Remapping Problematic LCSH in TRLN Discovery

Kristina M. Spurgin

2018-11-15

Introduction (About me)

- Library Data Strategist, Head of Library Data Strategy & Services
- TRLN Discovery Developer Team

Speaker notes

My title is Library Data Strategist and I am head of the Library Data Strategy & Services (LDSS) section in Davis Library Tech Services.

LDSS was created as a service point for our colleagues across the libraries. We help folks solve workflow issues and execute projects by using technology to leverage library data at scale. (If this means nothing to you, follow the LDSS link there when you download the slides to learn more.) By "we," I mainly mean jcm who is the data wizard to the "data magician" folks think I am. :-)

We also have 3 ongoing routine responsibilities:

- MARC metadata for monographic and integrating resource online resources. Connie Israel and jcm keep this running smoothly despite challenges you wouldn't believe.
- non-MARC metadata policy, creation, assessment, remediation, and consultation for the institutional repository, digital collections, and other projects across the Libraries. Anna Goslen and Julia Gootzeit are the powerhouse team on this.
- And the one that's relevant to this presentation: maintaining the processes that extract our entire catalog (Sierra) and transform it for ingest into our shared online catalog.

I'm also the data expert on the TRLN Discovery Developer Team. TRLN Discovery is the new shared online catalog we've been building for the past couple of years. Which brings me to the question of…

Introduction (About the shared online catalog)

- TRLN initiative since 2007
- Legacy to now – Endeca
  - http://search.trln.org/
  - https://search.lib.unc.edu/
  - https://search.lib.unc.edu/filmfinder/
- Coming soon – TRLN Discovery
  - https://disco-qa.lib.unc.edu/

Speaker notes

That it is a TRLN initiative means the catalog is shared between Duke, NC Central, NC State, and UNC Libraries.

The idea behind this is:

Each institution smooshes the data from their ILS-based catalog, digital collections, and other sources into a central, shared index.

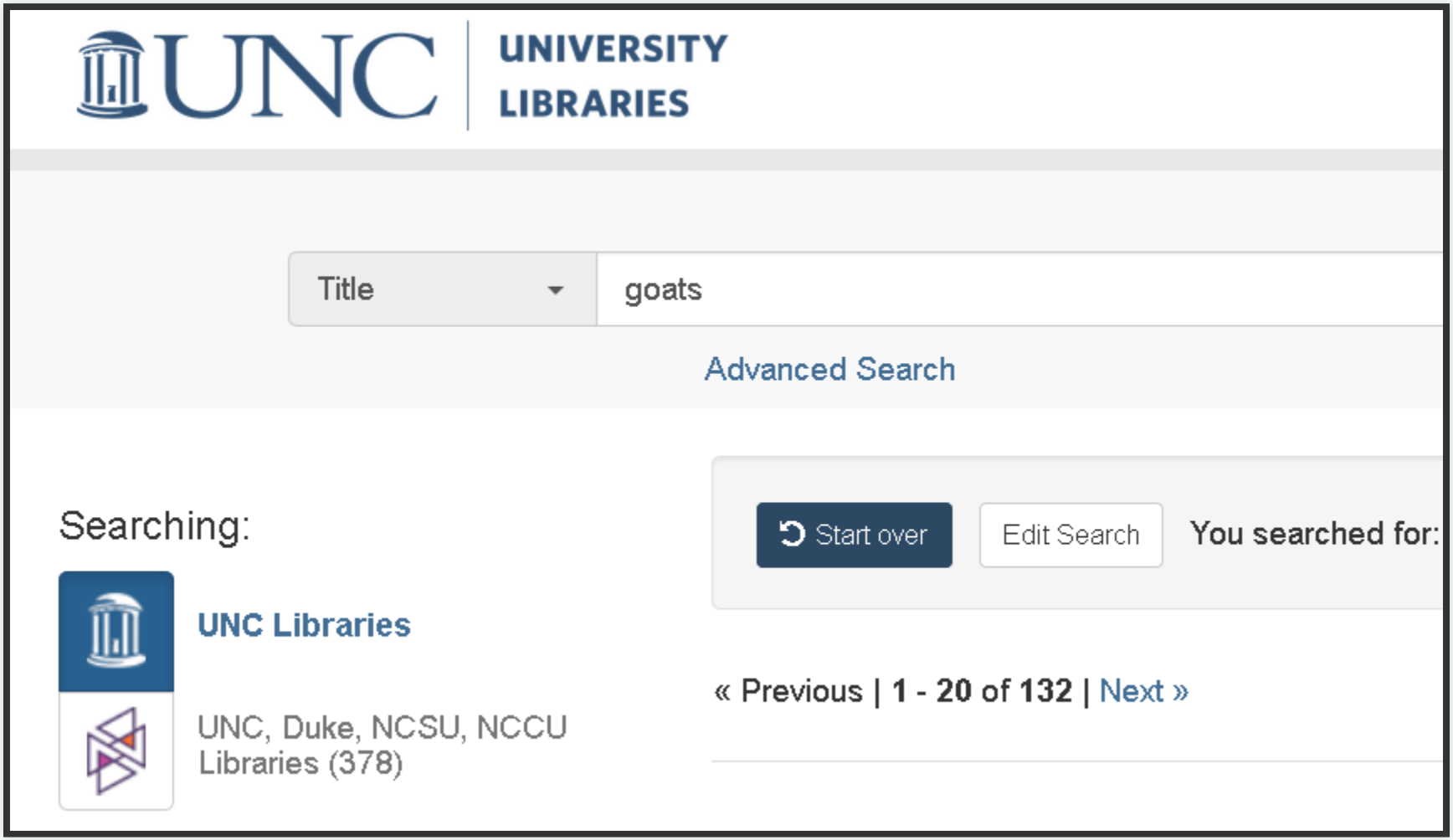
The online catalog applications (or catalog websites) are built using that shared index as the underlying data source.

