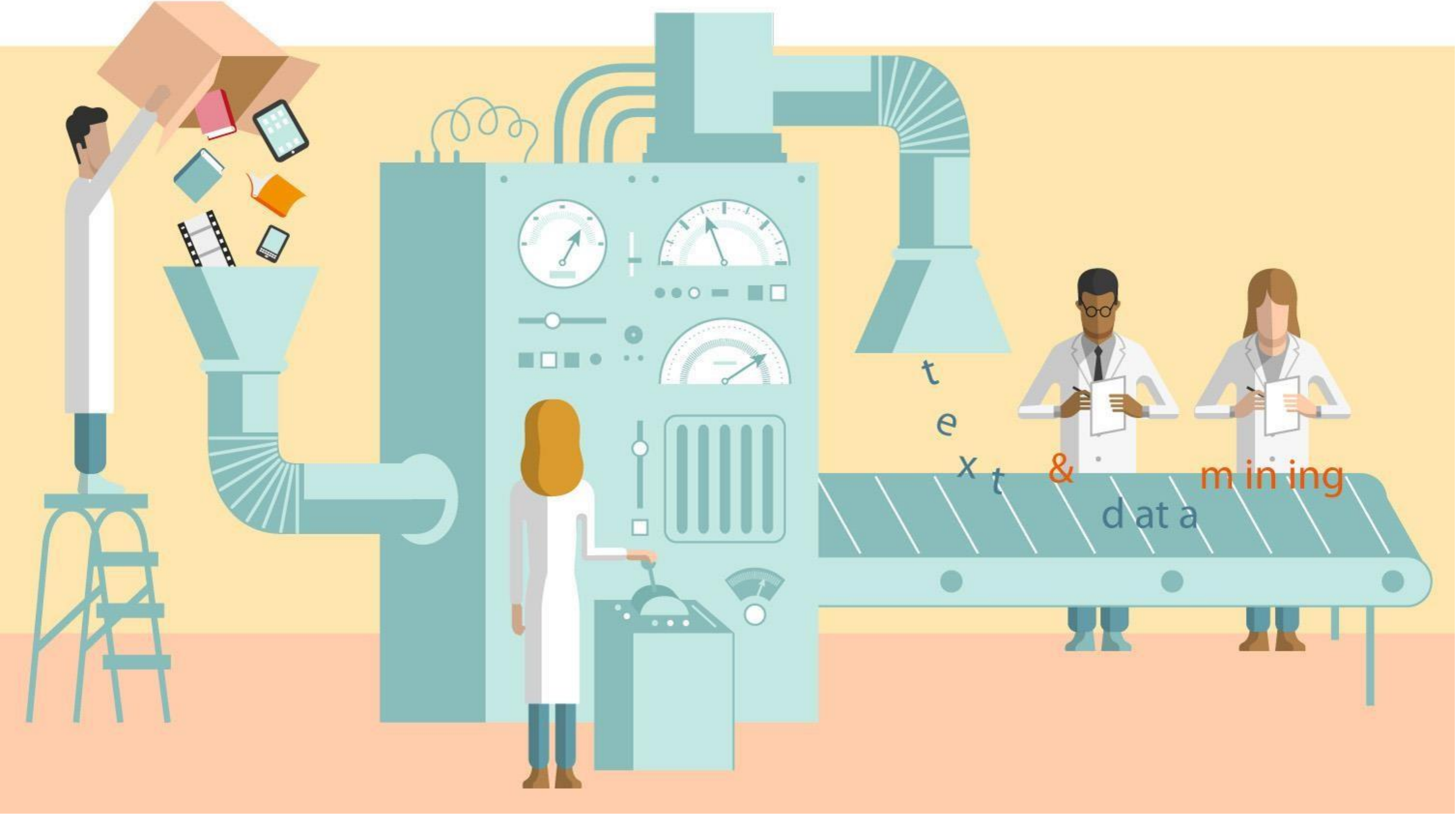
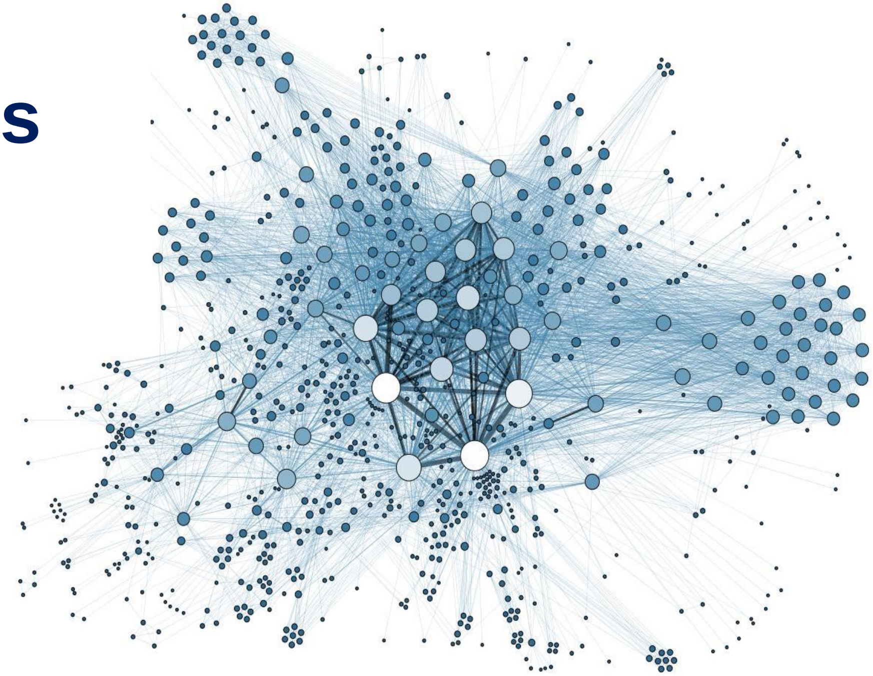
AUTOMATING CATALOGING AND METADATA



This Photo by Unknown Author is licensed under CC BY

Robin Fay 2024

Agenda



Data Management Principles

Dirty Data

Tools

- MARCEdit
- OpenRefine
- Google Sheets / Excel
- AI

ROI (Return on Investment)

Processes developed for QC workflows

- Data Templates: We use templates to standardize data (often without thinking about it)
- Validation Tools
- Batch Processing
- Our Subject Matter Expertise in Data and Our Collections
- AI*

Data management practices for Quality data

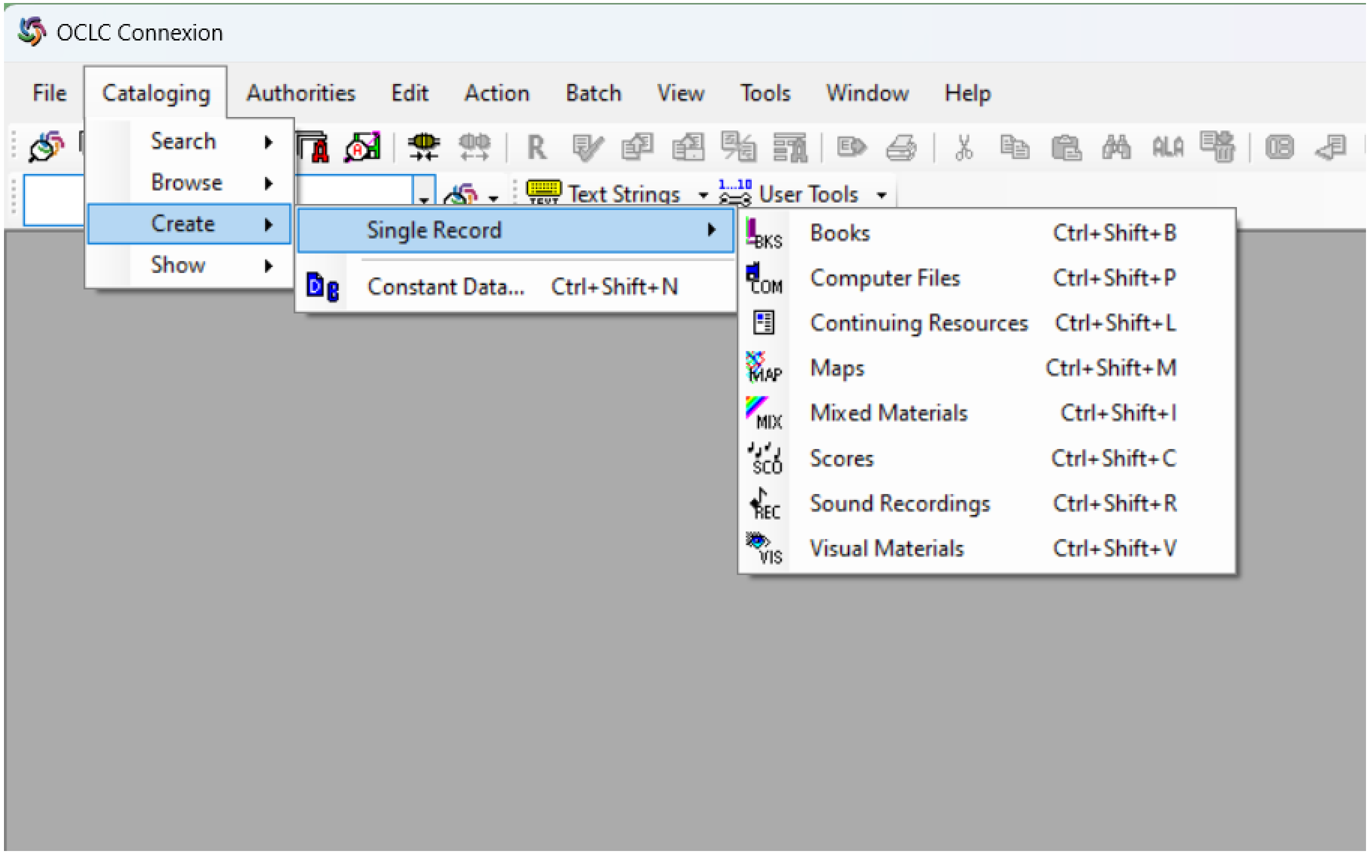
- As with all data work, we should develop a good data management practice, even if not documented*
  - Data should be complete to the extent possible
  - Data should accurately describe the resource that it represents
  - Data should not be insufficient to cause issues (editions unclear, lacking subject headings, etc.)
  - Data should not be conflicting or confusing (editions for the same title with different subject headings)

Data management practices for Quality data

- We use the following tools as part of our Quality Assurance Processes:
  - When creating or gathering data from other sources we ensure it matches our resource, is free from errors, and meets our data standards for quality.
  - We use templates, macros, batch processing, automated tools such as MARC Edit/OpenRefine and those within our systems such as “jobs” and AI to check/review our data.
  - We regularly run data analysis reports reviewing for specific patterns of issues – typos, old/outdated data, etc.

Data templates

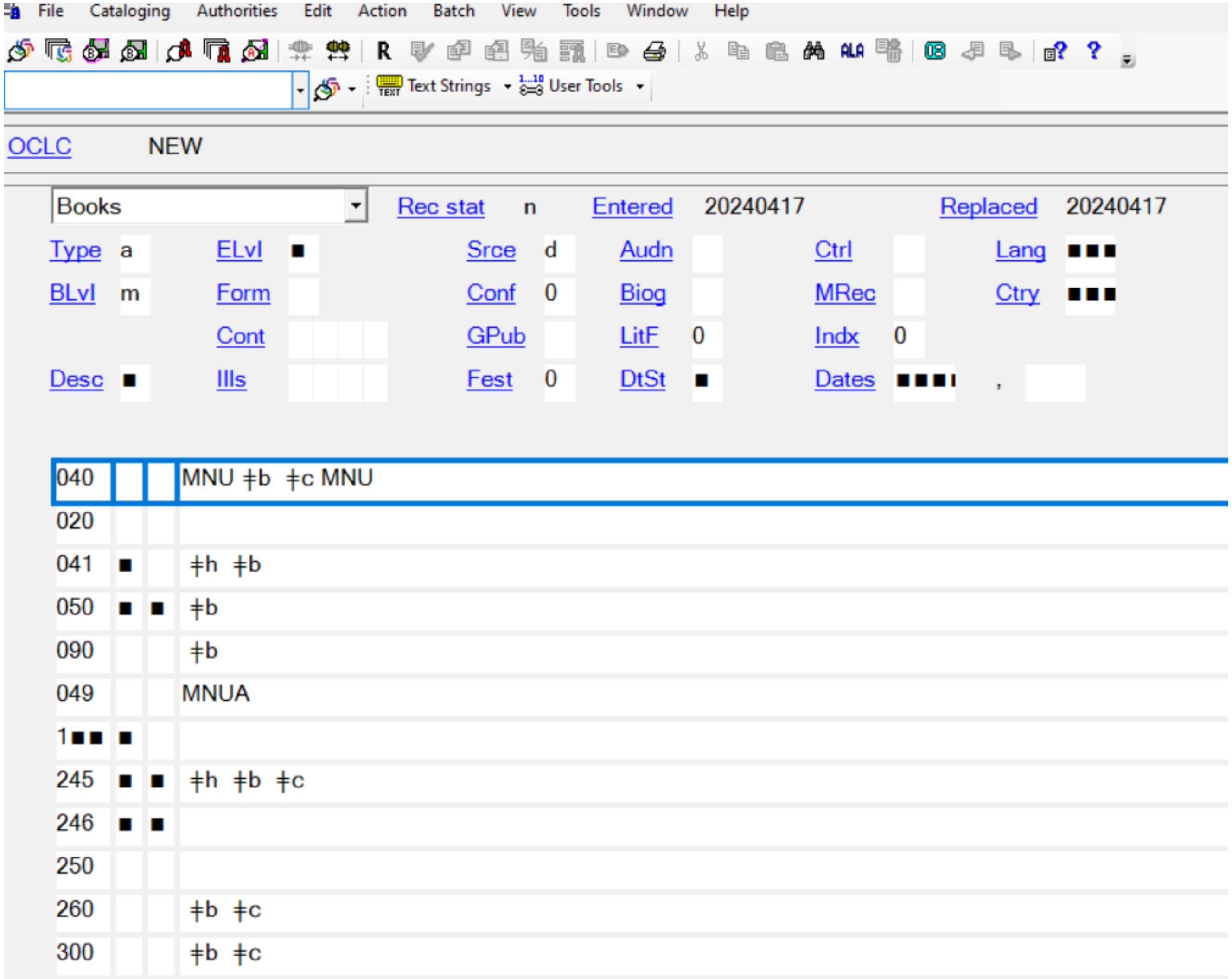
Data Templates: Prefill and format data in specific ways; standardizing how data appears and often providing data validation



Data templates

Applies constant, standard data

For example, all “Books” records have specific qualities / attributes / data fields



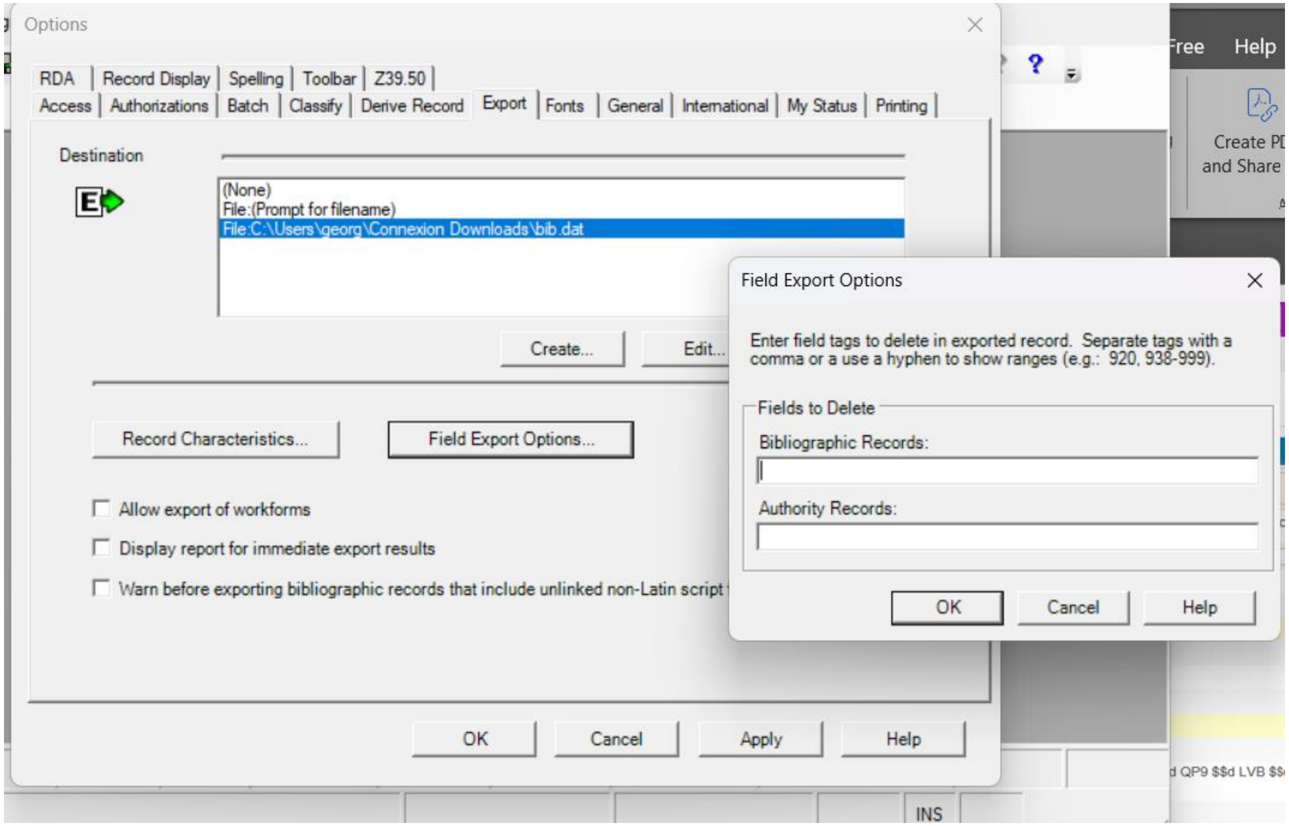
OCLC NEW

Books				Rec stat	n	Entered	20240417			Replaced	20240417
Type	a	ELvl	■	Srce	d	Audn		Ctrl		Lang	■■■
BLvl	m	Form		Conf	0	Biog		MRec		Ctry	■■■
		Cont		GPub		LitF	0	Indx	0		
Desc	■	Ills		Fest	0	DtSt	■	Dates	■■■■ ,		

040			MNU ǂb ǂc MNU
020			
041	■		ǂh ǂb
050	■	■	ǂb
090			ǂb
049			MNUA
1■■	■		
245	■	■	ǂh ǂb ǂc
246	■	■	
250			
260			ǂb ǂc
300			ǂb ǂc

Data templates

Exporting in OCLC Connexion Record Characteristics & Field Export Options both provide additional data templates for specific purposes



Destination

(None)
File:(Prompt for filename)
File:C:\Users\georg\Connexion Downloads\bib.dat

☐ Allow export of workforms

☐ Display report for immediate export results

☐ Warn before exporting bibliographic records that include unlinked non-Latin script

Field Export Options

Enter field tags to delete in exported record. Separate tags with a comma or a use a hyphen to show ranges (e.g.: 920, 938-999).

