

Issues in Advanced Search

With linked data



Overview

- Kinds of searching
- Searching at ID.loc.gov
- Resource-Centric Approach vs Bibliographic Record
- Not at my institution
- SPARQL vs Text
- Simplifying Structures
- Data Cleanup, Normalization



Kinds of searching

- Known Item retrieval (API or interactive)
 - machine verification, system to system updates
- Cataloger search
 - catalogers doing authority work, looking for errors
 - attaching new descriptions to existing ones
 - know the jargon, terms, rules
 - probably more likely to want advanced searching
- End user discovery
 - open ended search for a thing
 - don't know the jargon, terms, rules
 - may just submit text strings and hope for the best
 - ...or they might like to be guided progressively...

Searching at ID.LOC.GOV

Exact Match of known terms works well

The Library of Congress > [Linked Data Service](#)

Refine your results

Scheme ▾

- BIBFRAME Instances 3
- BIBFRAME Works 3
- Name Authority 1

Type ▾

- Instance 3
- Print 3

Results: 1-7 of 7 < 1 >


Label	Dataset	Type	Subdivision	Identifier
1. Traille, Kay	LC Name Authority File (LCNAF)	PersonalName SimpleType Name Authority		no2020002305

Searching at ID.LOC.GOV

- Simple search returns a lot of resources
- Need progressive elaboration or more fields
- What does “Jones” mean on an Instance? Not the creator.
- Current architecture is more for systems than humans, but facets help.

The Library of Congress > [Linked Data Service](#)

Refine your results

Scheme 

- BIBFRAME Instances 53,555
- BIBFRAME Works 43,475
- Name Authority 19,164
- BIBFRAME Hubs 2,117
- Providers 580
- LC Classification 130
- Subject Headings 107
- Cultural Heritage Orgs 29
- Children's Subject Headings 2
- Genre/Form Terms 2
- Roles 1

Search: **GO** **RESET**

Results: 1-20 of 119,236 < 1 2 3 4 5 6 ... >

Label	Dataset	Type	Subdivision	Identifier
1. Jones	BIBFRAME Works	Work Text		14209365
2. Jones	LC Name Authority File (LCNAF)	PersonalName SimpleType Name Authority		nb2011013551
3. Jones	BIBFRAME Works	Work Audio		19256068

ADVANCED SEARCH ISSUES. PRESENTED AT BIBFRAME WORKSHOP IN EUROPE 2022

Searching the Internal catalogers database

BIBFRAME Database

Search

Works Instances Items

Refinements

Format

Text [17,198,452]
StillImage [1,067,557]
Audio [664,600]
Multimedia [449,048]
Cartography [386,469]
MovingImage [368,171]
NotatedMusic [262,510]
Work [16,762]
MixedMaterial [14,084]
true [493]
Object [217]
Collection [71]
Monograph [14]
Serial [5]

Results 1 - 10 of about 20,428,456 (works)

Sort by **Relevance** Go

1. [Vivaldi, Antonio, 1678-1741](#)
NotatedMusic
Vivaldi, Antonio, 1678-1741.
M2076.V
2. [Delfosse, Heinrich P](#) (Wo)
Text
Delfosse, Heinrich P

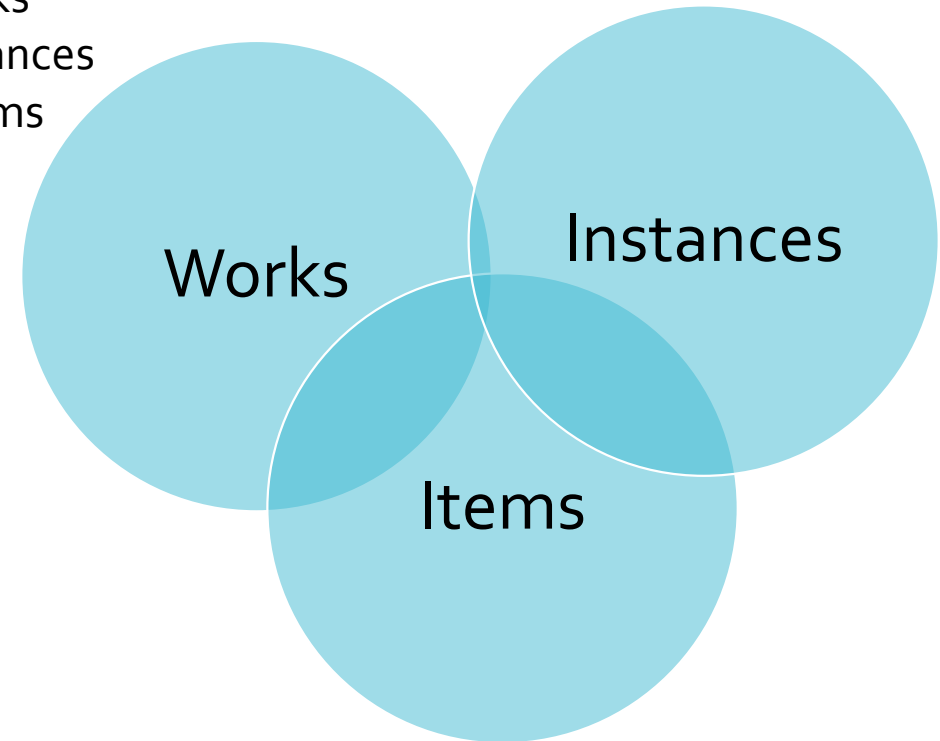


Resource-Centric vs Bibliographic Record Searching

- In BIBFRAME, (and WEMI, LRM) we've atomized the record into more coherent authoritative parts, but they are separate.
- If you are querying a particular term, you will get back the resources with that term.
- But if you want to search a catalog, you still need some notion of a record
 - Is that the Instance?
 - Can we build a search around returning Instances no matter what resource a hit was made on?
 - Yes, but it's complicated.