Currently we have a proprietary shared index built using an Oracle e-commerce product called Endeca. When you use SearchTRLN, you are searching and seeing all the data in the entire Endeca index. When you use search.lib.unc.edu, you are searching that same index, but it's pre-scoped to show only items held by UNC. When you use FilmFinder, it's searching that same index, but pre-scoping even narrower to show only UNC-held films, videos, etc.

We are excited to be building a new shared index and associated online catalog web applications based on open-source technologies from (and/or heavily used by) the library community: TRLN Discovery.

This is going to work a bit differently than the current catalog where you have to go to SearchTRLN to search across all institutions.

"Expand to TRLN"



Speaker notes

There won't be a separate equivalent to SearchTRLN any more. UNC users will, in general, remain in the UNC catalog. But the UNC catalog gives you the ability to seamlessly toggle back and forth between UNC holdings only, and all TRLN holdings.

Here you can see that we have 132 results at UNC. By clicking on "UNC, Duke, NCSU, NCCU Libraries," we can see all 378 items held across those institutions.

TRLN Discovery: Shared index: Solr

- Solr
  - open source
  - index + search

Speaker notes

With TRLN Discovery, the shared index is built with Solr, an extremely popular open-source indexing and search platform used for all kinds of applications.

Headers

rint


```
":{
:true,


main_t:goats AND access_type_f:Online AND institution_a:unc"}},
mFound":129,"start":0,"docs":[

b3547657",
"{\"href\":\"http://purl.access.gpo.gov/GPO/LPS4781\",\"type\":\"fulltext\",\"text\":\"Open Access resource -- Full text ava
uggest":["United States. Agricultural Statistics Board",
 States. National Agricultural Statistics Service"],
:["{\"name\":\"United States. Agricultural Statistics Board\"}",
e\":\"United States. National Agricultural Statistics Service\"}"],
:["unc"],
a":["LCCN: 00230086",
em Number: 0021-N (online)"],
_a":["English"],
":"b3547657",
e_a":["Available"],
y_current_a":["Annual"],
e_a":["eng"],
d":"OCLC44616828",
k_a":["{\"title\":[\"Sheep and goats (Online)\"]}"],
ggest":["Sheep and goats."],
in":"Sheep and goats.",
rt_ssort_single":"Sheep and goats.",
ype_a":["Online"],
ion_a":["unc",


,
ber":"44616828",
main_a":["{\"type\":\"imprint\",\"value\":\"Washington, D.C. : Agricultural Statistics Board\"}"],
eral_a":["\"MtAn 5-2,\" Jan. 29, 1999-",
ption based on: Jan. 1, 1995; title from screen caption.",
 issue consulted: Feb. 1, 2008."],
mary_a":["This full-text file contains the number of sheep, value and classes; Angora goat and all goat inventory and value;
ons keeping sheep; operations and inventory by size groups, regions, and U.S.; total number on feed for slaughter. Statistic
```

Speaker notes

You can search Solr directly but it ain't easy and it ain't pretty. To find online items with goats in the title at UNC, you've got to type in a query like this:

"q":"title_main_t:goats AND access_type_f:Online AND institution_a:unc"

Imagine if you want to search across all the different types of titles!

