Migrating XML resources into ArchivesSpace

ArchivesSpace Member Forum 2018
Celia Caust-Ellenbogen
Friends Historical Library of Swarthmore College
ccauste1@swarthmore.edu

Getting Ready

TABLE OF CONTENTS

Summary
SHORT HISTORY OF MEETING
BRIEF DESCRIPTION OF THE RECORDS
Restrictions
Searching in Our On-Line Catalogue for Related Materials
Administrative Information
DETAILED DESCRIPTION OF THE RECORDS
Ministers & Elders/ Ministry and Counsel
State of the Meeting Reports
Faith and Practice
Related Committees
Haverford College Quaker Collection
Friends Historical Library of Swarthmore College
Back to List of Meetings
Go to New York Yearly Meeting website

NEW YO
OF THE RELI

RECORDS OF NEW YO
A

A Finding Aid for its Record
Sv

SUMMARY

Meeting Name:	New York Yearl
Title	Records of Mini
Dates:	1672-
Quantity:	ca.46 linear ft.
Repository:	Friends Historic Swarthmore Col
Location:	For current infor online catalog:

SHORT HISTORY OF MEETING

New York Yearly Meeting was first hel
consisting essentially of the Flushing Q
Meeting. Meeting for Sufferings was es
business of the Yearly Meeting while it

Return to the Table of Contents

BRIEF DESCRIPTION OF THE RECORDS

Records of the Meeting of Ministry and Counsel and its Orthodox and Hicksite predecessors, the Meetings of Ministers and Elders of New York Yearly Meeting.

An Inventory of the Allen Family Papers, 1700-1954

(4 boxes; 2 linear ft.)

RG 5/001

© Friends Historical Library of Swarthmore College.
Swarthmore, Pennsylvania 19081-1399 U.S.A.

SWARTHMORE COLLEGE INCORPORATED 1864

Table of contents

- Background note
- Scope and content
  - Arrangement
- Administrative information
  - Restrictions
  - Preferred citation
  - Provenance
  - Separated materials
- Added entries
- Contact information
- Inventory
  - Ser. 1. Biographical and Genealogical Material, 1839-1935
  - Ser. 2. Correspondence, 1789-1954
  - Ser. 3. Journals and Diaries, 1831-1893
  - Ser. 4. Financial Papers, 1813-1851
  - Ser. 5. Legal Papers, 1849-1950, n.d.
  - Ser. 6. Miscellaneous Papers, 1823-1924, n.d.
- Return to Friends Historical Library
- Return to Tripod

Abstract

Contains the papers of the Allen Family, a Quaker family of Philadelphia and Delaware County, Pennsylvania, which traced its ancestry back to one of William Penn's commissioners. Includes papers relating to the Allen and related Breidenhart, Casdorp, Gibbons, Hubley, Jacobs, Matlack, and Samuel families. The Allen family papers contain extensive genealogical material, family correspondence, journals and memorials which provide excellent source material on Quaker family life in the 19th century in the Philadelphia area. Of particular interest are letters from Charles and Rebecca Allen to their children, several of whom were educated at Westtown School.

Background note:

The Allen family traces its ancestry to one of William Penn's commissioners, Nathaniel Allen, who landed at Chester Creek in 1681. The collection's main focus is on the 5th and 6th generations: Charles Allen (1776-1843) and his son George B. Allen (1804-1869).

Charles Allen, a druggist, first married Sarah Breidenhart (1778-1807) and later married Rebecca Jackson (1781-1860), a teacher. Their home was on Pine Street in Philadelphia, and Charles and Rebecca were overseers and elders of Philadelphia Monthly Meeting, Southern District. Charles frequently accompanied ministers on their visits to other meetings. Rebecca served on the committee in charge of Westtown Boarding School.

TABLE OF CONTENTS

Short History of the Meeting

bym
bym-rsf.org
Baltimore Yearly Meeting
of the Religious Society of Friends
Quakers

LTIMORE MONTHLY MEETING FOR THE EASTERN DISTRICT

ecords Deposited in Friends Historical Library of Swarthmore College

e Monthly Meeting for the Eastern District (Society of Friends : 1807-1819
re, Md.)

9
t.
n English
Historical Library of Swarthmore College

EETING

became the City of Baltimore in 1781 after moving from Patapsco, and in
g was formed. The meeting grew quickly, and in 1805 a second meeting
eet. In 1807, the large number of Friends in Baltimore led to the creation of
more Monthly Meeting changed its name to the Baltimore Monthly
, and the new meeting at Lombard Street became the Baltimore Monthly
t. Tensions over the Aisquith St. graveyard arose, and resulted in the laying
eting for the Eastern District by the Yearly Meeting; it became a
more Monthly Meeting for the Western District.

ORDS

Meeting of Friends for the Eastern District, and its predecessor Baltimore
1729-1819. Includes men's and joint minutes 1792-1819, women's minutes

Records

Includes minutes, reports, and other papers of the Arrangements and Hospitality Committee, 1964-86. Also included are records of the Meal Committee, 1962-86, and the Philadelphia... Full description

Archives and Manuscripts

Corporate Author: Philadelphia Yearly Meeting of the Religious Society of Friends. Arrangements and Hospitality Committee.

Other Authors: Philadelphia Yearly Meeting of the Religious Society of Friends. Planning Committee.; Philadelphia Yearly Meeting of the Religious Society of Friends. Lunch Committee.; Philadelphia Yearly Meeting of the Religious Society of Friends. Meal Committee.

Language: English

Subjects: Philadelphia Yearly Meeting of the Religious Society of Friends > Archives.
Society of Friends > Pennsylvania > Philadelphia.
Society of Friends > Pennsylvania.

Holdings | Description | Staff View

LEADER	02505npdaa2200397la 4500		
001	36551507		
005	19970314102439.0		
008	970314i19621986pau eng d		
040			|a PSH |e appm |c PSH
043			|a n-us-pa
049			|a PSHF
090			|a RG2/Phy/346
090	8	1	|b RG2/Phy/346
099			|a RG2/Phy/346
110	2		|a Philadelphia Yearly Meeting of the Religious Society of Friends. |b Arrangements and Hospitality Committee.
245	1	0	|a Records, |f 1962-1986.

Similar Items

Philadelphia Yearly Meeting Records
Published: (1668-2018)

Records concerning Administration of Yearly Meeting Funds for the Benefit of Elderly Friends
Published: (1968-1970)

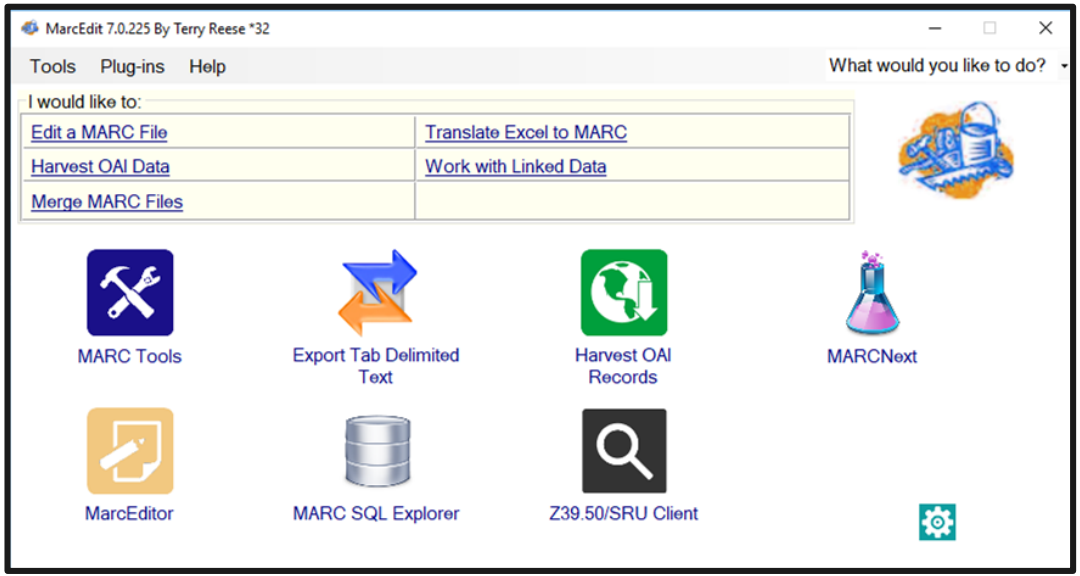
Memorials
Published: (1828-1898)

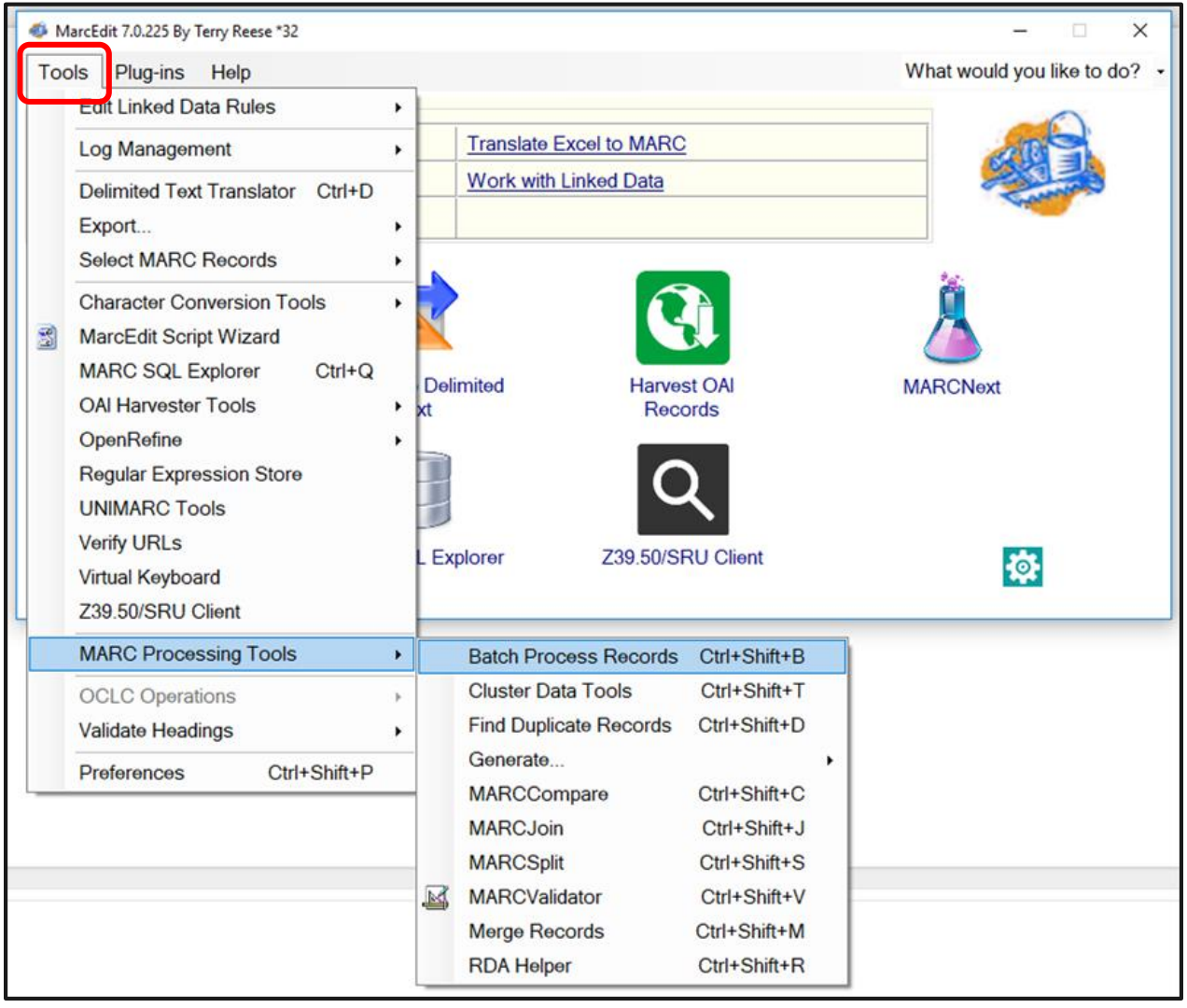
Name index to the minutes of Philadelphia Yearly Meeting (Society of Friends), 1681-1779
Published: (1994)