Constructing a complex search

- Search :
 - Title = cats 40,000 Works
 - creator = Jones 4,000 Works
 - topic= Korea 3,000 Works
 - pub year = 2001 13,000 Instances
 - Location = MotionPicture RR 100,000 Items
- Where to start the query?
- What if the Item search was for barcode or shelflist?



Index key terms on multiple resources?

Here we've added the pub date to the Work descriptions, so we can display all the shelflisting criteria

Browse Shelflist from Bibliographic Records

[← Back to home page](#)



[Previous](#)



[Next](#)

Jump to:

PN 4899.M9 M57

Found in bibliographic resources

PN009.K6 Y52	[1]	Yi, Chae-ch'öl, 1931-	Han'guk hyöndae adong munhaksa ūi yŏn'gu	1968	• Children's literature, Korean--History and criticism.
PN0231.I5 M4	[1]	Meyer, Jerome S. (Jerome Sydney), 1895-1975.	Advice on the care of babies,	1927	• Infants--Anecdotes, facetiae, satire, etc.
PN0465.B7 P3	[1]	Tardel, Hermann, 1869-	Bremen im sprichwort,	1929	• Proverbs, German.



Not at my Institution

- We have URLs and identifiers from each others resources.
- At LC, we have OCLC numbers, Getty vocabulary terms, FAST identifiers, etc.
- How much of an external resource do we store/index:
 - Is the URI and the label sufficient?
 - If we fetch it once, when do we go back and get it again?
 - How is it made available?
 - Can I just get a “label” change?
- (See current efforts at the Program for Cooperative Cataloging’s “BIBFRAME Interchange Group”)

SPARQL vs Text searching

- Do we have a triplestore with SPARQL or a SOLR/Lucene-type search system?
- SPARQL is great for relating between resources
 - for node-based display/browsing
 - refining searches to always return Instances on a complex query
- String searches are much more performant in a text based index
- Need both, but with improvements.
- Here's an example of a combination of text searching and SPARQL in play:
 - <https://id.loc.gov/search/?q=cs:http://id.loc.gov/authorities/names&q=Traille,%20Kay>
 - <https://id.loc.gov/authorities/names/no2020002305.html>



Simplifying Structures

- For SPARQL queries to work better, we need to simplify our triples
- BIBFRAME tries to accurately reflect the multi-node complexity of terms in bibliographic description, creating blank nodes and stuff that is not expressed *simply* in triple form, eg:
 - Qualified identifiers (“invalid”, “CD ROM only” etc.)
 - Notes with their types

(This is because of cataloging rules, not because it came from MARC.)

- At LC, we “manage” or index as triples only a small set of the triples from BIBFRAME

Simplifying Structures

Two examples:

- Index FAST topics instead of indexing LCSH MADSRDF
- Provision Activities:
 - If we really need to know :

```
<bf:provisionActivity><bf:Publication>
  <bf:place><bf:Place>
    <rdfs:label>Cairo</rdfs:label>
  </bf:Place>
</bf:place>
<bf:agent><bf:Agent>
  <rdfs:label>al Mokattam Print. Off.</rdfs:label>
</bf:Agent></bf:agent>
  <bf:date>1907</bf:date>
</bf:Publication>
</bf:provisionActivity>
```

- Can we just index it as:

```
publishedIn: Cairo
publishedOn: 1907
publishedBy: Mokattam Print. Off.
```

Data Cleanup, Normalization

- Need more URIs
- Data in bibliographic descriptions can be messy
 - from typos
 - from cataloging rules requiring transcription (can we just take a picture?)
 - just too hyper-specific to turn into a relationship:

Published in norwegian with title

<http://id.loc.gov/entities/relationships/publishedinnorwegianwithtitle>

Instance Of

[BFLC Relation](#)

Variants

Also published in Norwegian with title

- How about **bf:translation** instead of **rel:publishedinnorwegianwithtitle**?

Which kinds of search can we support?

Free form, fielded:

Picklist, guided:

Language	English	▼
Classification	A General Works	▼
Format	Electronic	▼
Location	Main RR	▼
title	creator	
<input type="button" value="Submit"/>		

API:

https://id.loc.gov/authorities/subjects/suggest2?q=History&memberOf=http://id.loc.gov/authorities/subjects/collection_Subdivisions



Marching Orders:

Clean up data

Simplify for indexing

Make more URIs

Exchange updates with each other

Thanks for listening!

- Nate Trail
- NDMSO
- Library of Congress
- <https://www.loc.gov/bibframe>
- <https://id.loc.gov/>