



REGULAR EXPRESSIONS

FOR DATA CLEAN UP IN SIERRA

Lloyd Chittenden
Union Catalog Coordinator
Marmot Library Network

WHAT ARE REGULAR EXPRESSIONS?

Combine literal characters and meta-characters to create complex wildcard type searches in Create Lists and Global Update

Regular expression searches are invoked with the “matches” condition in Create Lists

All searches in Create Lists and Global Update ignore case, including regular expressions

BASICS

. → match any single character

[a-z] → match any single letter

[0-9] → match any single numeral

[a-z0-9] → match any single letter or numeral

[aei] → match a single letter that is either a, e, or i
(not the string aei)

[-,] → match hyphen, comma, or space
(- must be first)

[\$|] → match dollar sign or pipe

BASICS

- * → match the preceding character 0 or more times
- + → match the preceding character 1 or more times
- ? → match the preceding character 0 or 1 times
- {4} → match the preceding character 4 times
- .^{*} → match any character 0 or more times
- [0-9]⁺ → match 1 or more numerals
- [0-9]^{*} → match 0 or more numerals
- [0-9].^{*} → match a single numeral followed by anything or nothing
- [0-9].⁺ → match a single numeral followed by one or more characters

NEGATING A CHARACTER

[^] → circumflex in brackets - negate the following character(s) in the brackets

[^a-z] → match any single character that is not a letter

[^.] → match any single character that is not a period

[^aei] → match any single character that is not a or e or i
(does not negate the string aei)

[^p][^m] → match a character that is not p followed by character that is not m

LITERALS

What if the character I'm searching for is a meta-character?

use [] around the character to search as a literal

[.] → match an actual . in the data

[?] → match an actual ? in the data

[.]b[0-9] → match any .b number

(\ will not work in Sierra to escape meta-characters)

ANCHORS

`^` → anchor to the start the field

`^1` → match every field starting with a 1

`^130` → match every field starting with 130

`^[^1]` → match every field starting with a character other than 1

`$` → anchor to the end the field

`8$` → match every field ending in 8

`[^.]$` → match every field not ending with a period

SEARCH PECULIARITIES

Searching on field group is different from searching on MARC tag
MARC Tag 245 matches “^245..to kill a mockingbird”

The screenshot shows a 'Boolean Search' window with the following details:

- Review File Name: Marmot MUG 3 LHC
- Store Record Type: BIBLIOGRAPHIC b
- Range: (dropdown)
- Start: b10000008
- Stop: b51225992
- Search Mode: Classic, Enhanced, JSON
- Search Query Table:

| Term | Operator | Type | Field | Condition | Value A | Value B |
|------|----------|--------------|--------------|-----------|-----------------------------|---------|
| 1 | | BIBLIOGRA... | MARC Tag 245 | matches | ^245..to kill a mockingbird | |

Below the table, the search criteria are displayed as: BIBLIOGRAPHIC MARC Tag 245 matches "^245..to kill a mockingbird"

At the bottom of the window, there are several buttons: Search, Use Existing Search, Retrieve Saved Query, Save, Save As, and Close.

SEARCH PECULIARITIES

Searching on field group is different from searching on MARC tag

TITLE matches “^245..|ato kill a mockingbird”

The screenshot shows a 'Boolean Search' window with the following details:

- Review File Name: Marmot MUG 3 LHC
- Store Record Type: BIBLIOGRAPHIC b
- Range: Start b10000008, Stop b51225992
- Search Mode: Classic (selected), Enhanced, JSON
- Search Query Table:

| Term | Operator | Type | Field | Condition | Value A | Value B |
|------|----------|--------------|-------|-----------|-------------------------------|---------|
| 1 | | BIBLIOGRA... | TITLE | matches | ^245.. ato kill a mockingbird | |

Below the table, the search criteria are displayed as: BIBLIOGRAPHIC TITLE matches "^245..|ato kill a mockingbird".

Control buttons on the right include: Group, Ungroup, Insert Line, Append Line, Delete, and Clear All.

Bottom navigation buttons include: Search, Use Existing Search, Retrieve Saved Query, Save, Save As, and Close.

SEARCH SPECIAL CHARACTERS

Sierra stores special characters in Unicode format.