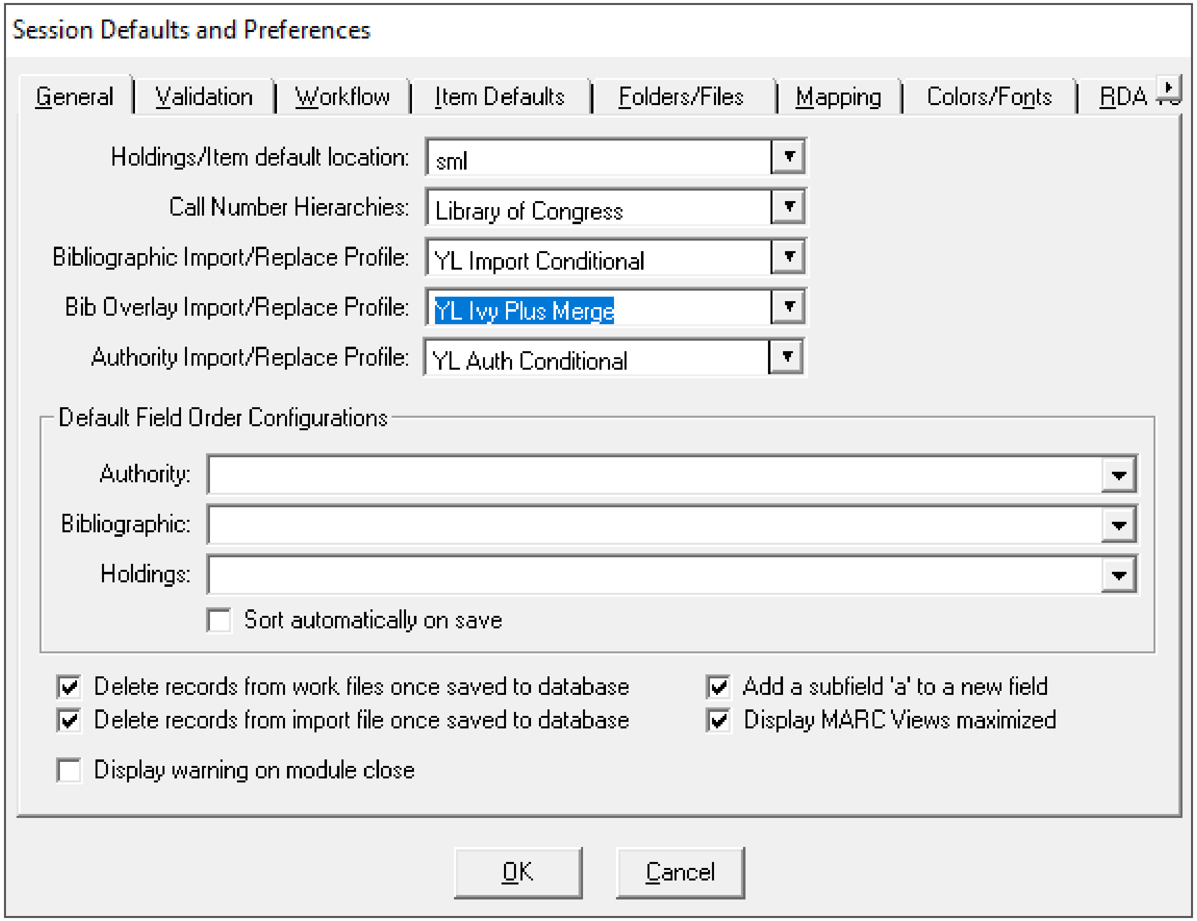
Fields to Delete

Bibliographic Records:

Authority Records:

Validation for quality assurance

May be built in tools in our systems or a separate process (MARCEdit, 3rd party, tools, QC requirements and agreements with vendors)



Session Defaults and Preferences

Holdings/Item default location: sml

Call Number Hierarchies: Library of Congress

Bibliographic Import/Replace Profile: YL Import Conditional

Bib Overlay Import/Replace Profile: YL Ivy Plus Merge

Authority Import/Replace Profile: YL Auth Conditional

Default Field Order Configurations

Authority:

Bibliographic:

Holdings:

□ Sort automatically on save

☑ Delete records from work files once saved to database

☑ Delete records from import file once saved to database

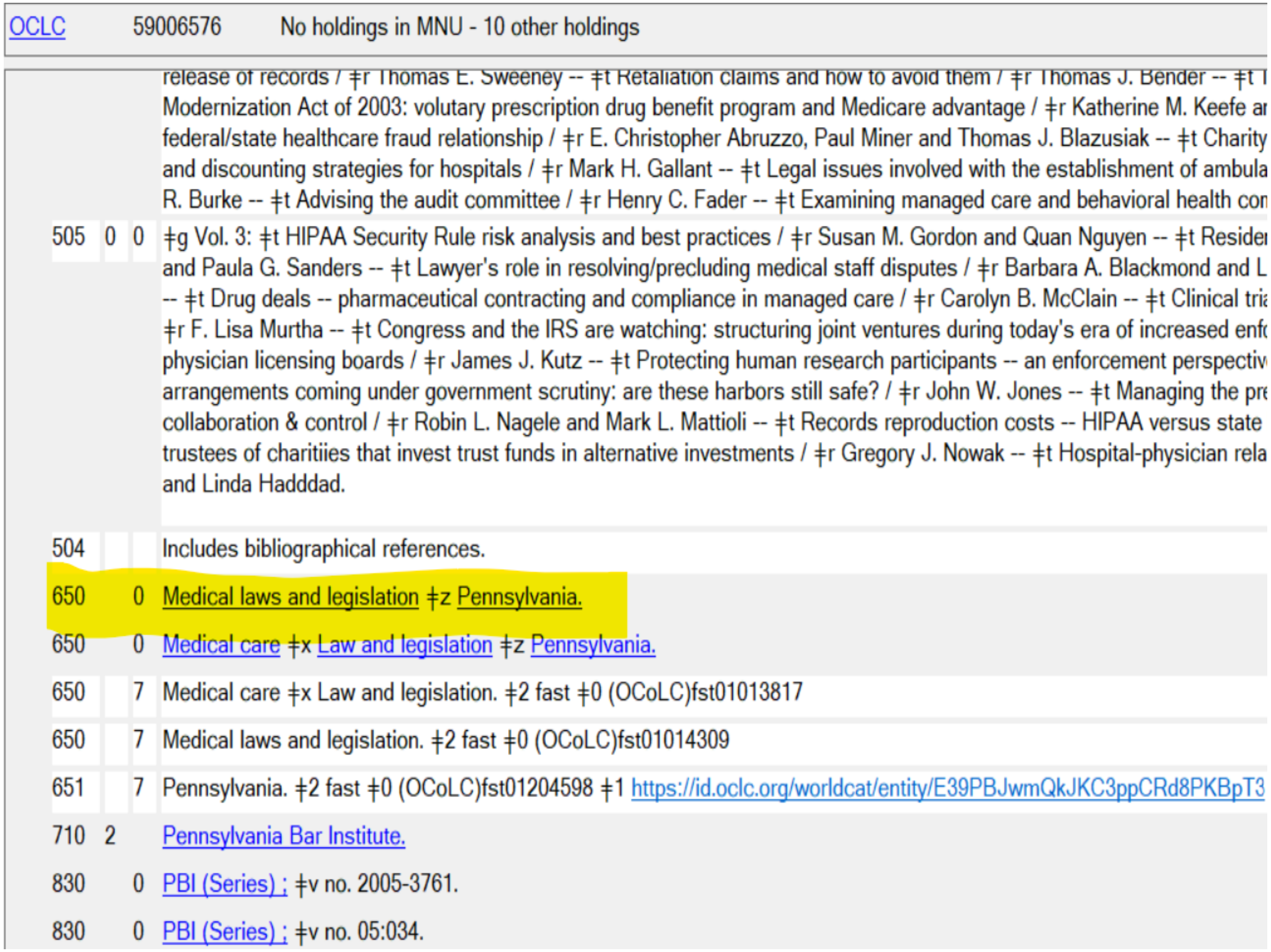
□ Display warning on module close

☑ Add a subfield 'a' to a new field

☑ Display MARC Views maximized

Validation for quality assurance

Authority work is a validation / QC process

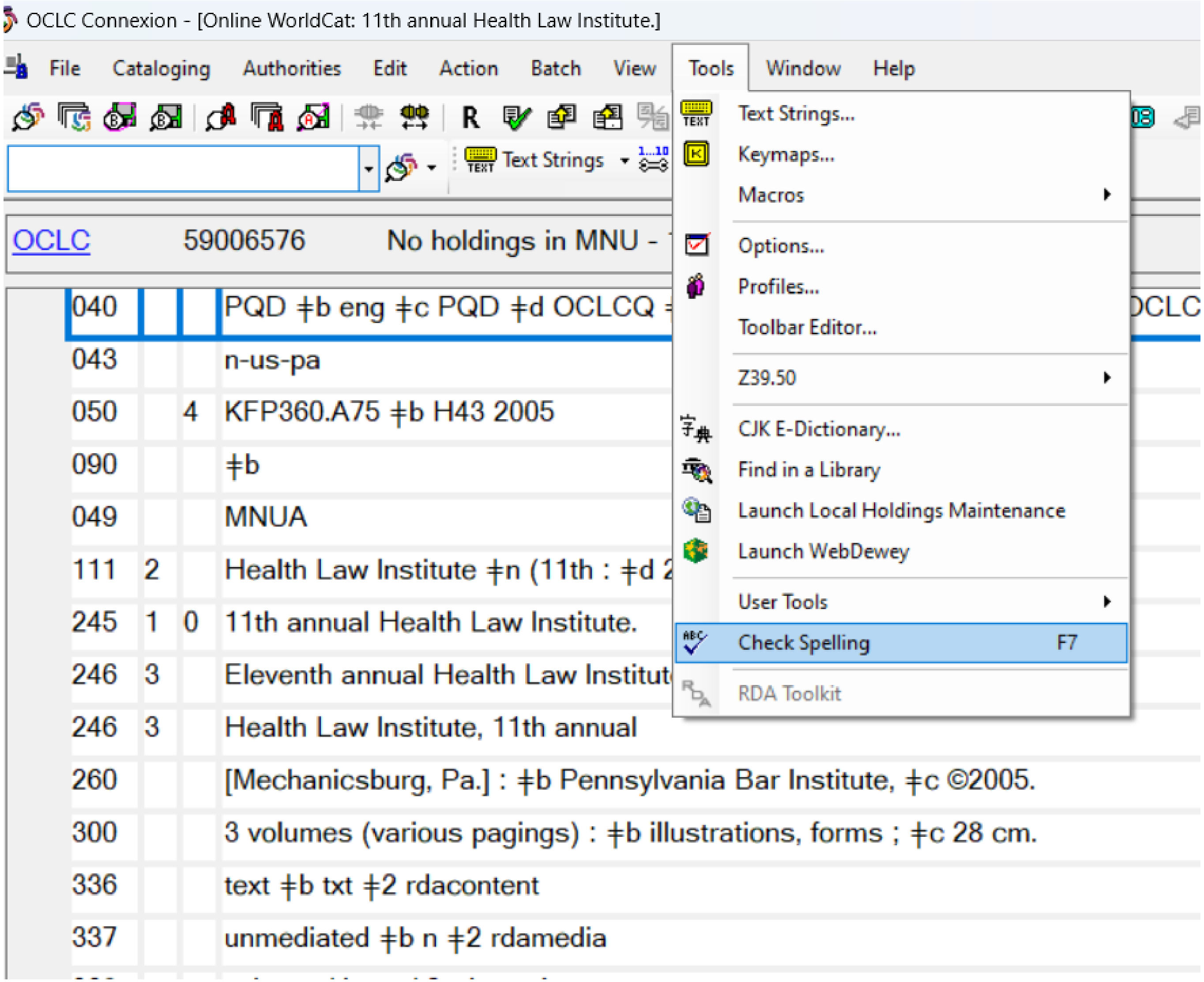


OCLC 59006576 No holdings in MNU - 10 other holdings

			release of records / ǂr Thomas E. Sweeney -- ǂt Retaliation claims and how to avoid them / ǂr Thomas J. Bender -- ǂt I Modernization Act of 2003: volutary prescription drug benefit program and Medicare advantage / ǂr Katherine M. Keefe ar federal/state healthcare fraud relationship / ǂr E. Christopher Abruzzo, Paul Miner and Thomas J. Blazusiak -- ǂt Charity and discounting strategies for hospitals / ǂr Mark H. Gallant -- ǂt Legal issues involved with the establishment of ambula R. Burke -- ǂt Advising the audit committee / ǂr Henry C. Fader -- ǂt Examining managed care and behavioral health con
505	0	0	ǂg Vol. 3: ǂt HIPAA Security Rule risk analysis and best practices / ǂr Susan M. Gordon and Quan Nguyen -- ǂt Reside and Paula G. Sanders -- ǂt Lawyer's role in resolving/precluding medical staff disputes / ǂr Barbara A. Blackmond and L -- ǂt Drug deals -- pharmaceutical contracting and compliance in managed care / ǂr Carolyn B. McClain -- ǂt Clinical tria ǂr F. Lisa Murtha -- ǂt Congress and the IRS are watching: structuring joint ventures during today's era of increased enf physician licensing boards / ǂr James J. Kutz -- ǂt Protecting human research participants -- an enforcement perspectiv arrangements coming under government scrutiny: are these harbors still safe? / ǂr John W. Jones -- ǂt Managing the pre collaboration & control / ǂr Robin L. Nagele and Mark L. Mattioli -- ǂt Records reproduction costs -- HIPAA versus state trustees of charitiies that invest trust funds in alternative investments / ǂr Gregory J. Nowak -- ǂt Hospital-physician rela and Linda Hadddad.
504			Includes bibliographical references.
650		0	Medical laws and legislation ǂz Pennsylvania.
650		0	Medical care ǂx Law and legislation ǂz Pennsylvania.
650		7	Medical care ǂx Law and legislation. ǂ2 fast ǂ0 (OCoLC)fst01013817
650		7	Medical laws and legislation. ǂ2 fast ǂ0 (OCoLC)fst01014309
651		7	Pennsylvania. ǂ2 fast ǂ0 (OCoLC)fst01204598 ǂ1 https://id.oclc.org/worldcat/entity/E39PBJwmQkJKC3ppCRd8PKBpT3
710	2		Pennsylvania Bar Institute.
830		0	PBI (Series) ; ǂv no. 2005-3761.
830		0	PBI (Series) ; ǂv no. 05:034.

