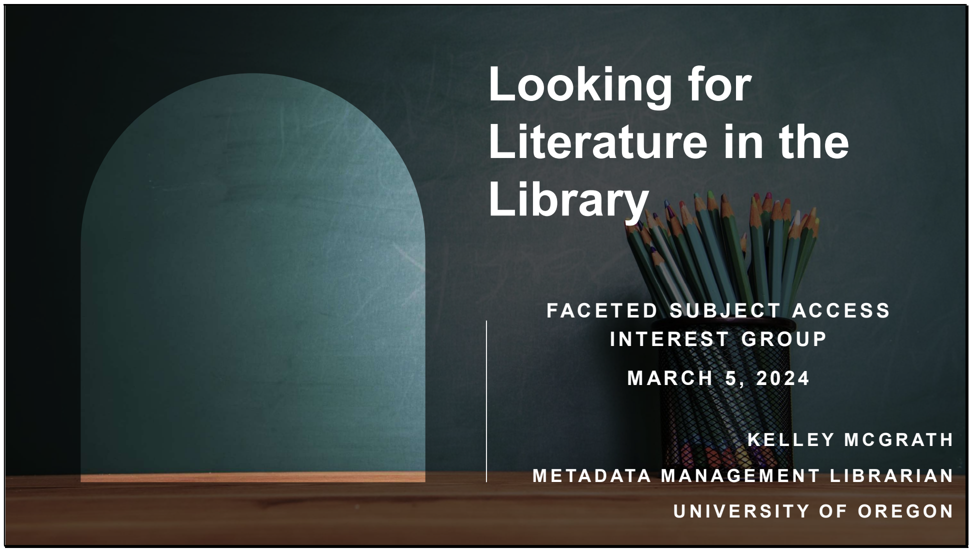


Looking for Literature in the Library

FACETED SUBJECT ACCESS INTEREST GROUP

MARCH 5, 2024

KELLEY MCGRATH

METADATA MANAGEMENT LIBRARIAN

UNIVERSITY OF OREGON

Looking for Literature in the Library

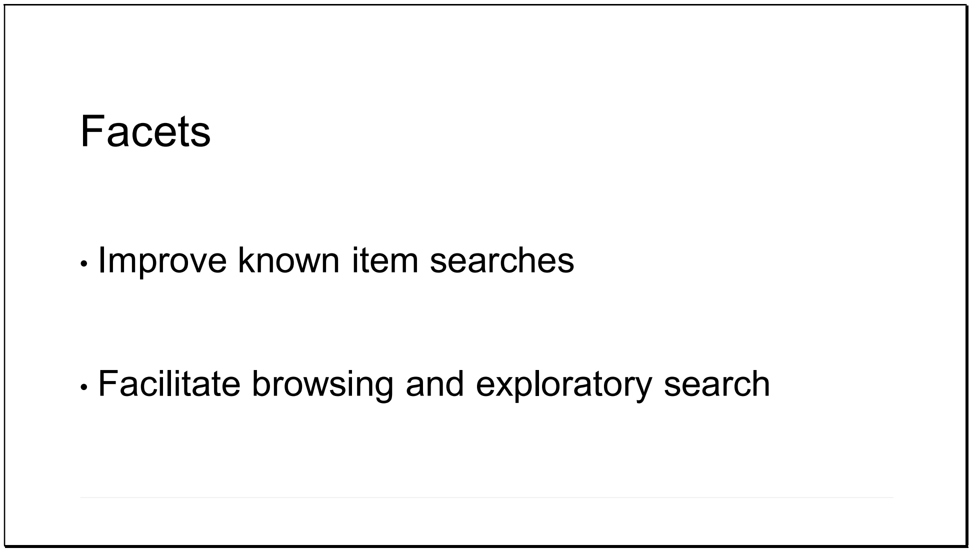
Faceted Subject Access Interest Group

March 5, 2024

Kelley McGrath

Metadata Management Librarian

University of Oregon



Facets

- Improve known item searches
- Facilitate browsing and exploratory search

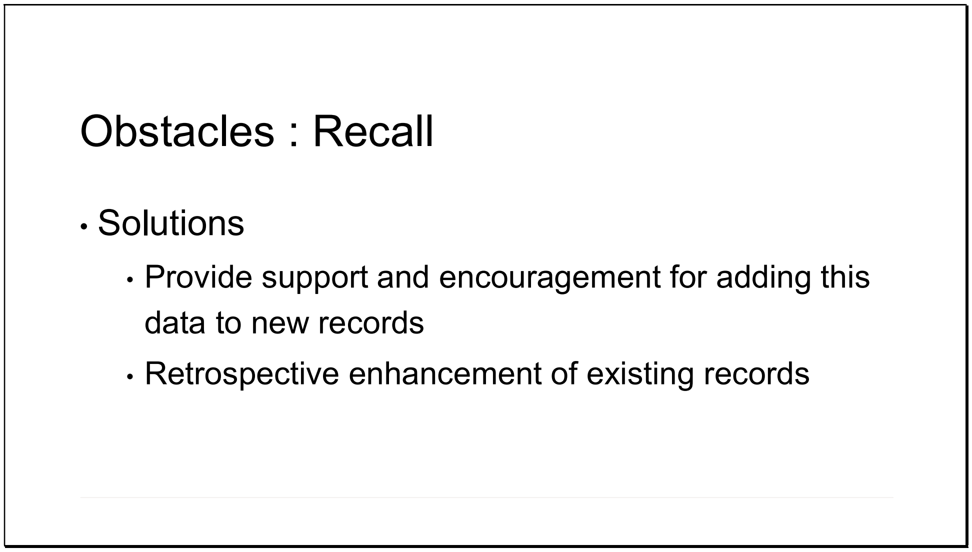
The concept of facets and faceted classification has been around in the library world for nearly a century. Facets can empower users by enabling them to more easily take advantage of structured metadata. Facets can speed up known-item searches and support browsing and exploratory search.

More recently the technological tools have been developed to enable a variety of online search engines to incorporate faceted search. Faceted search has become ubiquitous in e-commerce sites and began to be available in library catalogs starting in the mid-2000s.

Tools for faceted search

- Discovery interfaces
- Vocabularies: FAST, LCGFT, LCMPT, LCDGT
- MARC fields

After library catalogs began to widely support faceted search, the library world began to make changes to better support the use of facets. New faceted vocabularies have been developed. OCLC created FAST, which is based on deconstructed LCSH. LC led the way in creating several new vocabularies, which cover genre and form, physical medium of performance and demographic groups. New fields and subfields were added to the MARC format to give catalogers a place to record the values from these vocabularies.

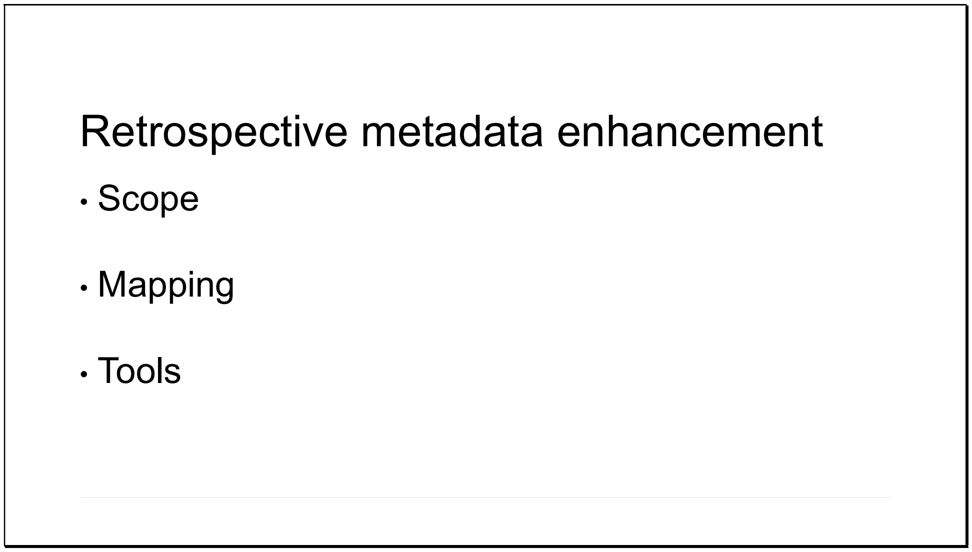


Obstacles : Recall

- Solutions
  - Provide support and encouragement for adding this data to new records
  - Retrospective enhancement of existing records

Despite the introduction of these new vocabularies and MARC fields, many of them are not in common use in library catalogs. One of the biggest obstacles to wider implementation is lack of sufficient recall. Recall will always fall short of perfection in any database of sufficient size. However, some minimum level of recall is necessary to produce reliable results that users find sufficient for their needs. It is not clear exactly where to draw the line for these facets. This is a fruitful area for potential research.

