



5-2010

Batch Editing MARC Records with MarcEdit and Regular Expressions

Kathryn Lybarger

University of Kentucky, kathryn.lybarger@uky.edu

Julene L. Jones

University of Kentucky, julene.jones@uky.edu

Right click to open a feedback form in a new tab to let us know how this document benefits you.

Follow this and additional works at: https://uknowledge.uky.edu/libraries_present

 Part of the [Library and Information Science Commons](#)

Repository Citation

Lybarger, Kathryn and Jones, Julene L., "Batch Editing MARC Records with MarcEdit and Regular Expressions" (2010). *Library Presentations*. 13.

https://uknowledge.uky.edu/libraries_present/13

This Presentation is brought to you for free and open access by the University of Kentucky Libraries at UKnowledge. It has been accepted for inclusion in Library Presentations by an authorized administrator of UKnowledge. For more information, please contact UKnowledge@lsv.uky.edu.

Batch Editing MARC Records with MarcEdit and Regular Expressions

Kathryn Lybarger and Julene Jones
May 20, 2010

MARC

- MARC = MAchine Readable Cataloging
- File format for exchange of cataloging information
- MARC has many advantages

MARC is old reliable

- Created in 1960's for Library of Congress
- ANSI standard (1971)
- ISO standard (1973)



MARC is popular

- Large collection, only growing
- OCLC has 183 million bibliographic records
- New record added every ten seconds



MARC is flexible



- Bibliographic
- Holdings
- Authority

MARC can be hard to work with.

What does MARC look like?

000 01373cam a2200433 a 4500

001 2237424

005 20070330085528.0

008 050107s2004 nyua b 000 1 eng

010__ | a 2004048210

020__ | a 0143039067

024__ | a 2126912

035__ | a (OCoLC)ocm55044526

040__ | a DLC | c DLC | d OCLCQ

049__ | a KUJY

05000 | a PS3545.E365 | b D3 2004

08200 | a 813/.52 | 2 22

1001_ | a Webster, Jean, | d 1876-1916.

24510 | a Daddy Long Legs ; | b and, Dear enemy / | c Jean
Webster ; edited with an introduction and notes by Elaine
Showalter.

Blank indicators

OPAC:

—

OCLC Bib Formats
documentation:

b

LC MARC Bibliographic
Documentation:

#

OCLC Connexion Client
(and actually in file):

(blank)

Subfield delimiters

OPAC:

|a

OCLC Connexion Browser:

\$a

Voyager:

⌘a

OCLC Connexion Client:

‡a

It is none of those!