Validation for quality assurance

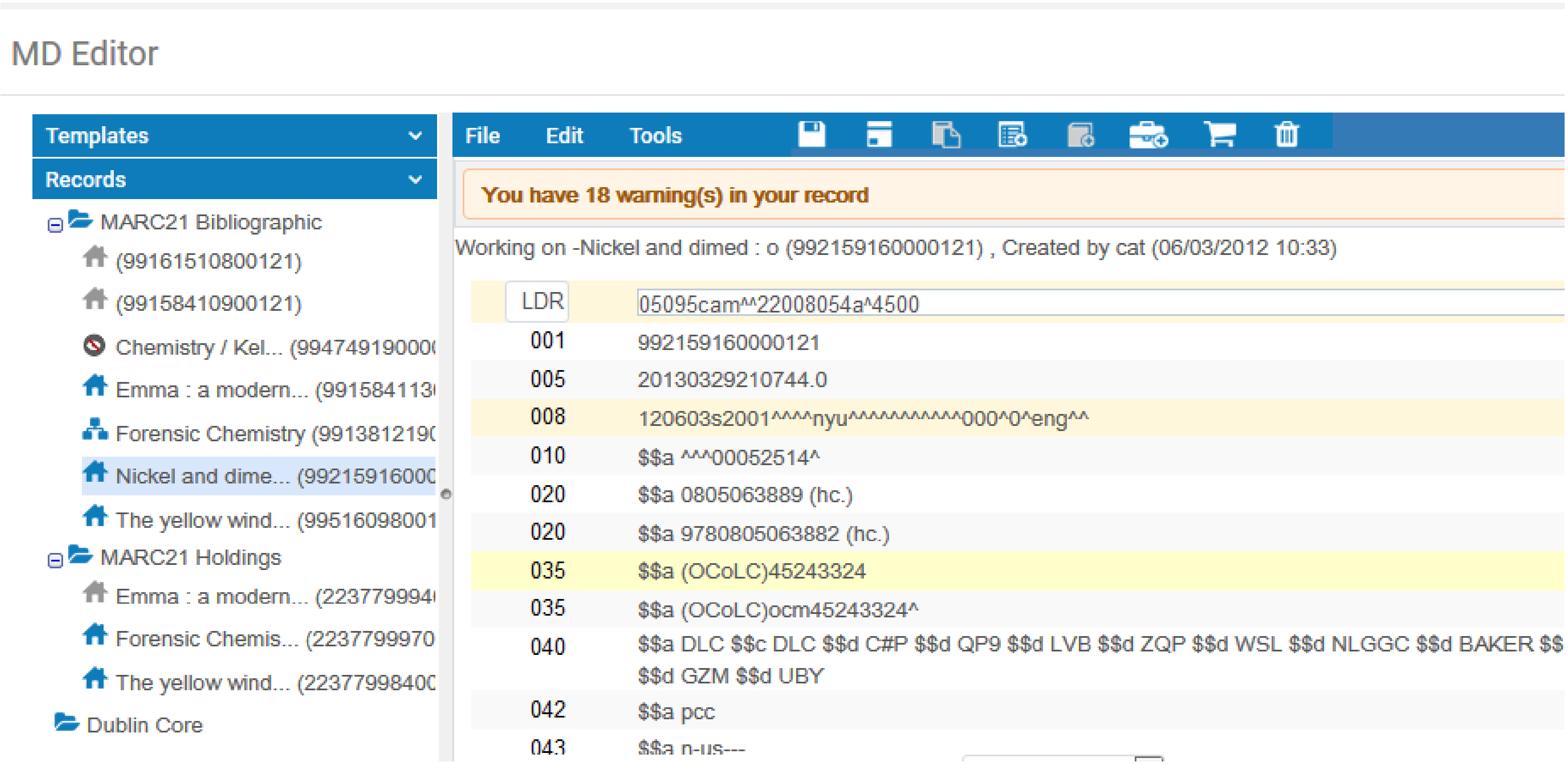
Checking spelling & grammar are as well



OCLC 59006576 No holdings in MNU -

040			PQD ǂb eng ǂc PQD ǂd OCLCQ ǂ
043			n-us-pa
050		4	KFP360.A75 ǂb H43 2005
090			ǂb
049			MNUA
111	2		Health Law Institute ǂn (11th : ǂd 2
245	1	0	11th annual Health Law Institute.
246	3		Eleventh annual Health Law Institut
246	3		Health Law Institute, 11th annual
260			[Mechanicsburg, Pa.] : ǂb Pennsylvania Bar Institute, ǂc ©2005.
300			3 volumes (various pagings) : ǂb illustrations, forms ; ǂc 28 cm.
336			text ǂb txt ǂ2 rdacontent
337			unmediated ǂb n ǂ2 rdamedia

Validation for quality assurance



MD Editor

Templates

Records

- MARC21 Bibliographic
  - (99161510800121)
  - (99158410900121)
  - Chemistry / Kel... (994749190000
  - Emma : a modern... (99158411130
  - Forensic Chemistry (99138121900
  - Nickel and dime... (99215916000
  - The yellow wind... (99516098001
- MARC21 Holdings
  - Emma : a modern... (22377999940
  - Forensic Chemis... (2237799970
  - The yellow wind... (22377998400
- Dublin Core

File Edit Tools

You have 18 warning(s) in your record

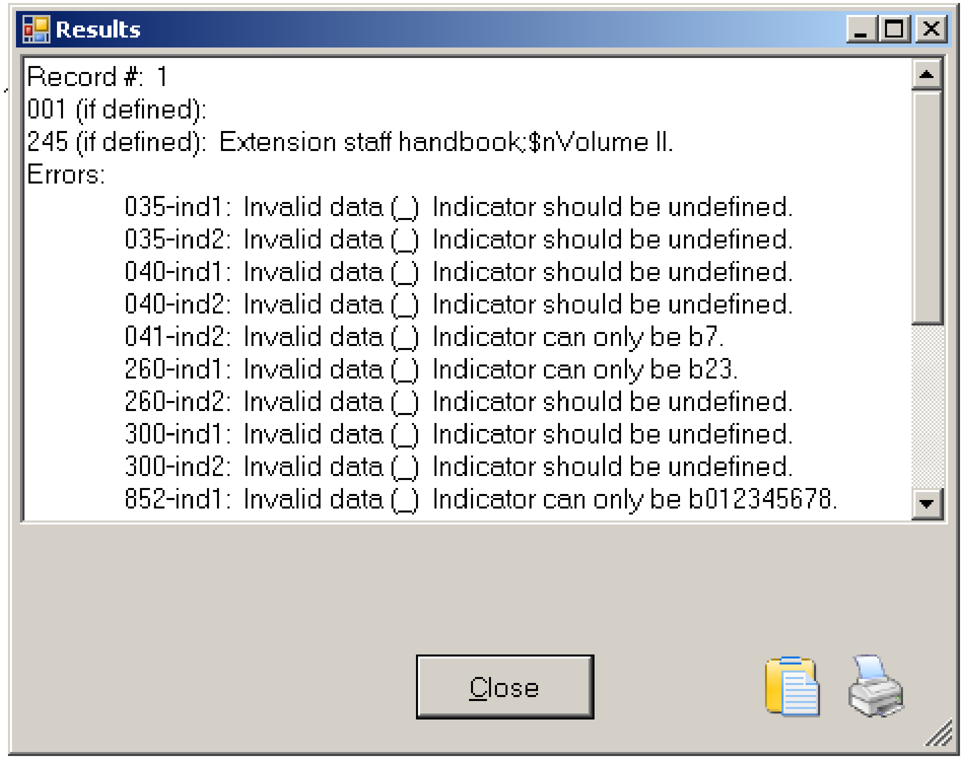
Working on -Nickel and dimed : o (992159160000121) , Created by cat (06/03/2012 10:33)

LDR	05095cam^^22008054a^4500
001	992159160000121
005	20130329210744.0
008	120603s2001^^^^nyu^^^^^^^^^^^000^0^eng^^
010	$$a ^^^00052514^
020	$$a 0805063889 (hc.)
020	$$a 9780805063882 (hc.)
035	$$a (OCoLC)45243324
035	$$a (OCoLC)ocm45243324^
040	$$a DLC $$c DLC $$d C#P $$d QP9 $$d LVB $$d ZQP $$d WSL $$d NLGGC $$d BAKER $$ $$d GZM $$d UBY
042	$$a pcc
043	$$a n-us---

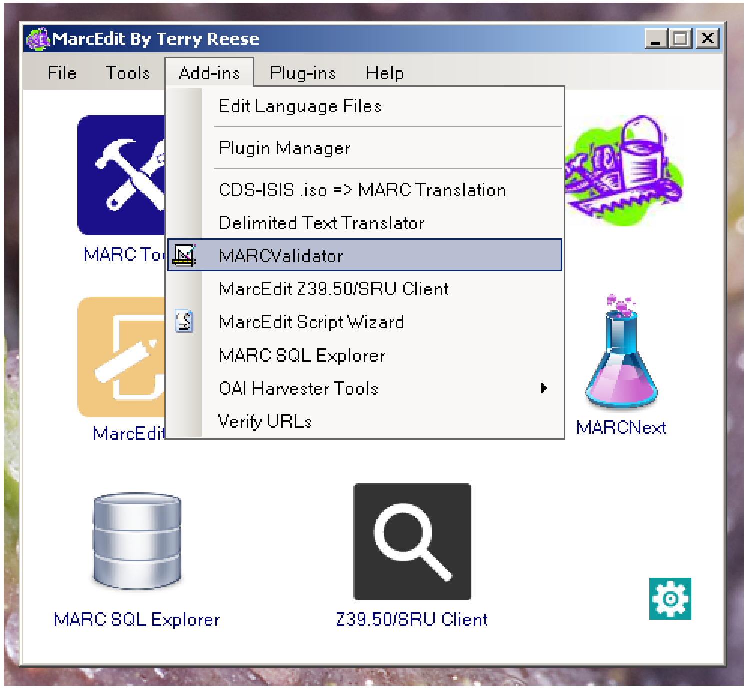
MARC Validation in Alma

Validation for quality assurance

MARC Validation Tools in MARC Edit

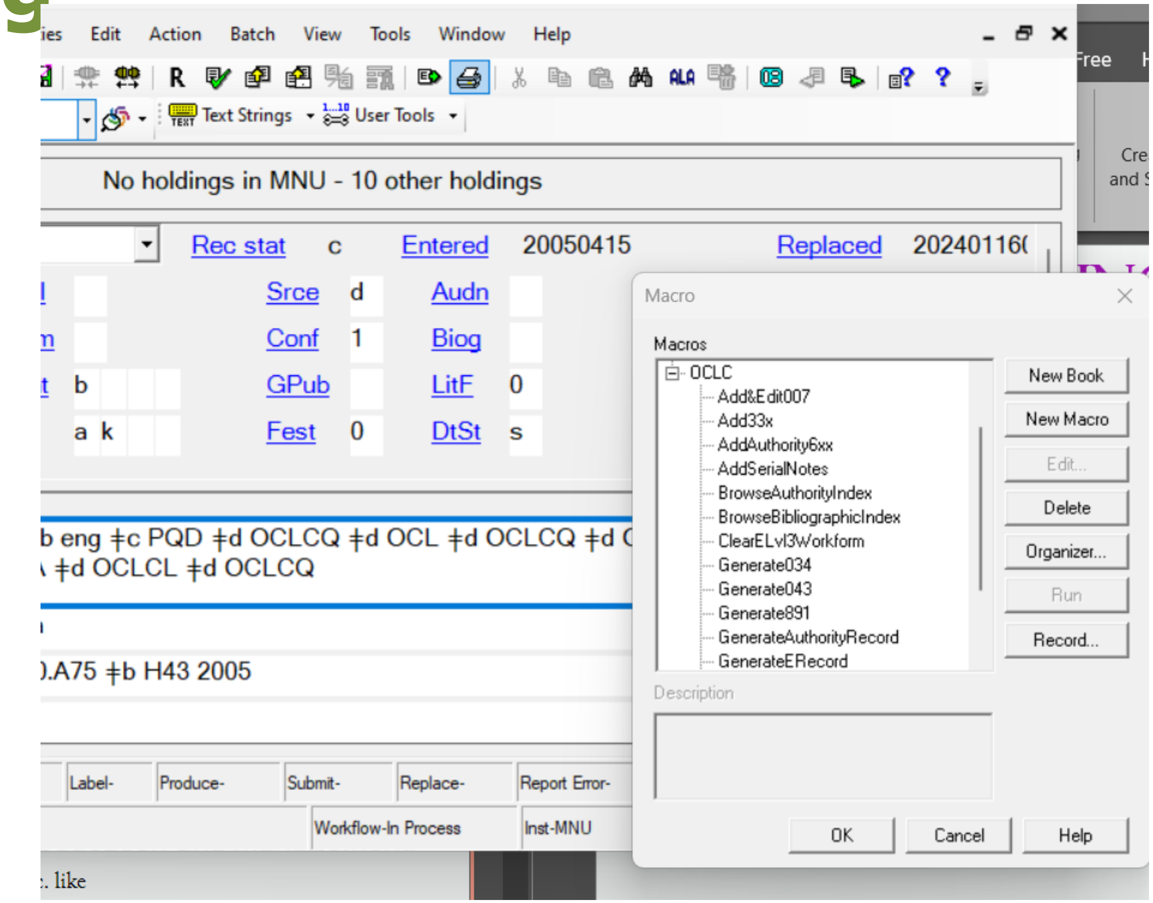


```
Record #:  1
001 (if defined):
245 (if defined):  Extension staff handbook;$nVolume II.
Errors:
          035-ind1:  Invalid data (_)  Indicator should be undefined.
          035-ind2:  Invalid data (_)  Indicator should be undefined.
          040-ind1:  Invalid data (_)  Indicator should be undefined.
          040-ind2:  Invalid data (_)  Indicator should be undefined.
          041-ind2:  Invalid data (_)  Indicator can only be b7.
          260-ind1:  Invalid data (_)  Indicator can only be b23.
          260-ind2:  Invalid data (_)  Indicator should be undefined.
          300-ind1:  Invalid data (_)  Indicator should be undefined.
          300-ind2:  Invalid data (_)  Indicator should be undefined.
          852-ind1:  Invalid data (_)  Indicator can only be b012345678.
```



Batch processing

- May be standalone (MARCEdit, OpenRefine, Excel, Scripting like Python, JSON)
- Built into a system, etc. like MACROS



No holdings in MNU - 10 other holdings

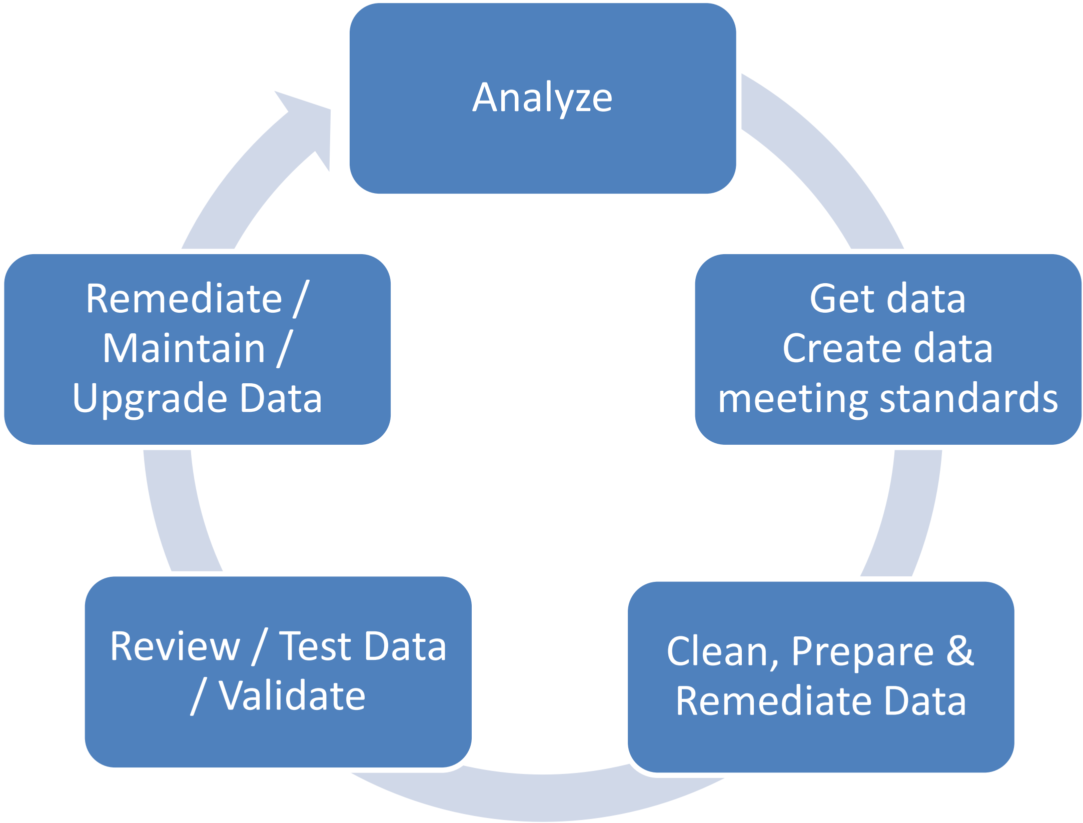
Rec stat	c	Entered	20050415	Replaced	20240116(
Srce	d	Audn			
Conf	1	Biog			
GPub		LitF	0		
Fest	0	DtSt	s		

b eng ‡c PQD ‡d OCLCQ ‡d OCL ‡d OCLCQ ‡d O
‡d OCLCL ‡d OCLCQ

.A75 ‡b H43 2005

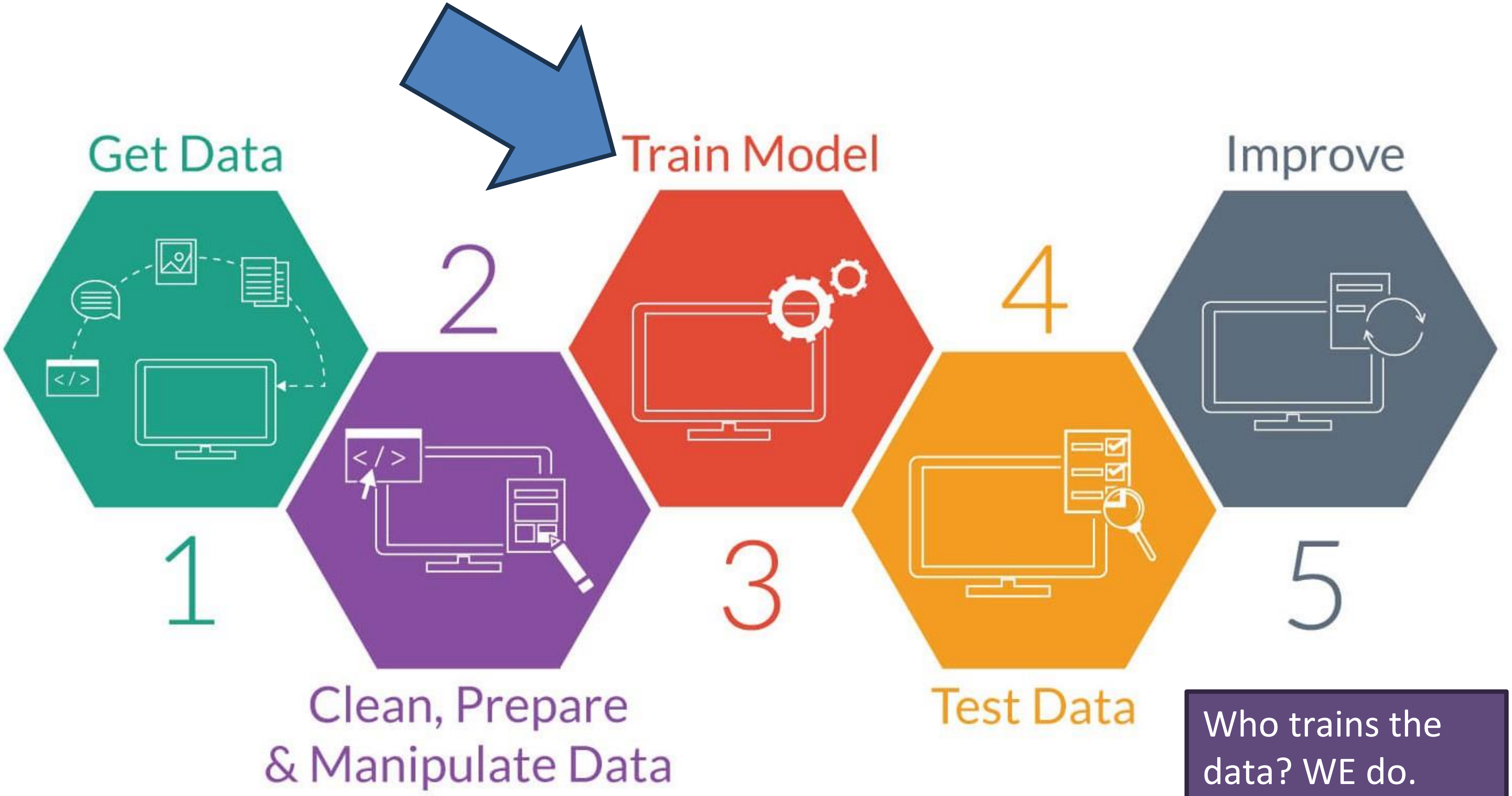
- OCLC
  - Add&Edit007
  - Add33x
  - AddAuthority6xx
  - AddSerialNotes
  - BrowseAuthorityIndex
  - BrowseBibliographicIndex
  - ClearELvl3Workform
  - Generate034
  - Generate043
  - Generate891
  - GenerateAuthorityRecord
  - GenerateERecord

QC process for data



Validation is built into our processes in many places ; as standards and workflows change (system migration, etc.) we often re-validate.

