

# DAM It: Migrating Images and Cleaning Metadata – Tools and Tips to Make it Easier

Michelle Sweetser  
Midwest Archives Conference Annual Meeting  
May 7, 2015



MARQUETTE  
UNIVERSITY

Be The Difference.

# Background

Scanning and Digital Image Status in the University Archives before summer 2014



MARQUETTE  
UNIVERSITY

Be The Difference.



# CONTENTdm at Marquette

- Primary means of sharing images with the public
- Thematically-based collections
- Limited in scope; topics and images selected by archives staff
  - Patron-driven scanning requests were not included
- Approaching our item-limit

# Extensis Portfolio for Patron-Driven Scans

- Archives had a license; several stand-alone licenses for Portfolio elsewhere on campus
- Beginning in 2008, obtained approval to begin to create, describe, and manage high resolution scans from patron-requests
- Employ approved keywords set by campus working group
- Extensis ended standalone sales December 2013; support for product ceased June 2014

	CONTENTdm	Portfolio
Pros	<ul style="list-style-type: none"> <li>• Online delivery and search functionality</li> <li>• Cross-collection functionality and ability to reduce silos</li> <li>• Improved ease in applying controlled vocabularies</li> <li>• Data can be exported as a .txt file</li> </ul>	<ul style="list-style-type: none"> <li>• Software tracks patron selections</li> <li>• Automatic collection of high resolution files from multiple server locations</li> <li>• Metadata travels with the image</li> <li>• Data can be exported as .txt file</li> </ul>
Cons	<ul style="list-style-type: none"> <li>• Selected images must be downloaded one at a time (low-res)</li> <li>• Patrons must track identifiers to request high-res versions</li> <li>• Staff time to pull and collect images from server</li> <li>• Metadata does not travel with the image</li> </ul>	<ul style="list-style-type: none"> <li>• No online presence for or discovery of images</li> <li>• Siloed collection</li> <li>• Clunky interface for applying controlled vocabulary</li> <li>• Uncertain future for the product</li> </ul>

# Opportunity in Disguise

- Move 10,000+ images to an online-discoverable platform while
  - Reviewing and editing keywords to gain more consistency in our records
  - Creating a cross-collection search interface online
- For the short-term: continue to actively maintain Portfolio database to facilitate workflow for internal clients

# Software Needed to Make it Happen

- Metadata analysis and clean-up
  - OpenRefine
- Write cleaned data back to master image files
  - VRA Panel Export-Import Tool
  - Portfolio

Plus,

- Update Portfolio to reflect new, improved metadata

# Workflow Overview

- Exported initial batch of metadata from Portfolio as .txt file
- Analysis and bulk editing of select fields in OpenRefine
- Export from OpenRefine as .txt file; additional message in Excel (concatenation, string substitutions)
- Upload to CONTENTdm
- Write changes to master files
- Update Portfolio catalog to reflect changes

# Metadata Clean-Up



MARQUETTE  
UNIVERSITY

Be The Difference.



# Open Refine

- Requires separation of multi-valued cells to analyze data
- Keywords stored in multi-valued cells
- Must atomize records

# Open Refine – Atomization

25 rows

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

Item ID	Keywords	Column	OriginalImageR	OriginalImageB	OriginalPhotoDa
389	AthleteAthleticsIntercollegiate AthleticsMalesMen's Basketball SportsTerrell Schlundt BB_1955_vs_Miami_of_Ohio_15_22	Athlete, Athletics, Intercollegiate Athletics, Males, Men's Basketball, Sports, Terrell Schlundt, BB_1955_vs_Miami_of_Ohio_15_22	Facet Text filter Edit cells Edit column Transpose Sort... View Reconcile	x 15A	1981-1982
390	AvalancheBarsStudentsMilwaukee	Avalanche, Bars, Students, Milwaukee	Hilltop Photo Collection		
391	AthletesAthleticsClub Sports FemalesStudentsWomen's Soccer - Club	Athletes, Athletics, Club Sports, Females, Students, Women's Soccer - Club			
392	CampusOrientationStudents	Campus, Orientation, Students	D-6 Series 2.1 - Hilltop Photo Collection	Box 15A	1981
401	Memorial LibraryStudentsStudying BuildingsInterior	Memorial Library, Students, Studying, Buildings, Interior	D-6 Series 2.1 - Hilltop Photo Collection	Box 15A	1982
393	CampusMalesOrientationStudents	Campus, Males, Orientation, Students	D-6 Series 2.1 - Hilltop Photo Collection	Box 15A	1981