- ◉ Binary: 0 0 0 1 1 1 1 1

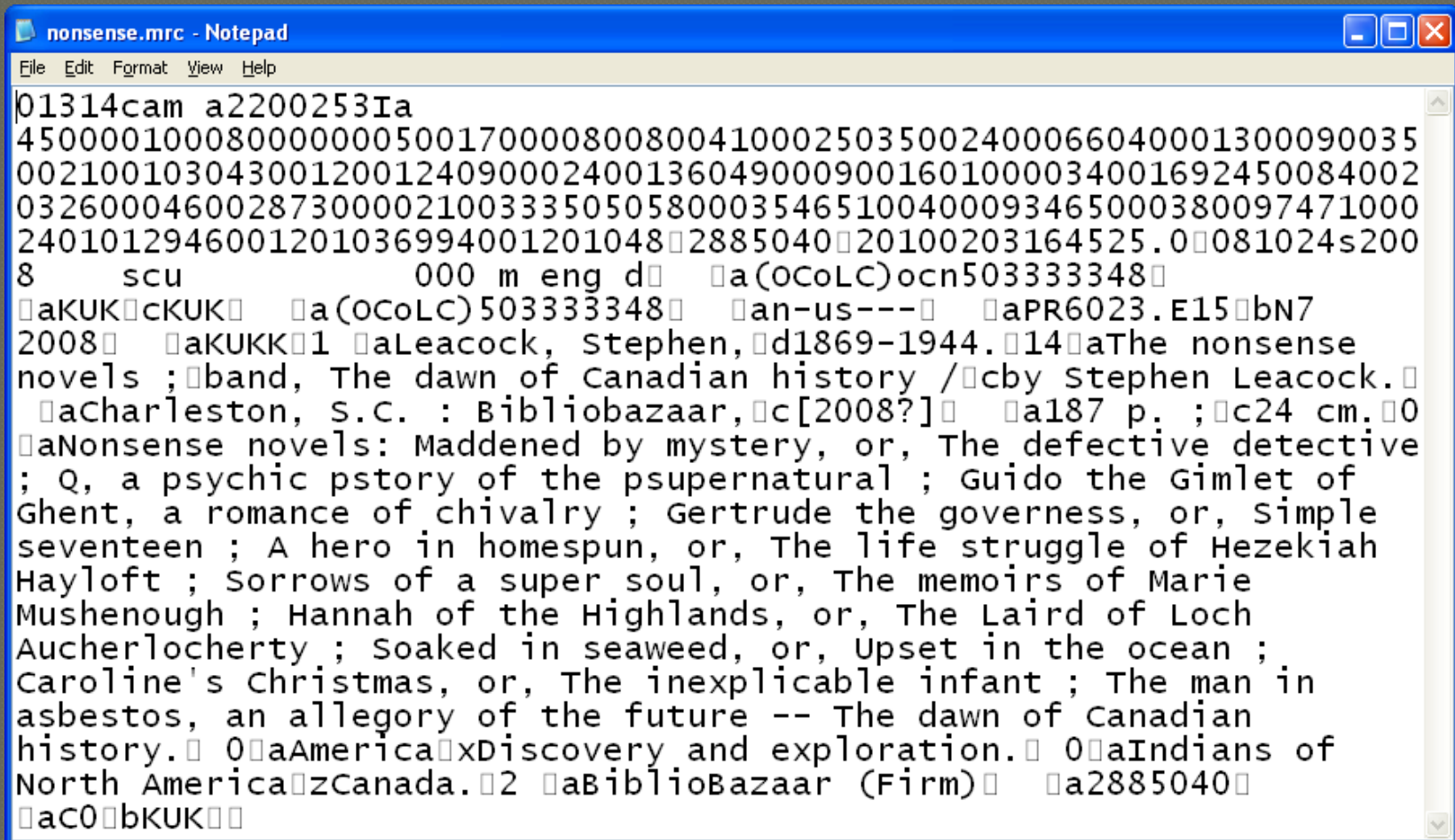
- ◉ Decimal: 31

- ◉ Hex: 1F

MARC is a Binary format

- ◉ MARC is often displayed as text
- ◉ MARC is not plain text
- ◉ You need a MARC editor to easily edit MARC

MARC in a text editor



nonsense.mrc - Notepad

File Edit Format View Help

01314cam a2200253Ia
45000010008000000005001700008008004100025035002400066040001300090035
0021001030430012001240900024001360490009001601000034001692450084002
0326000460028730000210033350505800035465100400093465000380097471000
2401012946001201036994001201048 2885040 20100203164525.0 081024s200
8 scu 000 m eng d a(OCOLC)ocn503333348
aKUKcKUK a(OCOLC)503333348 an-us--- aPR6023.E15bN7
2008 aKUKK1 aLeacock, Stephen,d1869-1944.14aThe nonsense
novels ;band, The dawn of Canadian history /cby Stephen Leacock.
aCharleston, S.C. : Bibliobazaar,c[2008?] a187 p. ;c24 cm.0
aNonsense novels: Maddened by mystery, or, The defective detective
; Q, a psychic pstory of the psupernatural ; Guido the Gimlet of
Ghent, a romance of chivalry ; Gertrude the governess, or, Simple
seventeen ; A hero in homespun, or, The life struggle of Hezekiah
Hayloft ; Sorrows of a super soul, or, The memoirs of Marie
Mushenough ; Hannah of the Highlands, or, The Laird of Loch
Aucherlocherty ; Soaked in seaweed, or, Upset in the ocean ;
Caroline's Christmas, or, The inexplicable infant ; The man in
asbestos, an allegory of the future -- The dawn of Canadian
history. 0aAmerica xDiscovery and exploration. 0aIndians of
North America zCanada.2 aBiblioBazaar (Firm) a2885040
ac0bKUK