Also, who wants to ever see a page that looks like this (except for a data nerd like me)?

Can you even see what the title of this first thing is? (Sheep and goats)

TRLN Discovery: Shared catalog application: Blacklight

- Blacklight
  - open source
  - library-created
  - discovery interface

Speaker notes

The new online catalog application that will use Solr index as a data source is a modified version of Blacklight. Blacklight is a generic library-centric discovery interface that can be used on top of an index to provide a good user experience.

UNC | UNIVERSITY LIBRARIES

HISTOR

Title | goats | Search

Advanced Search

earching:

UNC Libraries

UNC, Duke, NCSU, NCCU Libraries (378)

Limit your search

Available Online 132

Availability

Available 132

Library Location

UNC Chapel Hill 5

Davis Library 5

Start over | Edit Search | You searched for: Title > goats | Access Type > Online

« Previous | 1 - 20 of 132 | Next »

Sort by Relevance | 20 per page

1. Sheep and goats.

Washington, D.C. : Agricultural Statistics Board

Government publication, Journal, Magazine, or Periodical

Open Access resource -- Full text available

2. Rugged Rocky Mountain sheep & goats : rough and tough mountain climbers.

[Provo, UT] : USDA, Forest Service , [2000]

Book, Government publication

Open Access resource -- Full text available

Speaker notes

The new online catalog application that will use Solr index as a data source is a modified version of Blacklight. Blacklight is a generic library-centric discovery interface that can be used on top of an index to provide a good user experience.

This looks much better!

We are getting generally the same results (Sheep and goats is still the first result), but they look much nicer! We are getting a few more results because we are able to tell Blacklight that a title search should search a bunch of fields: main title, variant title, included title, related title, earlier title, etc.

We have the facets, and a checkbox to limit to online things – all the user interface stuff.

Headers

rint


```
":{
:true,



main_t:goats AND access_type_f:Online AND institution_a:unc"}},
mFound":129,"start":0,"docs":[

b3547657",
"{\"href\":\"http://purl.access.gpo.gov/GPO/LPS4781\",\"type\":\"fulltext\",\"text\":\"Open Access resource -- Full text ava
uggest":["United States. Agricultural Statistics Board",
 States. National Agricultural Statistics Service"],
:["{\"name\":\"United States. Agricultural Statistics Board\"}",
e\":\"United States. National Agricultural Statistics Service\"}"],
:["unc"],
a":["LCCN: 00230086",
em Number: 0021-N (online)"],
_a":["English"],
":"b3547657",
e_a":["Available"],
y_current_a":["Annual"],
e_a":["eng"],
d":"OCLC44616828",
k_a":["{\"title\":[\"Sheep and goats (Online)\"]}"],
ggest":["Sheep and goats."],
in":"Sheep and goats.",
rt_ssort_single":"Sheep and goats.",
ype_a":["Online"],
ion_a":["unc",


,
ber":"44616828",
main_a":["{\"type\":\"imprint\",\"value\":\"Washington, D.C. : Agricultural Statistics Board\"}"],
eral_a":["\"MtAn 5-2,\" Jan. 29, 1999-",
ption based on: Jan. 1, 1995; title from screen caption.",
 issue consulted: Feb. 1, 2008."],
mary_a":["This full-text file contains the number of sheep, value and classes; Angora goat and all goat inventory and value;
ons keeping sheep; operations and inventory by size groups, regions, and U.S.; total number on feed for slaughter. Statistic
```

Speaker notes

Ok, I know this feels like a bit of a deep dive without getting to the point, but I promise this is useful foundation.

Let's look at the data in the Solr index again. This is the bibliographic data that the shared online catalog knows about and has available to work with.

What do you NOT see here?

(MARC!)

TRLN Discovery: Shared data model: Argot

- Argot is our name for our shared data model
  - MARC -> Argot
  - Digital collections -> Argot
  - ICPSR DDI metadata -> Argot
- Argot specification is publicly available

Speaker notes

We call our shared data model Argot. It's an apt name because it's defined as "the jargon or slang of a particular group or class."