# Open Refine – Atomization

341 rows

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

Item ID	Keywords	Column	OriginalImageR	OriginalImageB	OriginalPhotoDa
389	AthleteAthleticsIntercollegiate AthleticsMalesMen's Basketball SportsTerrell Schlundt BB_1955_vs_Miami_of_Ohio_15_22	Athlete	D-6 Series 2.1 - Hilltop Photo Collection	Box 15A	1981-1982
		Athletics			
		Intercollegiate Athletics			
		Males			
		Men's Basketball			
		Sports			
		Terrell Schlundt			
		BB_1955_vs_Miami_of_Ohio_15_22			
390	AvalancheBarsStudentsMilwaukee	Avalanche	D-6 Series 2.1 - Hilltop Photo Collection	Box 15A	1981-1982
		Bars			

# Open Refine – Facets, Filters, Clusters

- For our purposes, power of Open Refine was in its faceting, filtering, and clustering capabilities

The screenshot shows the Open Refine interface with a table of 13424 records. The table has columns for Item ID, Keywords, Column, OriginalImageR, OriginalImageB, OriginalPhotoDe, and Other. A context menu is open over the first row, showing options for Facet, Text filter, Edit cells, Edit column, Transpose, Sort, View, and Reconcile. The Facet menu is expanded, showing options for Text facet, Numeric facet, Timeline facet, Scatterplot facet, Custom text facet, Custom numeric facet, and Customized facets.

Item ID	Keywords	Column	OriginalImageR	OriginalImageB	OriginalPhotoDe	Other
389	AthleteAthleticsIntercollegiate AthleticsMalesMen's Basketball SportsTerrell Schlundt BB_1955_vs_Miami_of_Ohio_15_22				1981-1982	
390	AvalancheBarsStudentsMilwaukee	Avalanche	D-6 Series 2.1 - Hilltop Photo Collection	Box 15A	1981-1982	

# Open Refine – Faceting

The screenshot shows the 'Facet / Filter' panel in Open Refine. At the top, there are buttons for 'Refresh', 'Reset All', and 'Remove All'. Below these is a tab labeled 'Column' with a 'change' link. The panel displays '6489 choices' and is sorted by 'name' with a 'count' column. A 'Cluster' button is visible. The list of items is as follows:

Item Name	Count
melvin Peterson	1
Melvin Shimek	22
Memorial Library	62
Men's Basetball	1
Men's Baskeball	1
Men's Basketball	2820
Men's Basketballl	3
Men's Basketttball	1
Men's Baskteball	1
Men's Golf	42
Men's Rugby	1
Men's Soccer	27

# Open Refine – Editing Facet

The screenshot shows the Open Refine interface. On the left, the 'Facet / Filter' panel is active, showing a 'Column' facet with 6489 choices. The facet is sorted by 'name' and 'count'. A modal dialog is open over the facet, allowing the user to edit the facet value. The dialog contains the text 'Men's Basketball' and has 'Apply' and 'Cancel' buttons. The main data table shows 14268 records. The table has columns for 'Item ID', 'Keywords', 'Column', 'Modified', and 'Num'. The first row shows item 389 with keywords 'AthleteAthletics Intercollegiate AthleticsMales Men's Basketball MUA\_000001.tif SportsTerrell Schlundt' and a modified date of '2009-10-30T10:33:55Z'. The 'Column' column for this row is 'Athletes'. The second row shows 'Athletics', the third 'Intercollegiate Athletics', and the fourth 'Males'. The modal dialog is currently editing the 'Column' facet value for the first row.

Facet / Filter    Undo / Redo 28

Refresh    Reset All    Remove All

14268 records

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

Column change

6489 choices Sort by: name count    Cluster

Melvin Maceau 1

Melvin Mochalski 1

Melvin Peterson 1

Melvin Shimek 22

Memorial Library 62

Men's Basetball 1

Men's Baskeball 1

Men's Basketball 2820

Men's Basketballl 3

Men's Basketttball 1

Men's Baskteball 1

Item ID    Keywords    Column    Modified    Num

389    AthleteAthletics Intercollegiate AthleticsMales Men's Basketball MUA\_000001.tif SportsTerrell Schlundt    Athletes    2009-10-30T10:33:55Z

Athletics

Intercollegiate Athletics

Males

Men's Basketballl

Apply    Cancel

Enter    Esc

390    Avalancherebars Avalanchre MUA\_000002.tif Students    2010-04-27T14:29:50Z

# Open Refine – Clustering

- Clustering tool facilitates in finding groups of like entries
- Options available to employ a variety of similarity methods
  - Found that each method revealed some new and useful clusterings
  - Useful for correcting errors in spacing, capitalization, pluralization, and spelling

# Open Refine – Clustering

## Cluster & Edit column "Column"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. [Find out more ...](#)

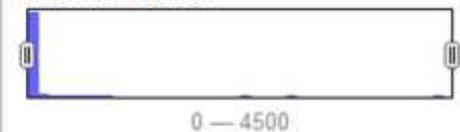
Method  Distance Function  Radius  Block Chars  187 clusters found