Tags in the directory

● 00100080000000500017000080080041
0002503500240006604000130009003
5002100103043001200124090002400
1360490009001601000034001692450
0840020326000460028730000210033
3505058000354651004000934650003
8009747100024010129460012010369
94001201048

MARC is out of order

- Tags are stored in the directory of the file
- Indicators and field contents are stored in the body of the record

MARC directory entry

- 0010008000000050017000080080041000250350
0240006604000130009003500210010304300120
0124090002400136049000900160100003400169
2450084002032600046002873000021003335050
5800035465100400093465000380097471000240
1012946001201036994001201048
- 001 2885040
- Field 001, 8 bytes long, starts at 0

Fixed (length) fields

- Fixed fields are always the same size
- In a bibliographic record, there are always four bytes for illustration data:
 - ____ (four blanks): no illustrations
 - abcd : illustrations, maps, portraits, charts
 - abcd : illustrations, maps, portraits, charts, plans, plates and music

Variable (length) fields

- ◉ Variable fields are not always the same size.
- ◉ The title field may vary in length:
 - The Iliad
 - Alexander and the Terrible, Horrible, No Good, Very Bad Day

MARC directory entry

- 0010008000000050017000080080041000250350
0240006604000130009003500210010304300120
0124090002400136049000900160100003400169
2450084002032600046002873000021003335050
5800035465100400093465000380097471000240
1012946001201036994001201048
- 245 14\$aThe nonsense novels ;\$band, The dawn of
Canadian history /\$cby Stephen Leacock.
- Field 245, 84 characters long, starts at position 203

MARC has structural metadata

- ◉ Variable fields make MARC flexible and compact
- ◉ Some structural metadata is needed to make this happen
- ◉ This book-keeping is usually (thankfully) invisible

Editing MARC is easy with a good MARC editor

- ◉ Binary characters represented as text
- ◉ Data in a reasonable order
- ◉ Structural metadata hidden

So what's the problem?

- Many MARC editors allow editing only one record at a time
- You may want to do batch editing
- You may want to do something nobody has ever considered before!

You may want to know...

- ◉ How many records in my file?
- ◉ Do they all have a field that I require?
- ◉ Does that field contain what I require?
- ◉ Are there any fields I don't want?

You may want to make changes...

- ◉ Remove a field in all records
- ◉ Add a field to all records
- ◉ Modify a field in all records
- ◉ Swap data between fields
- ◉ Change character encoding

You may want to do more...

- Create (one or more) holdings records based on bibliographic record content
- Extract some fields to make an RSS feed of new books in your catalog


MarcEdit

- Suite of tools for working with MARC
- Developed by Terry Reese at Oregon State
- Free download for Windows, Linux, Mac

MarcEdit MARC tools

- MarcBreaker – converts MARC to Mnemonic file format
- MarcMaker – converts Mnemonic file format to MARC
- MarcEditor – text editor for making common MARC edits
- MARC Spy – hex editor, good for finding problems in corrupted MARC records