Unicode uses a five character code to represent each character

u0040 → @

u00f1 → ñ

Sierra stores them in curly braces { }

{u0040} → @

{u00f1} → ñ

FYI, regular expressions will not work to search these:

{u....} will produce 0 results

Look up the codes here: https://en.wikipedia.org/wiki/List_of_Unicode_characters

SEARCH SPECIAL CHARACTERS

Search with the { }
(even though they are meta-characters, don't escape them)

The screenshot shows a 'Boolean Search' window with the following details:

- Review File Name: Marmot MUG 4 LHC
- Store Record Type: BIBLIOGRAPHIC b
- Range: Start b10000008, Stop b51227010
- Format: Classic (selected), Enhanced, JSON
- Search Query Table:

| Term | Operator | Type | Field | Condition | Value A | Value B |
|------|----------|--------------|-------|-----------|----------------------|---------|
| 1 | | BIBLIOGRA... | TITLE | matches | ^245.. ash{u014d}gun | |

The search results area displays: BIBLIOGRAPHIC TITLE matches "^245..|ash{u014d}gun"

At the bottom, there are buttons for: Search, Use Existing Search, Retrieve Saved Query, Save, Save As, and Close.

SEARCH SPECIAL CHARACTERS

You can also paste in the special character

The screenshot shows a 'Boolean Search' window with the following fields and options:

- Review File Name: Marmot MUG 4 LHC
- Store Record Type: BIBLIOGRAPHIC b
- Range: Start b10000008, Stop b51227058
- Options: Classic (selected), Enhanced, JSON

| Term | Operator | Type | Field | Condition | Value A | Value B |
|------|----------|--------------|-------|-----------|----------------|---------|
| 1 | | BIBLIOGRA... | TITLE | matches | ^245.. ashōgun | |

The search results area displays: BIBLIOGRAPHIC TITLE matches "^245..|ashōgun"

Control buttons on the right: Group, Ungroup, Insert Line, Append Line, Delete, Clear All

Bottom navigation buttons: Search, Use Existing Search, Retrieve Saved Query, Save, Save As, Close

THE MOST USEFUL REGULAR EXPRESSION

The regular expression I use most often is `.*`

`.*` will match anything, but not nothing

Use this to find if a field exists

DATA CLEAN UP

Search for MARC fields that should be non-MARC

ITEM BARCODE matches ^ [0-9] (5 spaces)

Marmot MUG 1 LHC

Store Record Type: ITEM i

Range Start i10000008 Stop i103961574

Classic Enhanced JSON

| Ter... | Operator | Type | Field | Condition | Value A | Value B |
|--------|----------|------|---------|-----------|---------|---------|
| 1 | | ITEM | BARCODE | matches | ^ [0-9] | |

ITEM BARCODE matches "^ [0-9]"

OK

DATA CLEAN UP

Find barcodes that start with a space

ITEM BAROCDE matches ^ [0-9]

Marmot MUG 2 LHC

Store Record Type: ITEM i

Range Start i10000008 Stop i103961586

Classic Enhanced JSON

| Ter... | Operator | Type | Field | Condition | Value A | Value B |
|--------|----------|------|---------|-----------|---------|---------|
| 1 | | ITEM | BARCODE | matches | ^ [0-9] | |

ITEM BARCODE matches "^ [0-9]"

OK

DATA CLEAN UP

Find barcodes with non-number characters

ITEM BARCODE matches `[^0-9]`

Marmot MUG barcode w letters

Store Record Type: ITEM i

Range Start i10000008 Stop i104109956

Classic Enhanced JSON

| Ter... | Operator | Type | Field | Condition | Value A | Value B |
|--------|----------|------|---------|-----------|---------|---------|
| 1 | | ITEM | BARCODE | matches | [^0-9] | |

ITEM BARCODE matches "[^0-9]"

OK

DATA CLEAN UP

Barcodes that don't start with the right number

ITEM BARCODE matches `^[^1]`

Marmot MUG bad barcodes

Store Record Type:

Range

Classic Enhanced JSON

| Ter... | Operator | Type | Field | Condition | Value A | Value B |
|--------|----------|------|---------|-----------|---------|---------|
| 1 | | ITEM | AGENCY | equal to | 170 | |
| 2 | AND | ITEM | BARCODE | matches | ^[^1] | |