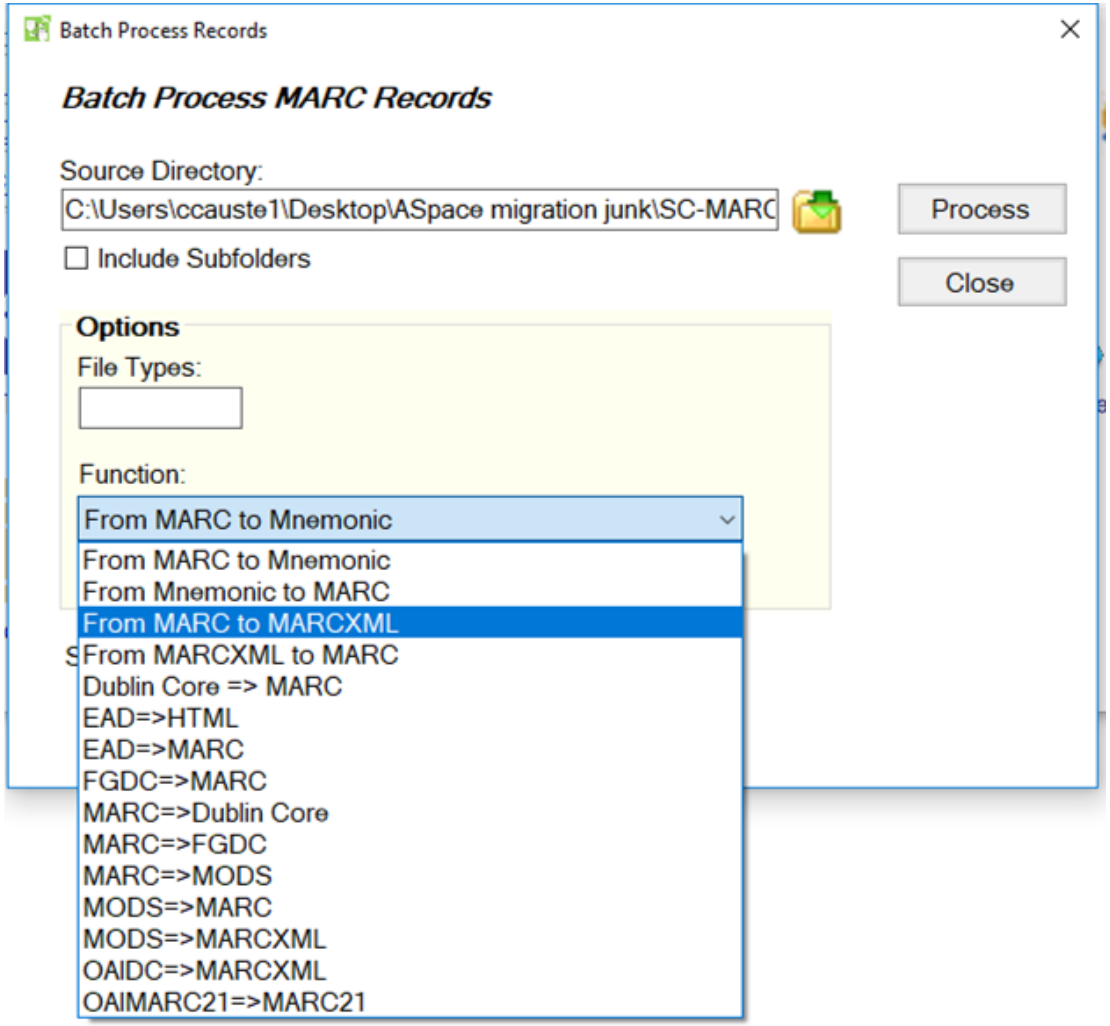
Philadelphia Quarterly Meeting Records
Published: (1828-2018)

MarcEdit

Free! https://marcedit.reeset.net/




MarcEdit 7.0.225 By Terry Reese *32
Tools
Plug-ins
Help
What would you like to do?
Edit Linked Data Rules
Log Management
Delimited Text Translator Ctrl+D
Export...
Select MARC Records
Character Conversion Tools
MarcEdit Script Wizard
MARC SQL Explorer Ctrl+Q
OAI Harvester Tools
OpenRefine
Regular Expression Store
UNIMARC Tools
Verify URLs
Virtual Keyboard
Z39.50/SRU Client
MARC Processing Tools
OCLC Operations
Validate Headings
Preferences Ctrl+Shift+P
Batch Process Records Ctrl+Shift+B
Cluster Data Tools Ctrl+Shift+T
Find Duplicate Records Ctrl+Shift+D
Generate...
MARCCompare Ctrl+Shift+C
MARCJoin Ctrl+Shift+J
MARCSplit Ctrl+Shift+S
MARCValidator Ctrl+Shift+V
Merge Records Ctrl+Shift+M
RDA Helper Ctrl+Shift+R
Translate Excel to MARC
Work with Linked Data
Harvest OAI Records
MARCNext
Z39.50/SRU Client


Batch Process Records
Batch Process MARC Records
Source Directory:
C:\Users\ccauste1\Desktop\ASpace migration junk\SC-MARC
Include Subfolders
Process
Close
Options
File Types:
Function:
From MARC to Mnemonic
From MARC to Mnemonic
From Mnemonic to MARC
From MARC to MARCXML
From MARCXML to MARC
Dublin Core => MARC
EAD=>HTML
EAD=>MARC
FGDC=>MARC
MARC=>Dublin Core
MARC=>FGDC
MARC=>MODS
MODS=>MARC
MODS=>MARCXML
OAIDC=>MARCXML
OAIMARC21=>MARC21

https://www.oxygenxml.com/

Type	Cost per new license (renewals are cheaper)
Academic	$99
Business	$688
Individual	$198
30-day trial	FREE

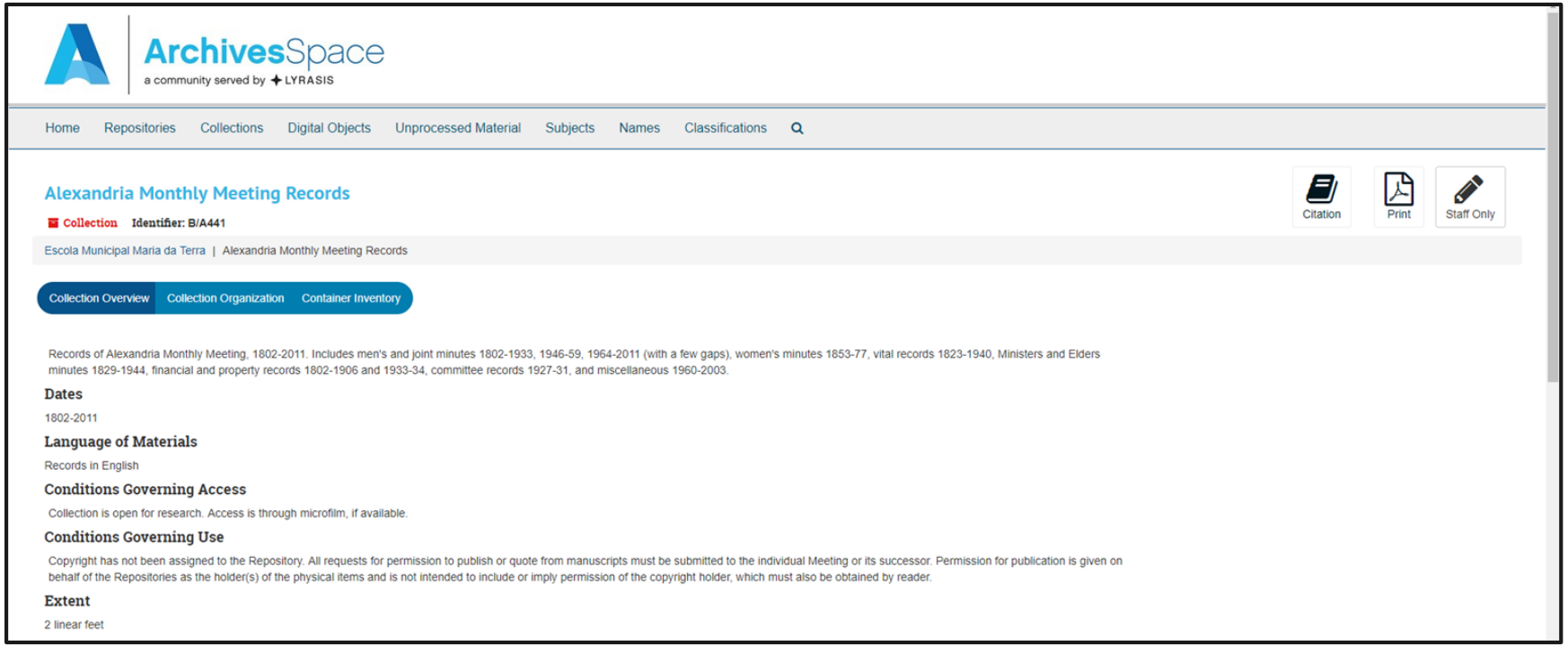
Blogs

- ArchivesSpace @ Yale
  http://campuspress.yale.edu/yalearchivesspace/, especially
  http://campuspress.yale.edu/yalearchivesspace/2015/07/22/validation-scenarios/
- Bentley
  http://archival-integration.blogspot.com/, especially http://archival-integration.blogspot.com/2015/05/legacy-ead-clean-up-getting-started.html
- Chaos → Order
  https://icantiemyownshoes.wordpress.com/, especially
  https://icantiemyownshoes.wordpress.com/category/eads-in-aggregate/
- Way More at **Awesome ArchivesSpace!**
  https://github.com/archivesspace/awesome-archivesspace#migrations

Getting started

Test imports in the Sandbox (or your dev server)

http://sandbox.archivesspace.org/



Test imports in the Sandbox (or your dev server)

http://sandbox.archivesspace.org/

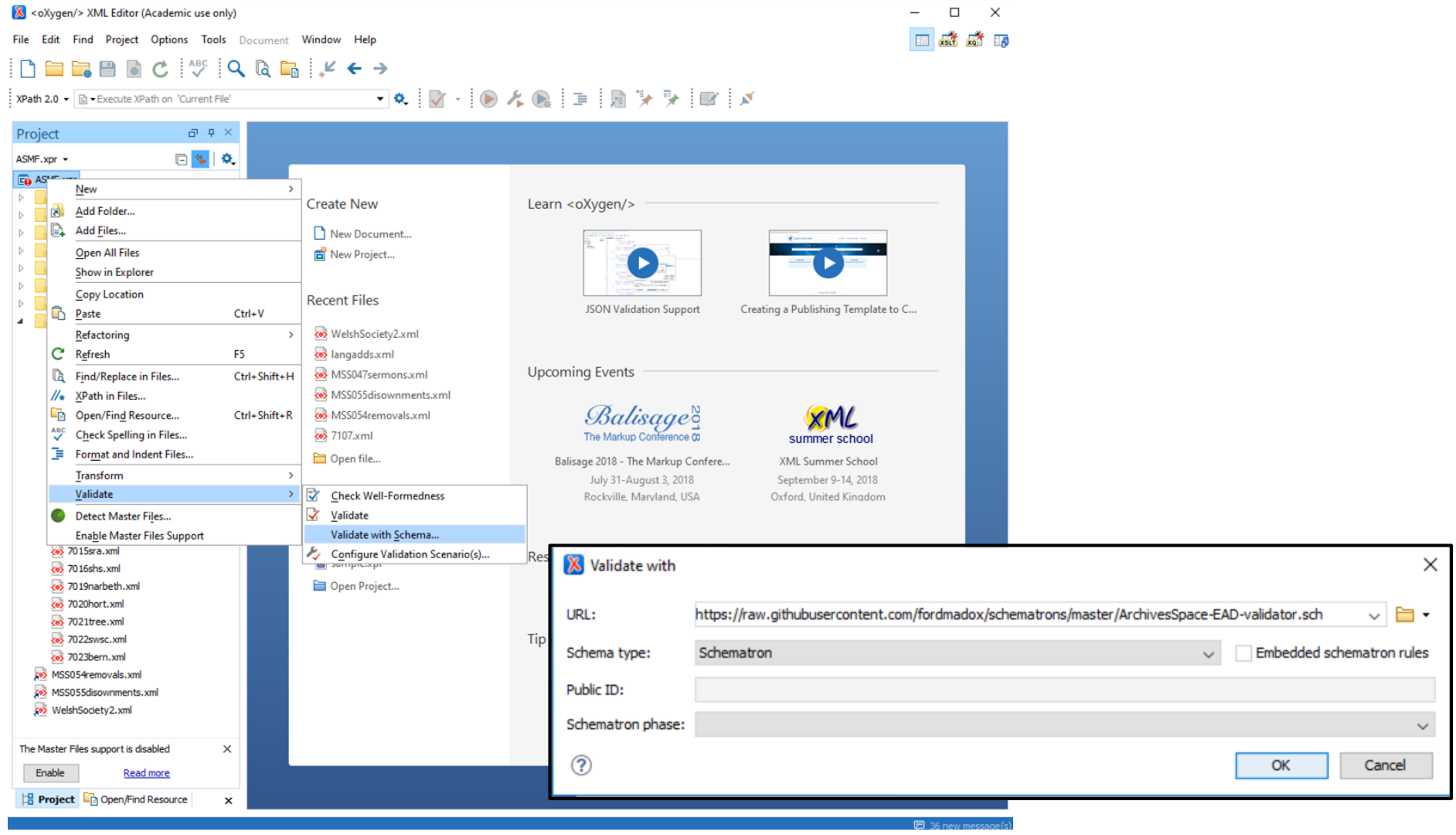
Log

```
!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
IMPORT ERROR
!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!


The following errors were found:
        dates/0/end : must not be before begin


 For JSONModel(:archival_object):
 #<JSONModel(:archival_object) {"jsonmodel_type"=>"archival_object", "external_ids"=>[], "subjects"=>[], "linked_events"=>[], "extents


In :
 &lt;c02 class=&quot;cdata&quot; level=&quot;file&quot;&gt; ... &lt;/c02&gt;
```