MARC Spy


 D080903.KNOVEL_unknown.mrc - Mini Hex Edit

File

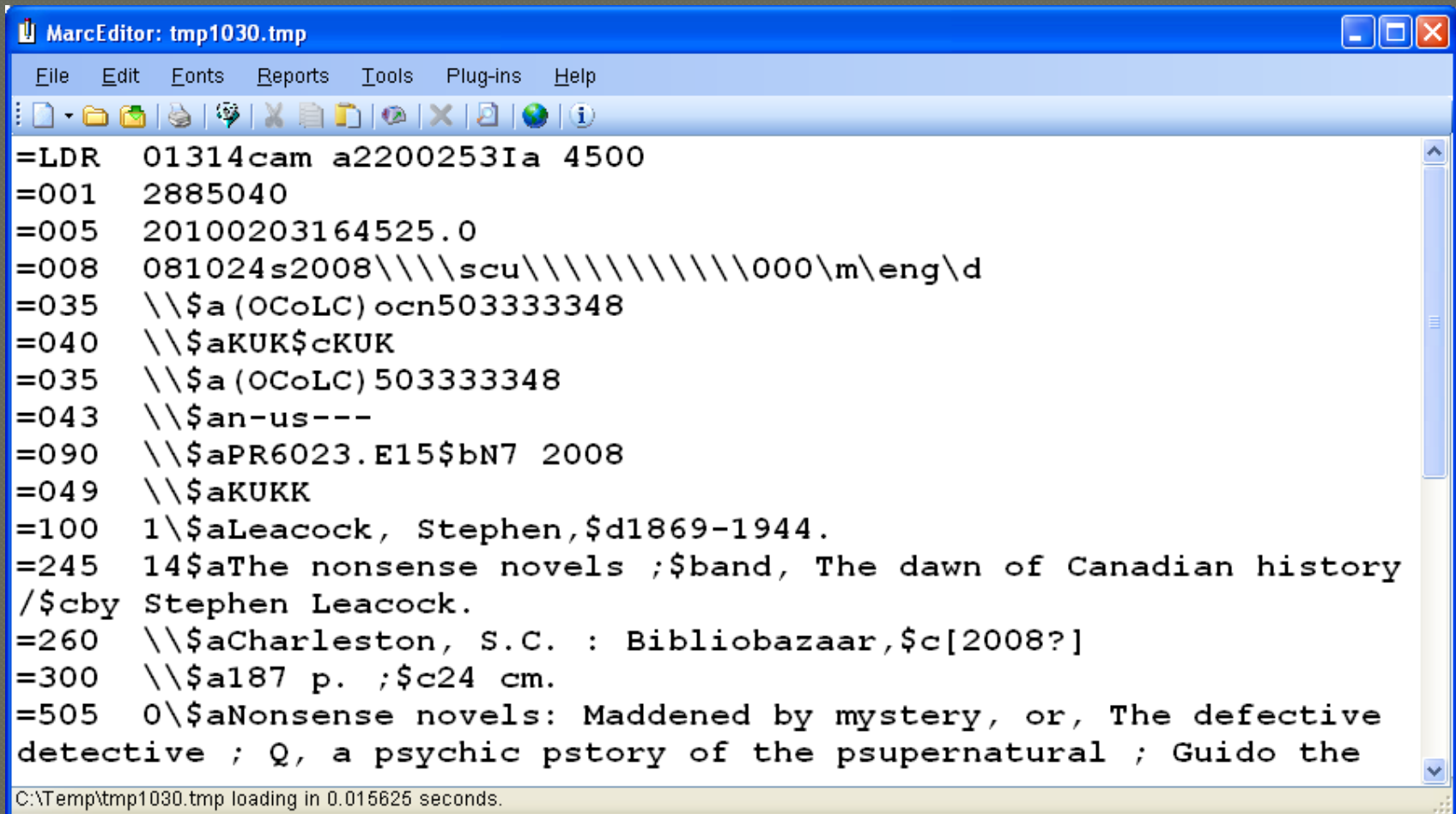
Position	Bytes	Text
00000220h	4E-4F-56-4C-1F-64-4B-55-4B-1E-20-20-1F-61-39-37	NOVL.dKUK. .a97
00000230h	38-31-36-30-31-31-39-36-31-34-39-20-28-65-6C-65	81601196149 (ele
00000240h	63-74-72-6F-6E-69-63-20-62-6B-2E-29-1E-20-20-1F	ctronic bk.). .
00000250h	61-31-36-30-31-31-39-36-31-34-38-20-28-65-6C-65	a1601196148 (ele
00000260h	63-74-72-6F-6E-69-63-20-62-6B-2E-29-1E-20-20-1F	ctronic bk.). .

