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

\[ ^{\text{[} \text{^]} \text{]} } \rightarrow \text{circumflex in brackets - negate the following character(s) in the brackets} \]

\[ ^{\text{[} \text{^a-z] } \rightarrow \text{match any single character that is not a letter} \]

\[ ^{\text{[} \text{.^} } \rightarrow \text{match any single character that is not a period} \]

\[ ^{\text{[} \text{^aei} } \rightarrow \text{match any single character that is not a or e or i} \]

(\text{does not negate the string aei})

\[ ^{\text{[} \text{^p}[\text{^m} ] \rightarrow \text{match a character that is not p followed by character that is not m} \]
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"
SEARCH PECULIARITIES

Searching on field group is different from searching on MARC tag
TITLE matches "^245..|ato kill a mockingbird"
Sierra stores special characters in Unicode format.

Unicode uses a five character code to represent each character

\[ u0040 \rightarrow @ \]
\[ u00f1 \rightarrow \text{n}\]

Sierra stores them in curly braces { }

\[ \{u0040\} \rightarrow @ \]
\[ \{u00f1\} \rightarrow \text{n} \]

FYI, regular expressions will not work to search these:
\[ \{u\ldots\} \] will produce 0 results

Search with the {} (even though they are meta-characters, don’t escape them)
SEARCH SPECIAL CHARACTERS

You can also paste in the special character
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)
DATA CLEAN UP

Find barcodes that start with a space
ITEM BARCODE matches ^ [0-9]
Find barcodes with non-number characters
ITEM BARCODE matches [0-9]
DATA CLEAN UP

Barcodes that don’t start with the right number
ITEM BARCODE matches ^[^1]
DATA CLEAN UP

Bad subfields
SUBJECT matches "650.*[^avxyz20]"
DATA CLEAN UP

Search for non-repeatable subfields
TITLE matches ^245.*|b.*|b
DATA CLEAN UP

Non-English cataloging
Wrong filing indicators
TITLE matches ^245.3.*[a..[^ "-]
DATA CLEAN UP

URL not starting with 8
ITEM URL matches ^[^8]
DATA CLEAN UP

Title starts with ?, followed by another ?
TITLE matches ^245.*[a]?.[+]?
THANKS TO RICHARD JACKSON

Many of the examples in this presentation were taken from the 2015 version of Richard’s “Playing with Matches” handout