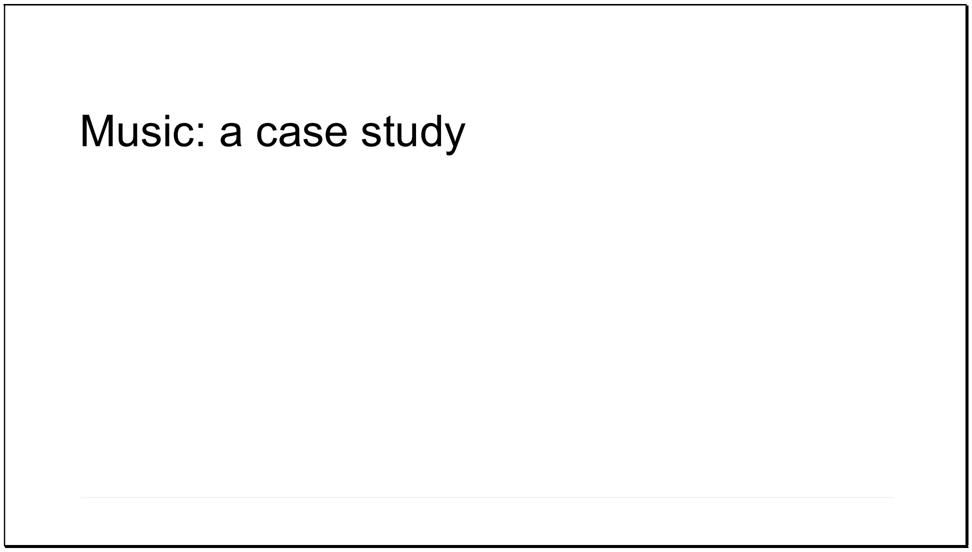
One way to improve recall is to increase the frequency with which these facet values are added to new records. This can be done through raising awareness among catalogers, developing guidance and best practices and creating tools to make adding this data easier. However, this leaves the far larger pool of existing records that lack this information untouched. In order to provide sufficiently comprehensive coverage for these new facets, we need strategies to add this information to existing records, ideally utilizing as many automated processes as possible.



Retrospective metadata enhancement

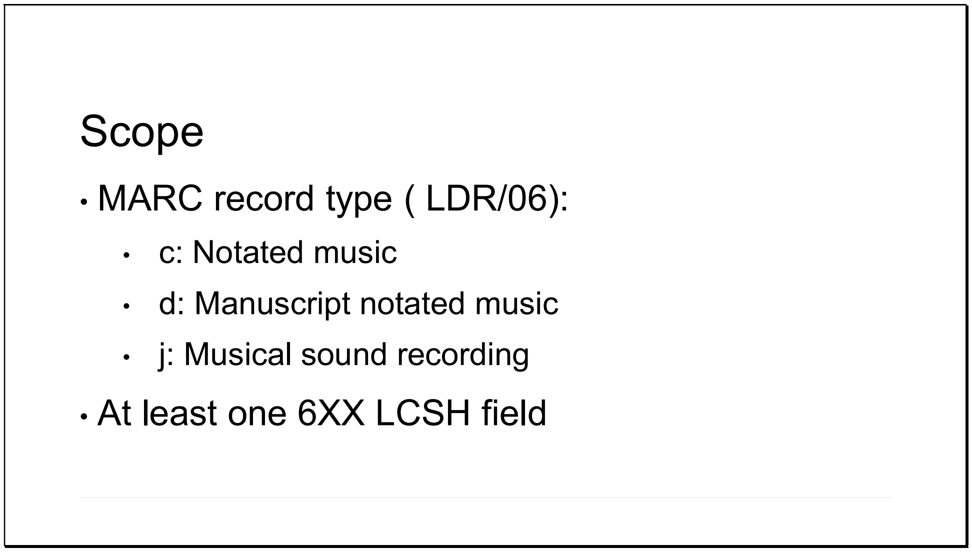
- Scope
- Mapping
- Tools

These types of retrospective metadata enhancement projects have three elements. The first is scope. It is necessary to be able to accurately identify the records that should be upgraded. The second is mapping from information that exists in a record to the new information that you wish to add. Finally, technological tools, such as macros and scripts, are needed to carry out the enhancement. These must be accompanied by thorough, clear documentation, especially when human verification is part of the workflow.



Music: a case study

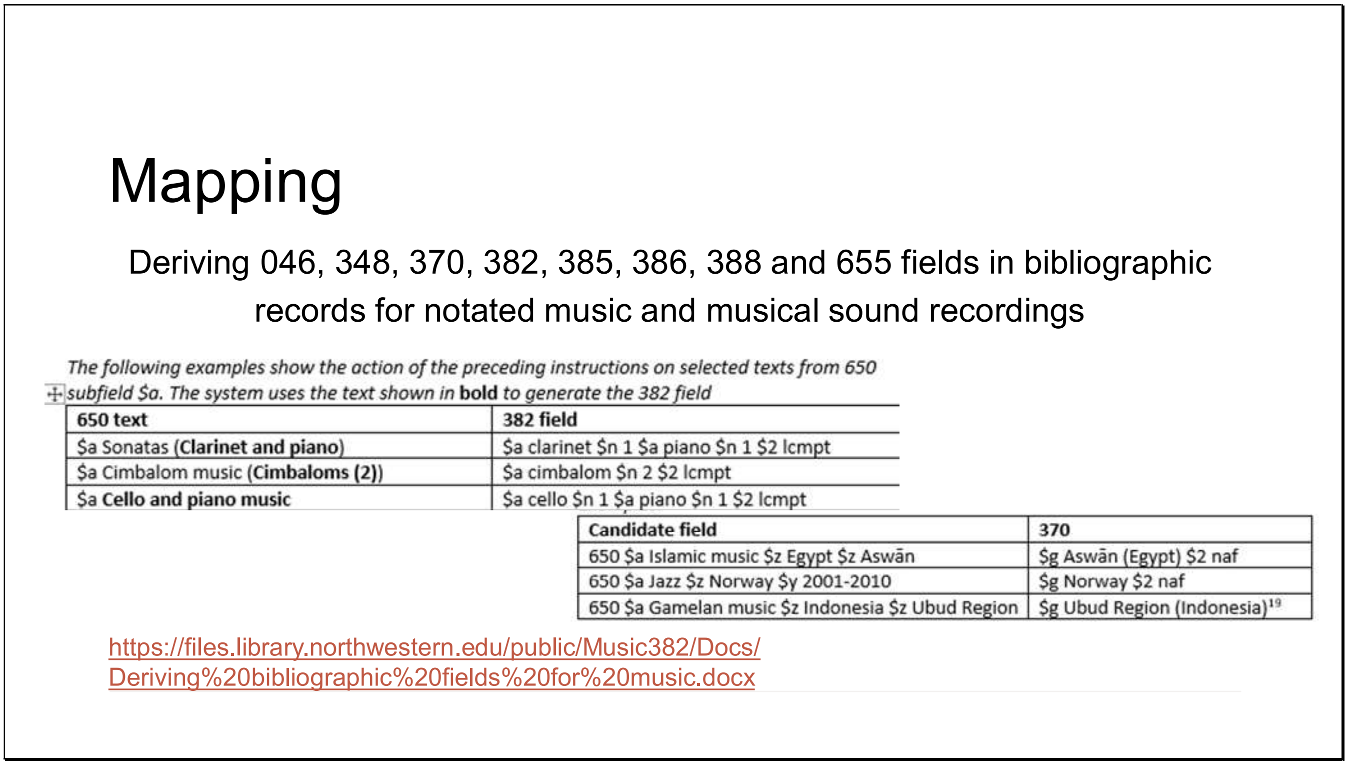
I am going to use work that the music cataloging community has done to deconstruct LCSH and map it to LC's newer genre/form and medium of performance vocabularies as an example of how this process can work. Then I will talk about some work that I am involved in to develop a similar process for literary works.