Bytes 00000240h - 0000024fh (576 - 591)

Hex	63	74	72	6f	6e	69	63	20	62	6b	2e	29	1e	20	20	1f
Decimal	99	116	114	111	110	105	99	32	98	107	46	41	30	32	32	31
iso-8859-1	c	t	r	o	n	i	c		b	k	.)				
Bits	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>



Mnemonic File Format (MarcEditor)



The screenshot shows the MarcEditor application window with the title bar "MarcEditor: tmp1030.tmp". The menu bar includes "File", "Edit", "Fonts", "Reports", "Tools", "Plug-ins", and "Help". The toolbar contains icons for file operations (new, open, save, print, delete, copy, paste, undo, redo) and a search icon. The main text area displays a MARC file in mnemonic format, with lines starting with tags like =LDR, =001, =005, =008, =035, =040, =043, =090, =049, =100, =245, =260, =300, and =505. The status bar at the bottom indicates "C:\Temp\tmp1030.tmp loading in 0.015625 seconds."

```
=LDR 01314cam a2200253Ia 4500
=001 2885040
=005 20100203164525.0
=008 081024s2008\\scu\\000\m\eng\d
=035 \\$a(OCOLC)ocn503333348
=040 \\$aKUK$cKUK
=035 \\$a(OCOLC)503333348
=043 \\$an-us---
=090 \\$aPR6023.E15$bN7 2008
=049 \\$aKUKK
=100 1\\$aLeacock, Stephen,$d1869-1944.
=245 14$aThe nonsense novels ;$band, The dawn of Canadian history
/$cby Stephen Leacock.
=260 \\$aCharleston, S.C. : Bibliobazaar,$c[2008?]
=300 \\$a187 p. ;$c24 cm.
=505 0\\$aNonsense novels: Maddened by mystery, or, The defective
detective ; Q, a psychic pstory of the psupernatural ; Guido the
```

C:\Temp\tmp1030.tmp loading in 0.015625 seconds.

MarcEdit Demo!

Regular Expressions

- “Regex” or “regexp”
- A more general (and powerful) search or search-and-replace
- A regular expression is a pattern which “matches” parts of your file

Regular expression support

- ◉ `grep`
- ◉ Powerful text editors: `MarcEditor`, `vim`, `emacs`
- ◉ Programming languages: `perl`, `php`

Many standard searches are also regular expressions

- Expression: Mar

- Matches:

Mark Twain

Steve Martin

Telemarketing

(case sensitive)

Anchored searches

- Expression: `^Mar`

- Matches:

Mark Twain ← only this one
Steve Martin
Telemarketing

Many standard searches are also regular expressions

- Expression: tin

- Matches:

Mark Twain

Steve Martin

Telemarketing

Anchored searches

- Expression: tin\$

- Matches:

Mark Twain

Steve Martin ← only this one

Telemarketing

Special characters

◉ So how do you search for \wedge or \$?

◉ Escape special characters with \

- $\backslash \$1$ matches $\text{\textcolor{yellow}{\$}1.35}$

- $2\backslash \wedge$ matches $2 + \text{\textcolor{yellow}{2}}^4 = 18$

◉ So how do you search for \ ?

You don't have to be an expert!

- You can modify searches just slightly for much more specific results
- You can use multiple simpler expressions
- You can ask other people (MARCEdit-L)

Meta-characters

^ \$. + ?

* () []

{ } | \

This or that?

- ◉ You can search for one of several phrases
(a | b | c | d)
- ◉ Example: (Bob | John | Joe) Smith
- ◉ Matches:
 - Bob Smith
- ◉ Does NOT match:
 - Robert Smith

Character classes

- ◉ Match not just one letter, but any of several
- ◉ Surround with []
- ◉ Example: [BR]ob
 - matches **Bob**, **Rob**, **Robert**
 - Does NOT match: Toby, bobbing, robbery

Character classes (negated)

- Match anything that is NOT in a specified list of characters
- Surround with [^]
- Example: [^aeiou]a
 - matches scu**ba**, Wild**cat**, **d**azzle, aard**v**ark
 - Does NOT match: each, toad, visual, antique