AI Data process



Who trains the data? WE do. What is training?

4 areas of metadata problems

- Missing data
- Incorrect data
- Confusing or inconsistent data
- Insufficient

We will see ALL of these problems in MARC records.

Naomi Dushay and Diane I. Hillman, “Analyzing Metadata for Effective Use and Re-use,” in Proceedings of the 2003 International Conference on Dublin Core and Metadata Applications: Supporting Communities of Discourse and Practice - Metadata Research & Applications (n.p.: Dublin Core Metadata Initative, 2003)

How did our data get so dirty?

- System error – data fragments, data junk, mapping incorrectly, migration, misunderstanding of information
- Unicode migration
- Batch importing of records without significant review (ebooks, digital resources, vendor records, etc.)
- Skeletal & incomplete records created by staff – “on the fly”, inprocess, circulation records

How did our data get so dirty?

- Changing standards – AACR records are skeletal compared to RDA an information is presented differently – 245/100
- Changing content – names and subject headings, call number classification (Controlled Vocabularies and classification evolve)
- Records in OCLC may get enhanced(or merged!) but your record is older (sometimes lesser, but not always)
- Human error – typos, mis-cataloging, wrong records (sent or used)

Where to start

Consider how maintenance will fit into your workflow

- Will you do some work daily?
- As a project?
- Will you have help?
- How will you prioritize work? (System migration in the mix?)

Good data practices

- Compatible: Data should facilitate access through being open, interoperable, actionable and readable. That means MARC records should use MARC fields correctly.
- Complete: Contain appropriate and comprehensive data reflecting the resource and its nature.
- Curated: Created and maintained over time.
- Current: Use current (RDA) practices for

data and current MARC fields.

Don’t catalog to the system limitations

- Changing the data because of how it displays to the end user/public is the WRONG approach – systems evolve and move forward.
- MARC will go away eventually (this process is underway at some libraries already).
- First, evaluate – is the data (MARC, formatting, etc.) correct? Can you change the DISPLAY will leaving the data intact?
- “Raw” data really should never been shown to our end users, except for advanced users, other librarians, etc.

Don’t catalog to the system limitations

- Instead, CHANGE the DISPLAY. Change the view, not the data.

Why?

- Data is actionable with some data specifically for machine instruction (such as linked data in $0 or $1)
- We are starting to work towards moving beyond MARC. That means all of our data needs enrichment to prepare for that.
- We need to share more data in the future. Consider standardization.

Good data takes work

- Consider interoperability. When evaluating new software consider its ability to export/import data – can you data easily migrate to the new system?
- Analyze data regularly, especially after upgrades or migrations. This may include link checking, reports to identify anomalies (even typos!), and more.
- Remediate and enhance data. Correct missing, outdated, or incorrect data.
- Data standardization. Authority work/Identity Management, standardization of dates, naming conventions on the web, site maps, & so much more.

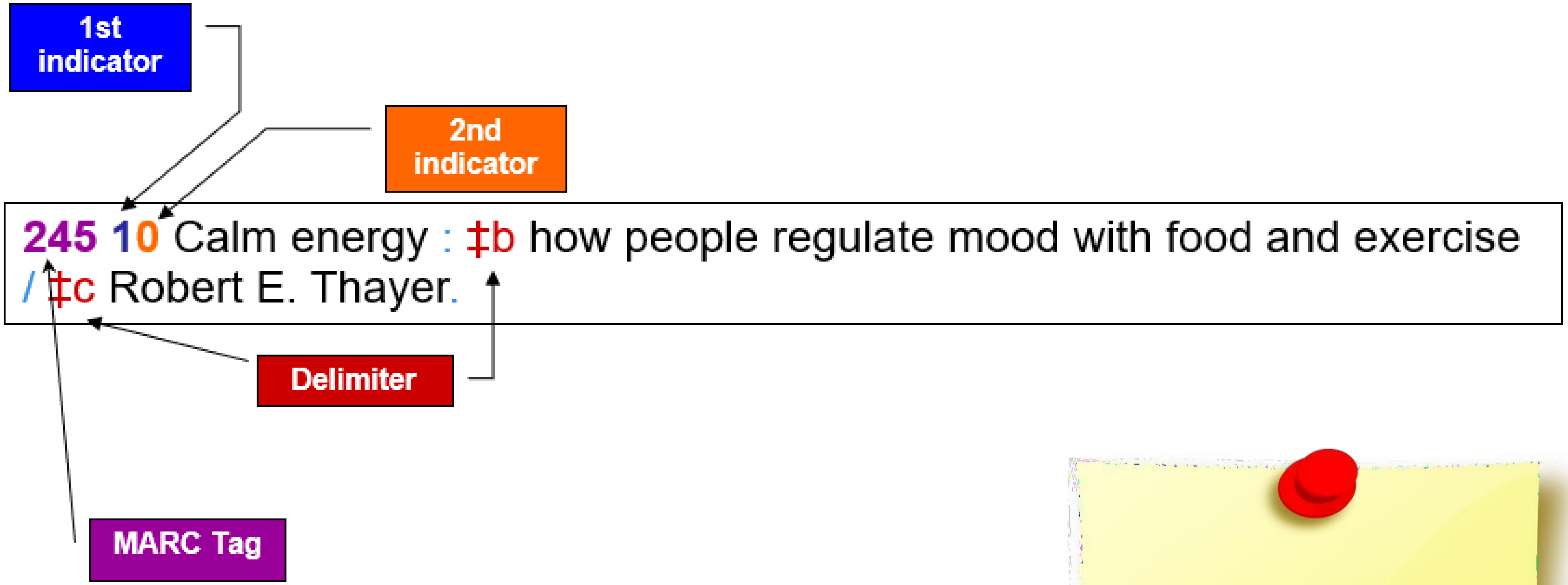
Tips for finding the “bad” data

- Look for anomalies in patterns – search for records without specific information – data missing mandatory fields, data with outdate information, typos, etc.
- Search for known potential issues things with the exact same title, things with relationships, changes in practice; etc.
- Inventory.
- Create a process to report errors or data cleanup projects.

Tips for finding the “bad” data

- Look for anomalies in patterns – search for records without specific information – data missing mandatory fields, data with outdate information, typos, etc.
- Search for known potential issues things with the exact same title, things with relationships, changes in practice; etc.
- Inventory.
- Create a process to report errors or data cleanup projects.

Quick MARC review



1st indicator

2nd indicator

245 10 Calm energy : ‡b how people regulate mood with food and exercise / ‡c Robert E. Thayer.

Delimiter

MARC Tag

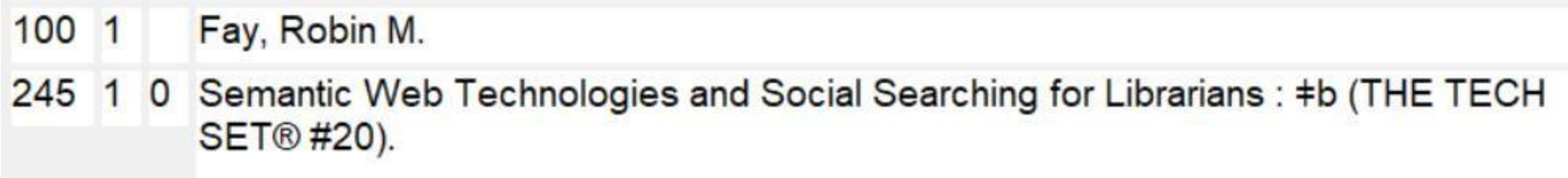
Tags represent textual names

They're divided by hundreds: e.g., 100, etc.

We also see some ISBD punctuation here. RDA is meant to be neutral of ISBD and MARC, which does impact how we enter some data. So much potential for invalid MARC….

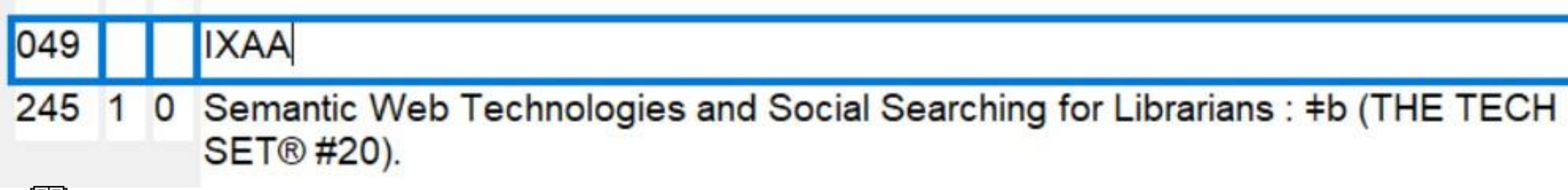
Miscoded indicators in 245s – First indicator

- 245 First indicator = 1 = there must be a 1XX field
- If you have MARC validation tools, you may catch them that way.
- MARCEdit also has MARC validation tools!
- You may be able to run reports or a search to find these.
- Serials often do not have a 1XX field, so that maybe a format to check.



100	1		Fay, Robin M.
245	1	0	Semantic Web Technologies and Social Searching for Librarians : ǂb (THE TECH SET® #20).

Wrong:



049			IXAA
245	1	0	Semantic Web Technologies and Social Searching for Librarians : ǂb (THE TECH SET® #20).

Miscoded indicators in 245s – second indicator

- 245 Second indicator – this field impacts title indexing and searching
- If you have MARC validation tools, you may catch them that way.
- You may be able to run reports or a search to find these. Search by incorrect indicator and phrase (e.g., The + second indicator=0) or search for title phrases “the”.
- Records which are more likely to have incorrect indicators – older records, brief order records, vendor records, records with diacritics in initial articles, and skeletal records.
- Note: Some systems are smart enough to return searches correctly regardless of indicator.

245 1 0 The black cat <- wrong ; should be 245 14 The black cat ; T+h+e+space = 4; start title search /indexing at black

Example: Modernizing or RDAizing your catalog – 260s to 264

While 260 is still valid (for now), if you would like to convert this field to the newer form, you have 2 choices

- Pull monographic pieces and do complete 264s for copyright, production,manufacturer, and distributor.
- Convert to 264_1 (Second indicator 1) for monographic records.

Work can be done manually, via a script or tool of your choice, or via MARCEdit.

RDAizing your catalog – Modernizing Your Data – GMD/336, 337, 338

- Use new fields only: OCLC, LC, and PCC – use new fields only do not use GMD. (Most newer ILS will as well!)

Use MARCEdit to add the new fields to older records and strip out GMDs. If new fields do not index in ILS, work with Systems/ILS to fix.

- Add GMD to records using MARCEdit without removing the 336-338. Retrospectively add the new fields in.
- We’ll look at MARCEdit in just a few.

Changes in authorities – names & subjects

- How to find new authorities – check LC's list of new headings. https://www.loc.gov/aba/cataloging/subject/weeklylists/
- Search manually.
- If your system will validate headings, compile a list.
- Use MARCedit to validate headings in a group of records (batch) – it can only fix names not subjects, but it will give you a list to review. (Then use MARCEdit's batch tools – edit field, add/delete field, etc. to do work in batch)
- Outsource.

Modernizing data projects & systems migrations

- GMD $h in a 245 is obsolete; replaced by 336, 337, 338
- 440 → Now 490 1 / 830 0
- Local fields → Any 9XX (900-999) or field ending X90 are local.
- Editions, serials, or translations: look at 7XX linking fields.
- Multiple formats on a single record, e.g., an ebook on a print record; eserials and print serials combined on one record
- Multiple versions of the same title from different vendors
- Outdated subjects
- Additional genres
- Missing 520 or 505 notes
- Literature missing subject headings
- Abbreviations in 300 fields v., p. etc
- Things with numbering or relationships to each other

More Things to look for

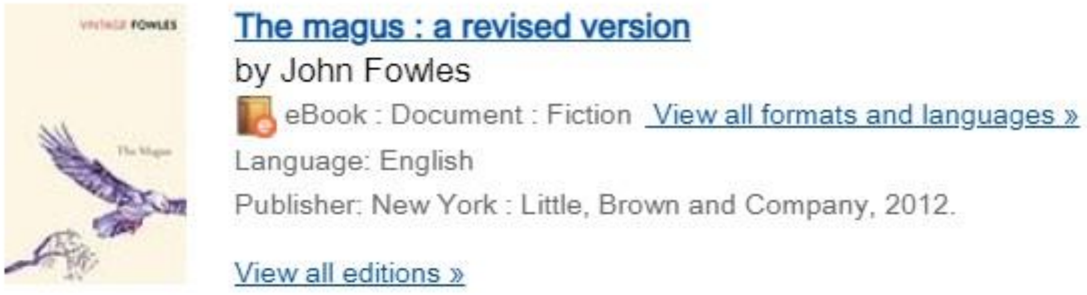
- Copies
- Editions
- Translations
- Things with Numbering
- Literature by the same author
- Serials, especially title changes

Editions

Different editions of the same work should have

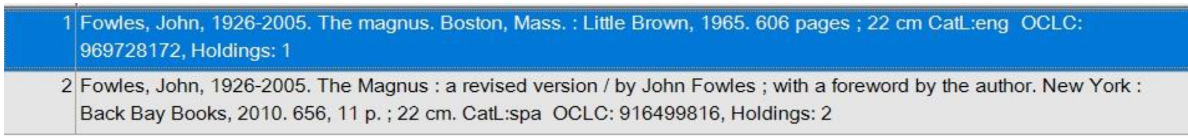
- The same call number with the edition separated by date.
- The same subject headings.

- Search for words like “revised”, “updated”, “revision”, “edition”, “version”
- Goal: Original first, editions next, translations beyond that.



The magus : a revised version
by John Fowles
eBook : Document : Fiction View all formats and languages »
Language: English
Publisher: New York : Little, Brown and Company, 2012.

View all editions »



1 Fowles, John, 1926-2005. The magnus. Boston, Mass. : Little Brown, 1965. 606 pages ; 22 cm CatL:eng OCLC: 969728172, Holdings: 1

2 Fowles, John, 1926-2005. The Magnus : a revised version / by John Fowles ; with a foreword by the author. New York : Back Bay Books, 2010. 656, 11 p. ; 22 cm. CatL:spa OCLC: 916499816, Holdings: 2

Primarily applies to monographs; serials calls different editions of a serial, “title changes”

Printings & facsimiles get a little more tricky, so you may need to do more investigation if you see the word “printing” or “reprinted”

Things that are More (or have numbering)

Things with numbering are often problematic.

Questions to consider – do you have more of the related things?

Do your users want to read/listen/view the things in a specific order (e.g., reading order?)

What does the data tell you?

Problemsolving Tools

- ✓ OCLC / WorldCat – verify information, Data Sync cleanup (first is typically free)
- ✓ LC Authorities – verify information especially subjects and names
- ✓ Your ILS – data reports, batch processing
- ✓ MARCEdit – for MARC but can work with non MARC data converted into MARCEdit files (Dublin Core, etc.) – check and update name authorities, check subject headings, validate MARC; enrich data with linked data or other data
- ✓ OpenRefine – works with CSV, XML – good for nonMARC metadata, data analysis, etc.
- ✓ Google Sheets & Excel -- works with CSV, XML – good for nonMARC metadata, data analysis (OpenRefine is more robust)

Problemsolving Tools – OCLC

OCLC can sometimes resolve questions such as:

does the library still own a copy of the title? (If no holdings in OCLC, did you ever own it? Was it withdrawn?)

questions about treatment (is it a monographic set? Serial?)