ITEM AGENCY equal to "170" AND ITEM BARCODE matches "^[^1]"

OK

DATA CLEAN UP

Bad subfields

SUBJECT matches "650.*|[^avxyz20]"

Marmot MUG bad subfields

Store Record Type: BIBLIOGRAPHIC b

Range: Start: b10000008 Stop: b51227241

Classic Enhanced JSON

| Ter... | Operator | Type | Field | Condition | Value A | Value B |
|--------|----------|-------------|---------|-----------|------------------|---------|
| 1 | | BIBLIOGR... | SUBJECT | matches | 650.* [^avxyz20] | |

BIBLIOGRAPHIC SUBJECT matches "650.*|[^avxyz20]"

OK

DATA CLEAN UP

Search for non-repeatable subfields

TITLE matches ^245.*|b.*|b

Marmot MUG non-repeatable

Store Record Type: BIBLIOGRAPHIC b

Range Start b10000008 Stop b51227253

Classic Enhanced JSON

| Ter... | Operator | Type | Field | Condition | Value A | Value B |
|--------|----------|-------------|-------|-----------|--------------|---------|
| 1 | | BIBLIOGR... | TITLE | matches | ^245.* b.* b | |

BIBLIOGRAPHIC TITLE matches "^245.*|b.*|b"

OK

DATA CLEAN UP

Non-English cataloging

Marmot MUG non-english

Store Record Type: BIBLIOGRAPHIC b

Range: Start: b10000008 Stop: b51227265

Classic Enhanced JSON

| Ter... | Operator | Type | Field | Condition | Value A | Value B |
|--------|----------|-------------|-------|-----------|----------------|---------|
| 1 | | BIBLIOGR... | MARC | matches | ^040.*[b[^e] | |
| 2 | OR | BIBLIOGR... | MARC | matches | ^040.*[be[^n] | |
| 3 | OR | BIBLIOGR... | MARC | matches | ^040.*[ben[^g] | |

BIBLIOGRAPHIC MARC matches "^040.*[b[^e]" OR BIBLIOGRAPHIC MARC matches "^040.*[be[^n]" OR BIBLIOGRAPHIC MARC matches "^040.*[ben[^g]"

OK

DATA CLEAN UP

Wrong filing indicators

TITLE matches ^245.3.*|a..[^ "'-]

Marmot MUG bad indicator

Store Record Type: BIBLIOGRAPHIC b

Range Start: b10000008 Stop: b51229092

Classic Enhanced JSON

| Ter... | Operator | Type | Field | Condition | Value A | Value B |
|--------|----------|-------------|-------|-----------|---------------------|---------|
| 1 | | BIBLIOGR... | TITLE | matches | ^245.3.* a..[^ "'-] | |

BIBLIOGRAPHIC TITLE matches "^245.3.*|a..[^ "'-]"

OK

DATA CLEAN UP

URL not starting with 8

ITEM URL matches $^{[8]}$

The screenshot shows a 'Boolean Search' window with the following configuration:

- Review File Name: Marmot MUG bad URL
- Store Record Type: ITEM i
- Range: (dropdown)
- Start: i10000008
- Stop: i104110752
- Search Mode: Classic, Enhanced, JSON
- Search Table:

| Term | Operator | Type | Field | Condition | Value A | Value B |
|------|----------|------|-------|-----------|----------|---------|
| 1 | | ITEM | URL | matches | $^{[8]}$ | |

ITEM URL matches " $^{[8]}$ "

