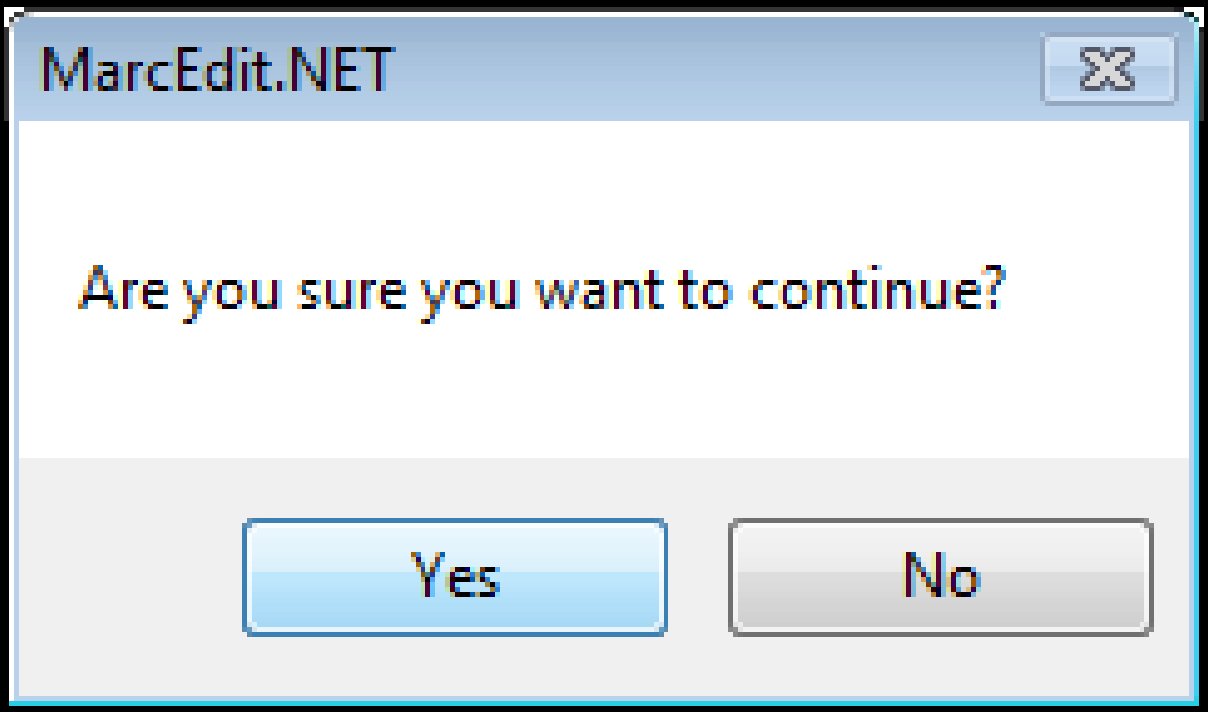
Express Yourself

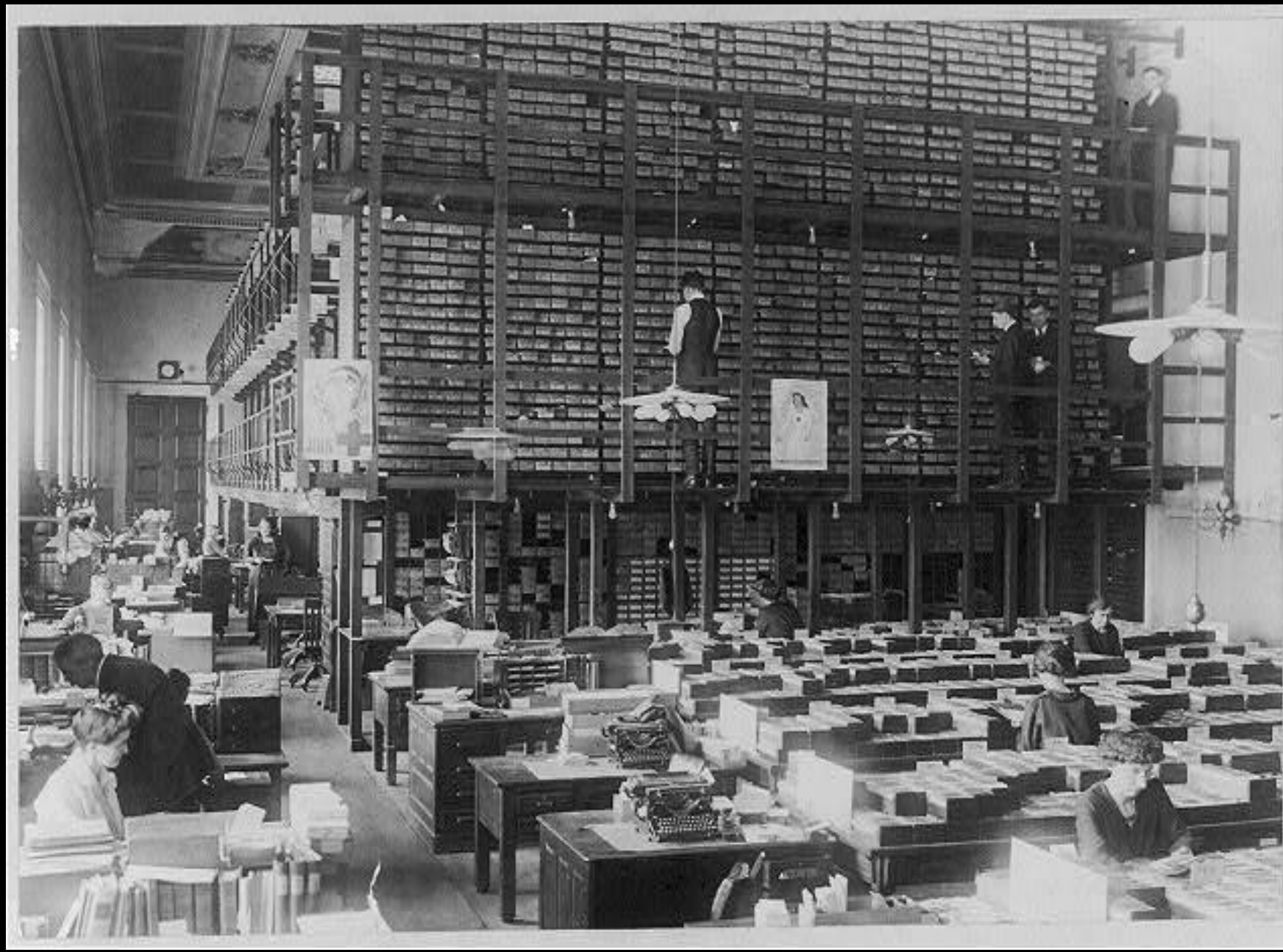
Using regular expressions
in MARCEdit

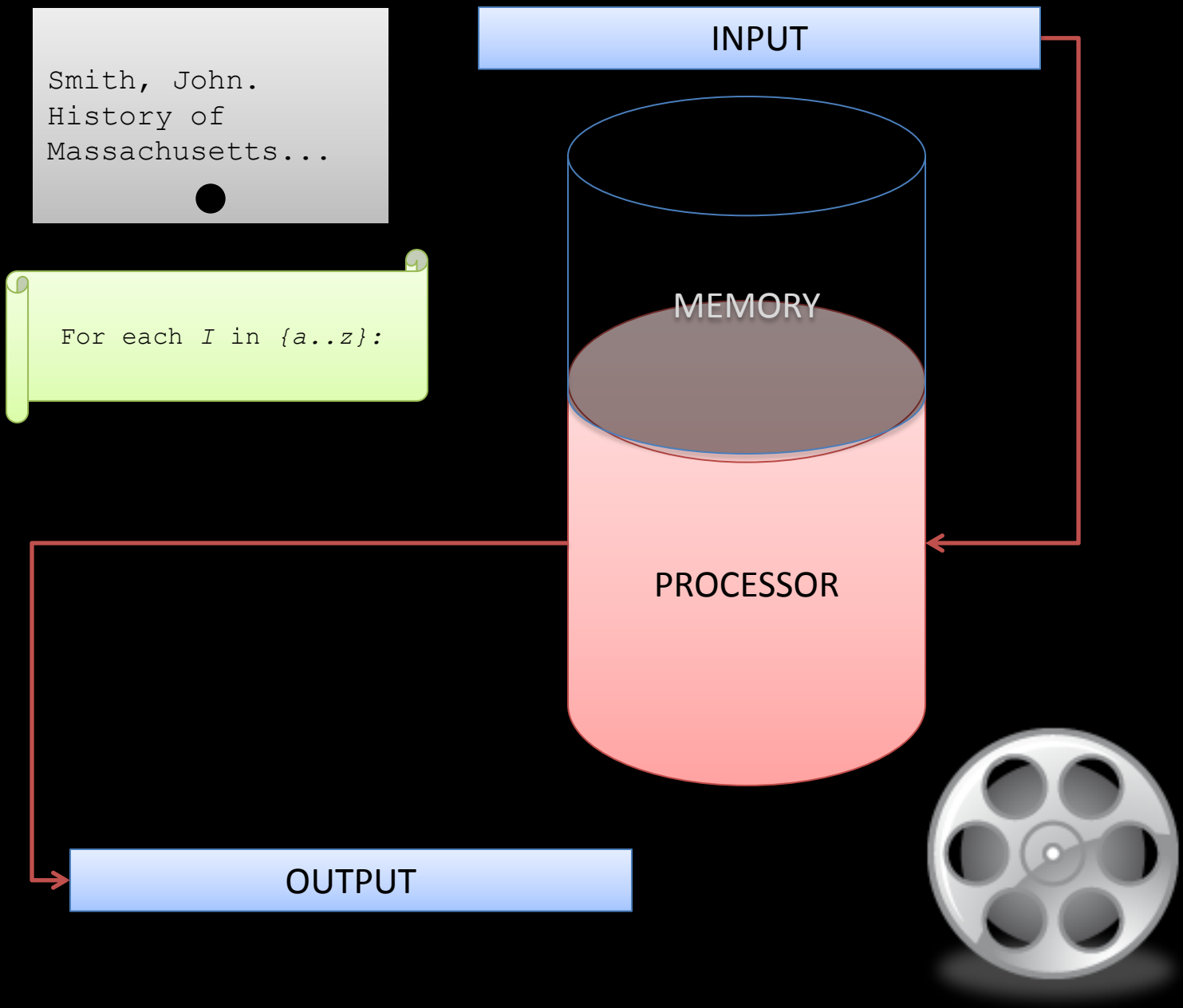
Ben Abrahamse
MLA 2012


MarcEdit.NET
Are you sure you want to continue?
Yes
No

AVRAM

THOMPSON




Smith, John.
History of
Massachusetts...
For each I in {a..z}:
INPUT
MEMORY
PROCESSOR
OUTPUT

Formal language

MAchine Readable Cataloging

245$a {Title statement}
245$c {Statement of responsibility}
300$a{Physical description—extent of material}

Regular Expressions

. {Any character}
.$ {Any character at the end of a line}
\w {Any "word character"}

grep

Lorem ipsum dolor sit amet, consectetur adipisicing elit,

.{3}

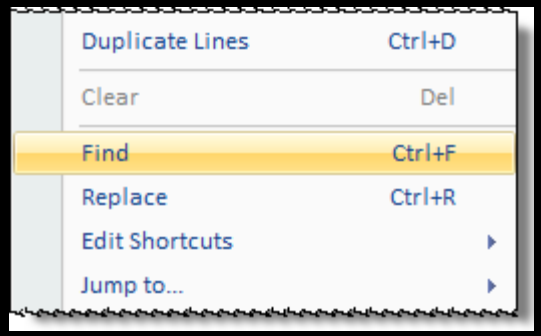
Lor|em |ips|um |dol|or |sit| am|et,| co|nse|cte|tur| ad|ipi|sic|ing| el|it,|