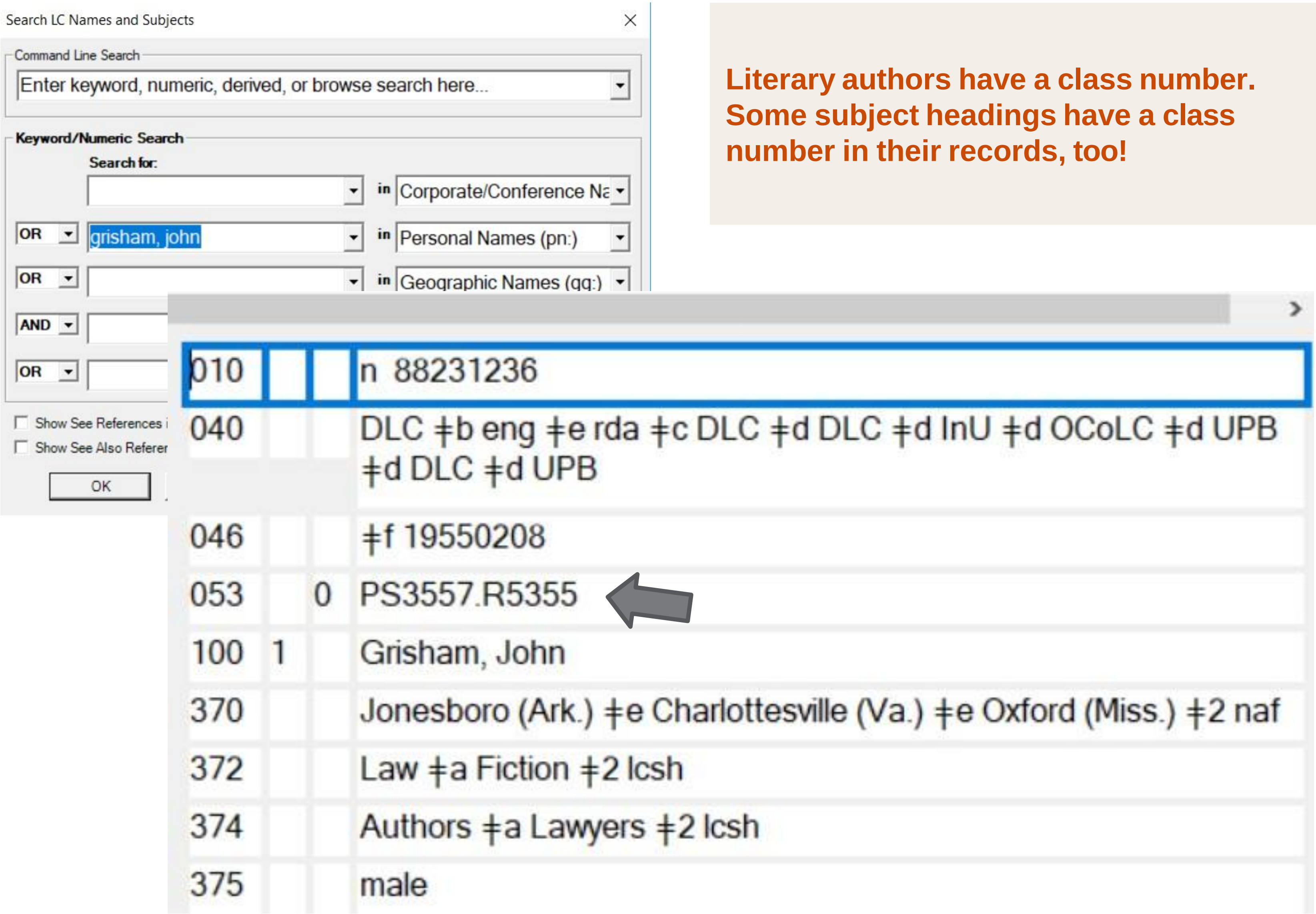
does it need a recataloging? Is the call number correct?

Problemsolving Tools – Authority records – SARs

Series Authority Records (SARs) are available via LC (authorities.loc.gov ← Search by title) and OCLC. They can sometimes resolve questions about treatment.

Use the series Cheatsheet to help decode these.

Name authority record



Search LC Names and Subjects

Command Line Search

Enter keyword, numeric, derived, or browse search here...

Keyword/Numeric Search

Search for:

in Corporate/Conference Na

OR grisham, john in Personal Names (pn:)

OR in Geographic Names (gg:)

AND

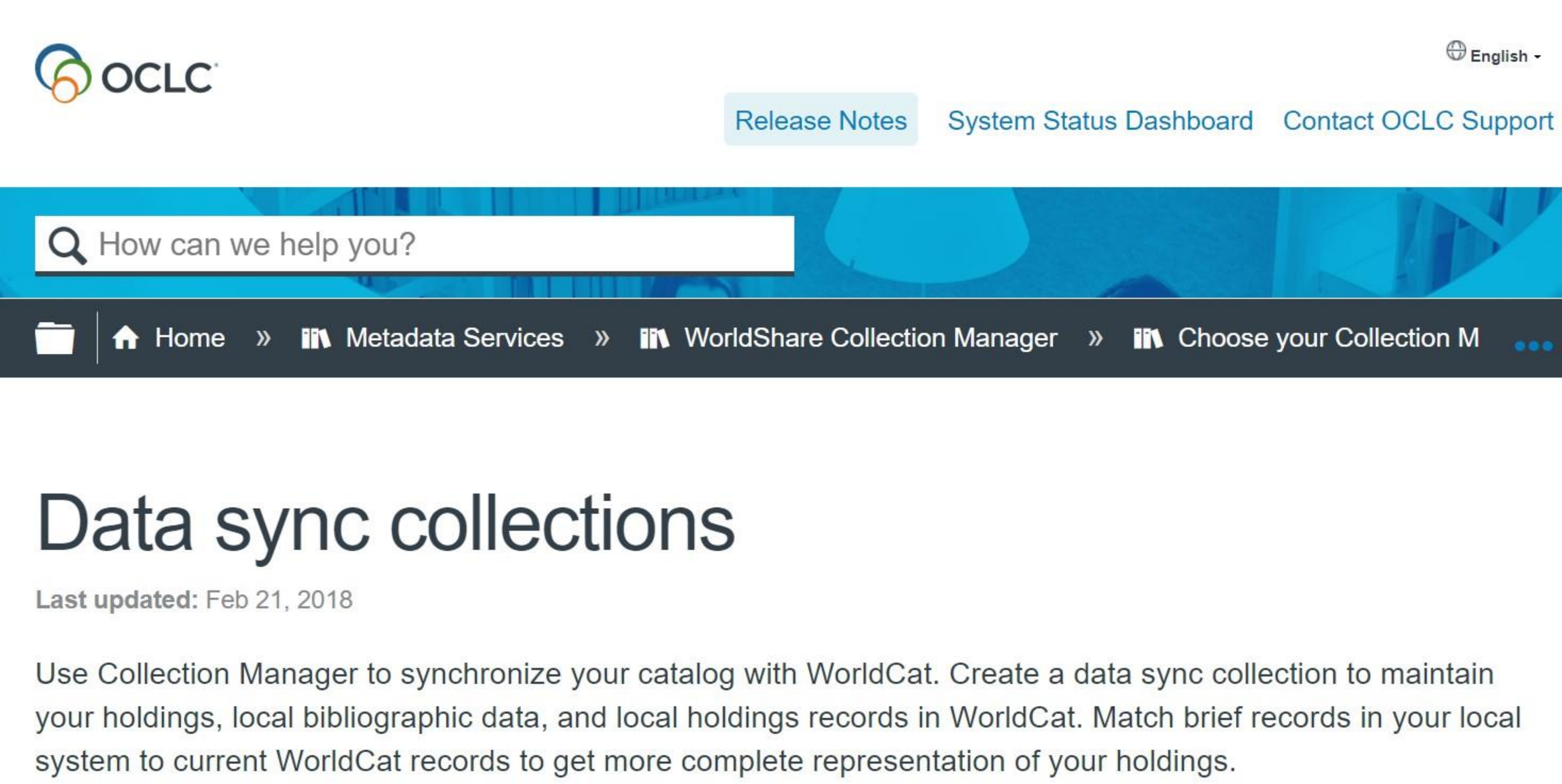
OR

□ Show See References i

□ Show See Also Referer

010			n 88231236
040			DLC ǂb eng ǂe rda ǂc DLC ǂd DLC ǂd InU ǂd OCoLC ǂd UPB ǂd DLC ǂd UPB
046			ǂf 19550208
053		0	PS3557.R5355
100	1		Grisham, John
370			Jonesboro (Ark.) ǂe Charlottesville (Va.) ǂe Oxford (Miss.) ǂ2 naf
372			Law ǂa Fiction ǂ2 lcsh
374			Authors ǂa Lawyers ǂ2 lcsh
375			male

Literary authors have a class number. Some subject headings have a class number in their records, too!


English
Release Notes
System Status Dashboard
Contact OCLC Support
How can we help you?
Home » Metadata Services » WorldShare Collection Manager » Choose your Collection M

Data sync collections

Last updated: Feb 21, 2018

Use Collection Manager to synchronize your catalog with WorldCat. Create a data sync collection to maintain your holdings, local bibliographic data, and local holdings records in WorldCat. Match brief records in your local system to current WorldCat records to get more complete representation of your holdings.

Free for one sync for OCLC members. Compares your ILS holdings against OCLC. Can set or delete holdings (auto or provide a report for review). https://help.oclc.org/Metadata_Services/WorldShare_Collection_Manager/Choose_your_Collection_Manager_workflow/Data_sync_collections

Break it to make it – thinking about data differently

Easiest process for working with MARC records is to start with MARCEdit.

MARCEdit can

- “break” the data structure
- has a built in connection to OpenRefine.
- Has its own built in clean up tools.
- is free software.

MARCEDIT

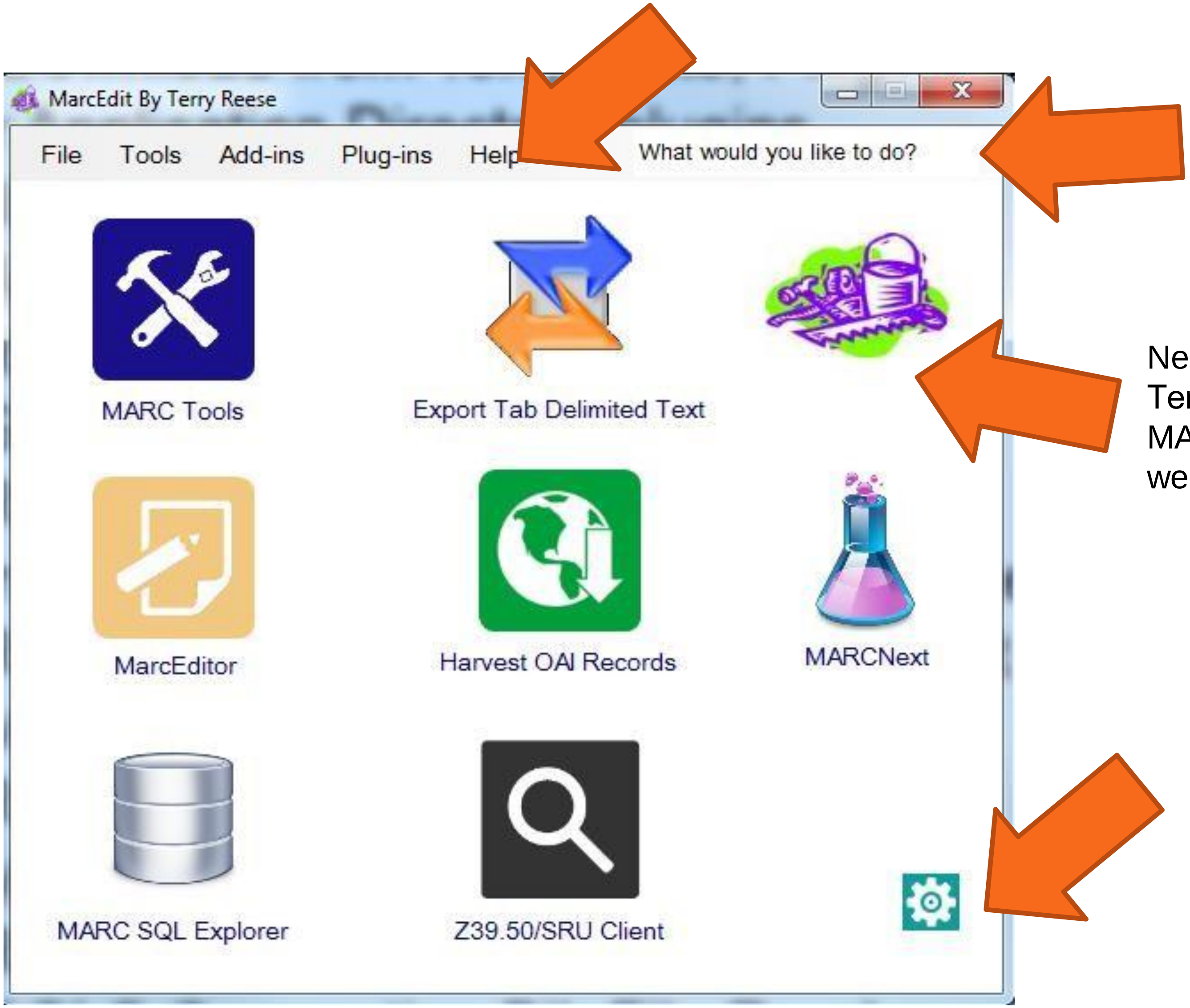
- Replace All
- Add New Field
- Delete Field
- Field Edit
- Edit Subfield
- Edit Indicator
- Swap Field
- Copy Field
- Add Task List
- RDA Helper
- Linked Data
- Build New Field
- Validate Headings
- BIBFRAME

TIP: In order to work with MARCEdit, you need 3 things:

1. The software installed on your machine
2. A group of MARC files
3. A blank file that you can “dump” your work in (when MARCEdit asks you where to save your work, the FILE must exist). See MARCEdit instructions for help.

MARCEdit Processes

- Identify (find) your records
- Sort/organize records (Group)
- Download them in an acceptable format (Export)
- Import them into MARCEDIT
- EDIT
- Validate
- Export out of MARCEDIT
- Import back into your system
- Review (Validate if possible)