Scope

- MARC record type (LDR/06):
  - c: Notated music
  - d: Manuscript notated music
  - j: Musical sound recording
- At least one 6XX LCSH field

Music is much easier to scope than literature. In most cases, it can be identified from the MARC record type, which is one of the more reliably coded fixed fields. The main exceptions are records where catalogers have incorrectly coded nonmusical sound recordings as musical sound recordings and music on video. Adding the RDA content type “performed music” to videos of musical performances would make these identifiable. However, since this has not been done historically, more logic would be required to retrospectively enhance video records with this information. A relevant LCSH heading must also already be present in the record.



Mapping

Deriving 046, 348, 370, 382, 385, 386, 388 and 655 fields in bibliographic records for notated music and musical sound recordings

The following examples show the action of the preceding instructions on selected texts from 650 subfield $a. The system uses the text shown in **bold** *to generate the 382 field*

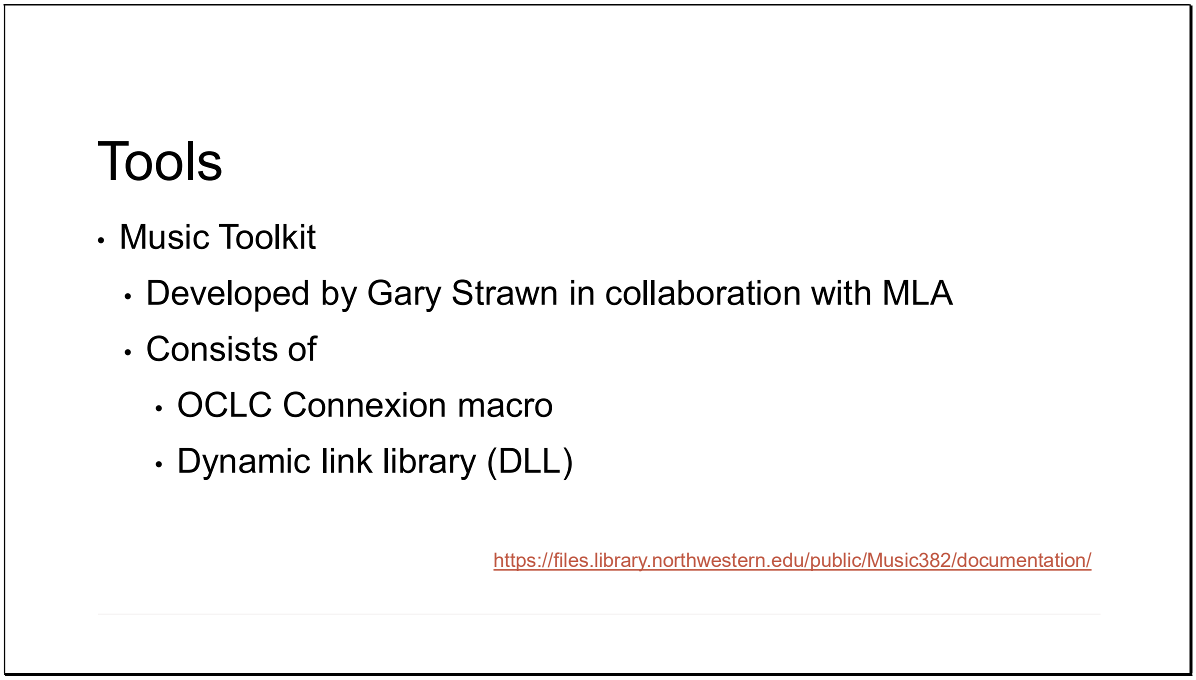
650 text	**382 field**
$a Sonatas (**Clarinet and piano**)	$a clarinet $n 1 $a piano $n 1 $2 lcmpt
$a Cimbalom music (**Cimbaloms (2)**)	$a cimbalom $n 2 $2 lcmpt
$a **Cello and piano music**	$a cello $n 1 $a piano $n 1 $2 lcmpt

Candidate field	**370**
650 $a Islamic music $z Egypt $z Aswān	$g Aswān (Egypt) $2 naf
650 $a Jazz $z Norway $y 2001-2010	$g Norway $2 naf
650 $a Gamelan music $z Indonesia $z Ubud Region	$g Ubud Region (Indonesia)[19]

https://files.library.northwestern.edu/public/Music382/Docs/Deriving%20bibliographic%20fields%20for%20music.docx

This slide shows a couple examples from the 52-page document describing the types of mapping that the music cataloging community has developed.[1] As you can imagine, this is a very complicated process.

[1] https://files.library.northwestern.edu/public/Music382/Docs/Deriving%20bibliographic%20fields%20for%20music.docx



Tools

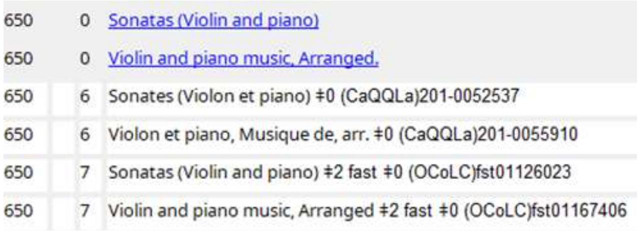
- Music Toolkit
  - Developed by Gary Strawn in collaboration with MLA
  - Consists of
    - OCLC Connexion macro
    - Dynamic link library (DLL)

https://files.library.northwestern.edu/public/Music382/documentation/

Gary Strawn of Northwestern University developed the Music Toolkit based on MLA's guidance.[2] The toolkit consists of an OCLC Connexion macro and a DLL library. It only works in the OCLC Connexion client. After installation, it runs like a regular OCLC macro. It analyzes the components of the LCSH headings in the record and adds new fields based on them.

[2] https://files.library.northwestern.edu/public/Music382/documentation/

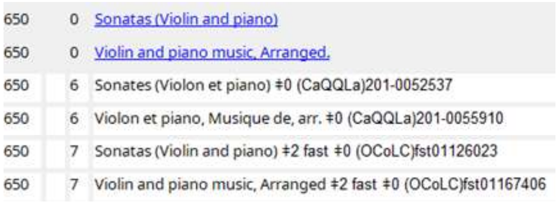
Before



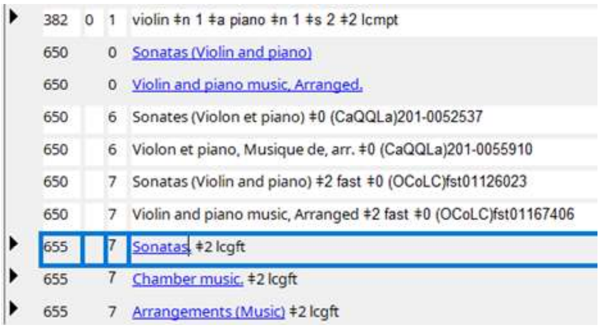
650		0	Sonatas (Violin and piano)
650		0	Violin and piano music, Arranged.
650		6	Sonates (Violon et piano) ‡0 (CaQQLa)201-0052537
650		6	Violon et piano, Musique de, arr. ‡0 (CaQQLa)201-0055910
650		7	Sonatas (Violin and piano) ‡2 fast ‡0 (OCoLC)fst01126023
650		7	Violin and piano music, Arranged ‡2 fast ‡0 (OCoLC)fst01167406

For example, this record has a couple of headings related to violin and piano music, but no LCGFT terms.