Mark Custer's EAD-validation schematron

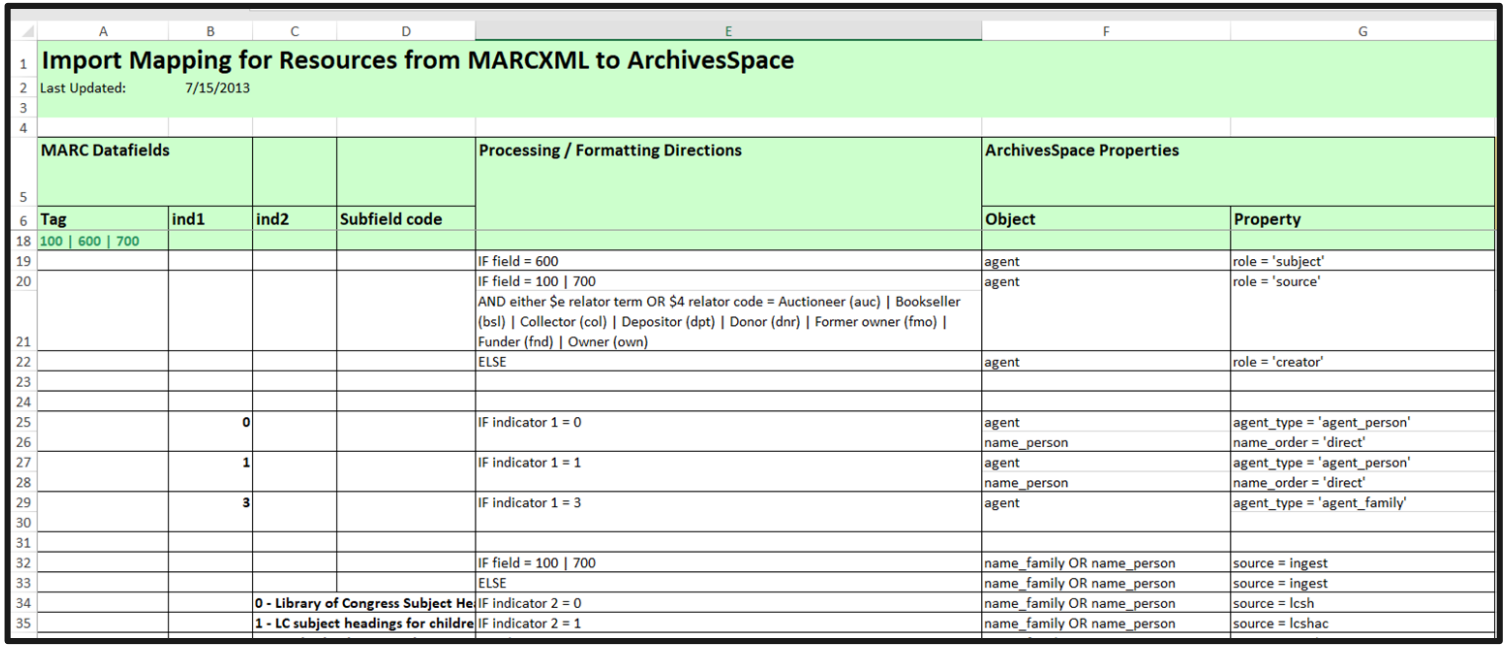
https://raw.githubusercontent.com/fordmadox/schematrons/master/ArchivesSpace-EAD-validator.sch

```
<?xml version="1.0" encoding="UTF-8"?>
<schema xmlns="http://purl.oclc.org/dsdl/schematron"
    xmlns:sqf="http://www.schematron-quickfix.com/validator/process"
    queryBinding="xslt2">
    <ns uri="http://www.w3.org/1999/xlink" prefix="xlink"/>
    <ns uri="urn:isbn:1-931666-22-9" prefix="ead"/>
    <!-- eventually, this file should test for all of the ASpace "EAD" requirements prior to import
        (still need to figure what all of those are, including all undocumented data model constraints, etc.)

for the time being, i removed namespace checks so that the same rules will work for DTD and/or schema-associated files.

        still to add:
            dao stuff (must have title and href attributes?)
           something about field lengths (also need to test for length of title / unit ids, etc.)... if something is too long for the database, etc., like the 65k character limit :(
           what else???

        still to do:
           re-write the error messages!
           make sure that the error messages include everything needed
           pray that i can use xpath 2.0, as i've done here
           group and arrange this file like a semi-decent schematron file should actually be structured (and learn how to do that!)

    -->

    <pattern>
        <rule context="*:archdesc">
            <assert test="@level">You must supply a level attribute at the resource level</assert>
        </rule>
        <rule context="*:archdesc/*:did">
            <assert test="*:unittitle[normalize-space()]">You must supply a
                title at the resource level</assert>
            <assert test="descendant::*:unitdate[normalize-space()] or descendant::*:unitdate[@normal]"
                >You must supply a date at the resource level (including as child of unittitle)</assert>
```


<oXygen/> XML Editor (Academic use only)
File Edit Find Project Options Tools Document Window Help
XPath 2.0
Execute XPath on 'Current File'
Project
ASMF.xpr
New
Add Folder...
Add Files...
Open All Files
Show in Explorer
Copy Location
Paste Ctrl+V
Refactoring
Refresh F5
Find/Replace in Files... Ctrl+Shift+H
XPath in Files...
Open/Find Resource... Ctrl+Shift+R
Check Spelling in Files...
Format and Indent Files...
Transform
Validate
Detect Master Files...
Enable Master Files Support
Check Well-Formedness
Validate
Validate with Schema...
Configure Validation Scenario(s)...
7015sra.xml
7016shs.xml
7019narbeth.xml
7020hort.xml
7021tree.xml
7022swsc.xml
7023bern.xml
MSS054removals.xml
MSS055disownments.xml
WelshSociety2.xml
The Master Files support is disabled
Enable
Read more
Project
Open/Find Resource
Create New
New Document...
New Project...
Recent Files
WelshSociety2.xml
langadds.xml
MSS047sermons.xml
MSS055disownments.xml
MSS054removals.xml
7107.xml
Open file...
Open Project...
Learn <oXygen/>
JSON Validation Support
Creating a Publishing Template to C...
Upcoming Events
Balisage 2018 The Markup Conference
Balisage 2018 - The Markup Confere...
July 31-August 3, 2018
Rockville, Maryland, USA
XML summer school
XML Summer School
September 9-14, 2018
Oxford, United Kingdom
Validate with
URL: https://raw.githubusercontent.com/fordmadox/schematrons/master/ArchivesSpace-EAD-validator.sch
Schema type: Schematron
Embedded schematron rules
Public ID:
Schematron phase:
OK
Cancel

ArchivesSpace MARCXML Import / Export Map

http://archivesspace.org/wp-content/uploads/2016/08/MARCXML-Import-Export-Mapping-20130715.xlsx

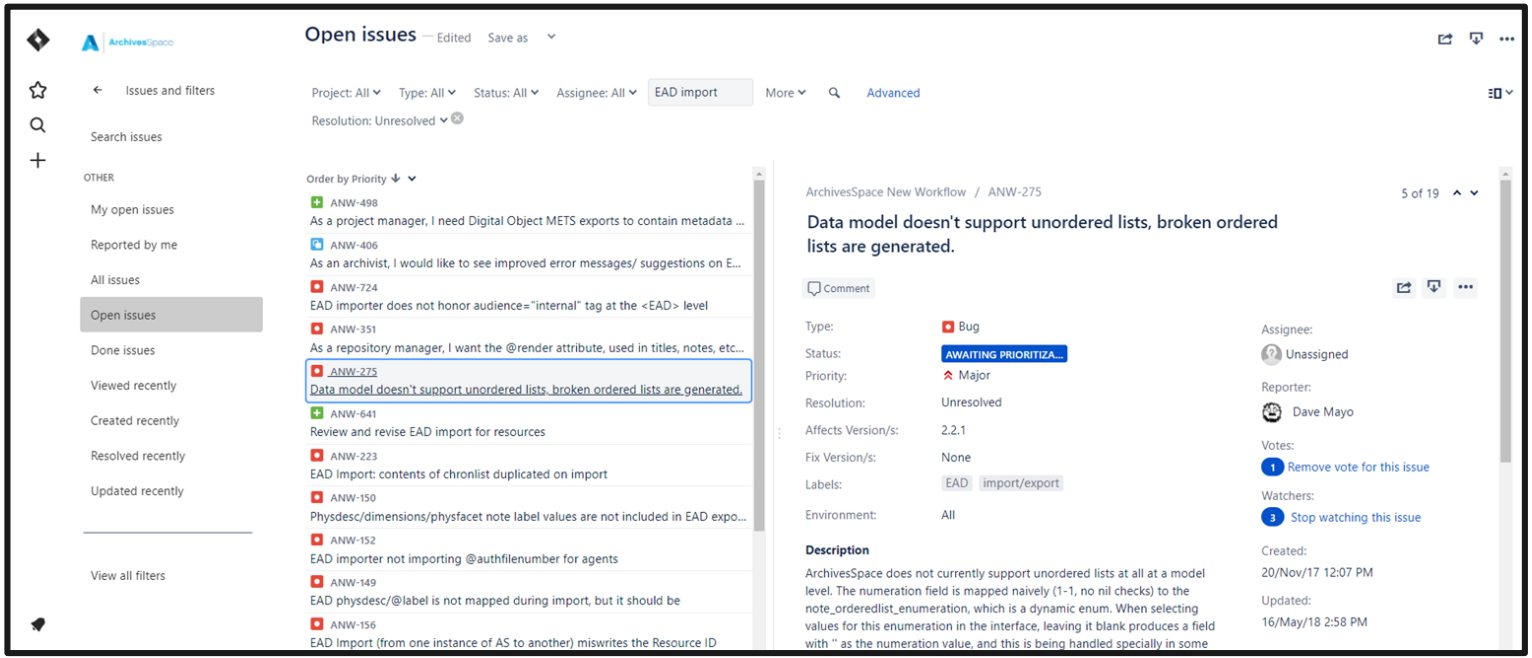


	A	B	C	D	E	F	G
1	**Import Mapping for Resources from MARCXML to ArchivesSpace**						
2	Last Updated:	7/15/2013					
3							
4							
5	**MARC Datafields**				**Processing / Formatting Directions**	**ArchivesSpace Properties**	
6	**Tag**	**ind1**	**ind2**	**Subfield code**		**Object**	**Property**
18	100 | 600 | 700						
19					IF field = 600	agent	role = 'subject'
20					IF field = 100 | 700	agent	role = 'source'
21					AND either $e relator term OR $4 relator code = Auctioneer (auc) | Bookseller (bsl) | Collector (col) | Depositor (dpt) | Donor (dnr) | Former owner (fmo) | Funder (fnd) | Owner (own)		
22					ELSE	agent	role = 'creator'
23							
24							
25		0			IF indicator 1 = 0	agent	agent_type = 'agent_person'
26						name_person	name_order = 'direct'
27		1			IF indicator 1 = 1	agent	agent_type = 'agent_person'
28						name_person	name_order = 'direct'
29		3			IF indicator 1 = 3	agent	agent_type = 'agent_family'
30							
31							
32					IF field = 100 | 700	name_family OR name_person	source = ingest
33					ELSE	name_family OR name_person	source = ingest
34			0 - Library of Congress Subject He		IF indicator 2 = 0	name_family OR name_person	source = lcsh
35			1 - LC subject headings for childre		IF indicator 2 = 1	name_family OR name_person	source = lcshac