Any data or records that will be included in TRLN Discovery must be transformed into Argot. This includes:

- ILS MARC bibliographic and holdings data
- ILS non-MARC item and order data
- Digital collections data (Dublin Core, MODS, RDF…)
- External record sets such as ICSPR datasets described in DDI XML

The fields and elements in Argot abstract away from any of these original data formats, translating the data into a common, generalized format **optimized for search and desired display behavior** in the end-result tool we are designing.

Designing Argot has been an iterative process, and the specification is currently a bit of a mess, but it IS publicly available and I'm working on producing more friendly views of it.

Representing subject data in Argot

- subject_headings
- subject_suggest
- subject_topical
- subject_chronological
- subject_geographic
- subject_headings_remapped

Speaker notes

Here's a good example of what I mean when I say we transform the original data source (MARC) to something optimized for TRLN Discovery's desired behavior (Argot).

In the MARC record, you record a subject heading or index terms in one of the many 6XX fields and you are done.

But each 6XX field (or parts of it) gets mapped to at least 3 of the fields shown here — sometimes a single 6XX gets mapped to all of them!

Subject data in Argot: subject_headings

- indexed for keyword and subject search
- displayed in full record view with adaptive hyperlinking

=650 \0$aAbolitionists$zNorth Carolina$xHistory$y19th century$vSources.


```
"subject_headings": [
    "Abolitionists -- North Carolina -- History -- 19th century -- Sources
  ]
```


Example mappings: subject_headings

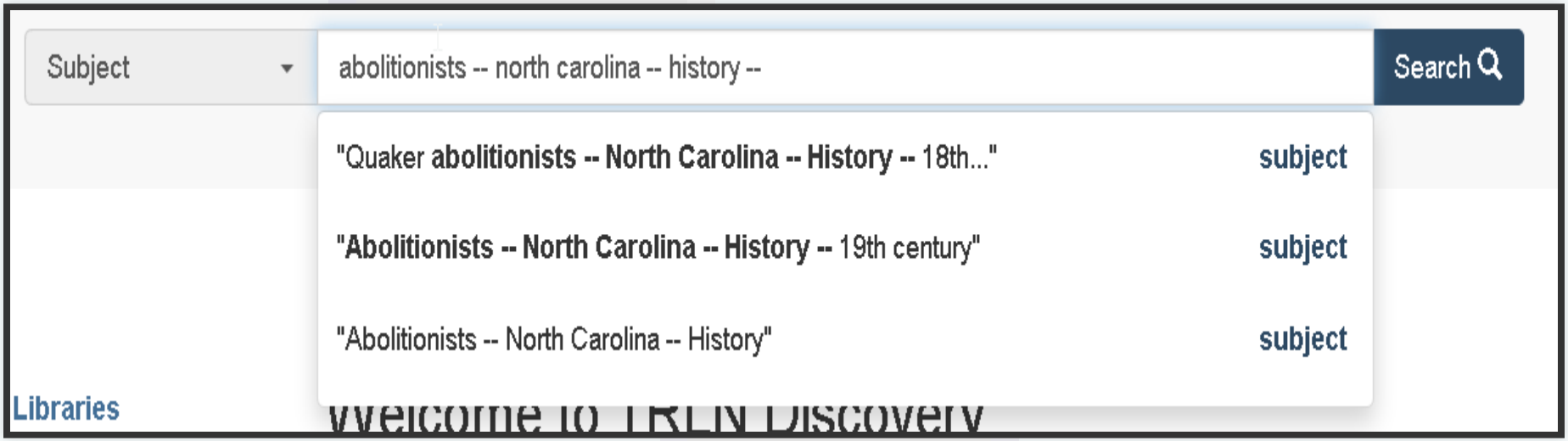
element/field	subelement/subfield(s)	constraints	processing_type
653	a	i2 =~ /[012345]/	subelement_to_value
662	abcdfgh	none	concat_subelements
656	akvxyz	none	concat_subelements
657	avxyz	none	concat_subelements
600	abcdfghjklmnopqrstuvxyz	none	concat_subelements
610	abcdfghklmnoprstuvxyz	none	concat_subelements
650	abcdgvxyz	none	concat_subelements
611	acdefghklnpqstuvxyz	none	concat_subelements
647	acdgvxyz	none	concat_subelements
630	adfghklmnoprstvxyz	none	concat_subelements
651	agvxyz	none	concat_subelements
648	avxyz	none	concat_subelements

Subject data in Argot: subject_suggest

- searched when you start typing in keyword or subject search box
- displayed in auto-suggest dropdown