After

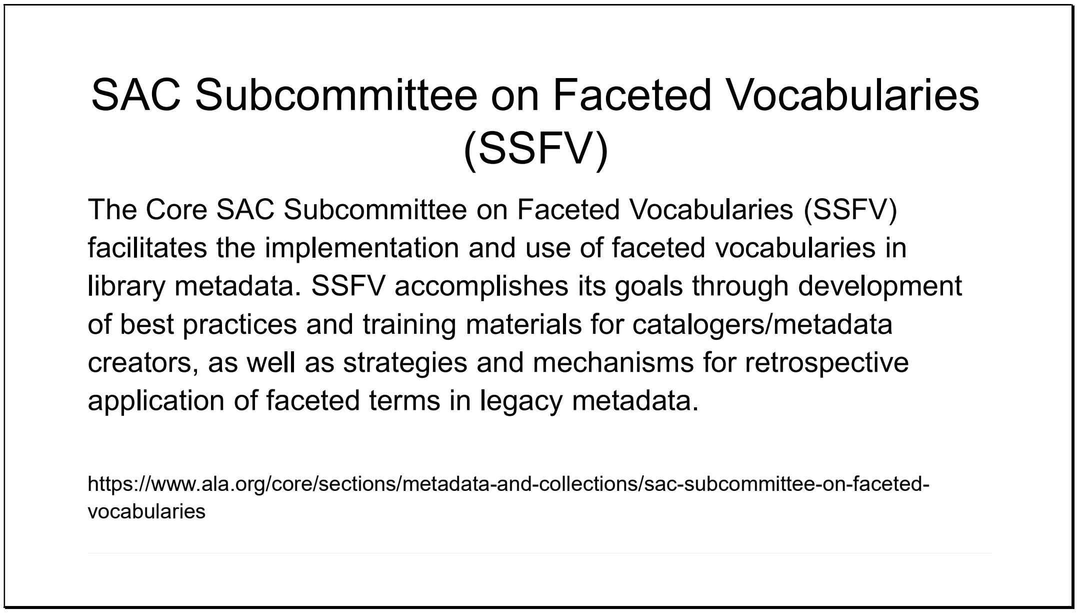


```
650    0  Sonatas (Violin and piano)
650    0  Violin and piano music, Arranged.
650    6  Sonates (Violon et piano) ‡0 (CaQQLa)201-0052537
650    6  Violon et piano, Musique de, arr. ‡0 (CaQQLa)201-0055910
650    7  Sonatas (Violin and piano) ‡2 fast ‡0 (OCoLC)fst01126023
650    7  Violin and piano music, Arranged ‡2 fast ‡0 (OCoLC)fst01167406
```



```
382  0  1  violin ‡n 1 ‡a piano ‡n 1 ‡s 2 ‡2 lcmpt
650     0  Sonatas (Violin and piano)
650     0  Violin and piano music, Arranged.
650     6  Sonates (Violon et piano) ‡0 (CaQQLa)201-0052537
650     6  Violon et piano, Musique de, arr. ‡0 (CaQQLa)201-0055910
650     7  Sonatas (Violin and piano) ‡2 fast ‡0 (OCoLC)fst01126023
650     7  Violin and piano music, Arranged ‡2 fast ‡0 (OCoLC)fst01167406
655     7  Sonatas. ‡2 lcgft
655     7  Chamber music. ‡2 lcgft
655     7  Arrangements (Music) ‡2 lcgft
```

The toolkit inserts the LCGFT terms sonatas, chamber music and arrangements (music), as well as a 382 field that gives the instrumentation.



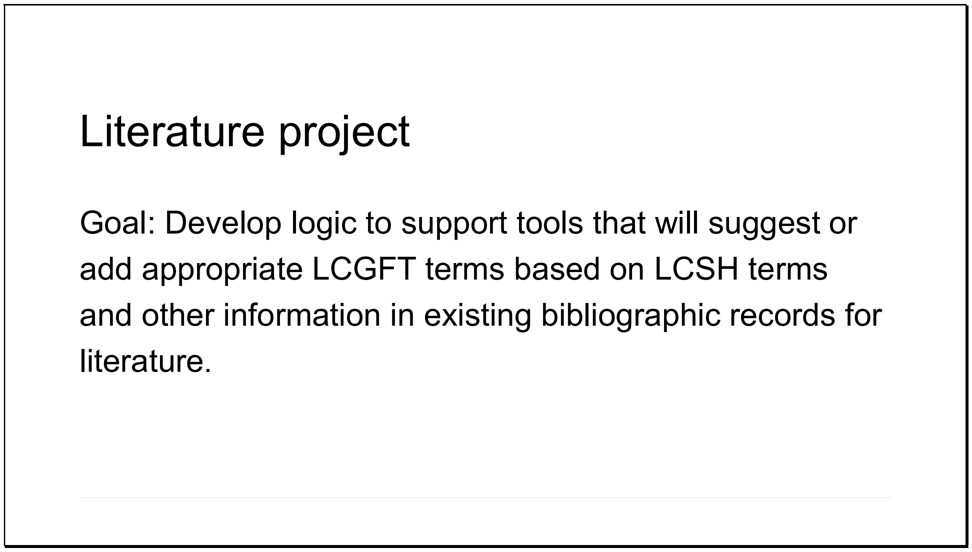
SAC Subcommittee on Faceted Vocabularies (SSFV)

The Core SAC Subcommittee on Faceted Vocabularies (SSFV) facilitates the implementation and use of faceted vocabularies in library metadata. SSFV accomplishes its goals through development of best practices and training materials for catalogers/metadata creators, as well as strategies and mechanisms for retrospective application of faceted terms in legacy metadata.

https://www.ala.org/core/sections/metadata-and-collections/sac-subcommittee-on-faceted-vocabularies

I am a member of the SAC Subcommittee on Faceted Vocabularies[3]. Our goal is to expand and improve the use of faceted vocabularies in library metadata. In addition to developing best practices and training materials, we are trying to develop and implement algorithms for enhancing existing records with metadata to support faceted search.

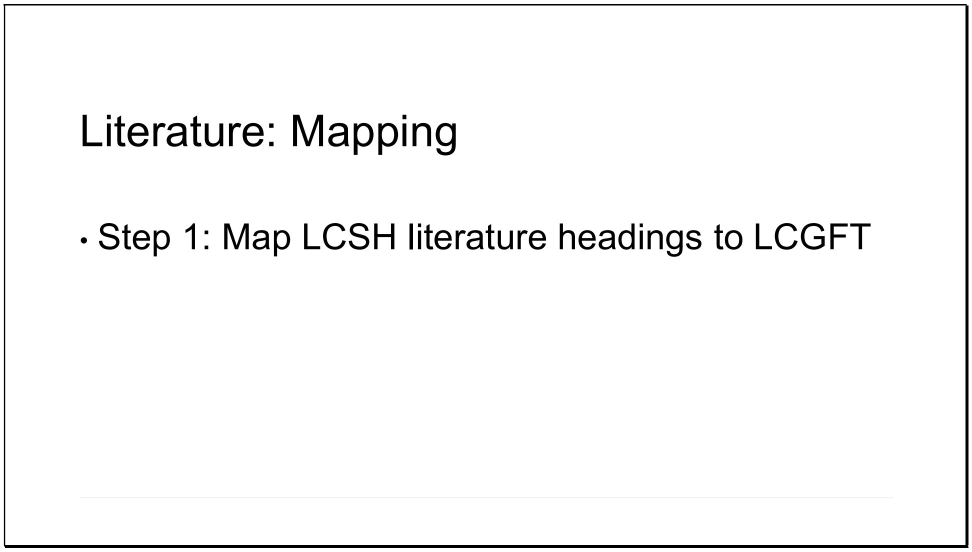
[3] https://www.ala.org/core/sections/metadata-and-collections/sac-subcommittee-on-faceted-vocabularies



Literature project

Goal: Develop logic to support tools that will suggest or add appropriate LCGFT terms based on LCSH terms and other information in existing bibliographic records for literature.