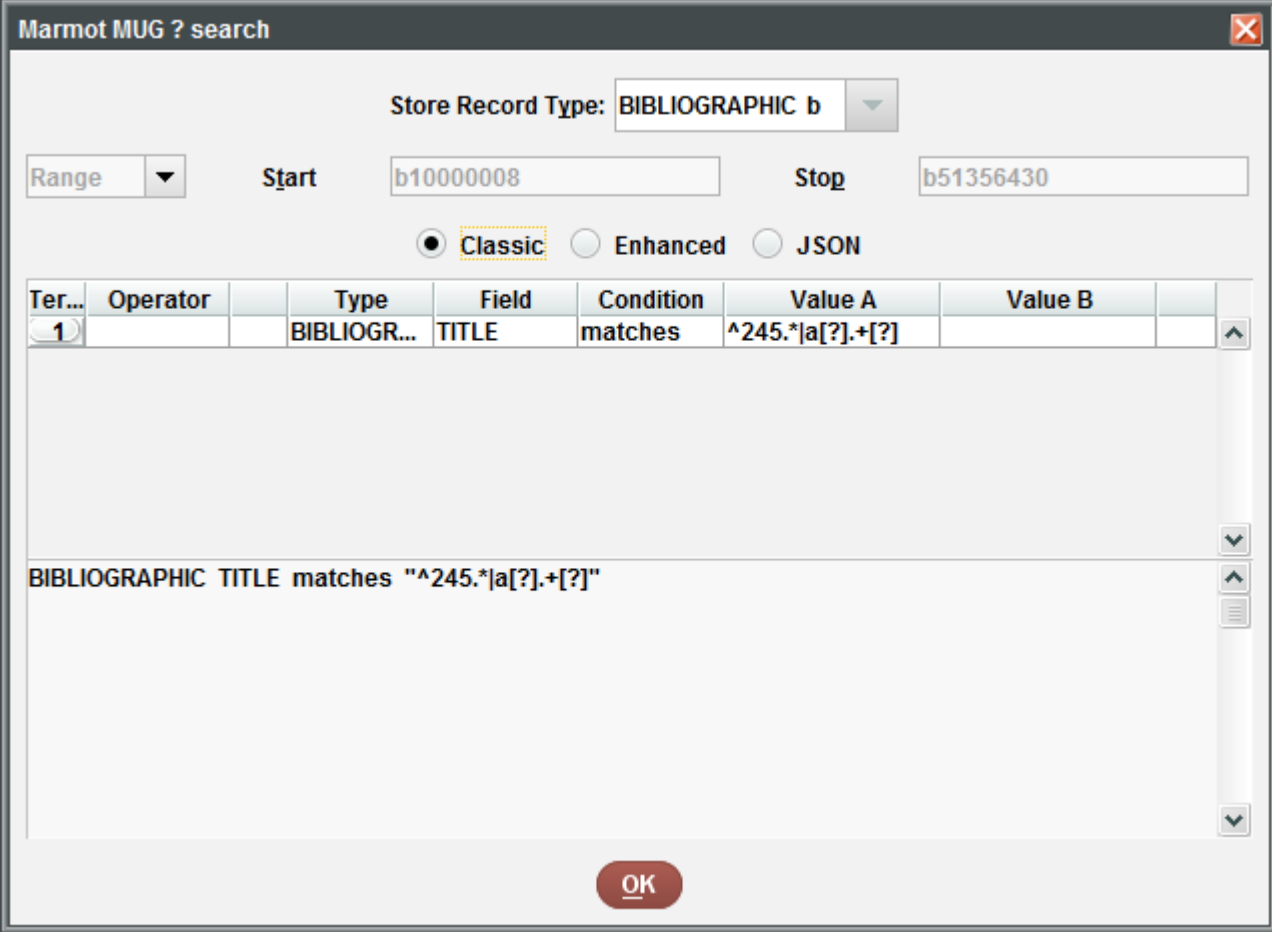
Buttons on the right: Group, Ungroup, Insert Line, Append Line, Delete, Clear All

Buttons at the bottom: Search, Use Existing Search, Retrieve Saved Query, Save, Save As, Close

DATA CLEAN UP

Title starts with ?, followed by another ?

TITLE matches `^245.*|a[?].+[?]`



Marmot MUG ? search

Store Record Type: BIBLIOGRAPHIC b

Range Start: b10000008 Stop: b51356430

Classic Enhanced JSON

| Ter... | Operator | Type | Field | Condition | Value A | Value B |
|--------|----------|-------------|-------|-----------|-------------------------------|---------|
| 1 | | BIBLIOGR... | TITLE | matches | <code>^245.* a[?].+[?]</code> | |

BIBLIOGRAPHIC TITLE matches "`^245.*|a[?].+[?]`"

OK

THANKS TO RICHARD JACKSON

Many of the examples in this presentation were taken from the 2015 version of Richard's "Playing with Matches" handout

<https://wiliug.files.wordpress.com/2015/02/matches-handout-2015.pdf>