=650 \0$aAbolitionists$zNorth Carolina$xHistory$y19th century$vSources.

```
"subject_suggest": [
  "Abolitionists -- North Carolina -- History -- 19th century -- Sources"
]
```



Subject data in Argot: subject_topical

- populates "About Topic" facet

=650 \0$aAbolitionists$zNorth Carolina$xHistory$y19th century$vSources.

```
"subject_topical": [
  "Abolitionists", "History"
]
```


About Topic
Abolitionists 47
History 35
Antislavery movements 28
Slavery 23
African Americans 11
Fugitive slaves 9
Quakers 9
Slaves 8
Hedrick, Benjamin Sherwood, 1827-1886 7
Politics and government 7
more »

Example mappings: subject_topical

element/field	subelement/subfield(s)	constraints	processing_type
600	abcdfghjklmnopqrstu	none	concat_subelements
600	x	none	subelement_to_value
610	abcdfghklmnoprstu	none	concat_subelements
610	x	none	subelement_to_value
611	acdefghklnpqstu	none	concat_subelements
611	x	none	subelement_to_value
630	adfghklmnoprst	none	concat_subelements
630	x	none	subelement_to_value
647	acdg	none	concat_subelements
647	x	none	subelement_to_value
648	x	none	subelement_to_value
650	abcdg	none	concat_subelements
650	x	none	subelement_to_value
651	x	none	subelement_to_value

653	a	i2 = 4	subelement_to_value
653	a	i2 =~ /[0123]/	subelement_to_value
656	a	none	subelement_to_value
656	x	none	subelement_to_value

Subject data in Argot: subject_chronological, subject_geographic

- populates "About Time Period" and "About Places" facets

=650 \0$aAbolitionists$zNorth Carolina$xHistory$y19th century$vSources.


```
"subject_chronological": [
  "19th century"
],
"subject_geographic": [
  "North Carolina"
]
```

About Time Period

19th century	22
Civil War, 1861-1865	7
1775-1865	4
1865-1950	3
1861 - 1865	2
18th century	2
1861-1865	1
John Brown's Raid, 1859	1
Revolution, 1775-1783	1

About Places

North Carolina	52
United States	27
Southern States	5
Ohio	4
Arkansas	2
Cabarrus County (N.C.)	2
Confederate States of America	2
Cottage Home Plantation (Lincoln County, N.C.)	2
Dallas County (Ala.)	2
Indiana	2

more »

Subject data in Argot: subject_headings_remapped

- NOT displayed in record, facets, or auto-suggest
- indexed for keyword and subject search

=650 \0$aIllegal aliens$zEurope.


```
"subject_topical": [
    "Undocumented immigrants"
  ],
  "subject_headings": [
    "Undocumented immigrants -- Europe"
  ],
  "subject_headings_remapped": [
    "Illegal aliens -- Europe"
  ]
```


Behavior in TRLN Discovery

- Search for subject: "illegal aliens"
  - 1993 results; not seen in About Topic facet, record; autosuggest issue
- Search for subject: "undocumented immigrants"
  - 1974 results; seen in About Topic facet, record

Not perfect yet... Proof of concept stage! Some weird glitches to look into

What to remap and what to map it to?

- Initial list for proof of concept
- Not complete
- Need TRLN-wide agreement
- Governance for this still undecided
- **Your input needed**

Speaker notes

My initial list was based on the following:

- CU Boulder Library's inclusive subject headings project
- Searching Twitter for #lcsh and complaints/suggestions
- Personal pet peeve headings

So what?

- This doesn't make our catalog unbiased or fully inclusive (c.f. Emily Drabinski)
- Can be seen as confusing to users – Is CU's approach better?
- Technically, not very difficult
- Small step = good step?

Speaker notes

Drabinski, Emily. "Queering the Catalog: Queer Theory and the Politics of Correction." The Library Quarterly: Information, Community, Policy 83, no. 2 (2013): 94-111. 10.1086/669547.

UX principle: don't return results that don't contain the user's query; they won't know why the result is in the set and that will be confusing. We are going to return results for "illegal aliens" where we've removed that phrase from the subject display. CU's approach added additional subject headings (searchable and displayed)

Technically, this is not very difficult. Most of the work is already done. There's some tweaking left and concerns about scaling up to large numbers of remapped headings.

The biggest concern is the decisionmaking/governance overhead. How will this be organized? Who approves these mappings? The developers want metadata folks to be the ones to maintain these mappings.