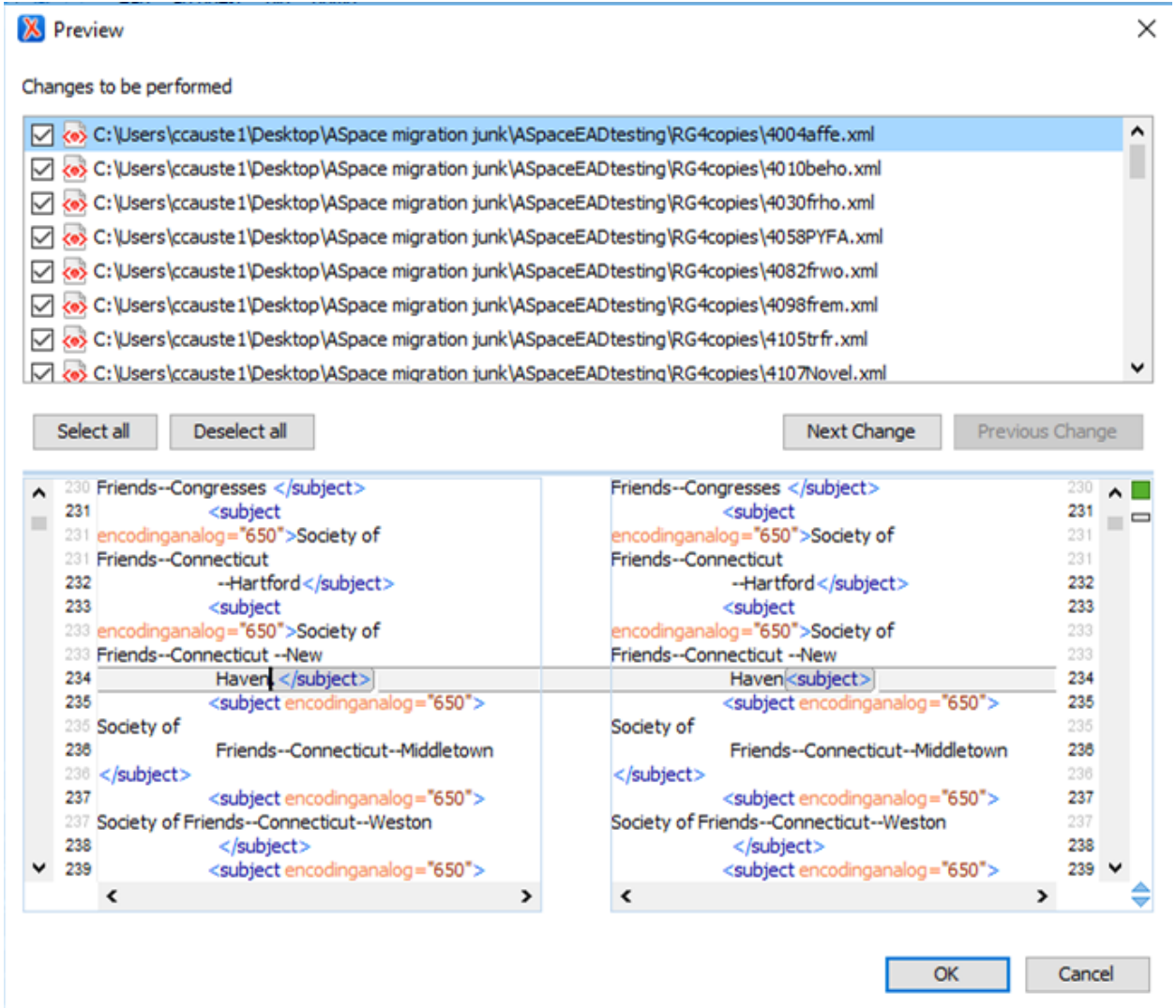
Known issues/bugs in JIRA

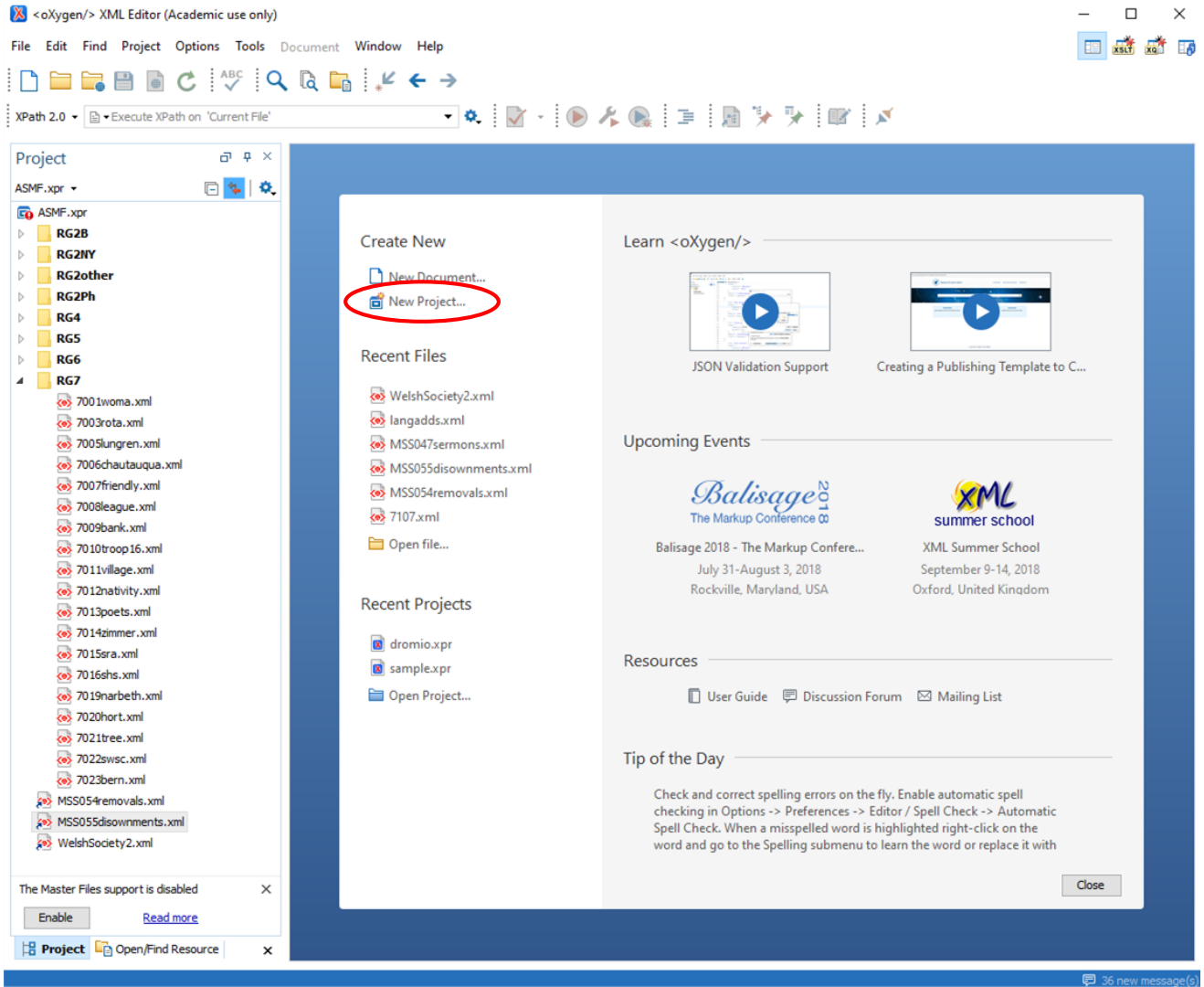
https://archivesspace.atlassian.net

Issues and filters > Open issues



Using Oxygen


Preview
Changes to be performed
C:\Users\ccauste1\Desktop\ASpace migration junk\ASpaceEADtesting\RG4copies\4004affe.xml
C:\Users\ccauste1\Desktop\ASpace migration junk\ASpaceEADtesting\RG4copies\4010beho.xml
C:\Users\ccauste1\Desktop\ASpace migration junk\ASpaceEADtesting\RG4copies\4030frho.xml
C:\Users\ccauste1\Desktop\ASpace migration junk\ASpaceEADtesting\RG4copies\4058PYFA.xml
C:\Users\ccauste1\Desktop\ASpace migration junk\ASpaceEADtesting\RG4copies\4082frwo.xml
C:\Users\ccauste1\Desktop\ASpace migration junk\ASpaceEADtesting\RG4copies\4098frem.xml
C:\Users\ccauste1\Desktop\ASpace migration junk\ASpaceEADtesting\RG4copies\4105trfr.xml
C:\Users\ccauste1\Desktop\ASpace migration junk\ASpaceEADtesting\RG4copies\4107Novel.xml
Select all
Deselect all
Next Change
Previous Change
Friends--Congresses </subject>
<subject
encodinganalog="650">Society of
Friends--Connecticut
--Hartford</subject>
<subject
encodinganalog="650">Society of
Friends--Connecticut --New
Haven</subject>
<subject encodinganalog="650">
Society of
Friends--Connecticut--Middletown
</subject>
<subject encodinganalog="650">
Society of Friends--Connecticut--Weston
</subject>
<subject encodinganalog="650">
OK
Cancel


<oXygen/> XML Editor (Academic use only)
File Edit Find Project Options Tools Document Window Help
XPath 2.0
Execute XPath on 'Current File'
Project
ASMF.xpr
ASMF.xpr
RG2B
RG2NY
RG2other
RG2Ph
RG4
RG5
RG6
RG7
7001woma.xml
7003rota.xml
7005lungren.xml
7006chautauqua.xml
7007friendly.xml
7008league.xml
7009bank.xml
7010troop16.xml
7011village.xml
7012nativity.xml
7013poets.xml
7014zimmer.xml
7015sra.xml
7016shs.xml
7019narbeth.xml
7020hort.xml
7021tree.xml
7022swsc.xml
7023bern.xml
MSS054removals.xml
MSS055disownments.xml
WelshSociety2.xml
The Master Files support is disabled
Enable
Read more
Project
Open/Find Resource
Create New
New Document...
New Project...
Recent Files
WelshSociety2.xml
langadds.xml
MSS047sermons.xml
MSS055disownments.xml
MSS054removals.xml
7107.xml
Open file...
Recent Projects
dromio.xpr
sample.xpr
Open Project...
Learn <oXygen/>
JSON Validation Support
Creating a Publishing Template to C...
Upcoming Events
Balisage 2018 - The Markup Confere...
July 31-August 3, 2018
Rockville, Maryland, USA
XML Summer School
September 9-14, 2018
Oxford, United Kingdom
Resources
User Guide
Discussion Forum
Mailing List
Tip of the Day
Check and correct spelling errors on the fly. Enable automatic spell checking in Options -> Preferences -> Editor / Spell Check -> Automatic Spell Check. When a misspelled word is highlighted right-click on the word and go to the Spelling submenu to learn the word or replace it with
Close
36 new message(s)

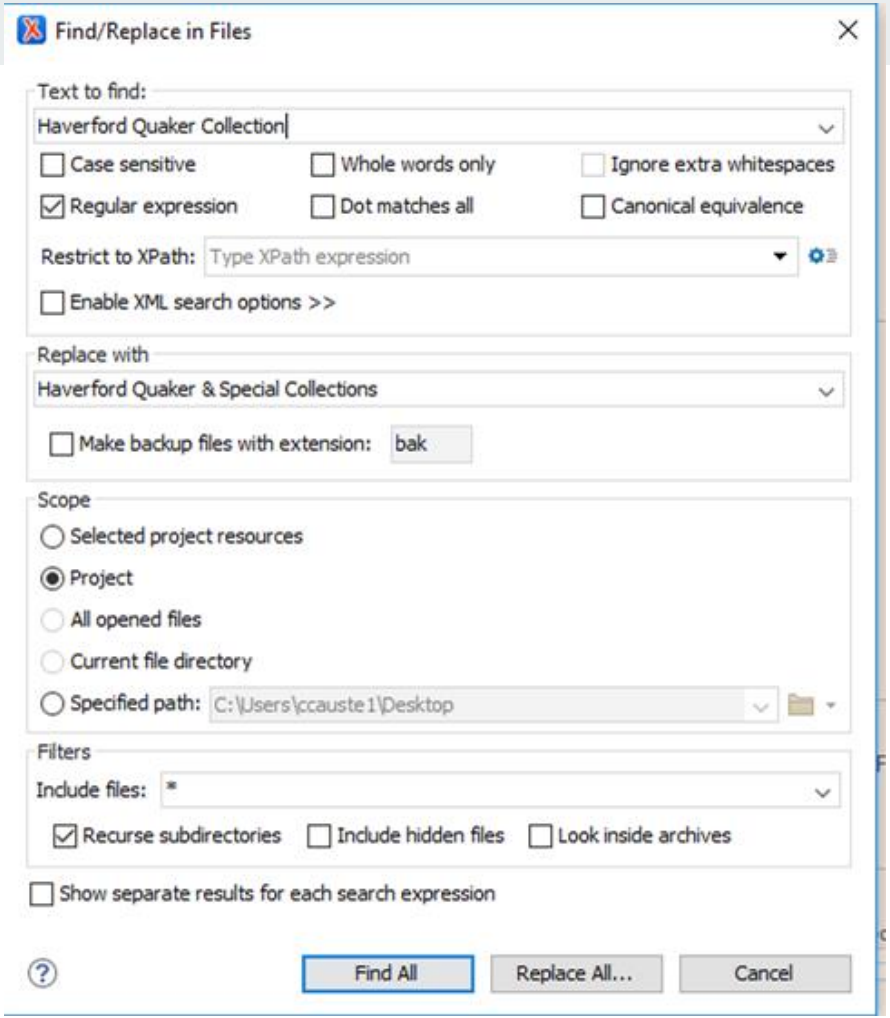
Problem

I need to find/replace text in a bunch of files.

Example: change “Haverford Quaker Collection” to “Haverford Quaker & Special Collections”

Easy find/replace

- Extra periods at end of subject
- - - instead of -- in subjects
- <langusage>English</langusage> instead of <langusage>ENG</langusage>
- Empty <p/> tags



Problem

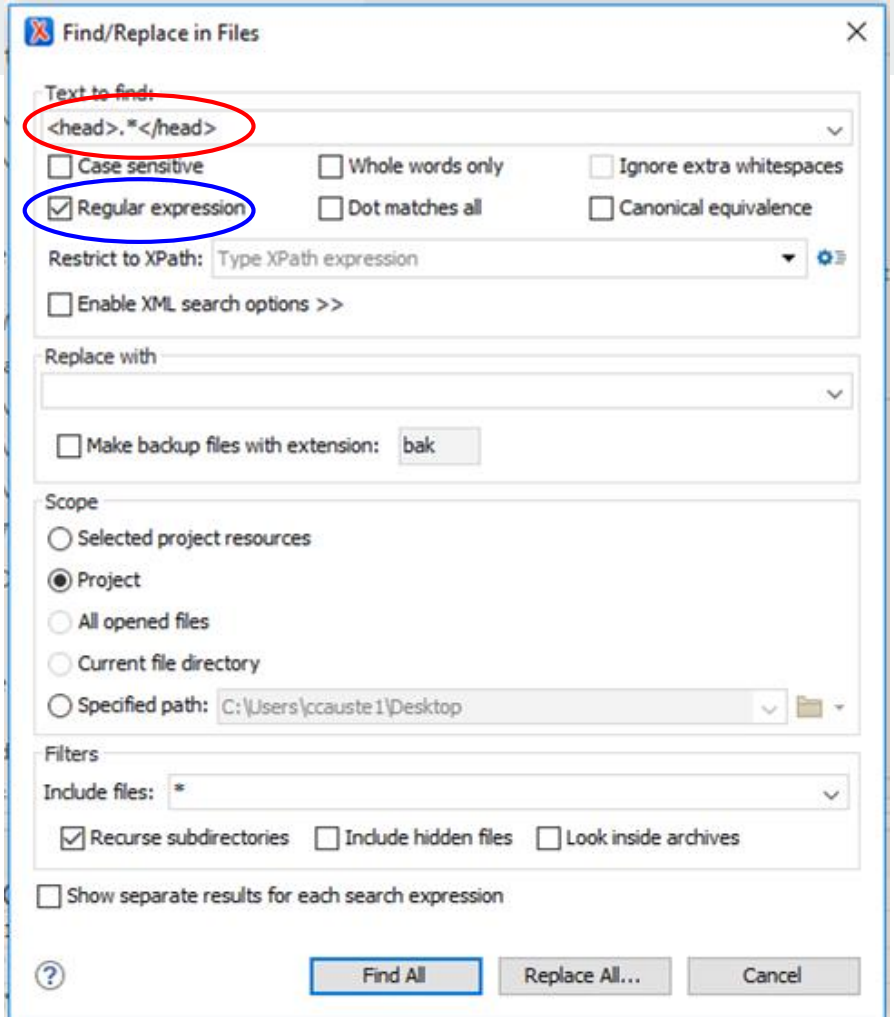
I need to delete a tag and anything it contains across all my files.