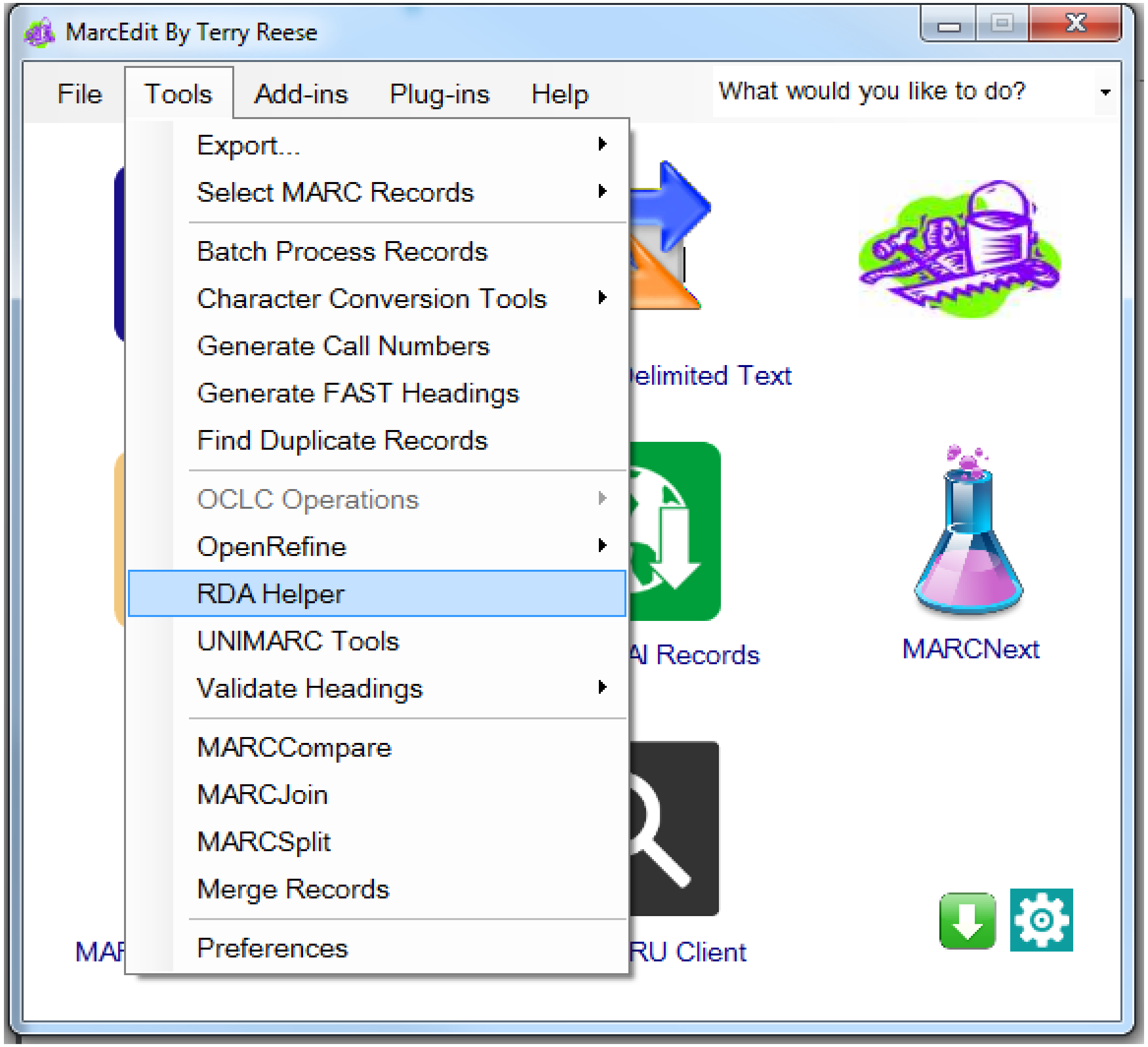


Need help?
Terry’s
MARCEdit
website

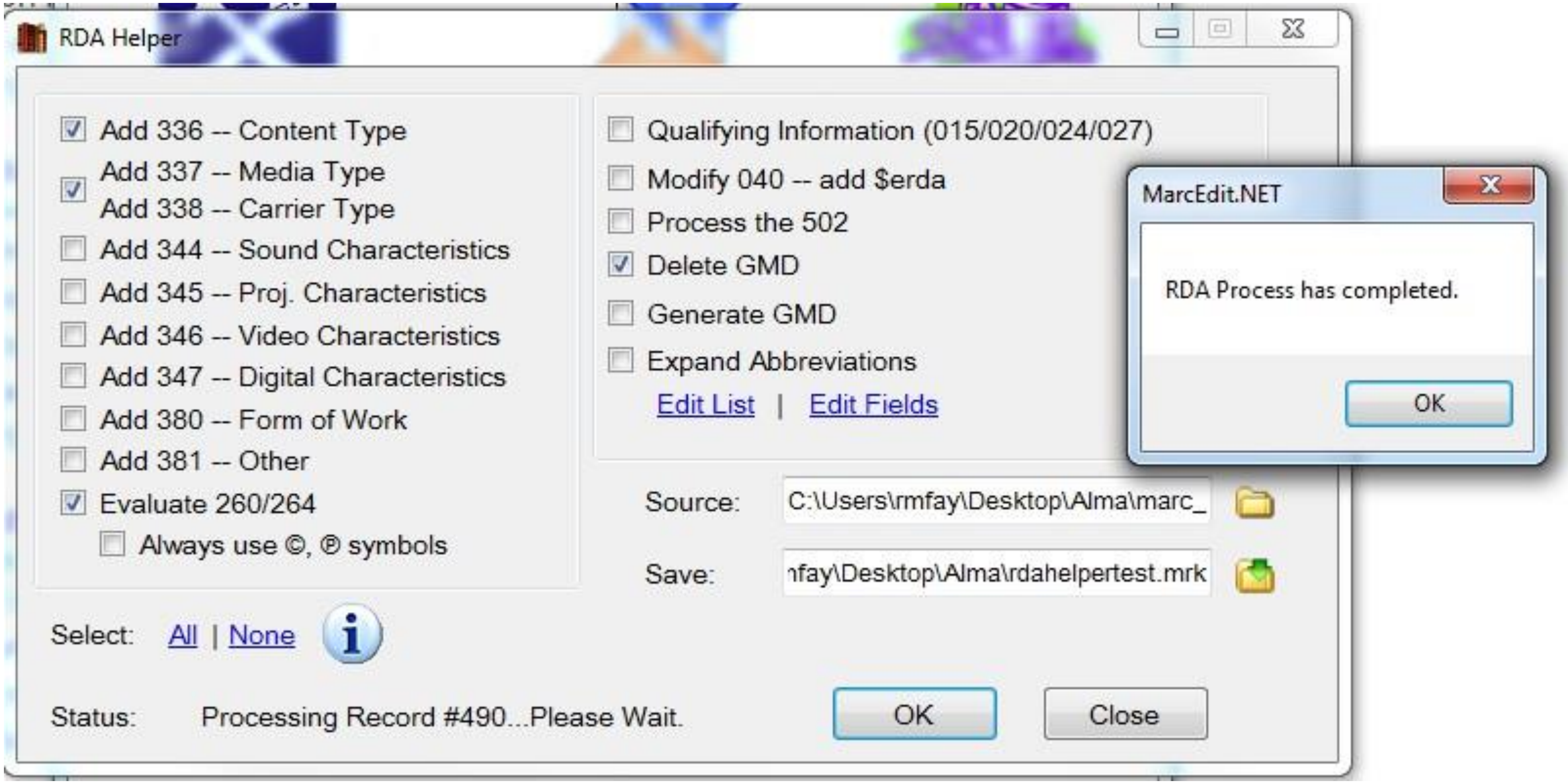
RDA Helper

Under Tools>RDA Helper

(or add to your default menu under the settings button)



RDA Helper



RDA Helper

☑ Add 336 -- Content Type

☑ Add 337 -- Media Type
Add 338 -- Carrier Type

□ Add 344 -- Sound Characteristics

□ Add 345 -- Proj. Characteristics

□ Add 346 -- Video Characteristics

□ Add 347 -- Digital Characteristics

□ Add 380 -- Form of Work

□ Add 381 -- Other

☑ Evaluate 260/264

□ Always use ©, ℗ symbols

□ Qualifying Information (015/020/024/027)

□ Modify 040 -- add $erda

□ Process the 502

☑ Delete GMD

□ Generate GMD

□ Expand Abbreviations

Edit List | Edit Fields

Source: C:\Users\rmfay\Desktop\Alma\marc_

Save: ıfay\Desktop\Alma\rdahelpertest.mrk

Select: All | None

Status: Processing Record #490...Please Wait.

OK Close

MarcEdit.NET

RDA Process has completed.

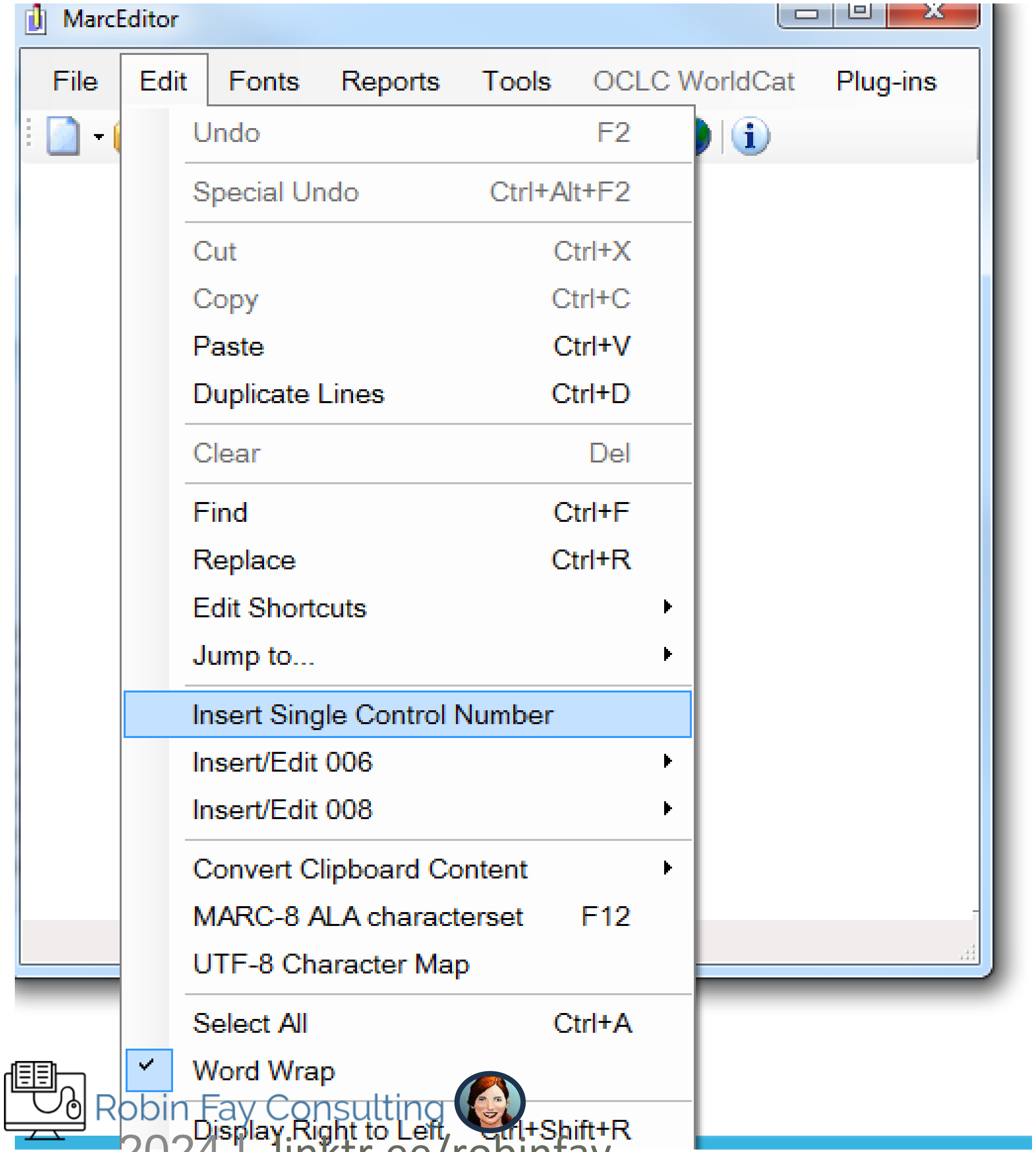
OK

MARCEdit

Add/Delete Field

- Can delete or add entire field
- Can be limited by MARC tag
- Some pre-built

Inserting/adding fields



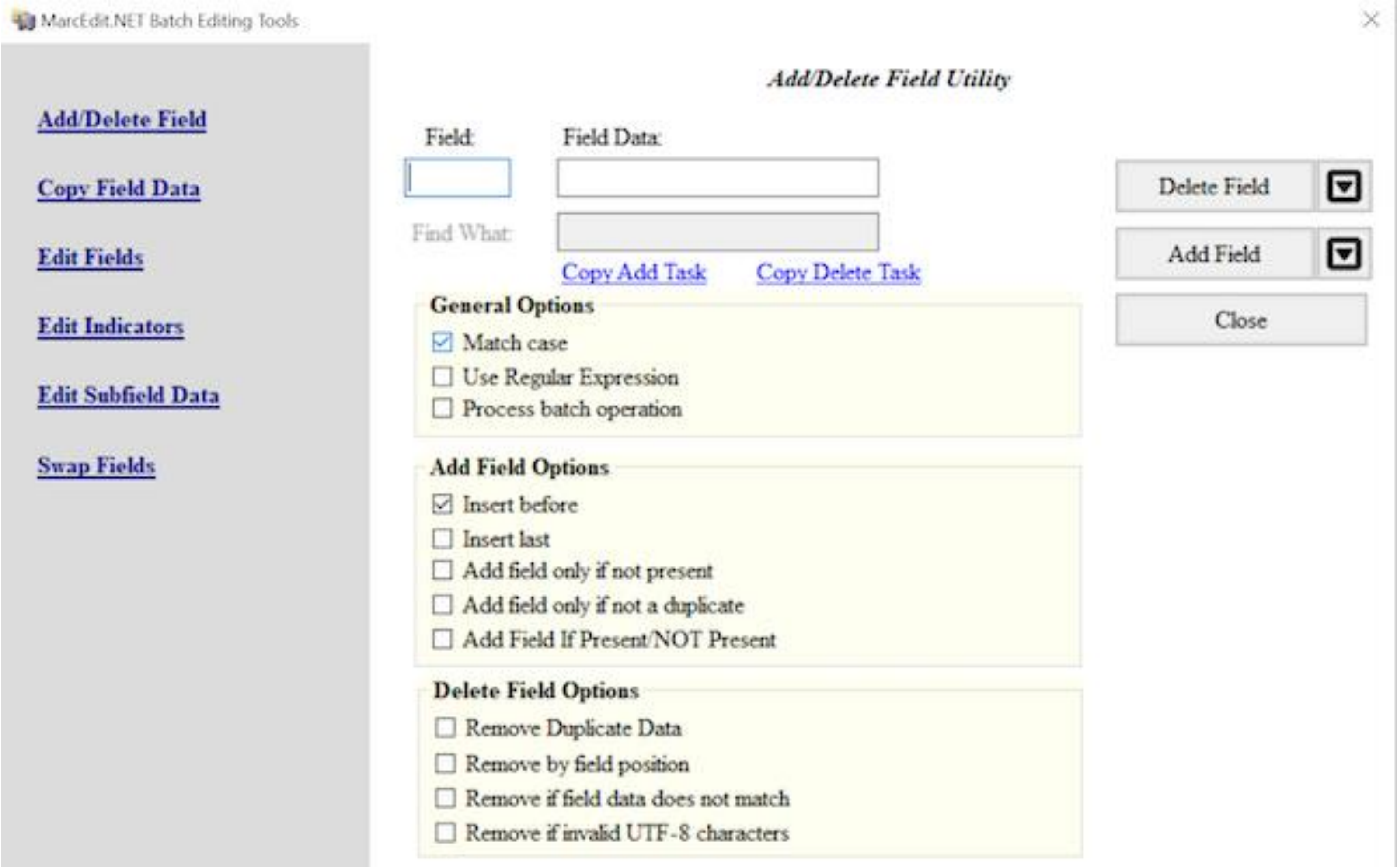
Adding/inserting fixed fields
Adding/inserting single #

Tip – don't forget to open your file

$ for delimiter

Don't forget your subfields

Inserting/adding fields



General Options

- ☑ Match case
- ☐ Use Regular Expression
- ☐ Process batch operation

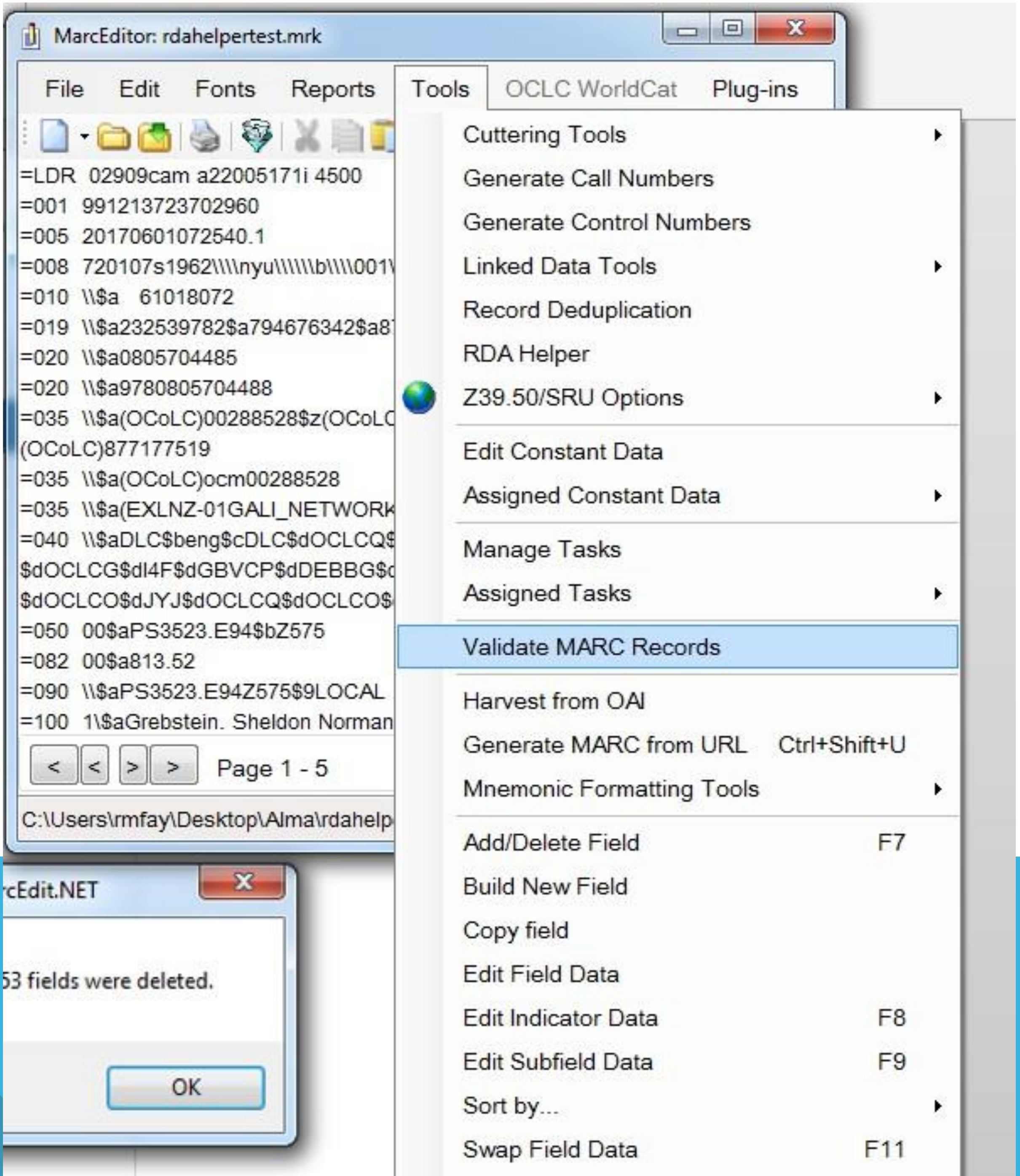
Add Field Options

- ☑ Insert before
- ☐ Insert last
- ☐ Add field only if not present
- ☐ Add field only if not a duplicate
- ☐ Add Field If Present/NOT Present

Delete Field Options

- ☐ Remove Duplicate Data
- ☐ Remove by field position
- ☐ Remove if field data does not match
- ☐ Remove if invalid UTF-8 characters