Match any character

- ◉ A period matches any character

- ◉ Example: ..**an**

- ◉ Matches: **woman**
 watchman

- ◉ Does NOT match: **man**

A handy regex

- Find all subject headings with second indicator other than 0 or 2

`^=6.. .[^02]`

- Matches: =650 \7\$aFilms\$xMontage.\$2
ram

Matching repetitions

- ⊙ * any number of what it follows
- ⊙ ? 0 or 1 of what it follows
- ⊙ + 1 or more of what it follows

⊙ Example:

• *

Matching repetitions

- * any number of what it follows
- ? 0 or 1 of what it follows
- + 1 or more of what it follows

● Example:

Joh?nathan

Search and replace

- Replace matched part with a static string

OR

- Capture parts of what you match with ()
- Use those captured parts in your replacement

Replacement string

- ◉ \$1 – contents of first parentheses
- ◉ \$2 – contents of second parentheses
- ◉ ...

- ◉ Search: (.*) (.*)
- ◉ Replace: \$2, \$1

- ◉ Bob Jones → Jones, Bob

Replacement string

- \$0 – whole string matched
- Example: make all URLs hyperlinks
- Search: `http://[^]+`
- Replace: `$0`

Counting matches (MarcEdit)

- Make sure all 245 fields have \$h
[electronic resource]

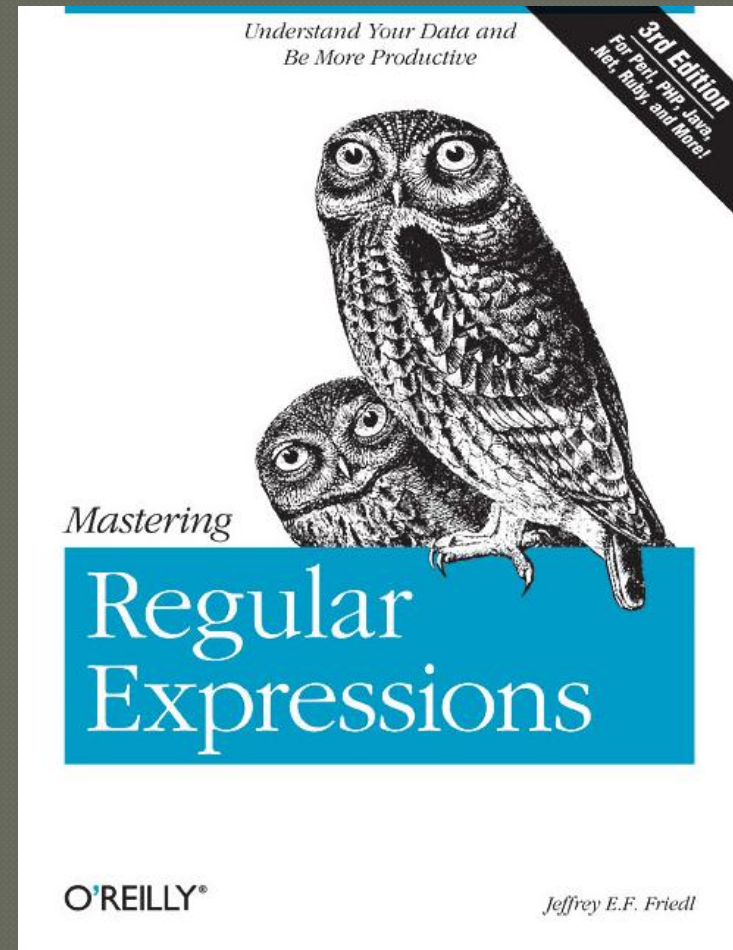
Search: ^=245.*\\$h\[electronic resource\]

Replace all: \$0

This makes no real changes to your file, but
lets you know how many matches it found

Learn more!

- Many books and websites
- <http://www.regular-expressions.info/>
- Unix man pages:
man perlretut



Any questions?