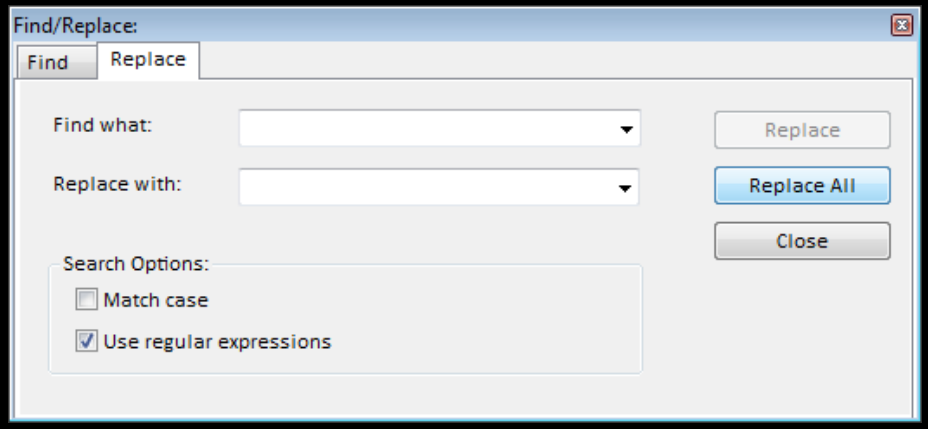
/g/re/p:

global/regex/print

\b\w+

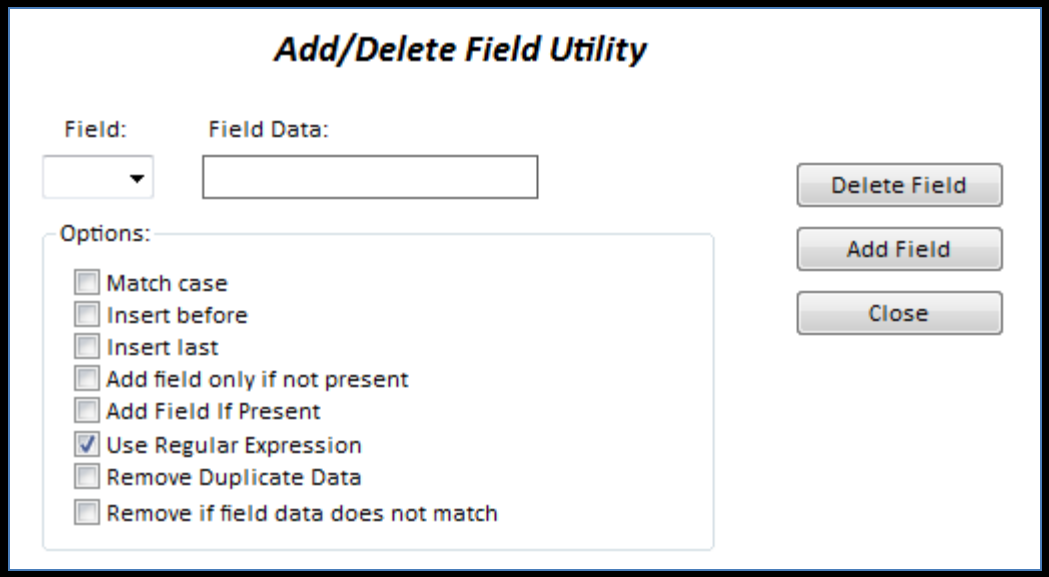
Lorem|ipsum|dolor|sit|amet|consectetur|adipisicing|elit





RegEx find/replace in the MARCEditor

RegEx functionality in MARC editing tools



FORMING REGULAR EXPRESSIONS

Characters and Character Classes

Look for...	Symbol	E.g. ("lor 123")
a specific character(s)	the character(s)	find: lor finds: **lor**
any character	.	find: ... finds **lor,or•,r•1, •12,...**

any word character (incl. digits, but not symbols)	\w	find: \w\w\w or \w{3}) finds **lor**
any digit character	\d	find: \d\d\d or \d{3} finds: **123**
any nondigit character (letters, spaces, and symbols)	\D	find: \D\D\D\D or \D{4} finds: **lor•**
any space	\s	find: .{3}\s.{3} finds: **lor•123**
UNICODE character range	\p{IsBasicLatin}	find: \p{IsBasicLatin}{7} finds: **lor•123**

Custom Groups

Look for...	Symbol	E.g. ("lorem.ipsum")
one of these characters ("Positive Character Group")	[...]	find: [l,o,r,e,m]{5} finds: **lorem**
not one of these characters ("Negative Character Group")	[^...]	find: [^l,o,r,e]{4}m finds**: ipsum**
any reserved character	\n	find: .+\. finds: **lorem.**
alternating ("or") constructs	[...|...]	find: [l,o,r,e|i,p,s,u]{4}m finds: **lorem, ipsum**

Anchors

Look for...	Symbol	E.g. ("lorem ipsum dolor sit amet")
character at the beginning of a line	^	find: ^\w finds: **l**
character at the end of a line	$	find: \w$ finds: **t**
word boundary	\b	find:\b\w+ finds: **lorem,ipsum,dolor,sit,...**