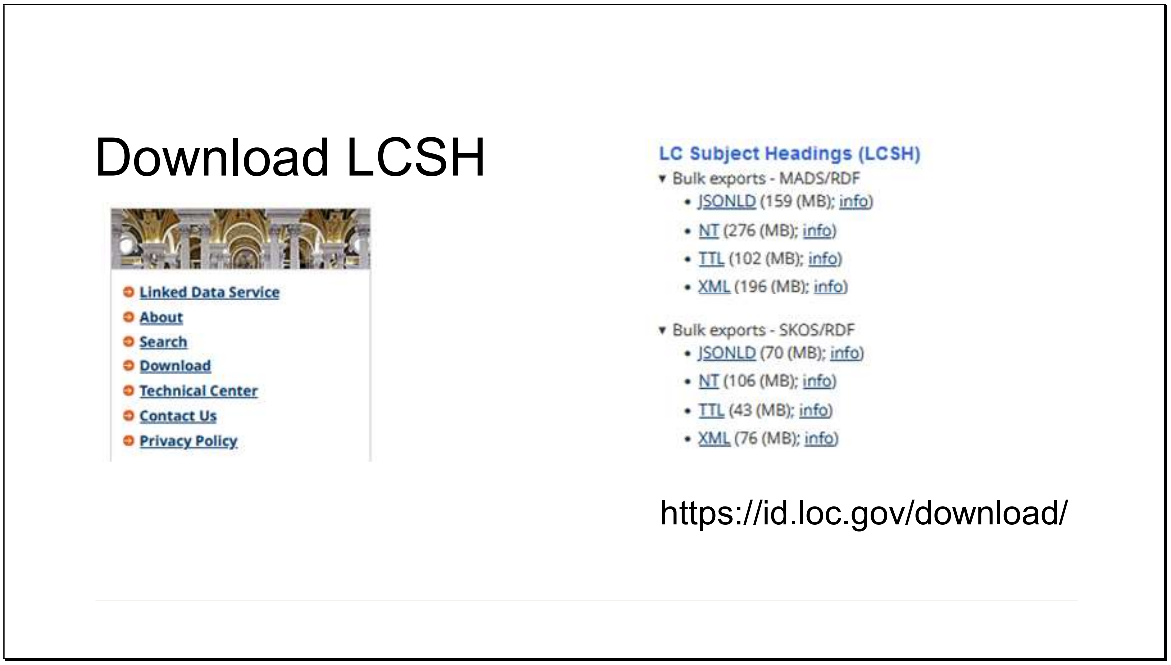
Right now, one of our projects is to develop a process to enhance bibliographic records for literature in a similar fashion to the method for enhancing music records that I just discussed. The overarching goal is to develop logic to support tools that will suggest or add appropriate LCGFT and other facet terms based on LCSH terms and other information in existing bibliographic records for literature.



Literature: Mapping

- Step 1: Map LCSH literature headings to LCGFT

We are beginning by identifying LCSH terms that are sometimes used to describe what something is and could be mapped to LCGFT terms. I was in charge of identifying the candidate terms.



Download LCSH

LC Subject Headings (LCSH)

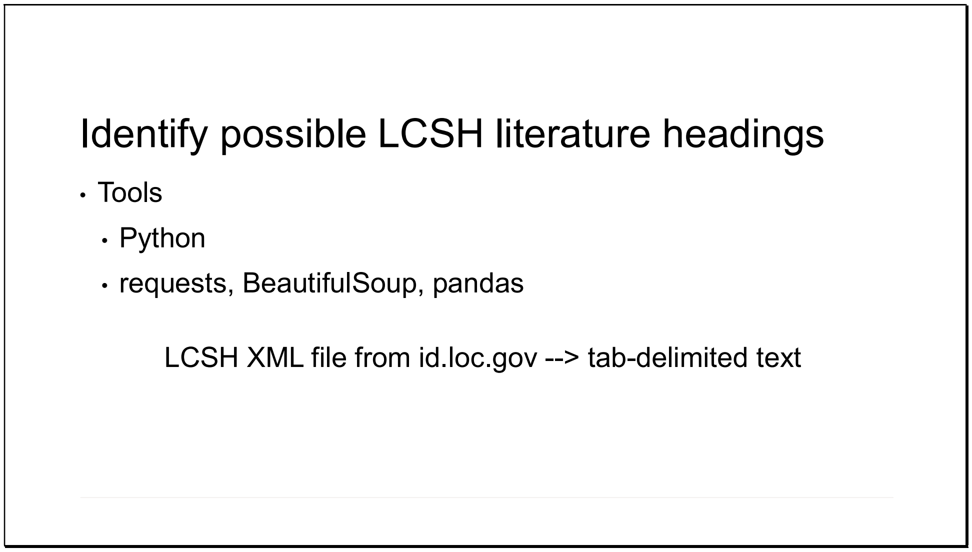
▾ Bulk exports - MADS/RDF
- JSONLD (159 (MB); info)
- NT (276 (MB); info)
- TTL (102 (MB); info)
- XML (196 (MB); info)

▾ Bulk exports - SKOS/RDF
- JSONLD (70 (MB); info)
- NT (106 (MB); info)
- TTL (43 (MB); info)
- XML (76 (MB); info)

https://id.loc.gov/download/

I started by downloading all of LCSH. LCSH can be downloaded in various formats from id.loc.gov[4].

[4] https://id.loc.gov/download/



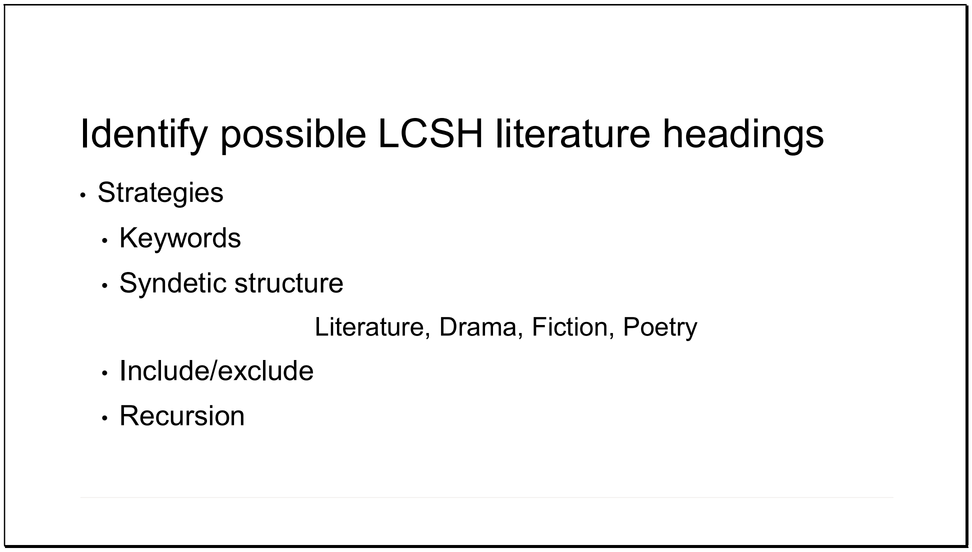
Identify possible LCSH literature headings

- Tools
  - Python
  - requests, BeautifulSoup, pandas

LCSH XML file from id.loc.gov --> tab-delimited text

I used Python and some of its libraries to download and manipulate the LCSH file. There are many ways that this could be done. I chose approaches that I was already familiar with. One nice thing about Python is that there are many libraries that make it easier to do various things. Requests is a library that helps you get information from the web. I downloaded the XML version of LCSH and used BeautifulSoup to parse it. BeautifulSoup is a library that will extract information from HTML and XML documents. Finally, pandas is one of my favorite Python libraries. It enables you to put data into tabular format and then manipulate it in many ways.

Once I had downloaded the LCSH file, I used a Python script to evaluate each record and decide if it met certain conditions that I'd set. If the record was a match, I added the fields that I was interested in to a pandas DataFrame, which is a tabular format like Excel. Pandas DataFrame's can be exported in various formats, including tab-delimited text and Excel.