Example: delete all the note head labels

```
<arrangement encodinganalog="351$a">
  <head id="a4">Organization of the Records</head>
  <p>Divided into series: Minutes (men's &amp;amp; joint): 1.1 1840-1894, 1.2
    1894-1915; women's minutes: 2.1 1840-1869; vital records:, 3.1 1840-1915;
    Ministers and Elders records: 4.1 1841-87, 4.2 1887-94; Property: 5.1 Deeds and
    Property Records 1898-1908; Misc.: 6.1 Correspondence 1856-1915.</p>
</arrangement><descgrp>
```

Do a lot more with a tiny bit of RegEx...

.* - anything



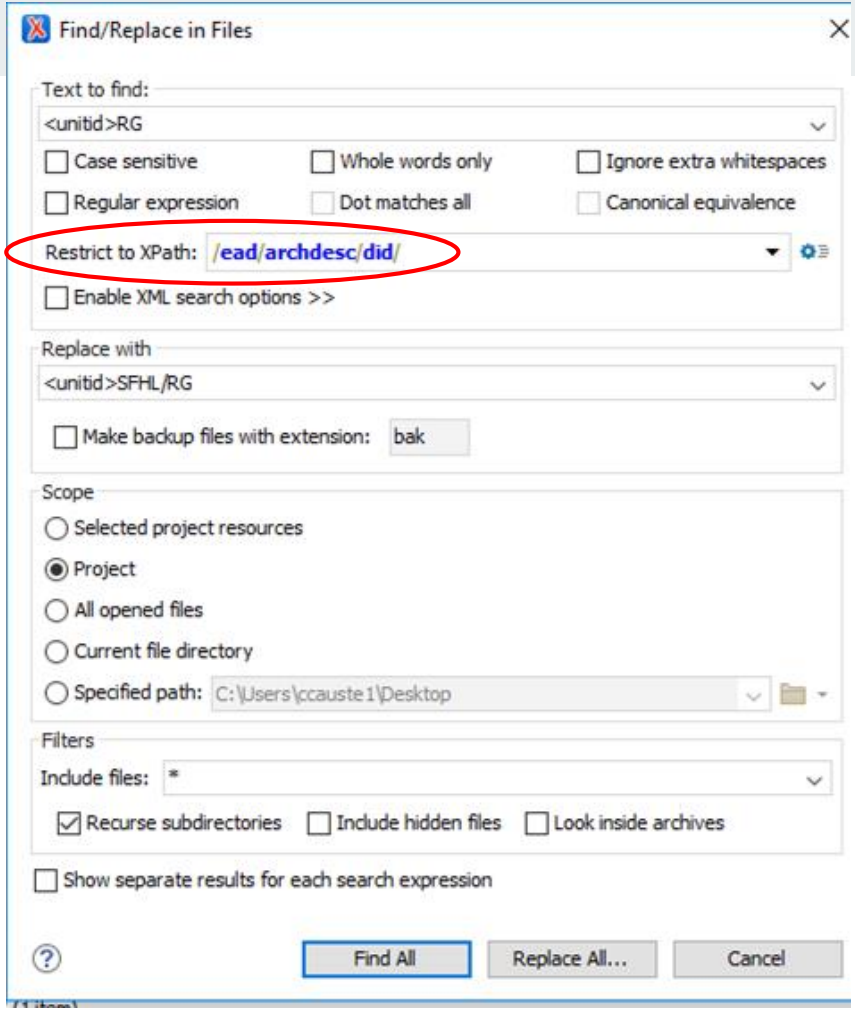
Problem

I need to make a change, but only in a certain field

Example: add a prefix to identifier at collection-level only

Do a lot more with a tiny bit of XPath...

/node/node/...etc



How to find the XPath

/ead

/archdesc[1]

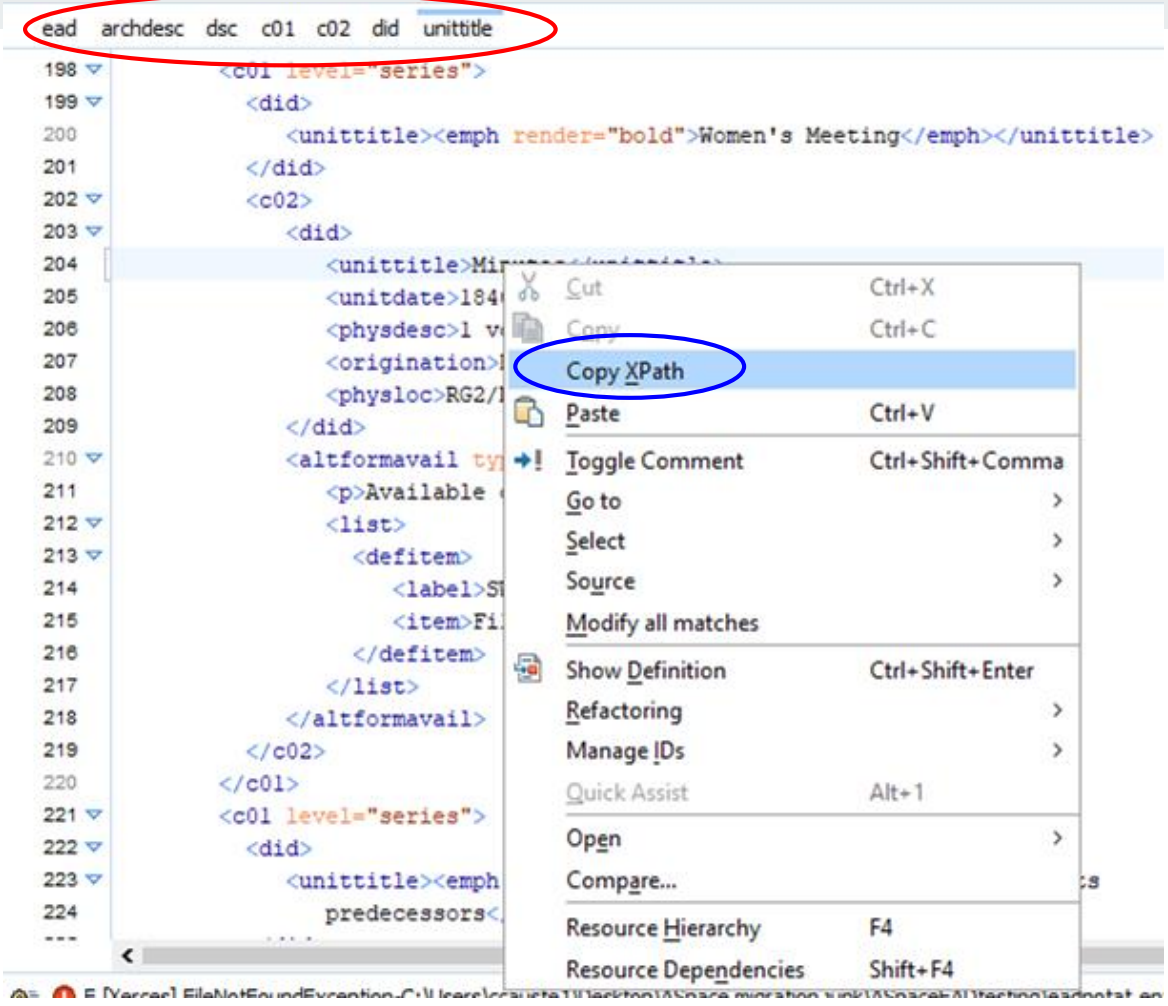
/dsc[1]

/c01[3]

/c02[1]

/did[1]

/unittitle[1]



msplit00000043.xml*

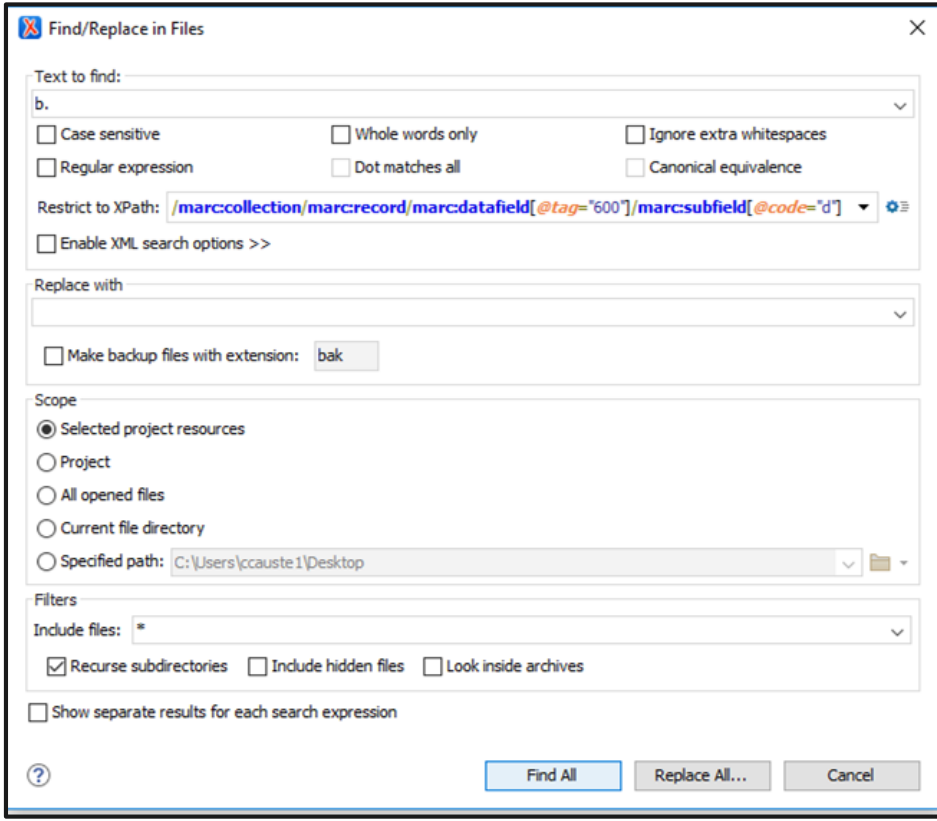
marc:collection marc:record marc:datafield marc:subfield

```
57  <marc:datafield tag="600" ind1="1" ind2="0">
58      <marc:subfield code="a">Ogden, Charles Smith,</marc:subfield>
59      <marc:subfield code="d">b. 1822.</marc:subfield>
60  </marc:datafield>
```

Do a lot more with a tiny bit of XPath...

select attributes with [@tag="XXX"]

or ignore with [not(@tag="XXX")]



Problem

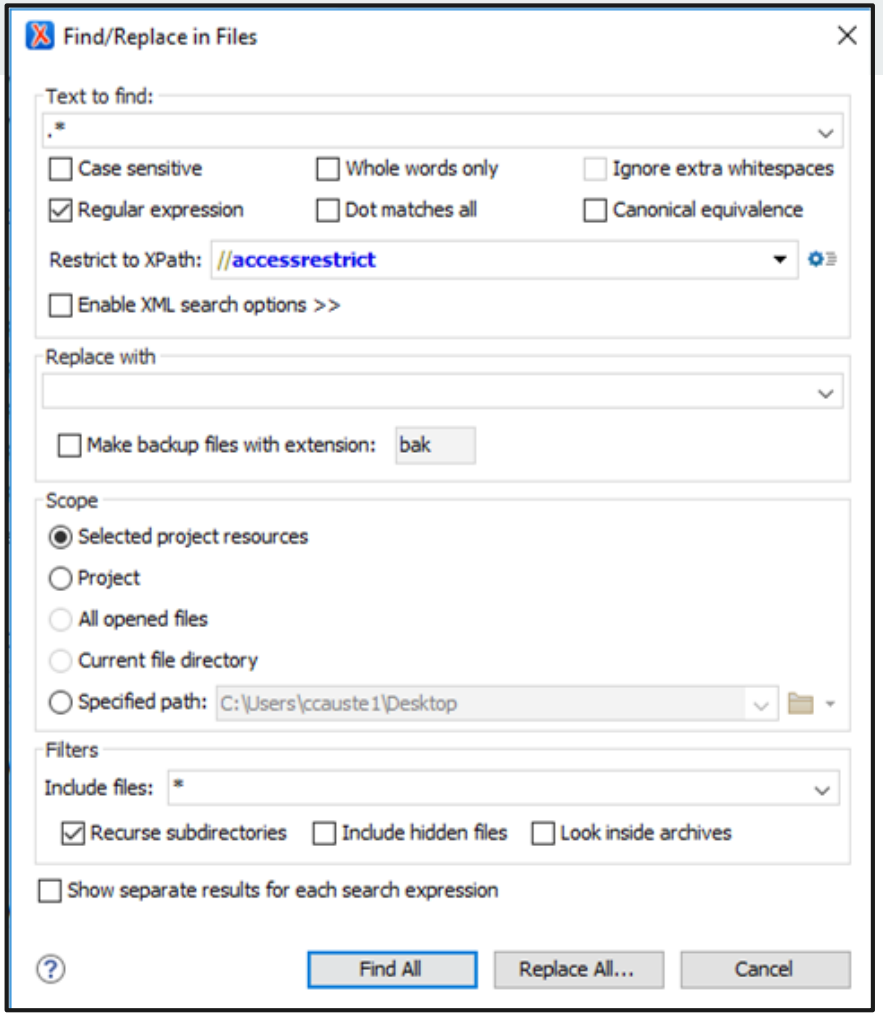
I want to review the contents of a tag in all my files