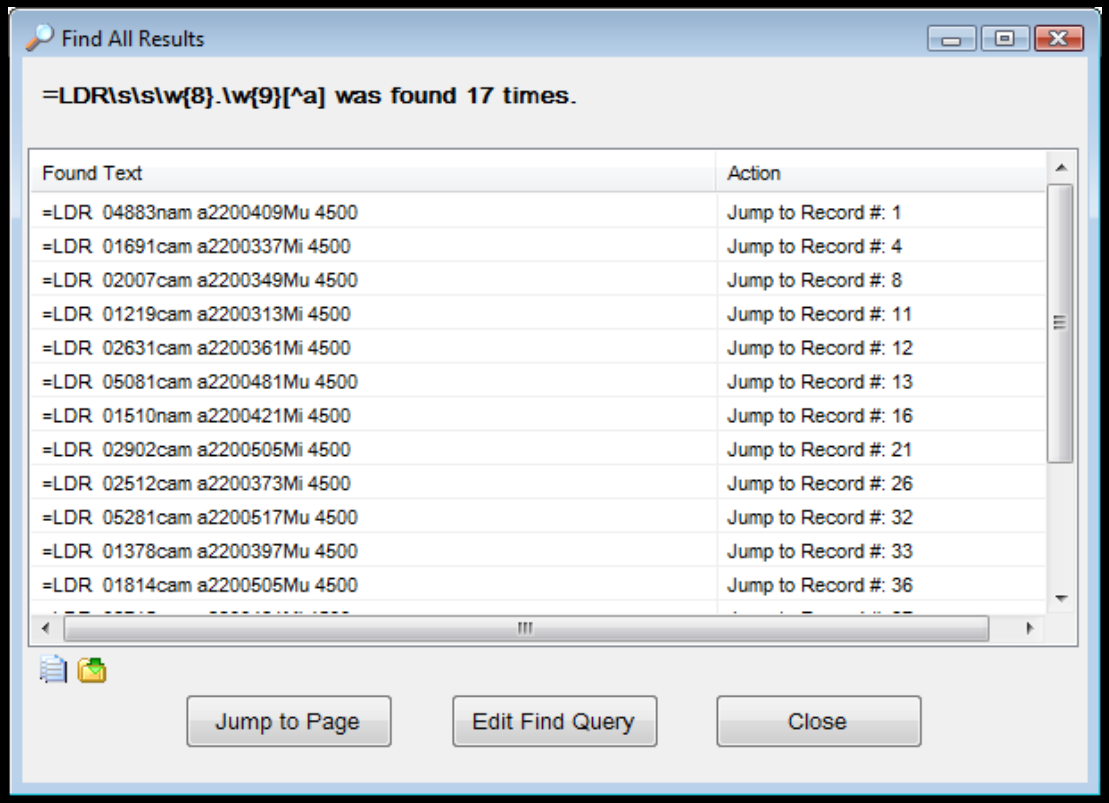
Quantifiers

Look for...	Symbol	E.g. ("be been being")
character zero or more times	*	find: \w* finds: **b,be,b,be,bee,been,...**
character one or more	+	find: e+ finds: **e,e,ee,e,e**
character exactly *n* times	{n}	find:e{2} finds: **ee**
character *n*-*m* times	{n,m}	find:e{1,2} find: **e,ee,e,e**

Example 1:

Use regex to sort a MARC file
by rule set.

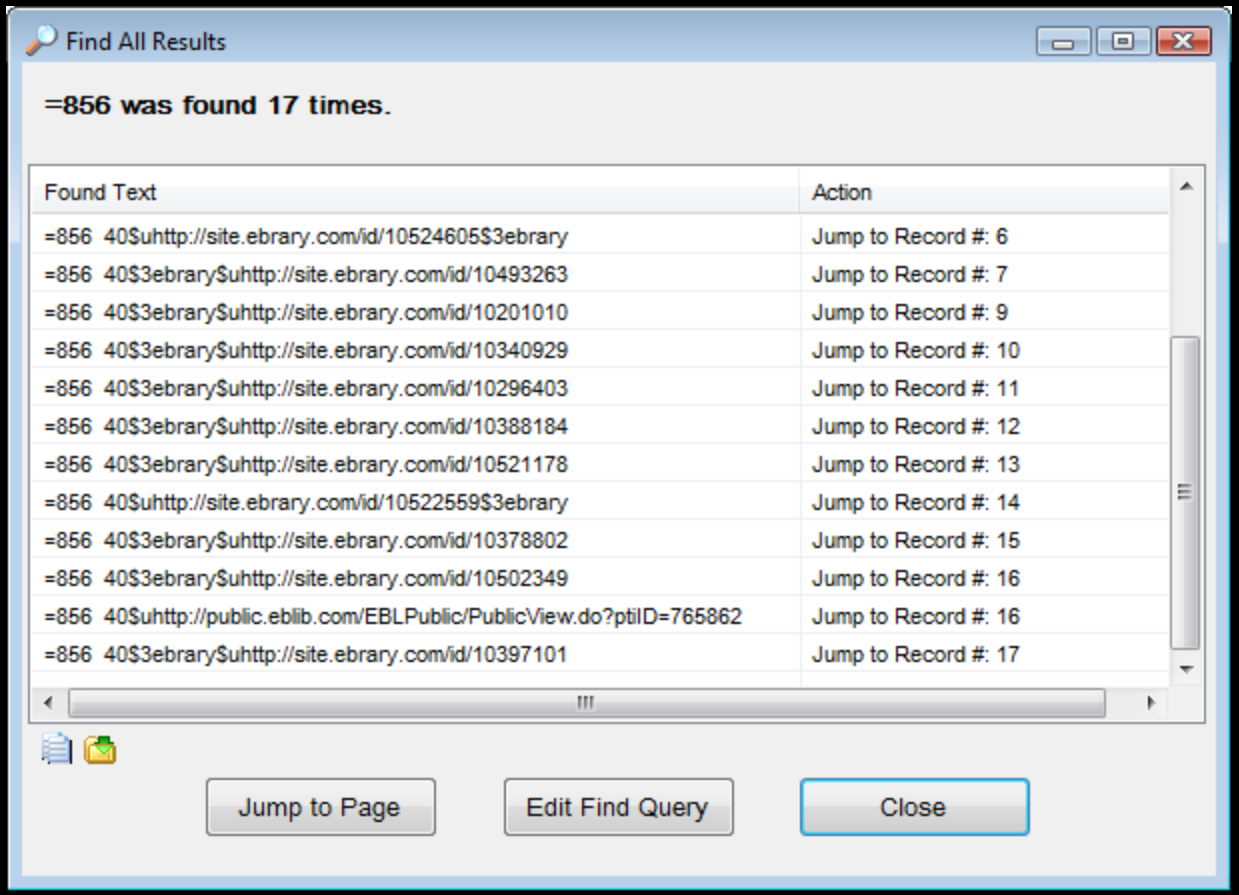

Find/Replace:
Find
Replace
Find what:
=LDR\s\s\w{8}.\w{9}[^a]
Find
Find All
Close
Search Options:
Match case
Use regular expressions