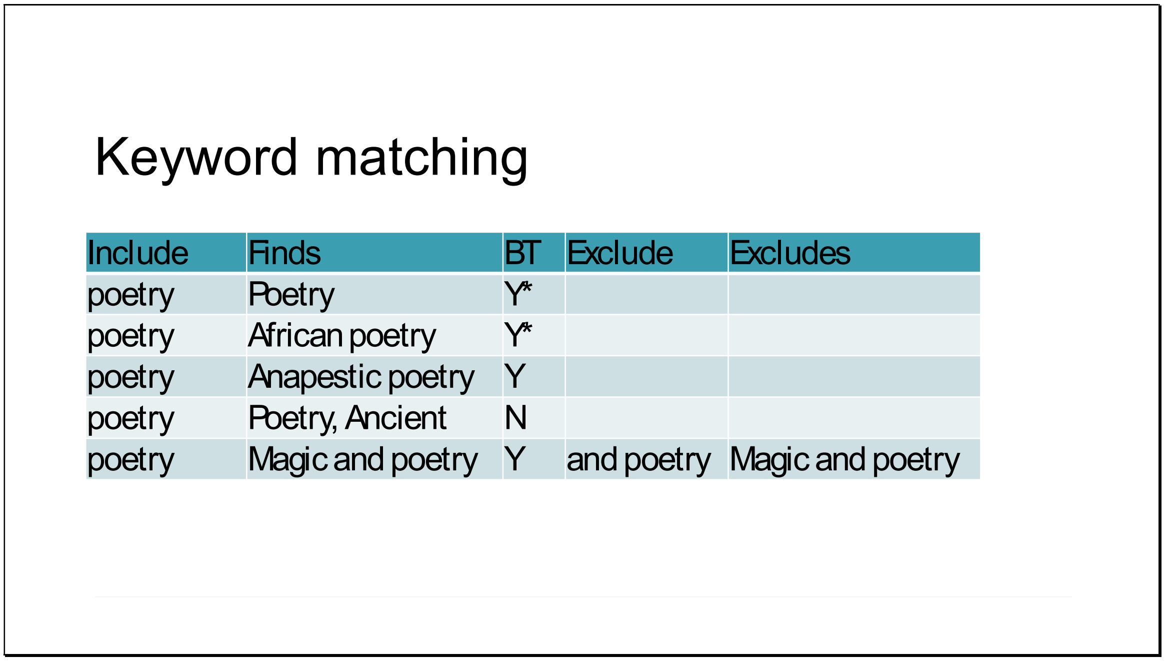
Identify possible LCSH literature headings

- Strategies
  - Keywords
  - Syndetic structure

    Literature, Drama, Fiction, Poetry
  - Include/exclude
  - Recursion

I employed two main strategies for finding literature-related subject headings. The first was looking for keywords in the subject heading or in cross-references. I started with some obvious basic headings for literary forms like the one shown below. The second was looking for narrower terms under broader terms of interest. Again I started with some basic headings for literary forms.

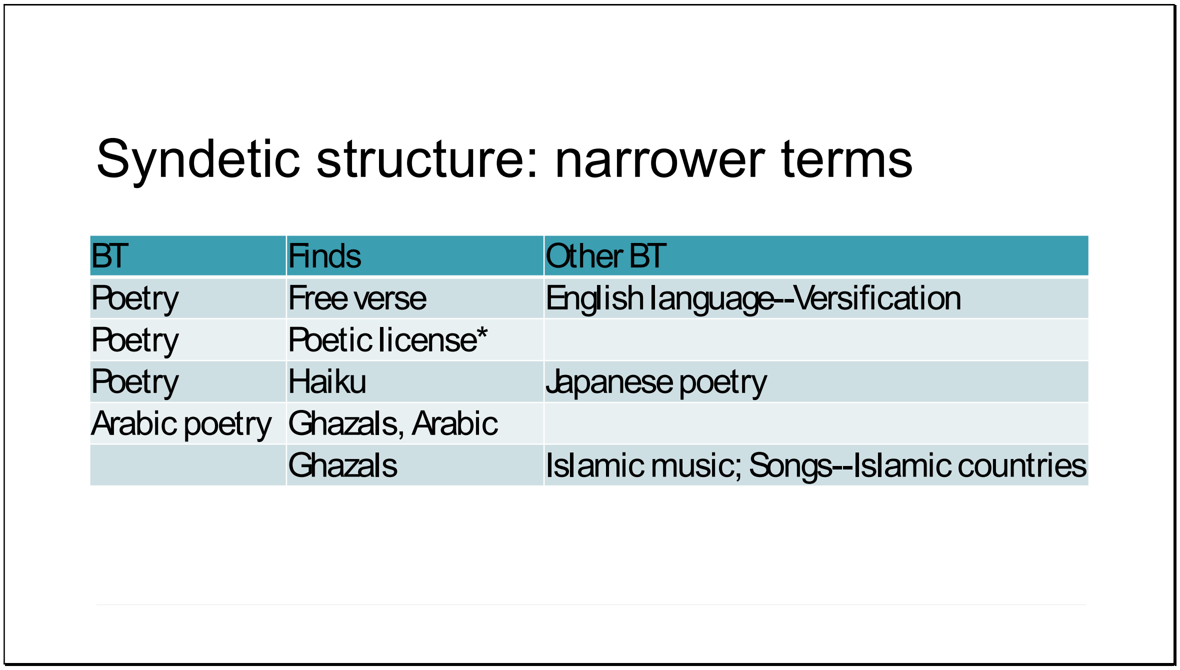
As I talk about examples, I will also discuss two techniques that were necessary: incorporating both inclusion and exclusion and recursion.



Keyword matching

Include	Finds	BT	Exclude	Excludes
poetry	Poetry	Y*		
poetry	African poetry	Y*		
poetry	Anapestic poetry	Y		
poetry	Poetry, Ancient	N		
poetry	Magic and poetry	Y	and poetry	Magic and poetry

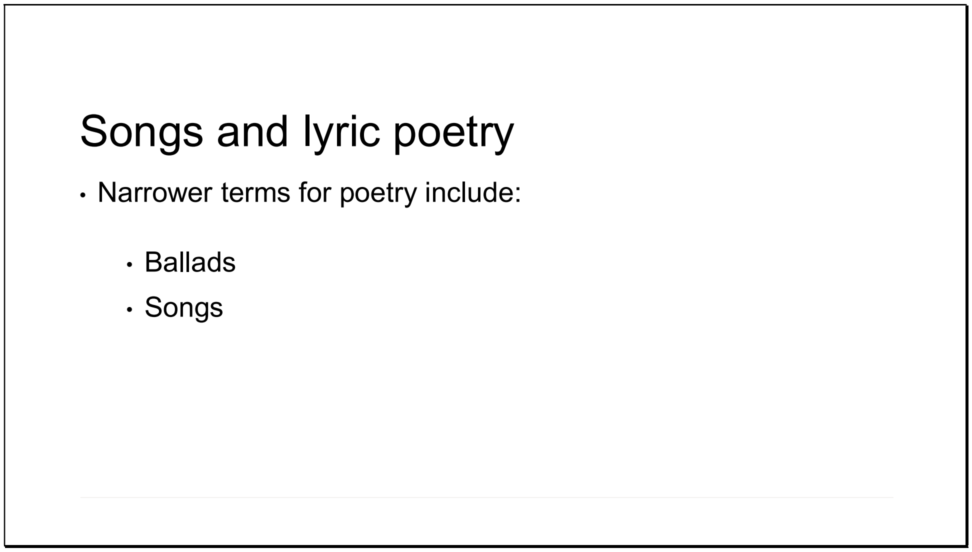
Here are some examples of how I approached keyword matching using the keyword "poetry." The search, of course, finds the base heading of "poetry." It also finds headings for forms of poetry that have been traditionally used as a genre/form in LCSH for anthologies, such as African poetry and anapestic poetry. Both of these would also have been found through the syndetic structure, although African poetry is a narrower term of African literature and not poetry. Using only the hierarchical structure would not be sufficient though, as there are headings which do not have broader terms or do not have relevant broader terms. For example "poetry, ancient" has no broader terms at all. The keyword search is also subject to false drops, such as things like "poetry and magic." To fix this, I developed a list of keywords and broader terms to exclude. After a heading met the criteria to include it as a match, I then evaluated it against the exclude criteria and removed the unwanted headings.