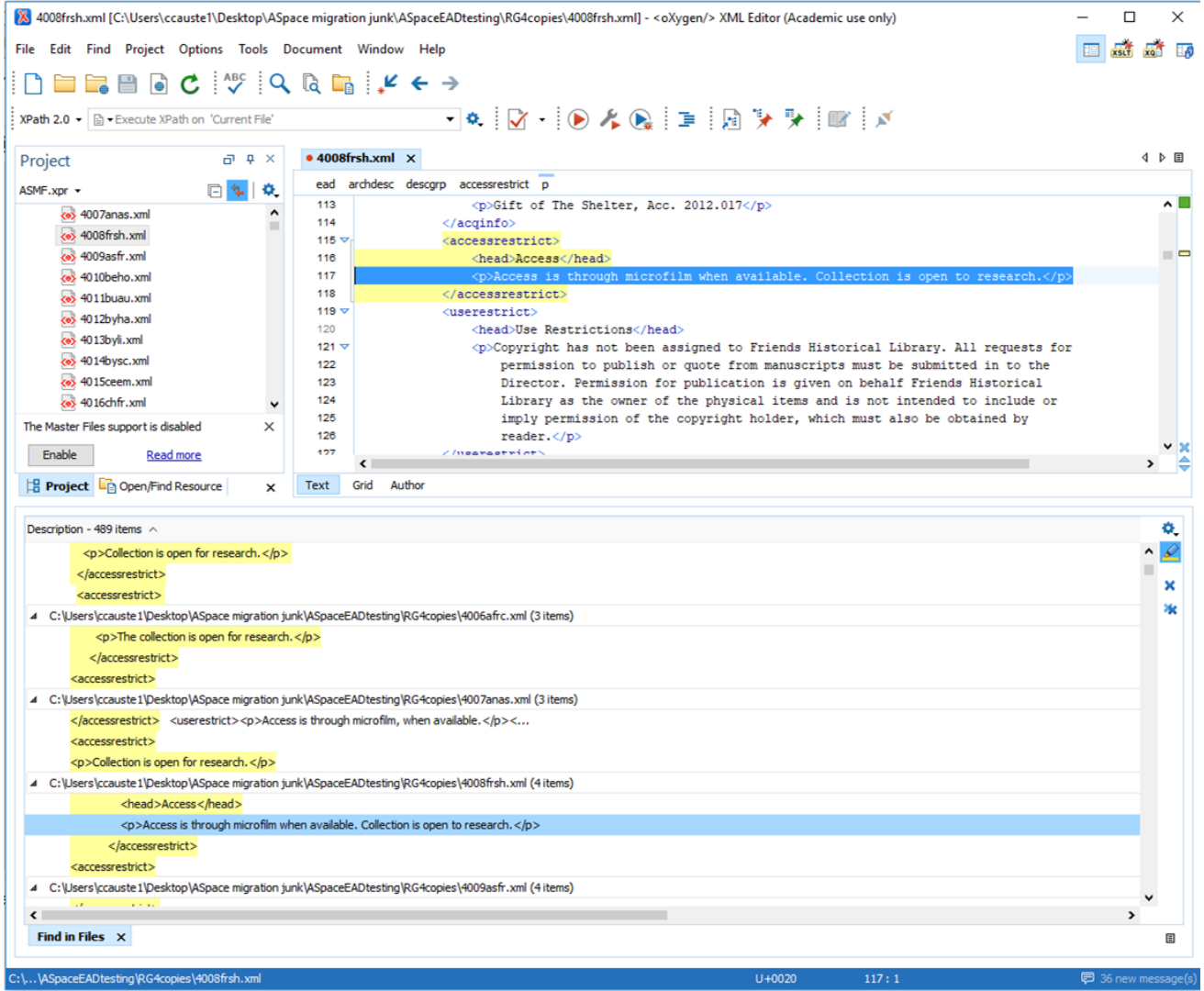
Example: verify access restrictions are consistent

Do a lot more with a tiny bit of XPath...

//tag

find tag wherever it is




4008frsh.xml [C:\Users\ccauste1\Desktop\ASpace migration junk\ASpaceEADtesting\RG4copies\4008frsh.xml] - <oXygen/> XML Editor (Academic use only)
File Edit Find Project Options Tools Document Window Help
XPath 2.0
Execute XPath on 'Current File'
Project
ASMF.xpr
4007anas.xml
4008frsh.xml
4009asfr.xml
4010beho.xml
4011buau.xml
4012byha.xml
4013byli.xml
4014bysc.xml
4015ceem.xml
4016chfr.xml
The Master Files support is disabled
Enable
Read more
Project
Open/Find Resource
4008frsh.xml
ead archdesc descgrp accessrestrict p
Gift of The Shelter, Acc. 2012.017
</acqinfo>
<accessrestrict>
<head>Access</head>
Access is through microfilm when available. Collection is open to research.
</accessrestrict>
<userestrict>
<head>Use Restrictions</head>
Copyright has not been assigned to Friends Historical Library. All requests for permission to publish or quote from manuscripts must be submitted in to the Director. Permission for publication is given on behalf Friends Historical Library as the owner of the physical items and is not intended to include or imply permission of the copyright holder, which must also be obtained by reader.
Text Grid Author
Description - 489 items
Collection is open for research.
</accessrestrict>
<accessrestrict>
C:\Users\ccauste1\Desktop\ASpace migration junk\ASpaceEADtesting\RG4copies\4006afrc.xml (3 items)
The collection is open for research.
</accessrestrict>
<accessrestrict>
C:\Users\ccauste1\Desktop\ASpace migration junk\ASpaceEADtesting\RG4copies\4007anas.xml (3 items)
</accessrestrict> <userestrict> Access is through microfilm, when available. <...
<accessrestrict>
Collection is open for research.
C:\Users\ccauste1\Desktop\ASpace migration junk\ASpaceEADtesting\RG4copies\4008frsh.xml (4 items)
<head>Access</head>
Access is through microfilm when available. Collection is open to research.
</accessrestrict>
<accessrestrict>
C:\Users\ccauste1\Desktop\ASpace migration junk\ASpaceEADtesting\RG4copies\4009asfr.xml (4 items)
Find in Files
C:\...\ASpaceEADtesting\RG4copies\4008frsh.xml
U+0020
117 : 1
36 new message(s)

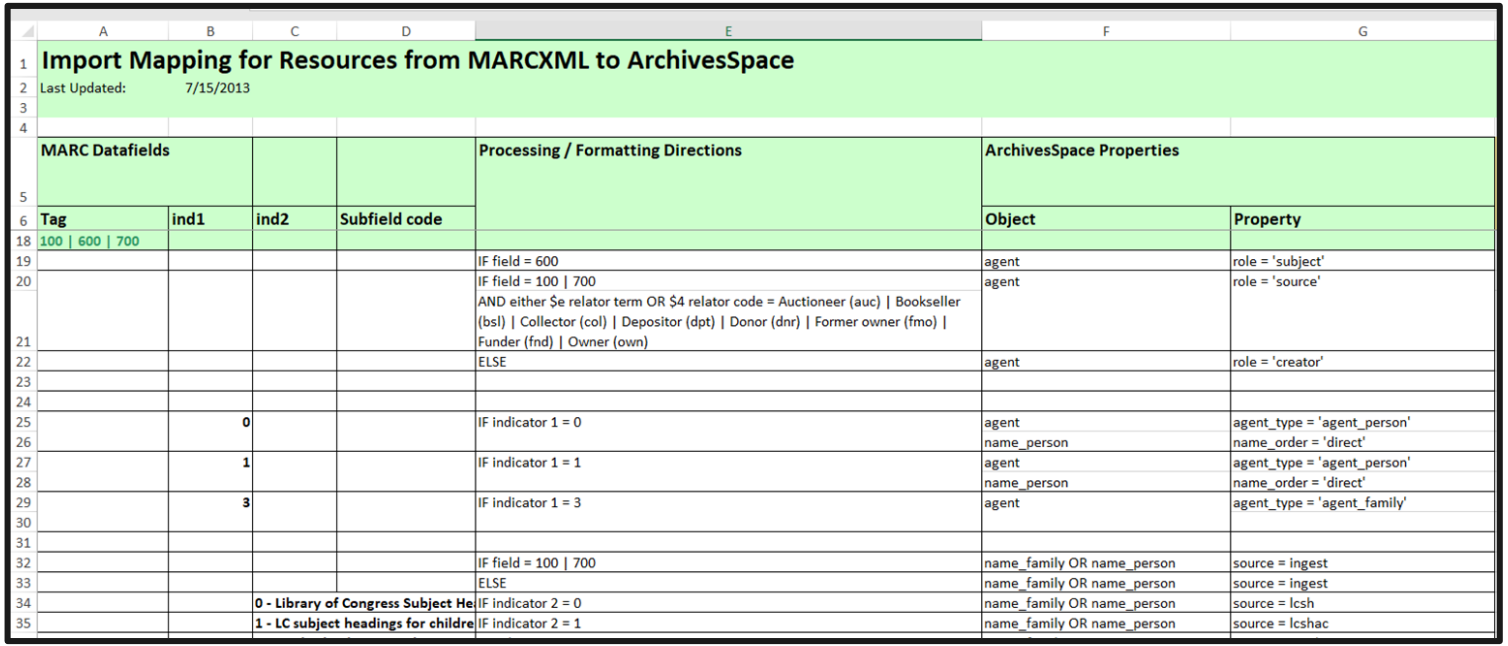
Extra Credit
Problem

I need a list of all the tags I'm using

Beware: some MARC fields are silently dropped from import

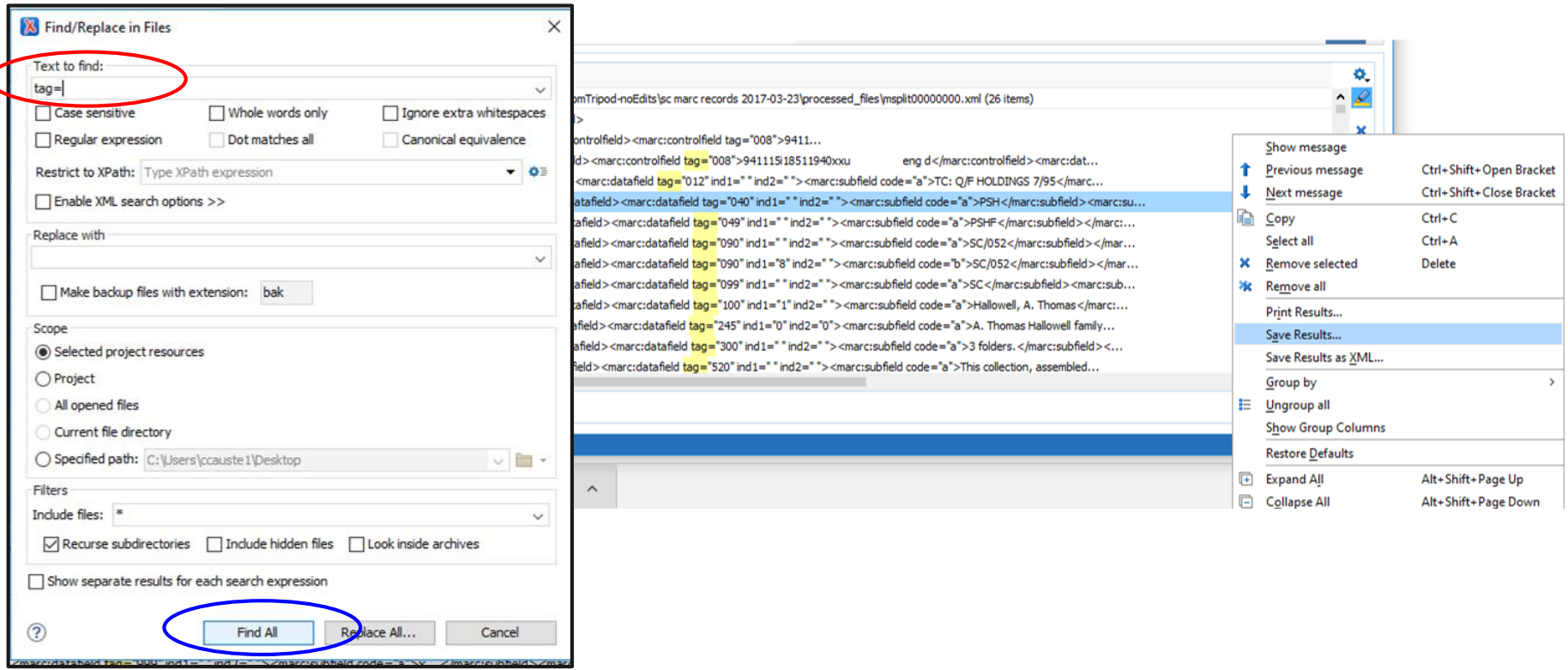
ArchivesSpace MARCXML Import / Export Map

http://archivesspace.org/wp-content/uploads/2016/08/MARCXML-Import-Export-Mapping-20130715.xlsx