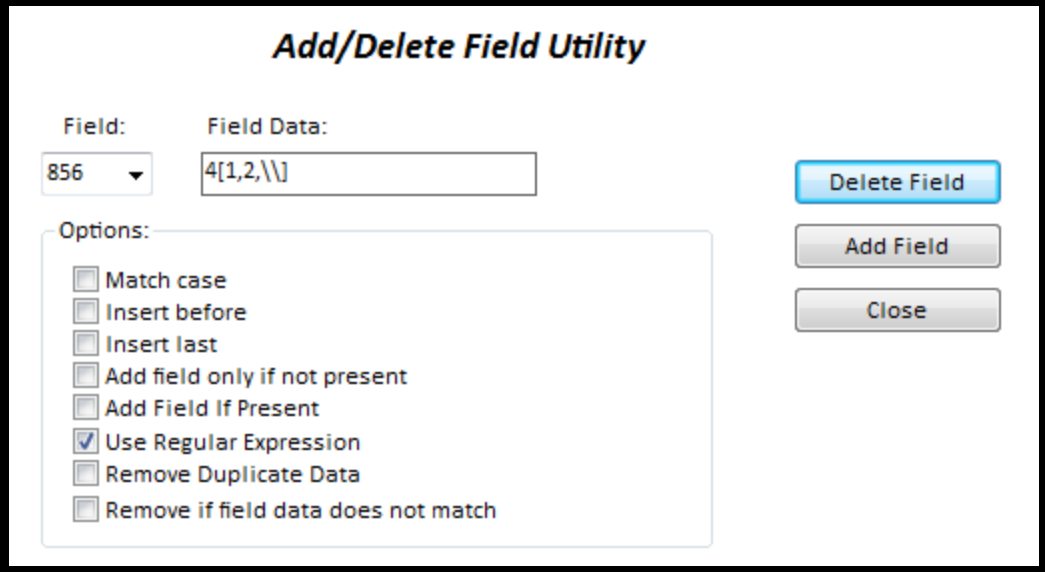

Find All Results
=LDR\s\s\w{8}.\w{9}[^a] was found 17 times.
Found Text
Action
=LDR 04883nam a2200409Mu 4500
Jump to Record #: 1
=LDR 01691cam a2200337Mi 4500
Jump to Record #: 4
=LDR 02007cam a2200349Mu 4500
Jump to Record #: 8
=LDR 01219cam a2200313Mi 4500
Jump to Record #: 11
=LDR 02631cam a2200361Mi 4500
Jump to Record #: 12
=LDR 05081cam a2200481Mu 4500
Jump to Record #: 13
=LDR 01510nam a2200421Mi 4500
Jump to Record #: 16
=LDR 02902cam a2200505Mi 4500
Jump to Record #: 21
=LDR 02512cam a2200373Mi 4500
Jump to Record #: 26
=LDR 05281cam a2200517Mu 4500
Jump to Record #: 32
=LDR 01378cam a2200397Mu 4500
Jump to Record #: 33
=LDR 01814cam a2200505Mu 4500
Jump to Record #: 36
Jump to Page
Edit Find Query
Close

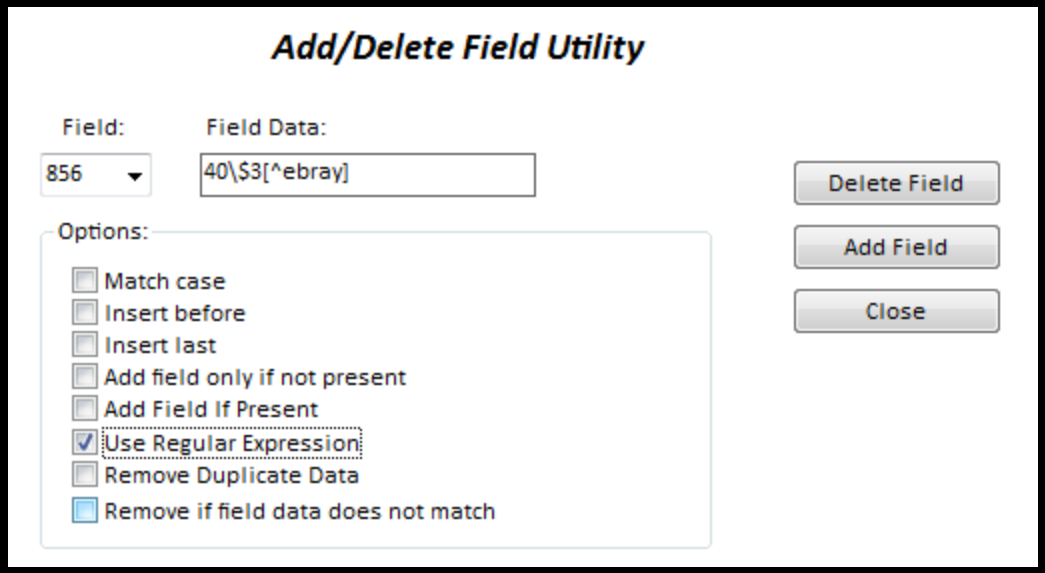
- Use "Find" function to search for leader (LDR) field.
- Replace variable characters with character class ("\w")
- Use quantifiers (if you want) for clarity
- Use a "negative character group" to show all records *not* a.

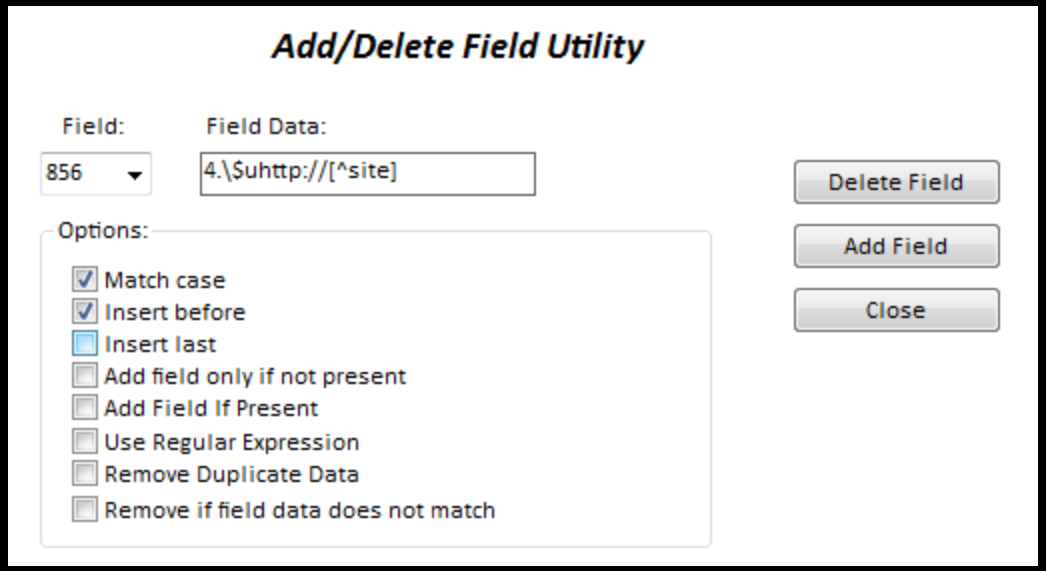
Example 2:

Use regex to remove URLs from other providers


Find All Results
=856 was found 17 times.
Found Text
Action
=856 40$uhttp://site.ebrary.com/id/10524605$3ebrary
Jump to Record #: 6
=856 40$3ebrary$uhttp://site.ebrary.com/id/10493263
Jump to Record #: 7
=856 40$3ebrary$uhttp://site.ebrary.com/id/10201010
Jump to Record #: 9
=856 40$3ebrary$uhttp://site.ebrary.com/id/10340929
Jump to Record #: 10
=856 40$3ebrary$uhttp://site.ebrary.com/id/10296403
Jump to Record #: 11
=856 40$3ebrary$uhttp://site.ebrary.com/id/10388184
Jump to Record #: 12
=856 40$3ebrary$uhttp://site.ebrary.com/id/10521178
Jump to Record #: 13
=856 40$uhttp://site.ebrary.com/id/10522559$3ebrary
Jump to Record #: 14
=856 40$3ebrary$uhttp://site.ebrary.com/id/10378802
Jump to Record #: 15
=856 40$3ebrary$uhttp://site.ebrary.com/id/10502349
Jump to Record #: 16
=856 40$uhttp://public.eblib.com/EBLPublic/PublicView.do?ptiID=765862
Jump to Record #: 16
=856 40$3ebrary$uhttp://site.ebrary.com/id/10397101
Jump to Record #: 17
Jump to Page
Edit Find Query
Close


Add/Delete Field Utility
Field:
Field Data:
856
4[1,2,\\]
Delete Field
Add Field
Close
Options:
Match case
Insert before
Insert last
Add field only if not present
Add Field If Present
Use Regular Expression
Remove Duplicate Data
Remove if field data does not match


Add/Delete Field Utility
Field:
Field Data:
856
40\$3[^ebray]
Delete Field
Add Field
Close
Options:
Match case
Insert before
Insert last
Add field only if not present
Add Field If Present
Use Regular Expression
Remove Duplicate Data
Remove if field data does not match


Add/Delete Field Utility
Field:
Field Data:
856
4.\$uhttp://[^site]
Delete Field
Add Field
Close
Options:
Match case
Insert before
Insert last
Add field only if not present
Add Field If Present
Use Regular Expression
Remove Duplicate Data
Remove if field data does not match

- Use "Find All" function to see how different URLs formed.
- Use "Add/Remove Fields" + RegEx to
  - Remove 856's with bad indicators
  - Remove 856's based on $3, $z, etc.
  - Remove 856's based on URL

Capturing and substituting values

- RegEx can be used to capture a value on "Find", and reproduce that value on "Replace".
- Use parentheses to capture.
- Use $*n* to replace, where *n* is the number of the captured expression (counting left-to-right)
- Can name captured expressions

Lorem ipsum dolor sit amet

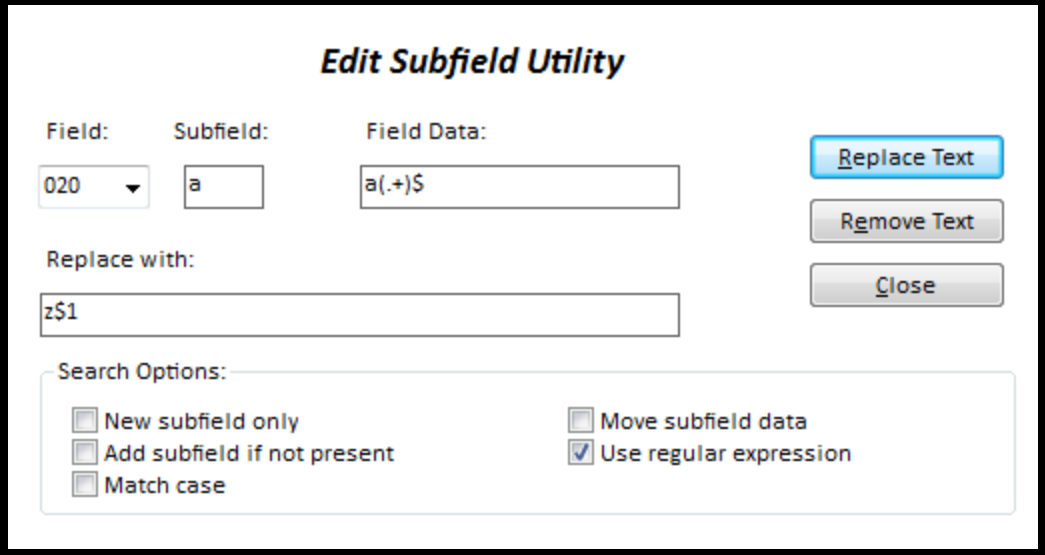
find: (\w)\s(\w)\s(\w)\s(\w)\s(\w)
replace: $5 $4 $3 $2 $1