Syndetic structure: narrower terms

BT	Finds	Other BT
Poetry	Free verse	English language--Versification
Poetry	Poetic license*	
Poetry	Haiku	Japanese poetry
Arabic poetry	Ghazals, Arabic	
	Ghazals	Islamic music; Songs--Islamic countries

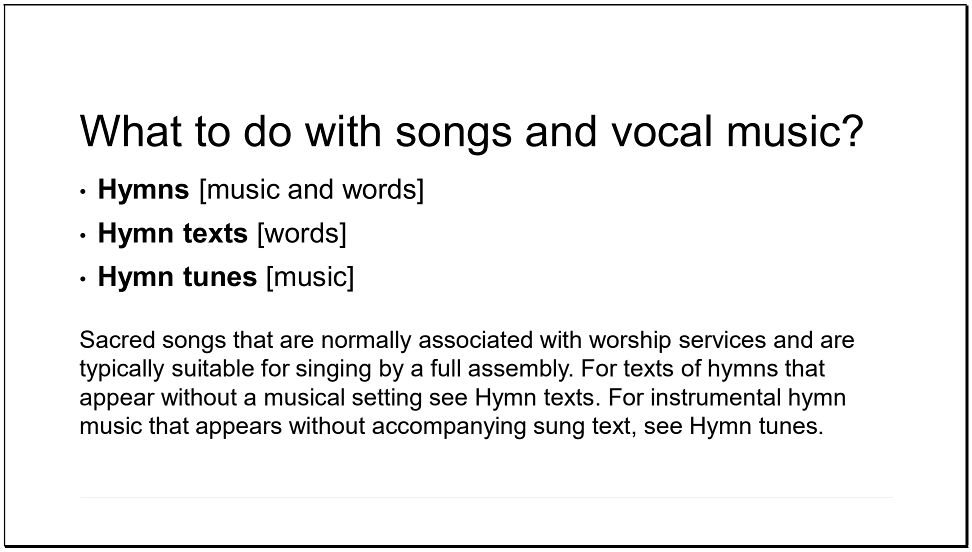
Here are some examples of using the syndetic structure to look for narrower terms of the heading “poetry” and some related headings. The heading “poetry” has numerous forms of poetry under it that do not include the word poetry, such as “free verse.” This also enabled me to flag “verse” as a potential new term for keyword searching for the next pass. Some of the narrower terms under poetry, such as “poetic license,” are not forms of poetry. These were manually excluded. Some of the narrower terms under “poetry” have additional broader terms that I added as additional broader terms to search for their narrower terms. The broader term Japanese poetry on the haiku record is an example of this. Unlike haiku, many culture- or language-specific forms of poetry are listed as narrower terms only under general forms like poetry qualified by language, nationality or ethnic group. For example, the heading “ghazals, Arabic” can only be found by searching for narrower terms under “Arabic poetry.” I then added the plain term “ghazals” to the keyword inclusion list. This elicited the plain heading “ghazals,” which turns out to have no literature-related broader terms at all. It is listed only under “Islamic music” and “songs--Islamic countries.” I can give you many more examples, but hopefully these demonstrate why it is important to use both keyword and broader/narrower term approaches, to use both include and exclude lists for matching and to iterate multiple times after adding new inclusion and exclusion criteria.



Songs and lyric poetry

- Narrower terms for poetry include:
  - Ballads
  - Songs

One of the more interesting things that came up during this process that I did not anticipate is the syndetic relationship between songs and lyric poetry. Both ballads and songs are narrower terms of poetry, as well as of vocal music. Confusingly, ballads are also a narrower term of songs. Through the syndetic structure, it seems like you can link all of vocal music to lyric poetry.

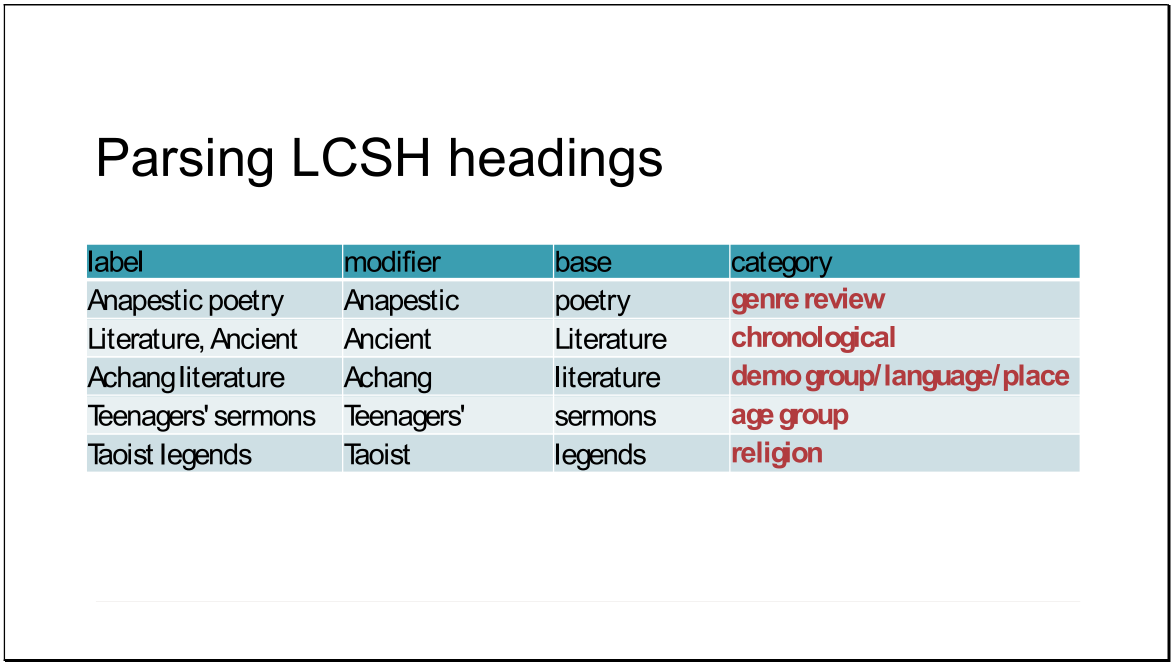


What to do with songs and vocal music?

- **Hymns** [music and words]
- **Hymn texts** [words]
- **Hymn tunes** [music]

Sacred songs that are normally associated with worship services and are typically suitable for singing by a full assembly. For texts of hymns that appear without a musical setting see Hymn texts. For instrumental hymn music that appears without accompanying sung text, see Hymn tunes.

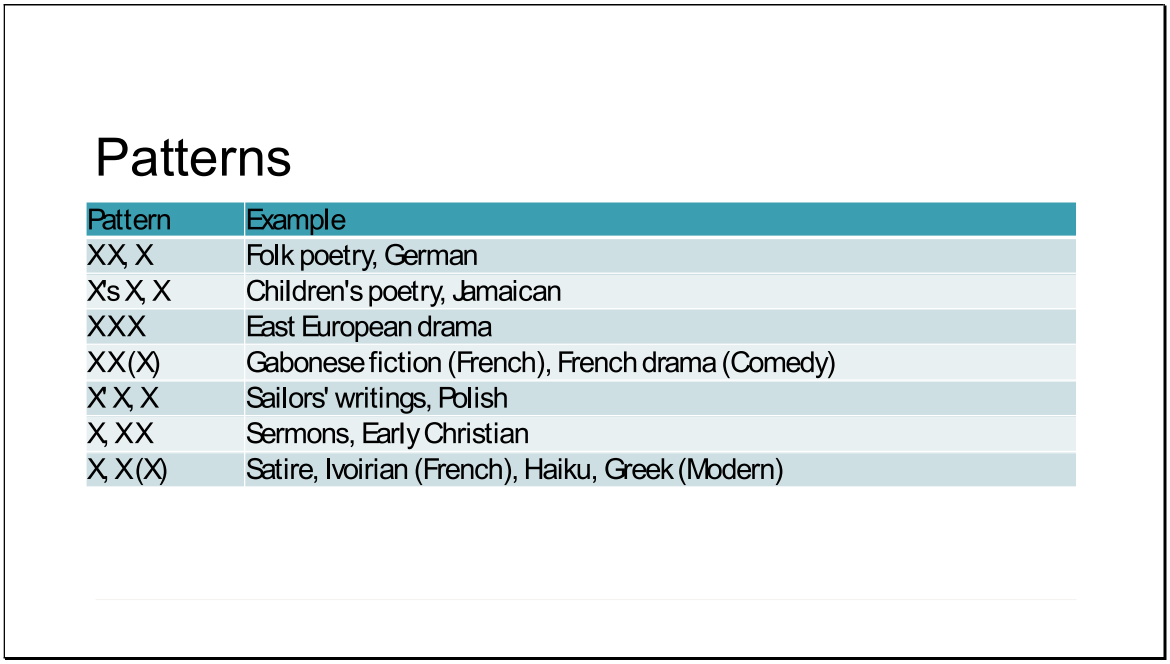
We’re not completely sure how to handle these headings, but are inclined to follow the pattern established in LCGFT for hymns. The plain term hymns is used for scores and recordings with words and music. Hymn tunes is used when there is only the music. Before the literature project, the relevant heading is hymn texts for the words without music. The broader terms for hymn texts are song texts and sacred music texts. There are a great many LCSH headings for various types of songs and most of these probably do not appear in textual form without their music. Do we map those to broader terms like song texts or try to anticipate more specific terms? There are also LCSH headings for specific types of hymns, such as Advent hymns. Do these need equivalent LCGFT text terms? Or just some combination of broader terms?