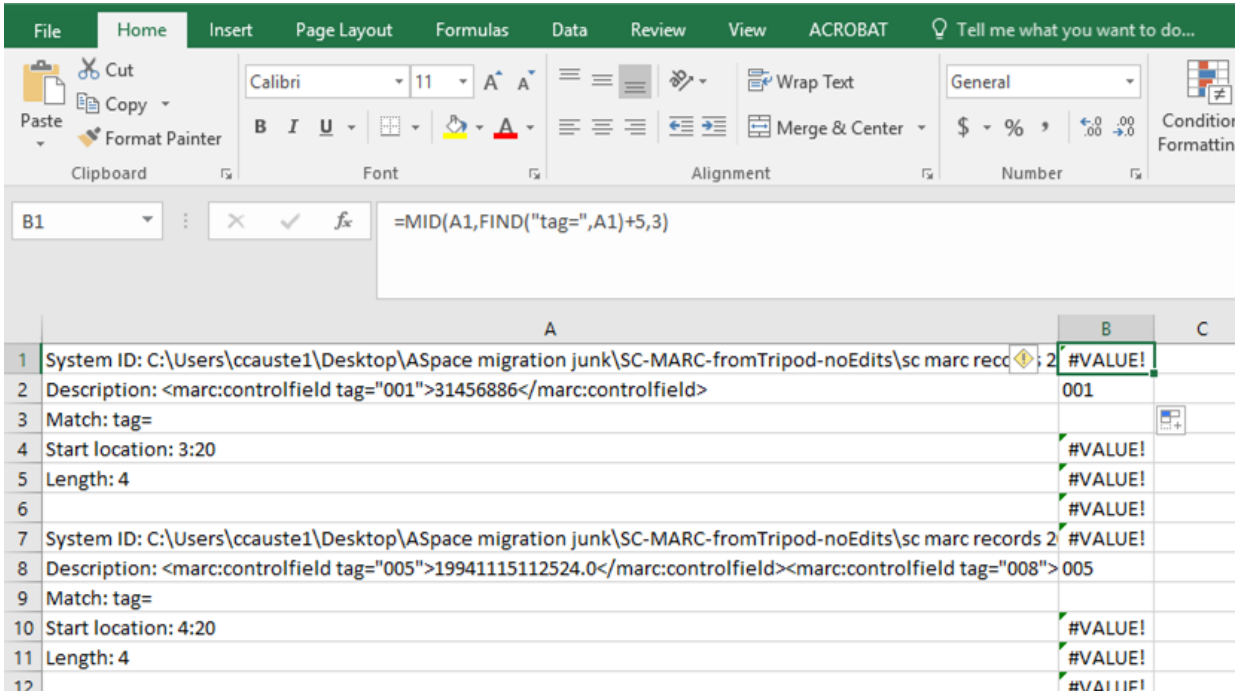


	A	B	C	D	E	F	G
1	**Import Mapping for Resources from MARCXML to ArchivesSpace**						
2	Last Updated:	7/15/2013					
3							
4							
5	**MARC Datafields**				**Processing / Formatting Directions**	**ArchivesSpace Properties**	
6	**Tag**	**ind1**	**ind2**	**Subfield code**		**Object**	**Property**
18	100 | 600 | 700						
19					IF field = 600	agent	role = 'subject'
20					IF field = 100 | 700	agent	role = 'source'
21					AND either $e relator term OR $4 relator code = Auctioneer (auc) | Bookseller (bsl) | Collector (col) | Depositor (dpt) | Donor (dnr) | Former owner (fmo) | Funder (fnd) | Owner (own)		
22					ELSE	agent	role = 'creator'
23							
24							
25		0			IF indicator 1 = 0	agent	agent_type = 'agent_person'
26						name_person	name_order = 'direct'
27		1			IF indicator 1 = 1	agent	agent_type = 'agent_person'
28						name_person	name_order = 'direct'
29		3			IF indicator 1 = 3	agent	agent_type = 'agent_family'
30							
31							
32					IF field = 100 | 700	name_family OR name_person	source = ingest
33					ELSE	name_family OR name_person	source = ingest
34			0 - Library of Congress Subject He		IF indicator 2 = 0	name_family OR name_person	source = lcsh
35			1 - LC subject headings for childre		IF indicator 2 = 1	name_family OR name_person	source = lcshac

Get all tags

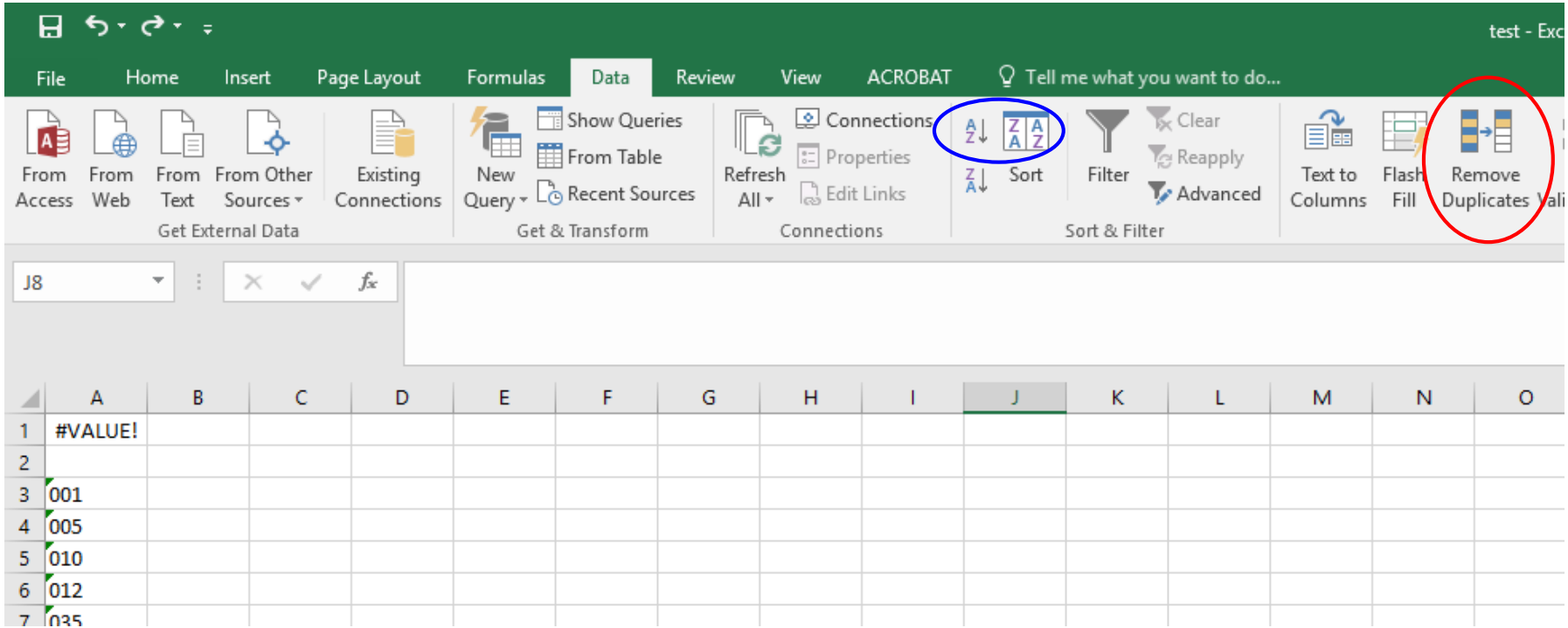

Find/Replace in Files
Text to find:
tag=
Case sensitive
Whole words only
Ignore extra whitespaces
Regular expression
Dot matches all
Canonical equivalence
Restrict to XPath:
Type XPath expression
Enable XML search options >>
Replace with
Make backup files with extension:
bak
Scope
Selected project resources
Project
All opened files
Current file directory
Specified path:
C:\Users\ccauste1\Desktop
Filters
Include files:
*
Recurse subdirectories
Include hidden files
Look inside archives
Show separate results for each search expression
Find All
Replace All...
Cancel
Show message
Previous message Ctrl+Shift+Open Bracket
Next message Ctrl+Shift+Close Bracket
Copy Ctrl+C
Select all Ctrl+A
Remove selected Delete
Remove all
Print Results...
Save Results...
Save Results as XML...
Group by
Ungroup all
Show Group Columns
Restore Defaults
Expand All Alt+Shift+Page Up
Collapse All Alt+Shift+Page Down

Clean in Excel



Formula =**MID(A1,FIND("tag=",A1)+3,LEN(A1))**

Clean in Excel



Thank you!

ccauste1@swarthmore.edu

ArchivesSpace and CSV metadata

Bria Parker
Head, Discovery and Metadata Services
University of Maryland Libraries

Where we were...

Accessions

- OpenRefine
  - Normalize date formats (yyyy-mm-dd, etc.)
  - Normalize extents (lin ft → linear_feet)
  - Apply controlled vocabularies
  - Make other metadata consistent (e.g. contacts)

OpenRefine is...

A free tool for cleaning your data.

This includes:

- normalization
- parsing/splitting
- batch transformation
- reconciliation
- data fetching

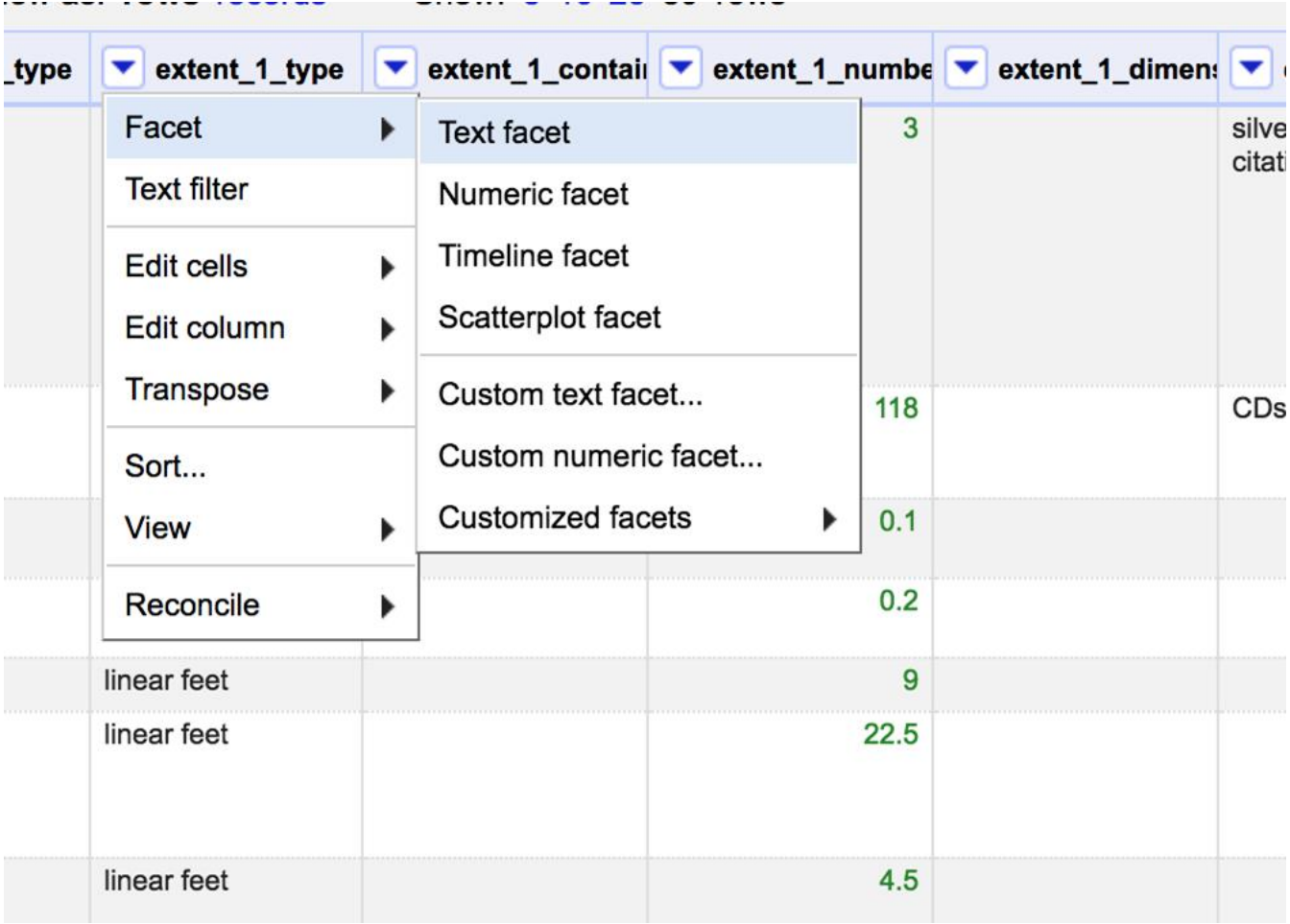
(Formerly known as Google Refine)

OpenRefine...