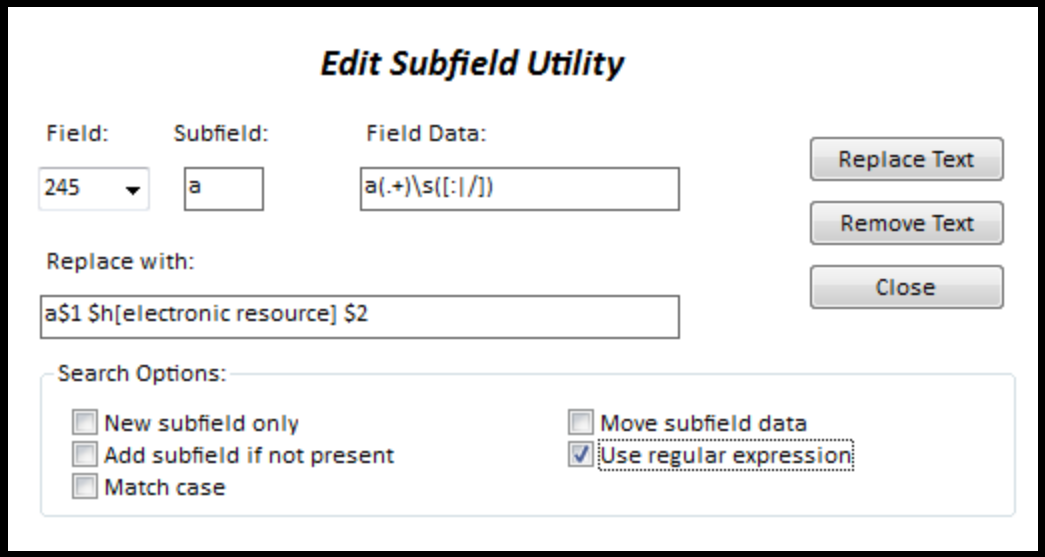
amet sit dolor ipsum Lorem

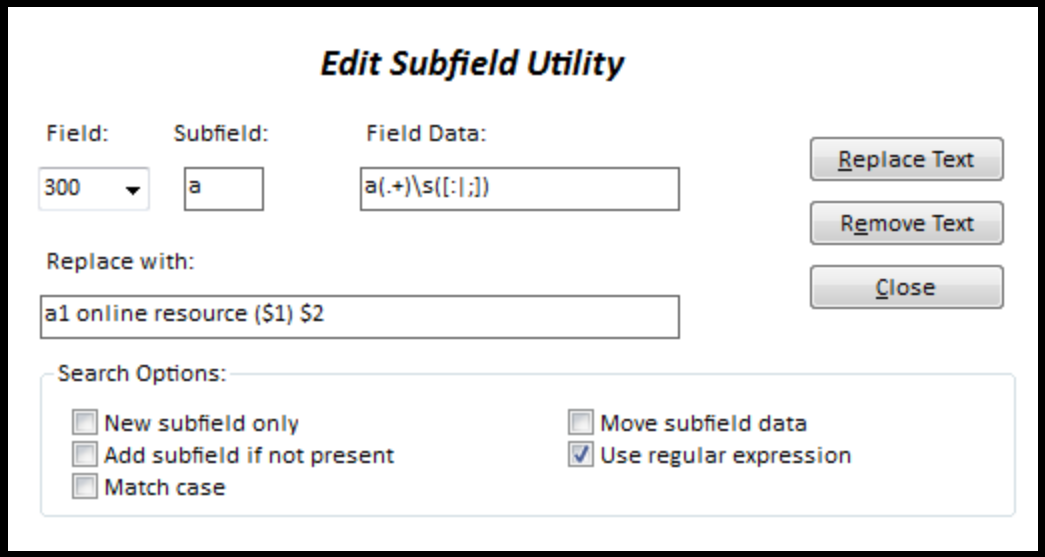
Capture...	Symbol	E.g. ("lorem ipsum")
a set of characters (assigned a number)	(...) $*n* $+	find: (\w)\s\w+ replace: ipsum $1 **ipsum lorem**
a set of characters (assigned a name)	(?<name>...) ${name}	find: (?<x>\w)\s\w+ replace: ipsum ${x} **ipsum lorem**

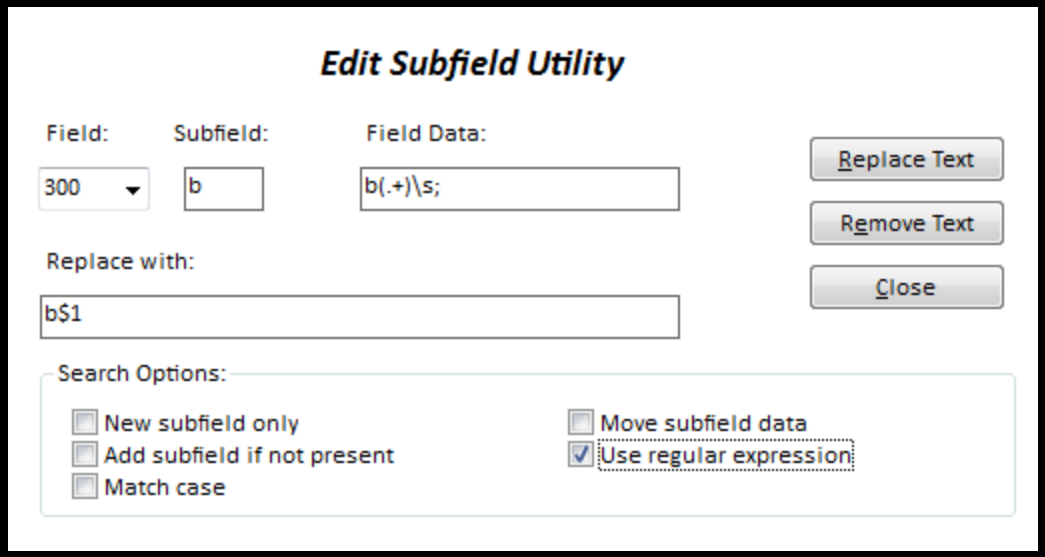
Example 3:

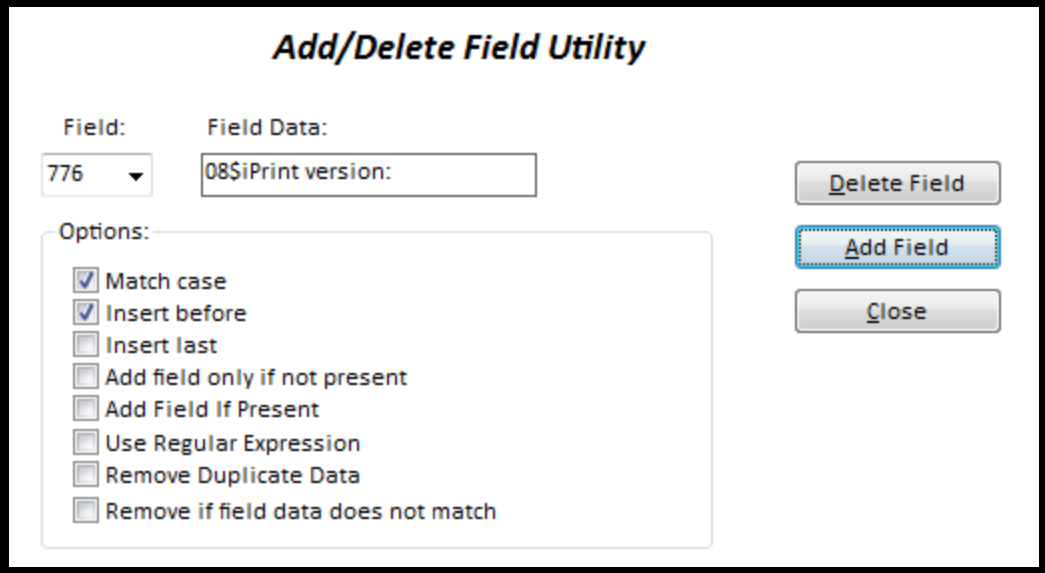
Use regex to convert print to electronic

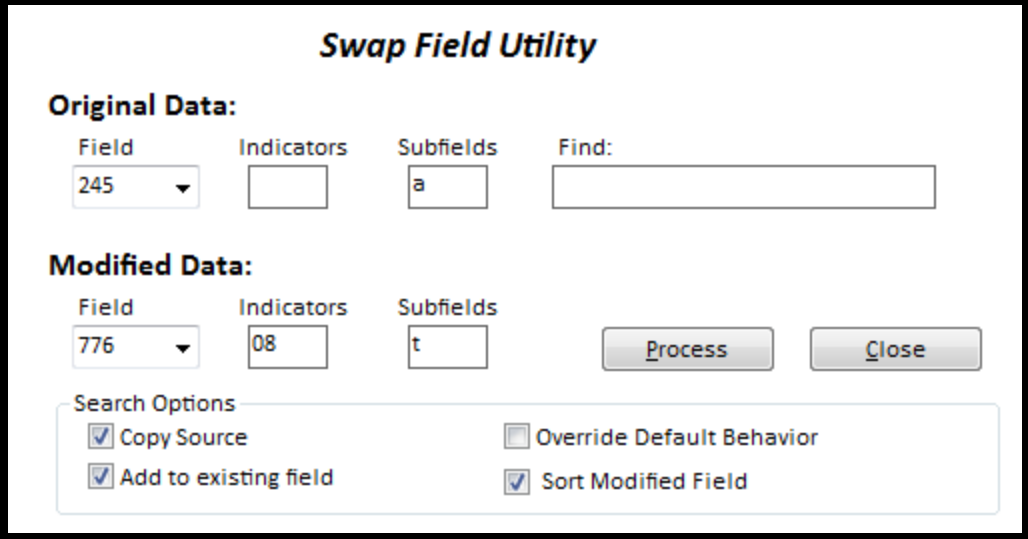

Edit Subfield Utility
Field:
Subfield:
Field Data:
020
a
a(.+)$
Replace Text
Remove Text
Close
Replace with:
z$1
Search Options:
New subfield only
Add subfield if not present
Match case
Move subfield data
Use regular expression


Edit Subfield Utility
Field:
Subfield:
Field Data:
245
a
a(.+)\s([:|/])
Replace Text
Remove Text
Close
Replace with:
a$1 $h[electronic resource] $2
Search Options:
New subfield only
Add subfield if not present
Match case
Move subfield data
Use regular expression


Edit Subfield Utility
Field:
Subfield:
Field Data:
300
a
a(.+)\s([:|;])
Replace Text
Remove Text
Close
Replace with:
a1 online resource ($1) $2
Search Options:
New subfield only
Add subfield if not present
Match case
Move subfield data
Use regular expression


Edit Subfield Utility
Field:
Subfield:
Field Data:
300
b
b(.+)\s;
Replace Text
Remove Text
Close
Replace with:
b$1
Search Options:
New subfield only
Add subfield if not present
Match case
Move subfield data
Use regular expression


Add/Delete Field Utility
Field:
Field Data:
776
08$iPrint version:
Delete Field
Add Field
Close
Options:
Match case
Insert before
Insert last
Add field only if not present
Add Field If Present
Use Regular Expression
Remove Duplicate Data
Remove if field data does not match


Swap Field Utility
Original Data:
Field
Indicators
Subfields
Find:
245
a
Modified Data:
Field
Indicators
Subfields
776
08
t
Process
Close
Search Options
Copy Source
Add to existing field
Override Default Behavior
Sort Modified Field

A few final suggestions

- Save often – special undo only goes 1 step.
- Measure twice, cut once:
  - use the "find" and "find all" features to test out your Regex
  - when it works the way you want, cut/paste the expression into the desired tool
- Save your expression!
  - use Notepad or other tool
  - MARCEdit has a tool (in development?)
- Regular Expression Quick Reference:
  http://msdn.microsoft.com/en-us/library/az24scfc.aspx

ACKNOWLEDGEMENTS

[slide 4]

Henriette Avram, http://en.wikipedia.org/wiki/Henriette_Avram

Ken Thompson, http://en.wikipedia.org/wiki/Ken_Thompson

[slide 5]

Library of Congress Cataloging Distribution Service, ca. 1900, http://blogs.loc.gov/picturethis/2011/09/a-closer-look-dating-a-photo/

[slide 6]

Ken Thompson, Dennis Ritchie, and the PDP-11, http://nushackers.org/why/

All screenshots from: MARCEdit v. 5.

The speaker would like to thank the Mass. Library Association, for providing the opportunity to present; and Daniel Saulean for providing logistical assistance.