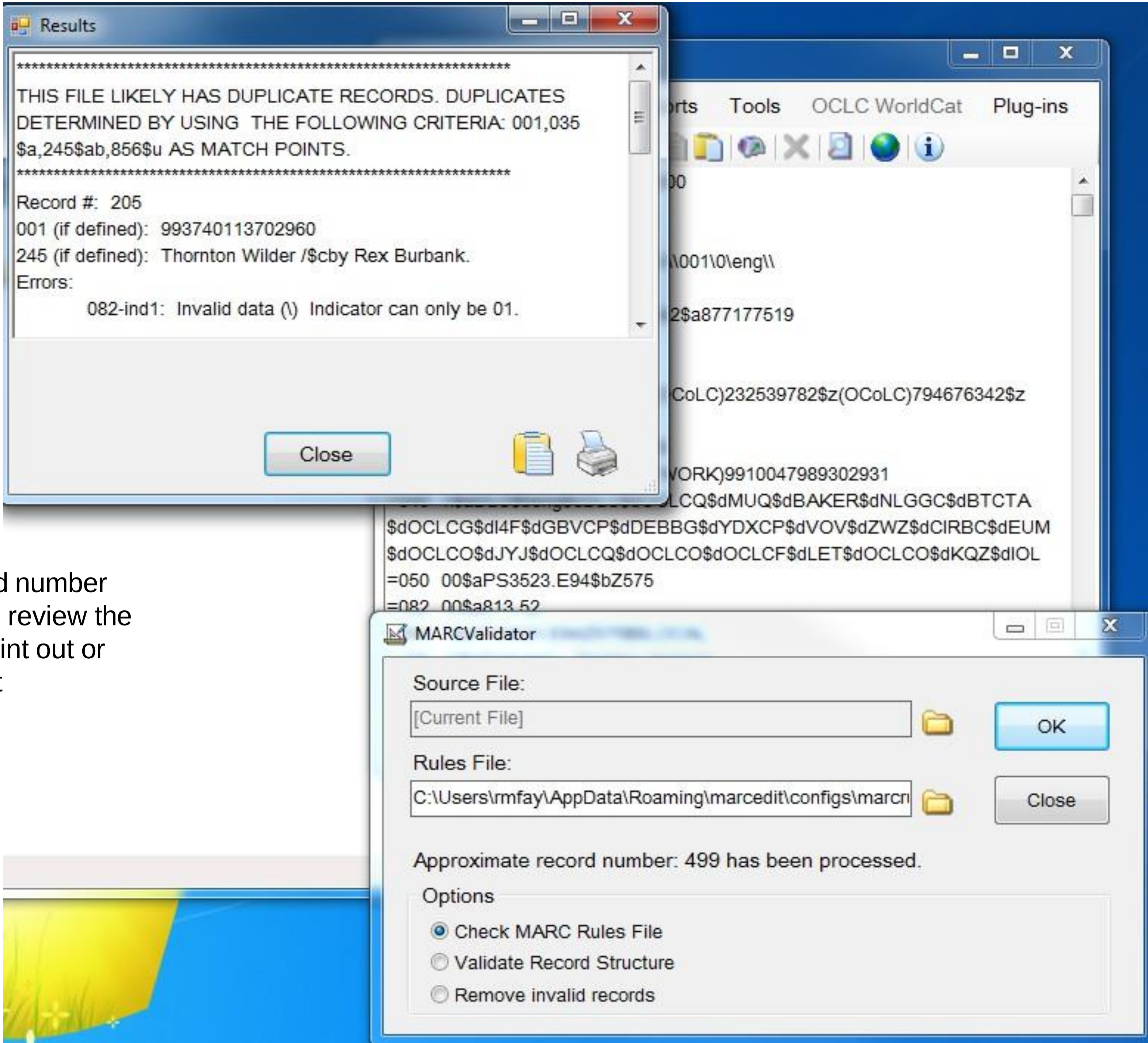
MARCEDIT – validate



```
=LDR  02909cam a22005171i 4500
=001  991213723702960
=005  20170601072540.1
=008  720107s1962\\\\nyu\\\\\\b\\\\001\
=010  \\$a   61018072
=019  \\$a232539782$a794676342$a8
=020  \\$a0805704485
=020  \\$a9780805704488
=035  \\$a(OCoLC)00288528$z(OCoLC
(OCoLC)877177519
=035  \\$a(OCoLC)ocm00288528
=035  \\$a(EXLNZ-01GALI_NETWORK
=040  \\$aDLC$beng$cDLC$dOCLCQ$
$dOCLCG$dI4F$dGBVCP$dDEBBG$d
$dOCLCO$dJYJ$dOCLCQ$dOCLCO$
=050  00$aPS3523.E94$bZ575
=082  00$a813.52
=090  \\$aPS3523.E94Z575$9LOCAL
=100  1\$aGrebstein. Sheldon Norman
```

C:\Users\rmfay\Desktop\Alma\rdahelp

53 fields were deleted.



Results

```
**********************************************************************
THIS FILE LIKELY HAS DUPLICATE RECORDS. DUPLICATES
DETERMINED BY USING  THE FOLLOWING CRITERIA: 001,035
$a,245$ab,856$u AS MATCH POINTS.
**********************************************************************
Record #:  205
001 (if defined):  993740113702960
245 (if defined):  Thornton Wilder /$cby Rex Burbank.
Errors:
          082-ind1:  Invalid data (\)  Indicator can only be 01.
```

```
00
\001\0\eng\\
2$a877177519
CoLC)232539782$z(OCoLC)794676342$z
ORK)9910047989302931
LCQ$dMUQ$dBAKER$dNLGGC$dBTCTA
$dOCLCG$dI4F$dGBVCP$dDEBBG$dYDXCP$dVOV$dZWZ$dCIRBC$dEUM
$dOCLCO$dJYJ$dOCLCQ$dOCLCO$dOCLCF$dLET$dOCLCO$dKQZ$dIOL
=050  00$aPS3523.E94$bZ575
=082  00$a813.52
```

MARCValidator

Source File: [Current File]

Rules File: C:\Users\rmfay\AppData\Roaming\marcedit\configs\marcr

Approximate record number: 499 has been processed.

Options

● Check MARC Rules File
○ Validate Record Structure
○ Remove invalid records

Provides record number so that you can review the errors – also print out or save this report

Excel

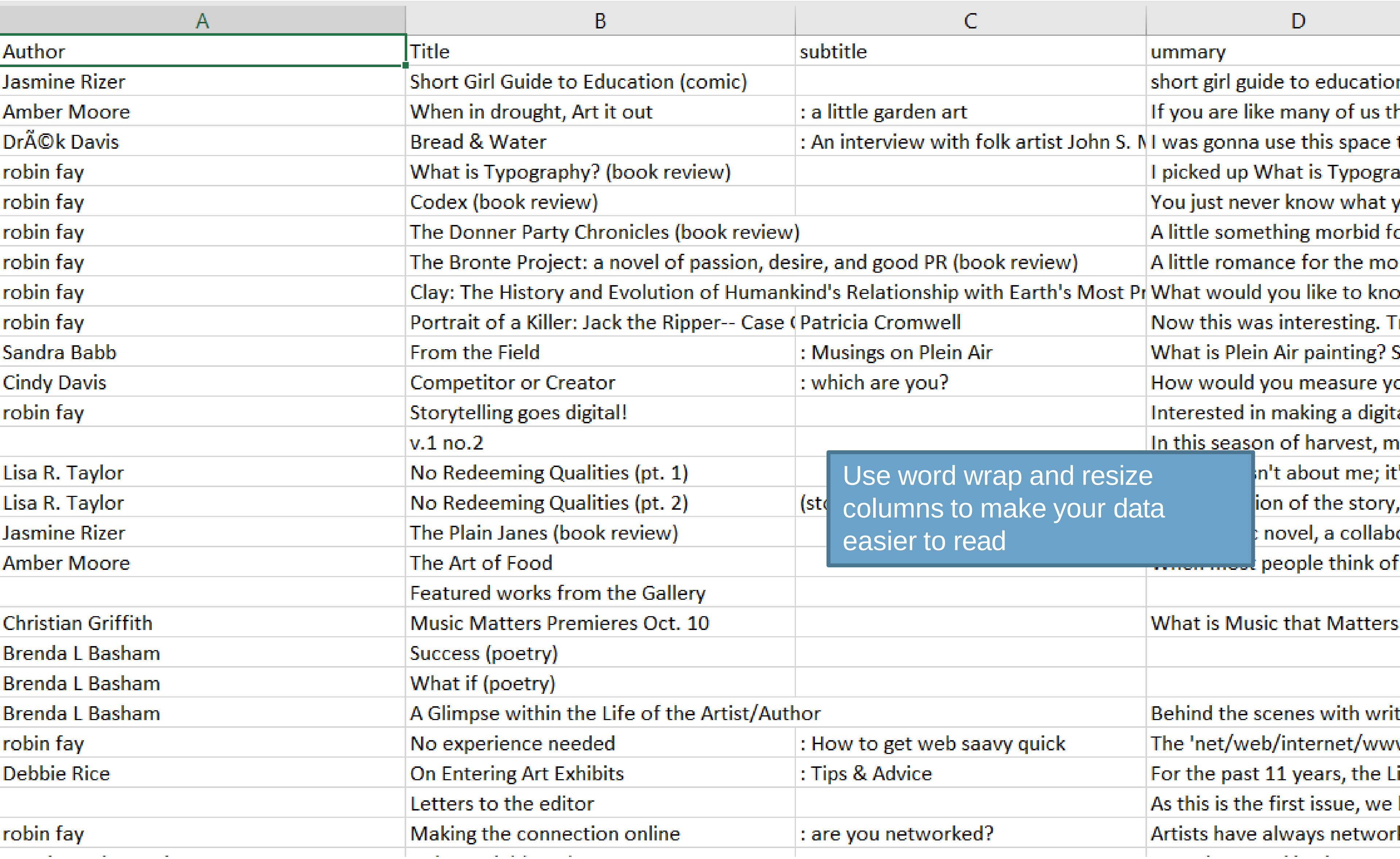
- CSV, XML or MARC will all work with OpenRefine.For MARC, you need to “break it” (MARCEdit) or another MARCXML tool.
- OpenRefine is more robust than Excel, but both can be used for data cleanup.

EXCEL
quick data-cleaning tips

Created by Miranda Lee
January 2020

This resource provides strategies for cleaning data in Microsoft Excel. Below is a brief overview of five situations you may find yourself in ("What") and corresponding solutions ("How"), followed by detailed instructions to implement the solutions.

	What?	How?
1	Identify all cells that contain a specific word or (short) phrase in a column with open-ended text	Use **Conditional Formatting**
2	Identify and remove duplicate data	Use **Remove Duplicates** function or **Conditional Formatting**
3	Identify the outliers within a data set (e.g., dates or grades)	Use **Data Validation** function
4	Separate data from a single column into two or more columns (e.g., first and last names)	Use **Flash Fill**
5	Categorize data in a column, such as class assignments or subject groups	Use **Formula** to fill in the category column



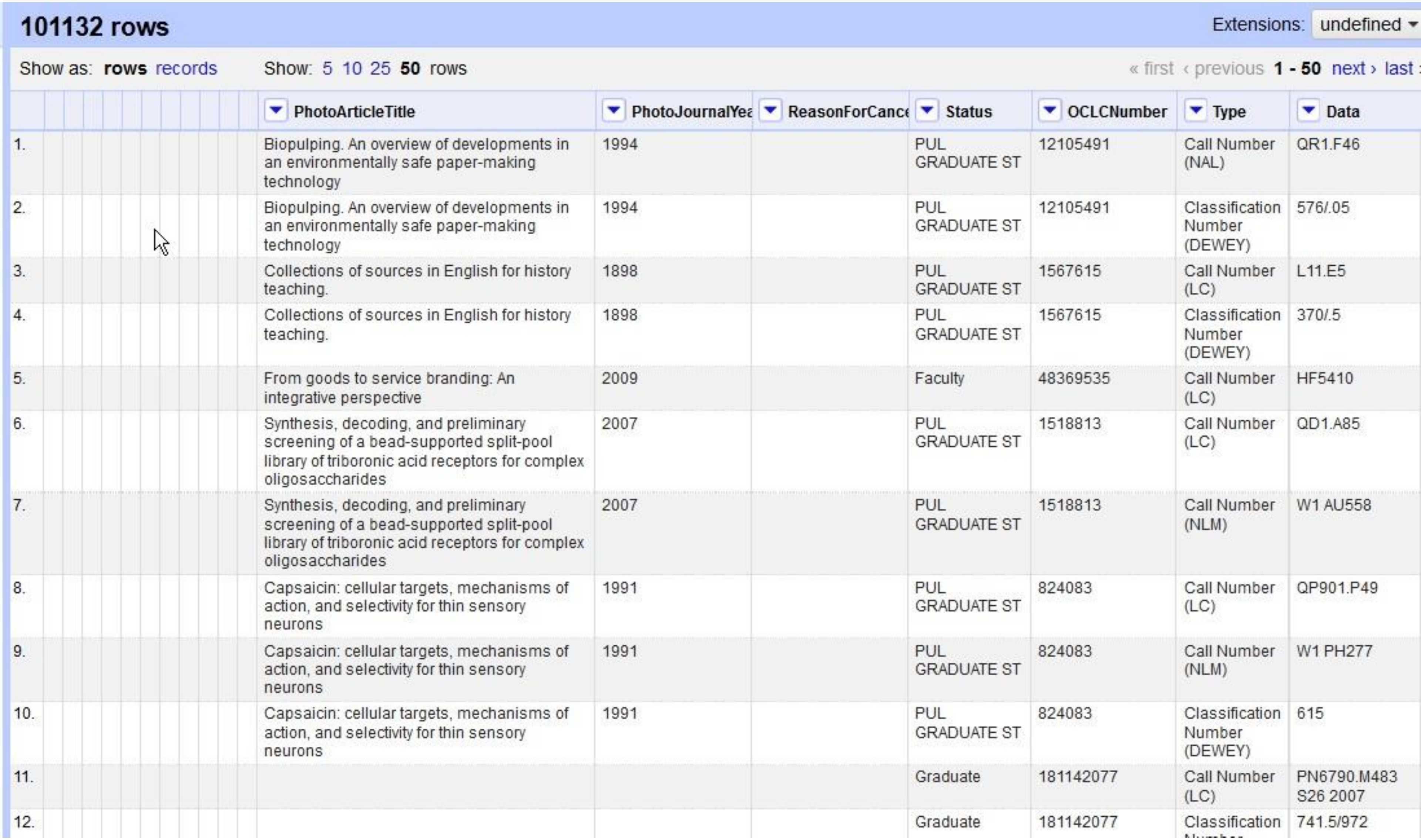
A	B	C	D
Author	Title	subtitle	ummary
Jasmine Rizer	Short Girl Guide to Education (comic)		short girl guide to educatio
Amber Moore	When in drought, Art it out	: a little garden art	If you are like many of us th
DrÃ©k Davis	Bread & Water	: An interview with folk artist John S. M	I was gonna use this space t
robin fay	What is Typography? (book review)		I picked up What is Typogra
robin fay	Codex (book review)		You just never know what y
robin fay	The Donner Party Chronicles (book review)		A little something morbid fo
robin fay	The Bronte Project: a novel of passion, desire, and good PR (book review)		A little romance for the mo
robin fay	Clay: The History and Evolution of Humankind's Relationship with Earth's Most Pr		What would you like to kno
robin fay	Portrait of a Killer: Jack the Ripper-- Case (	Patricia Cromwell	Now this was interesting. Tl
Sandra Babb	From the Field	: Musings on Plein Air	What is Plein Air painting? S
Cindy Davis	Competitor or Creator	: which are you?	How would you measure yo
robin fay	Storytelling goes digital!		Interested in making a digita
	v.1 no.2		In this season of harvest, m
Lisa R. Taylor	No Redeeming Qualities (pt. 1)		n't about me; it'
Lisa R. Taylor	No Redeeming Qualities (pt. 2)	(st	ion of the story,
Jasmine Rizer	The Plain Janes (book review)		novel, a collabo
Amber Moore	The Art of Food		people think of
	Featured works from the Gallery		
Christian Griffith	Music Matters Premieres Oct. 10		What is Music that Matters
Brenda L Basham	Success (poetry)		
Brenda L Basham	What if (poetry)		
Brenda L Basham	A Glimpse within the Life of the Artist/Author		Behind the scenes with writ
robin fay	No experience needed	: How to get web saavy quick	The 'net/web/internet/www
Debbie Rice	On Entering Art Exhibits	: Tips & Advice	For the past 11 years, the Li
	Letters to the editor		As this is the first issue, we l
robin fay	Making the connection online	: are you networked?	Artists have always networ

Use word wrap and resize columns to make your data easier to read

OpenRefine

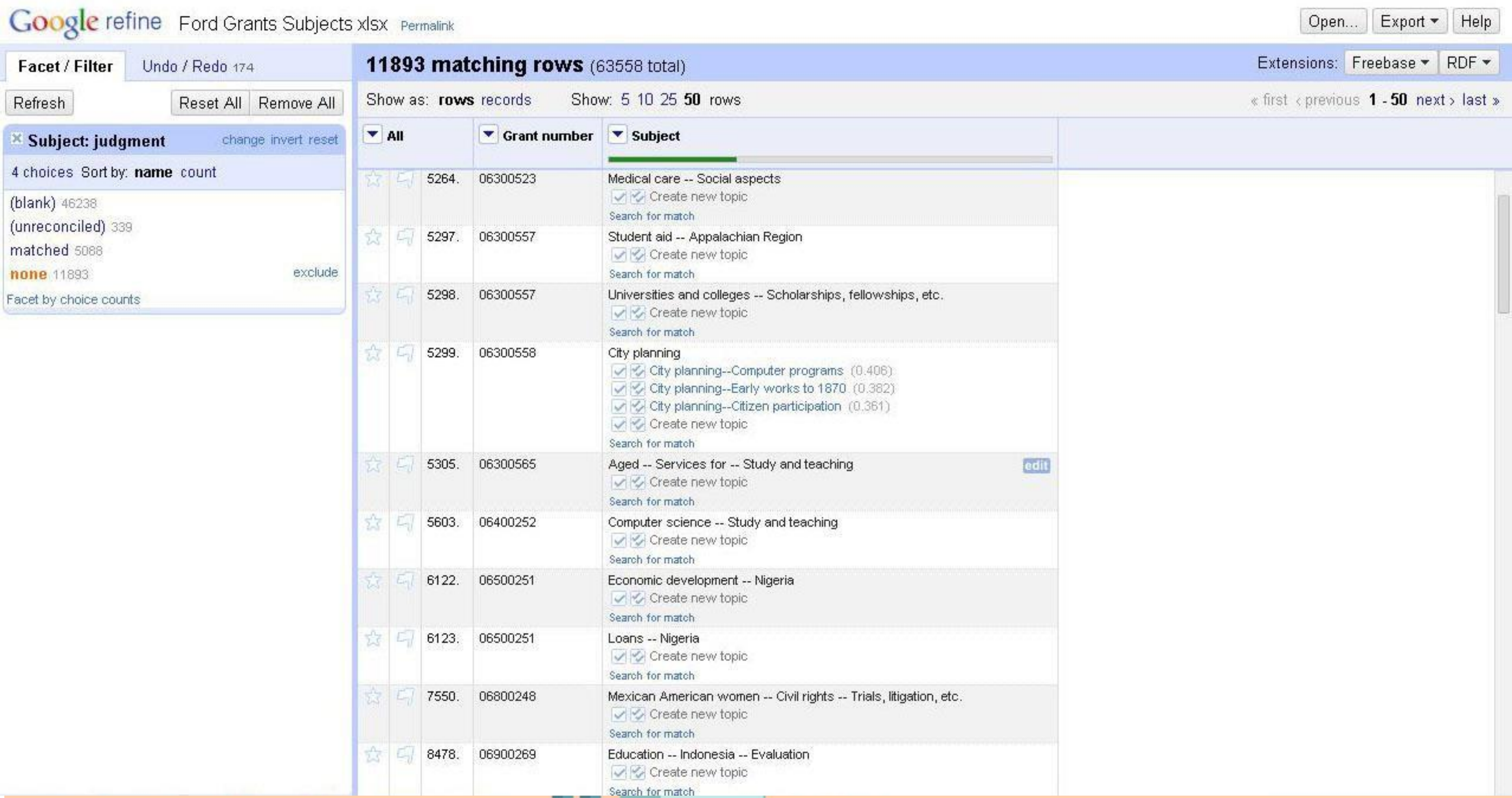
- Open Refine (at one point, Google Refine)
- Opensource http://openrefine.org/download.html
- Two basic functions
  - Visualizing your data (being able to see the type – date, number, text? Entries the same?)
  - Manipulating data (searching and replacing, update entries by batch or individually, add data, delete data, change date types, more)

Looks a lot like Excel (and works with Excel, too!)