Parsing LCSH headings

label	modifier	base	category
Anapestic poetry	Anapestic	poetry	genre review
Literature, Ancient	Ancient	Literature	chronological
Achang literature	Achang	literature	demo group/language/place
Teenagers' sermons	Teenagers'	sermons	age group
Taoist legends	Taoist	legends	religion

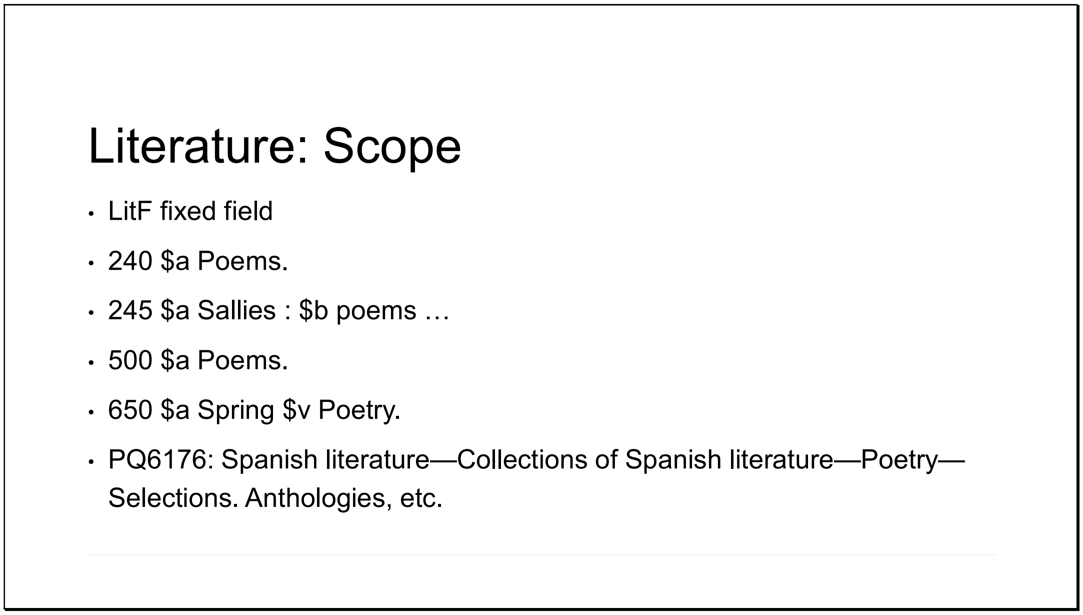
The task group is currently working on parsing candidate LCSH terms to disentangle the parts that refer to genre or form from information that needs to be mapped somewhere else. Most of this additional information refers to one or more of the following: a demographic group, a language or a place. At this point, we are not trying to tease these apart, but are just throwing them into a bucket for later review. There are a smaller number of headings that included information about a chronological period or an age group. Finally, we have tried to separate out some potentially problematic headings related to religious groups that could be interpreted as demographic groups, but not necessarily. These headings may also need to be mapped to some sort of genre/form term that brings out the religious aspect.



Patterns

Pattern	Example
XX, X	Folk poetry, German
X's X, X	Children's poetry, Jamaican
XXX	East European drama
XX(X)	Gabonese fiction (French), French drama (Comedy)
X' X, X	Sailors' writings, Polish
X, XX	Sermons, Early Christian
X, X(X)	Satire, Ivoirian (French), Haiku, Greek (Modern)

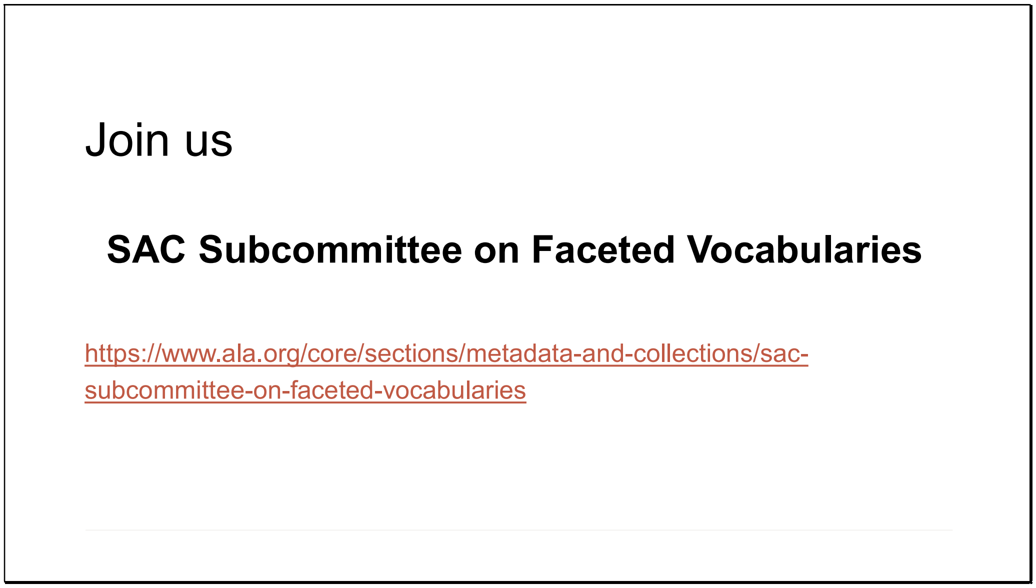
Once we have identified all the genre/form terms and their LCGFT equivalents, as well as the various terms that are used to modify LCSH headings for literary forms, we will need to analyze how these are put together. For example, in our current pool of LCSH headings potentially related to literary forms, there are fifteen patterns for headings that contain three words. This slide shows the most common patterns, as well as some examples. There are a great many building blocks, which are put together in a multitude of ways.



Literature: Scope

- LitF fixed field
- 240 $a Poems.
- 245 $a Sallies : $b poems …
- 500 $a Poems.
- 650 $a Spring $v Poetry.
- PQ6176: Spanish literature—Collections of Spanish literature—Poetry—Selections. Anthologies, etc.

Identifying records for literature to enhance with faceted metadata will be tricky. The low hanging fruit is the literary form fixed field. This has limitations, though. It is often not coded in older records. Even newer records often lack the specific codes for literary forms beyond the binary 1 fiction and 0 nonfiction. The single byte is also incapable of recording information for works that contain more than one literary form, such as an anthology that includes both poetry and short stories. Other possible strategies include collective uniform titles, subtitles, notes, subject subdivisions and classification numbers. Even with all these approaches, the results are likely to be error-prone, due to changes in cataloging practice over time and incorrect cataloging. Resources that include both literature and criticism may also be problematic. Many critical editions of literary works have subject headings for criticism, but nothing to indicate the literary form or forms contained within them. If the scoping is done badly, we will end up with the worst of all worlds in terms of separating topical and genre form terms. Unfortunately, there seems to be someone running some sort of automated process in WorldCat that is incorrectly generating genre headings from topical headings. For example, the overwhelming majority of records with the LCGFT genre/form term “free verse” in our catalog appear to actually be books about free verse. Ultimately, we may need to rely on the development of machine learning techniques that can identify literary forms and genres from the full text of cataloged works, rather than the metadata.



Join us

SAC Subcommittee on Faceted Vocabularies

https://www.ala.org/core/sections/metadata-and-collections/sac-subcommittee-on-faceted-vocabularies

Finally, if you find this kind of work interesting and are a member of ALA Core, consider volunteering to join our subcommittee. You can find more information at https://www.ala.org/core/sections/metadata-and-collections/sac-subcommittee-on-faceted-vocabularies.