- Get started by going to http://openrefine.org/
- Tutorials:
  - https://blogs.library.duke.edu/bitstreams/2015/05/01/metadata-adventures/
  - http://thomaspadilla.org/dataprep/
  - http://journal.code4lib.org/articles/11179
  - http://archival-integration.blogspot.com/2015/06/normalizing-dates-with-openrefine.html

Extent normalization

lin feet
lin foot
lin ft
linear ft
lin. feet
linear feet
linear foot
lin. ft.


_type
extent_1_type
extent_1_contai
extent_1_numbe
extent_1_dimens
Facet
Text filter
Edit cells
Edit column
Transpose
Sort...
View
Reconcile
Text facet
Numeric facet
Timeline facet
Scatterplot facet
Custom text facet...
Custom numeric facet...
Customized facets
3
118
0.1
0.2
9
22.5
4.5
linear feet
linear feet
linear feet
silve
citati
CDs

extent_1_type change

33 choices Sort by: **name** count Cluster

file 10
film reels 16
floppy disks 1
folders 78
item [composed of two parts] 1
items 1813
items 2 edit include
items approximately 1
leaves 2
linear feet 1060
linear feet 1
linear feet; oversized items 1
linear inch 2
negatives (photographic) 8
photographic prints 202
poster 4

Cluster & Edit column "Column 6"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For exa
york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably r

Method key collision Keying Function fingerprint

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
5	197	• b&w (178 rows) • b&w. (12 rows) • b&w. (5 rows) • b&w : (1 rows) • b&w, (1 rows)	☐	b&w
5	1667	• 16 mm. (1509 rows) • 16 mm. + (148 rows) • 16 mm + (8 rows) • 16 mm . (1 rows) • ;16 mm. (1 rows)	☐	16 mm.
4	201	• col. (196 rows) • col. (2 rows) • col., (2 rows) • col (1 rows) Browse this cluster	☐	col.
3	38	• 16mm. (31 rows) • 16mm + (4 rows) • 16mm. + (3 rows)	☐	16mm.
2	6	• color. (5 rows) • color, (1 rows)	☐	color.

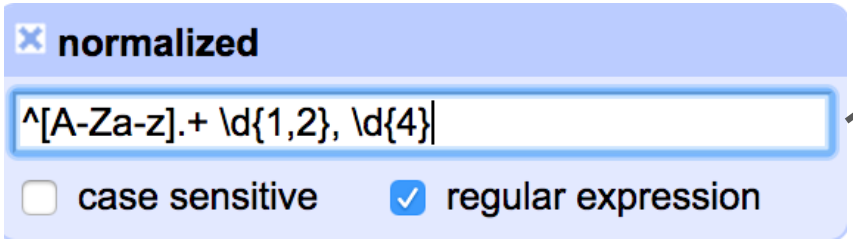
Date normalization

Borrowed and adapted scripts and process from University of Michigan Bentley Library (https://github.com/bentley-historical-library)

- Programmatically identify common date formats and insert the normal attribute in the correct date format in the EAD (for XML)
- Also provided guidance for extracting dates from EAD in .csv format and working with them in OpenRefine
- The same tools worked for more than EAD XML (hint: csv!)

Regular Expressions

- Very powerful tool for dealing with dates, or any alphanumeric data that needs to be structured
- Way of representing alpha-numeric patterns.



November 11, 1994	edit
November 28, 1991	
November 14, 1932	
January 18, 1954	
May 25, 1943	
December 15, 1947	
January 9, 1955	
January 9, 1955	
October 25, 1955	
June 13, 1989	
October 31, 2004	

Custom text transform on column normalized

Expression

Language: Google Refine Expression Language (GREL)

```
value.replace(/(\d{4})-\d{4}, (\d{4})/, '$1/$2')
```

No syntax error.

Preview | History | Starred | Help

row	value	value.replace(/(\d{4})-\d{4}, (\d{4})/, '$1/$2')
599.	1905-1938, 1947	1905/1947
834.	1976-1977, 1983	1976/1983
925.	1948-1959, 1974	1948/1974
945.	1990-1991, 1993	1990/1993
1022.	1999-2000, 2002	1999/2002
1112.	1988-1996, 2001	1988/2001

value.replace(/(\d{4})-\d{4}, (\d{4})/, '$1/$2')

More regular expressions

[A-Za-z] - any letter, upper or lower case

[A-Za-z]{6} - six letters

[A-Za-z]+ - any number of letters in a row

\d - any date

\d{4} - 4-digit date

[A-Za-z]+ \d{1,2}, \d{4} - date format like July 17, 1979

^\d{4} - begins with a 4-digit date

More regular expressions

- see regexr.com for more examples, and you can even test out expressions in their interface.

Woman at International Ladies Garment Workers Union march - http://hdl.handle.net/1903.1/32553

Challenges

- Spreadsheets not in the same format as those from The Beast
  - (plus, we had a converter we could use to get from The Beast to EAD)
- No .csv import of finding aid data function native in ArchivesSpace
  - Plugins available, but that required developer time that was not available

Steady converts CSV into EAD XML

Upload

Choose a CSV file to upload. The file must have headers which conform to the stead schema. If you have your container list data in Excel, you will need to export it as CSV before uploading it.

You can try it out with a sample CSV file.

Name*

Email*

Organization

CSV*

Choose File No file chosen

To import into ArchivesSpace and have instances created properly, you may need to make some other changes to the EAD. You can select and fill out the following options to try to do that for you. This is currently an experimental feature.

Add identifiers to containers for parent/child relationships?

archdesc/did/unitid

About

Use Cases

In most cases tools like ArchivesSpace are preferred because they centralize and normalize the creation of archival description and arrangement. We have two main use cases for maintaining container lists in spreadsheets instead of working directly in ArchivesSpace.

1. Sometimes donors provide container lists for their collections as a spreadsheet. In other cases hired appraisers will provide detailed spreadsheets which could also be used as container lists. We want to be able to edit those as little as possible to allow them to form the basis of a preliminary inventory. In this way in MPLP-style we can make the collection accessible as quickly as possible.
2. NCSU Libraries stores some of its architectural collections like drawings in off-site storage. With poor wireless coverage in some areas it is more convenient to enter container information in a spreadsheet off-line. Also this allows several people to be working on a single collection resource at one time. The spreadsheets can then be merged, exported to CSV, and transformed to EAD XML.

Code

Steady is a Ruby on Rails application and is available as open source software. See the full source code.

Contact

This little utility was developed by Jason Ronallo as part of his work at NCSU Libraries.

Email: jronallo@gmail.com

http://steady2.herokuapp.com/

Stead schema

These are the allowable values for the CSV header row. All values apply at the container level except the values that apply to the series or subseries level.

series title
series number
series dates
subseries number
subseries title
subseries dates
file id
file title
file dates
extent
physdesc
internal only
restrictions
conditions governing access
scopecontent
note1
note2

instance type
container 1 type
container 1 number
container 2 type
container 2 number
container 3 type
container 3 number
geogname
geogname_source
corpname
corpname_source
famname
famname_source
name
name_source
persname
persname_source
subject
subject_source

by owlpacino (CC-BY-ND 2.0)
https://flic.kr/p/6vwycW

Additional arguments

- Collection level
  - Resource id
  - Extent
  - Scope
  - Normalized dates

```ruby
#!/usr/bin/env ruby

$LOAD_PATH.unshift File.join(File.dirname(__FILE__), '..', 'lib')
require 'pp'
require 'stead'
require 'trollop'

opts = Trollop::options do
  banner <<-EOS
This script takes a csv file with a name in the format <eadid>_container_list.csv
and creates a stub EAD XML document.

Usage:
  csv2ead --csv /path/to/<eadid>_container_list.csv [options]

where options are:
EOS

  opt :csv, "A CSV file", :required => true, :type => String
  opt :baseurl, 'Base URL for adding on the eadid', :type => String
  opt :url, 'Full URL for this collection guide', :type => String
  opt :template, 'Specify using a different EAD XML template', :type => String
  opt :ncsu, 'Use NCSU specific template'
  opt :extra, 'Full path to a Stead::Extra file to add in other data', :type => String
  opt :output, 'Save the file by specifying the filename', :type => String
  opt :pretty, 'If --output is specified this will pretty indent the container list.'
  opt :stdout, 'Output full EAD to terminal'
  opt :idcontainers, 'Add id attributes to containers to show parent child relationships among containers in same com
  opt :unittitle, 'Add a unittitle to ead/eadheader/archdesc/did/unittitle', :type => String
  opt :unitid, 'Add a unitid to ead/eadheader/archdesc/did/unitid', :type => String
  opt :extent, 'Add an extent to ead/eadheader/archdesc/did/physdesc/extent', :type => String
  opt :unitdate, 'Add a unitdate to ead/eadheader/archdesc/did/unitdate', :type => String
  opt :begindate, 'Add a start date to ead/eadheader/archdesc/did/unitdate/@normal', :type => String
  opt :enddate, 'Add an end date to ead/eadheader/archdesc/did/unitdate/@normal', :type => String
  opt :scope, 'Add a scopecontent note to ead/archdesc/scopecontent/p', :type => String
```

```
def collectiondate_node
  @ead.xpath('//xmlns:archdesc/xmlns:did/xmlns:unitdate').first
end

def add_normal_date
  if @begindate
    collectiondate_node['normal'] = @begindate + "/" + @enddate
  end
end

def scope_node
  @ead.xpath('//xmlns:archdesc/xmlns:scopecontent/xmlns:p').first
end

def add_scope
  if @scope
    scope_node.content = @scope
  end
end

def to_ead
  @ead = template.dup
  add_eadid
  add_eadid_url
  add_archdesc_unittitle
  add_title_proper
  add_archdesc_unitid
  add_eadid
  add_archdesc_extent
  add_normal_date
  add_archdesc_unitdate
  add_scope
  @dsc = @ead.xpath('//xmlns:archdesc/xmlns:dsc')[0]
  if series?
    add_series
    add_subseries if subseries?
  end
    @component_parts.each do |cp|
    c = node(file_component_part_name(cp))
    c['level'] = 'file'
```

```
def add_series
  add_arrangement
  series = @component_parts.map do |cp|
    [cp['series number'], cp['series title'], cp['series dates']]
  end.uniq
  series.each do |ser|
    add_arrangement_item(ser)
    # create series node and add to dsc
    series_node = node('c01')
    @dsc.add_child(series_node)
    series_node['level'] = 'series'
    # create series did and add to series node
    series_did = node('did')
    series_node.add_child(series_did)
    unitid = node('unitid')
    unitid.content = ser[0]
    unittitle = node('unittitle')
    unittitle.content = ser[1]
    unitdate = node('unitdate')
    unitdate.content = ser[2]
    series_did.add_child(unitid)
    series_did.add_child(unittitle)
    series_did.add_child(unitdate)
  end
end
```

```
def add_series
  add_arrangement
  series = @component_parts.map do |cp|
    [cp['series number'], cp['series title'], cp['series start'], cp['series end'],
     cp['extent'], cp['scopecontent'], cp['persname'], cp['corpname'], cp['subject']]
  end.uniq
  series.each do |ser|
    add_arrangement_item(ser)
    # create series node and add to dsc
    series_node = node('c01')
    @dsc.add_child(series_node)
    series_node['level'] = 'series'
    series_node['id'] = 'series' + ser[0]
    # create series did and add to series node
    series_did = node('did')
    series_node.add_child(series_did)
    unitid = node('unitid')
    unitid.content = ser[0]
    unittitle = node('unittitle')
    unittitle.content = ser[1]
    series_did.add_child(unitid)
    series_did.add_child(unittitle)

    unitdate= node('unitdate')
    if ser[2]
      unitdate.content = ser[2] +'-'+ser[3]
      unitdate['normal'] = ser[2] + '/' +ser[3]
      series_did.add_child(unitdate).attr('normal')
      end
```

```
physdesc = node('physdesc')
extent= node('extent')
if ser[4]
    extent.content = ser[4]
    series_did.add_child(physdesc)
    physdesc.add_child(extent)
    end

scopecontent = node('scopecontent')
series_node.add_child(scopecontent)
if ser[5]
   serieshead = node('head')
   scopecontent.add_child(serieshead)
   serieshead.content= 'Scope and Contents'
   p = node('p')
   scopecontent.add_child(p)
   p.content = ser[5]
   end

controlaccess = node('controlaccess')
series_node.add_child(controlaccess)
if ser[6]
    ser[6].split(';').each do |c|
    persname = node('persname')
    controlaccess.add_child(persname)
    persname << c
    end

end
if ser[7]
    ser[7].split(';').each do |c|
    corpname = node('corpname')
    controlaccess.add_child(corpname)
    corpname << c
    end
end
if ser[8]
    ser[8].split(';').each do |c|
    subject = node('subject')
    controlaccess.add_child(subject)
    subject << c
    end
```

Lessons Learned

- Steady/Stead fits wide range of skill
- Reuse, reuse, reuse

Wrap-up

- OpenRefine
  - Date normalization!
  - Extent normalization!
  - Controlled vocabs!
  - Regular Expressions!
- Steady/Stead
  - template for converting .csv to EAD XML!

Questions?

Thanks!

blparker@umd.edu