	PhotoArticleTitle	PhotoJournalYea	ReasonForCance	Status	OCLCNumber	Type	Data
1.	Biopulping. An overview of developments in an environmentally safe paper-making technology	1994		PUL GRADUATE ST	12105491	Call Number (NAL)	QR1.F46
2.	Biopulping. An overview of developments in an environmentally safe paper-making technology	1994		PUL GRADUATE ST	12105491	Classification Number (DEWEY)	576/.05
3.	Collections of sources in English for history teaching.	1898		PUL GRADUATE ST	1567615	Call Number (LC)	L11.E5
4.	Collections of sources in English for history teaching.	1898		PUL GRADUATE ST	1567615	Classification Number (DEWEY)	370/.5
5.	From goods to service branding: An integrative perspective	2009		Faculty	48369535	Call Number (LC)	HF5410
6.	Synthesis, decoding, and preliminary screening of a bead-supported split-pool library of triboronic acid receptors for complex oligosaccharides	2007		PUL GRADUATE ST	1518813	Call Number (LC)	QD1.A85
7.	Synthesis, decoding, and preliminary screening of a bead-supported split-pool library of triboronic acid receptors for complex oligosaccharides	2007		PUL GRADUATE ST	1518813	Call Number (NLM)	W1 AU558
8.	Capsaicin: cellular targets, mechanisms of action, and selectivity for thin sensory neurons	1991		PUL GRADUATE ST	824083	Call Number (LC)	QP901.P49
9.	Capsaicin: cellular targets, mechanisms of action, and selectivity for thin sensory neurons	1991		PUL GRADUATE ST	824083	Call Number (NLM)	W1 PH277
10.	Capsaicin: cellular targets, mechanisms of action, and selectivity for thin sensory neurons	1991		PUL GRADUATE ST	824083	Classification Number (DEWEY)	615
11.				Graduate	181142077	Call Number (LC)	PN6790.M483 S26 2007
12.				Graduate	181142077	Classification	741.5/972

Use Facets



Google refine Ford Grants Subjects xlsx Permalink

Subject: judgment change invert reset

4 choices Sort by: **name** count

(blank)	46238
(unreconciled)	339
matched	5088
none	11893

exclude

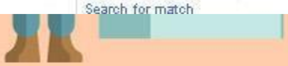
Facet by choice counts

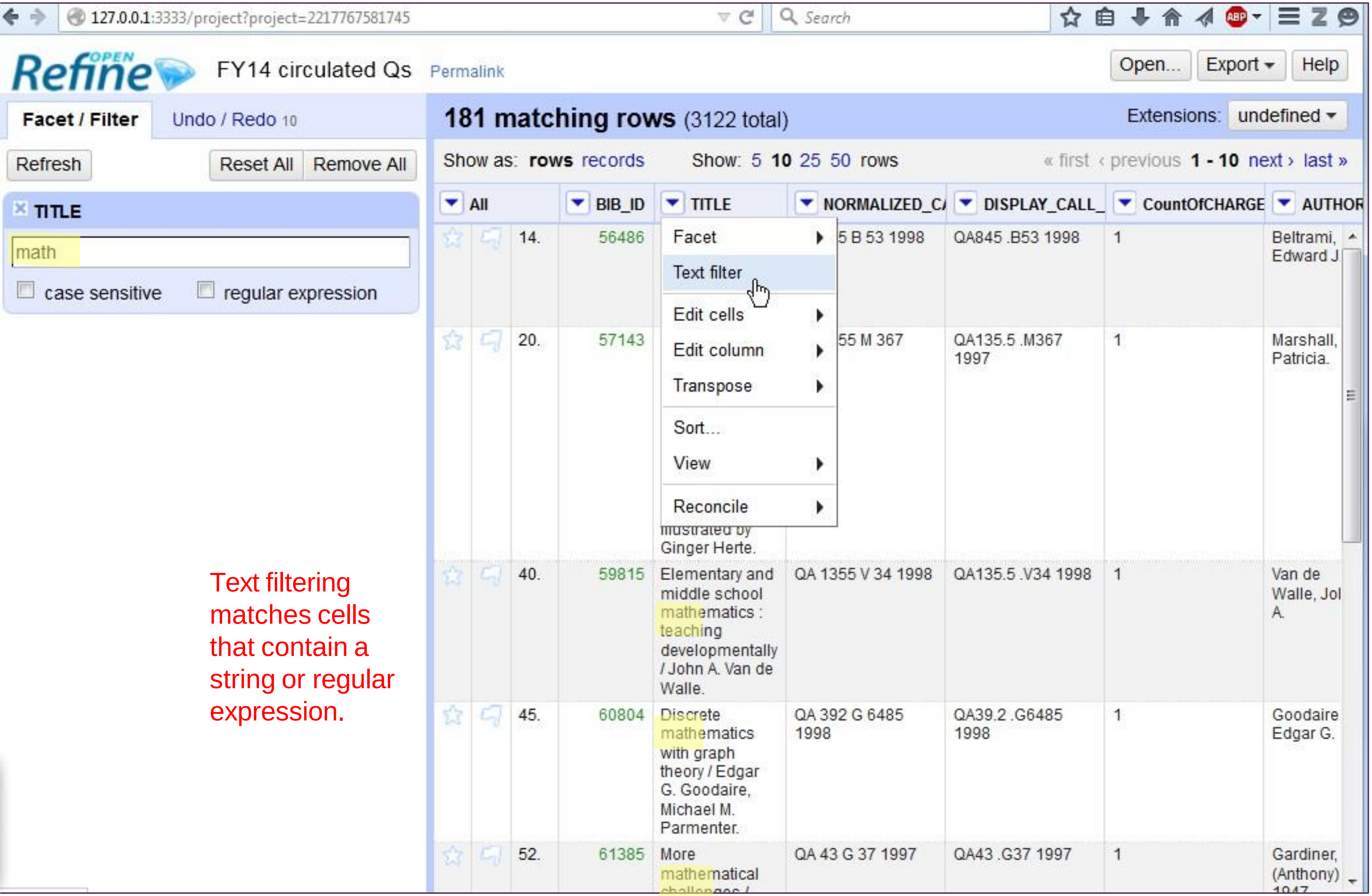
11893 matching rows (63558 total)

Show as: **rows** records Show: 5 10 25 **50** rows

« first ‹ previous **1 - 50** next › last »

All	Grant number	Subject
5264.	06300523	Medical care -- Social aspects ☑ ☑ Create new topic Search for match
5297.	06300557	Student aid -- Appalachian Region ☑ ☑ Create new topic Search for match
5298.	06300557	Universities and colleges -- Scholarships, fellowships, etc. ☑ ☑ Create new topic Search for match
5299.	06300558	City planning ☑ ☑ City planning--Computer programs (0.406) ☑ ☑ City planning--Early works to 1870 (0.382) ☑ ☑ City planning--Citizen participation (0.361) ☑ ☑ Create new topic Search for match
5305.	06300565	Aged -- Services for -- Study and teaching edit ☑ ☑ Create new topic Search for match
5603.	06400252	Computer science -- Study and teaching ☑ ☑ Create new topic Search for match
6122.	06500251	Economic development -- Nigeria ☑ ☑ Create new topic Search for match
6123.	06500251	Loans -- Nigeria ☑ ☑ Create new topic Search for match
7550.	06800248	Mexican American women -- Civil rights -- Trials, litigation, etc. ☑ ☑ Create new topic Search for match
8478.	06900269	Education -- Indonesia -- Evaluation ☑ ☑ Create new topic Search for match





FY14 circulated Qs Permalink

181 matching rows (3122 total)

Show as: **rows** records   Show: 5 **10** 25 50 rows   « first ‹ previous **1 - 10** next › last »

TITLE

math

□ case sensitive   □ regular expression

All	BIB_ID	TITLE	NORMALIZED_C/	DISPLAY_CALL_	CountOfCHARGE	AUTHOR
14.	56486		5 B 53 1998	QA845 .B53 1998	1	Beltrami, Edward J
20.	57143	illustrated by Ginger Herte.	55 M 367	QA135.5 .M367 1997	1	Marshall, Patricia.
40.	59815	Elementary and middle school mathematics : teaching developmentally / John A. Van de Walle.	QA 1355 V 34 1998	QA135.5 .V34 1998	1	Van de Walle, Jol A.
45.	60804	Discrete mathematics with graph theory / Edgar G. Goodaire, Michael M. Parmenter.	QA 392 G 6485 1998	QA39.2 .G6485 1998	1	Goodaire Edgar G.
52.	61385	More mathematical challenges /	QA 43 G 37 1997	QA43 .G37 1997	1	Gardiner, (Anthony) 1947

Text filtering matches cells that contain a string or regular expression.

What about ChatGPT / AI

ChatGPT is generative CONTENT creator. It is meant to write/create content or summarize information.

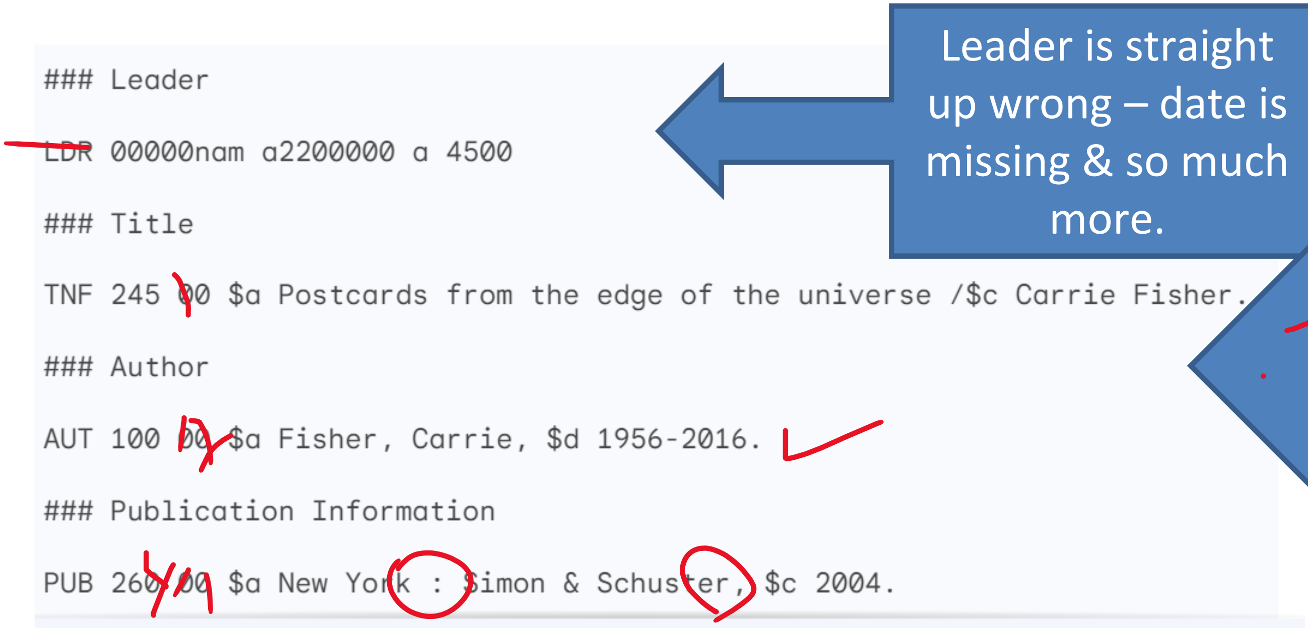
It does not have strong data analytic skills (it is built to be general; it is not currently a data tool).

However, ChatGPT and other AI tools can be used for finding typos in data and some dedicated data/AI tools may be useful for our data.

Typically, with data cleanup and remediation, we look to batch processing tools that are very clearly defined, e.g., change all entries in the contributor field to the creator field.

ChatGPT – be careful with creating data

This book does not exist.



### Leader

~~LDR~~ 00000nam a2200000 a 4500

### Title

TNF 245 00 $a Postcards from the edge of the universe /$c Carrie Fisher.

### Author

AUT 100 00 $a Fisher, Carrie, $d 1956-2016.

### Publication Information

PUB 260 00 $a New York : Simon & Schuster, $c 2004.

Leader is straight up wrong – date is missing & so much more.

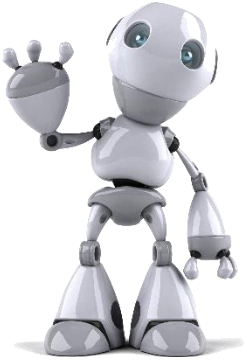
Transcription field –recorded as it appears on the resource (LOL!)
Order of fields is incorrect as indicator coding.
Subdivisions are missing.
260 is an older field and less used.

Opportunities – create starting points

Use AI it for writing -- draft or start a press release, instructions, letters, etc. Do NOT upload IP (Intellectual Property to AI – once it has that information there is no retraction – at this point).

Use it to start a record – we do have other ways of doing this which may be better. For example, a cataloger can clone or copy a record of their choosing (meaning: human cognition factored in the choice). AI may choose a record that has been used the most but is not "good" quality OR it may find the first record that seems to fit or it may just steal from another library (like LC).

Assign high level subject headings based upon data analysis – AI gets lost in the granularity of our work and the nuances of rules – 3-6 subject headings are preferable, but no more than 10*; exceptions abound!

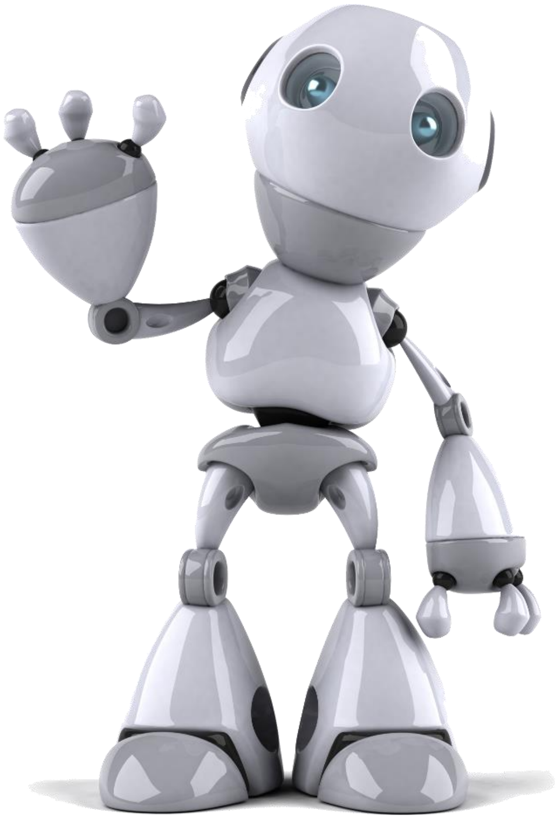


Opportunities – Enrich Data

Use AI to add DOIs or linked data (yes, we have other ways of doing this such as MARCEdit and OpenRefine).

Use AI to identify records missing/lacking info or wrong info (outdated fields, etc.)

Uses it to create a summary/abstract from full text articles. These summaries are then reviewed by catalogers, making corrections as needed.



This Photo by Unknown Author is licensed under CC BY-NC

System migration Prep

- Deviations from common practice, standards, etc. could be a problem
- For MARC records, OCLC Data Sync can help resolve issues (&potentially find issues, too!)
- Fix what you can before ingest/import/migration>> Cleanup data before import, harvest, etc (if possible)
- Don’t map what you don’t need – outdated data can go

System Migration prep

- Is each MARC field in your records have a “place” in the new system? If not, where will that data go? Did you misuse fields or create local fields?
- Did you combine formats onto one record?
- Do you have the ability to create a mapping on import ?

Considerations

- Can you re-label materials if needed? If not, then do not take on projects involving re-labeling (or perhaps, work on them at a slower pace).
- Can you identify a set of records, that can be batch exported and batch replaced? If so, consider using MARCEdit for work.
- What tools do you have access to?
- Can you do a shelfreading project?
- What tools do you have and what help do you have?
- What are the priorities?

This Photo by Unknown Author is licensed under CC BY-SA-NC

ROI of work

- Use tools where you can
- Prioritize work but do it
- Keep samples of before and after – document your work

Clean Up Your Data

Think in patterns

Experiment with tools

to help you.

Make a plan.

janitor::clean_names
convert all column names to *_case!
(not so awesome) column names
FUN(units)!
Sampleǝ!
Such.Mess!
ooo.OOF-.
MESS
Mm.,OOPS
clean_names((your data frame + case choice here))
CHOOSE:
✓ snake
lower_camel
big_camel
screaming_snake
+ more!
WAY MORE DEALWITHABLE COLUMN NAMES!
so_much_nicer
_yay_ya_
hope_rests
_than
good_feel_
_cool
teddy
HORST 2019

This Photo by Unknown Author is licensed under CC BY

Questions?

thank you!