2	2	<ul style="list-style-type: none"><li>Glenn Elliott (1 rows)</li><li>Glen Elliott (1 rows)</li></ul>	<input type="checkbox"/>	Glenn Elliott
2	257	<ul style="list-style-type: none"><li>Classrooms (229 rows)</li><li>Classroom (28 rows)</li></ul>	<input type="checkbox"/>	Classrooms
2	11	<ul style="list-style-type: none"><li>Sandy Pavlic (10 rows)</li><li>Sandy Pavlie (1 rows)</li></ul>	<input type="checkbox"/>	Sandy Pavlic
2	17	<ul style="list-style-type: none"><li>Naval ROTC - Naval Science (14 rows)</li><li>Naval ROTC - Naval Science (3 rows)</li></ul>	<input type="checkbox"/>	Naval ROTC - Naval Scienc
2	167	<ul style="list-style-type: none"><li>Freshman Frontier Program (164 rows)</li><li>Freshmen Frontier Program (3 rows)</li></ul>	<input type="checkbox"/>	Freshman Frontier Program
2	104	<ul style="list-style-type: none"><li>Johnston Hall (103 rows)</li><li>Johnton Hall (1 rows)</li></ul>	<input type="checkbox"/>	Johnston Hall
2	17	<ul style="list-style-type: none"><li>M. Thomas Kolba (16 rows)</li><li>M.Thomas Kolba (1 rows)</li></ul>	<input type="checkbox"/>	M. Thomas Kolba
2	7	<ul style="list-style-type: none"><li>Lectures (6 rows)</li></ul>	<input type="checkbox"/>	Lectures

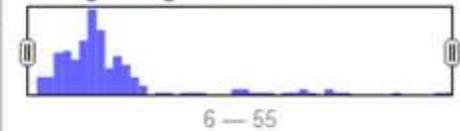
### # Choices in Cluster



### # Rows in Cluster



### Average Length of Choices



### Length Variance of Choices



Select All Deselect All

Merge Selected & Re-Cluster

Merge Selected & Close

Close

# Export for Ingest into CONTENTdm Project Client

The screenshot displays a software interface with a table and an open 'Export' menu. The table has columns for 'Horizontal Reso', 'Item ID', and file names. The 'Export' menu is open, showing various export options.

Horizontal Reso	Item ID	File Name
600	392	Campus MUA_000006.tif, Orientation Students
360	393	Campus MUA_000006.tif, Orientation Students
360	394	Campus MUA_000007.tif, Orientation

Export menu options:

- Export project
- Tab-separated value
- Comma-separated value
- HTML table
- Excel
- ODF spreadsheet
- Triple loader
- MQLWrite
- Custom tabular exporter...
- Templating...

# Updating Embedded Metadata



MARQUETTE  
UNIVERSITY

Be The Difference.

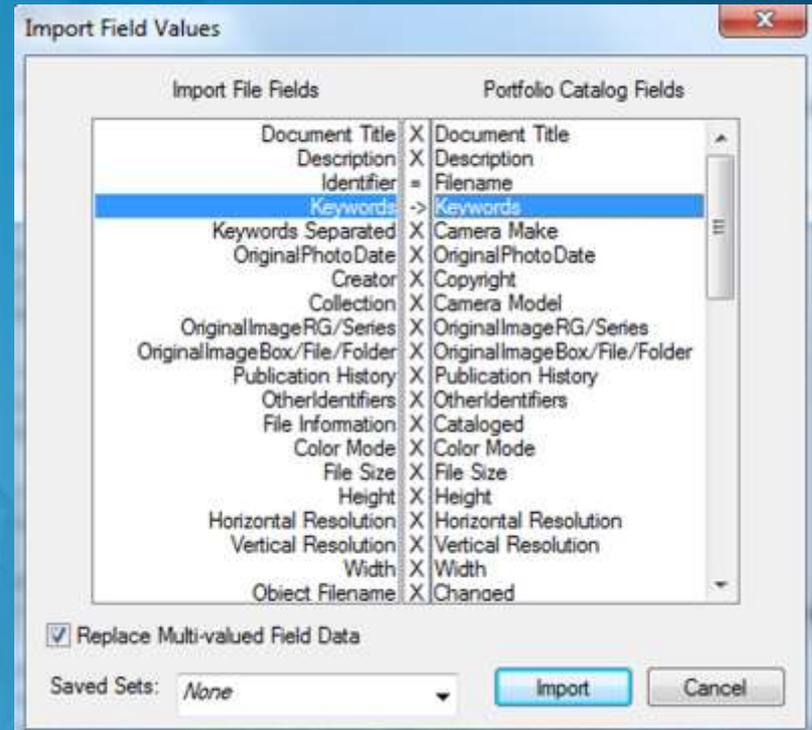


# Writing Metadata to Image Files

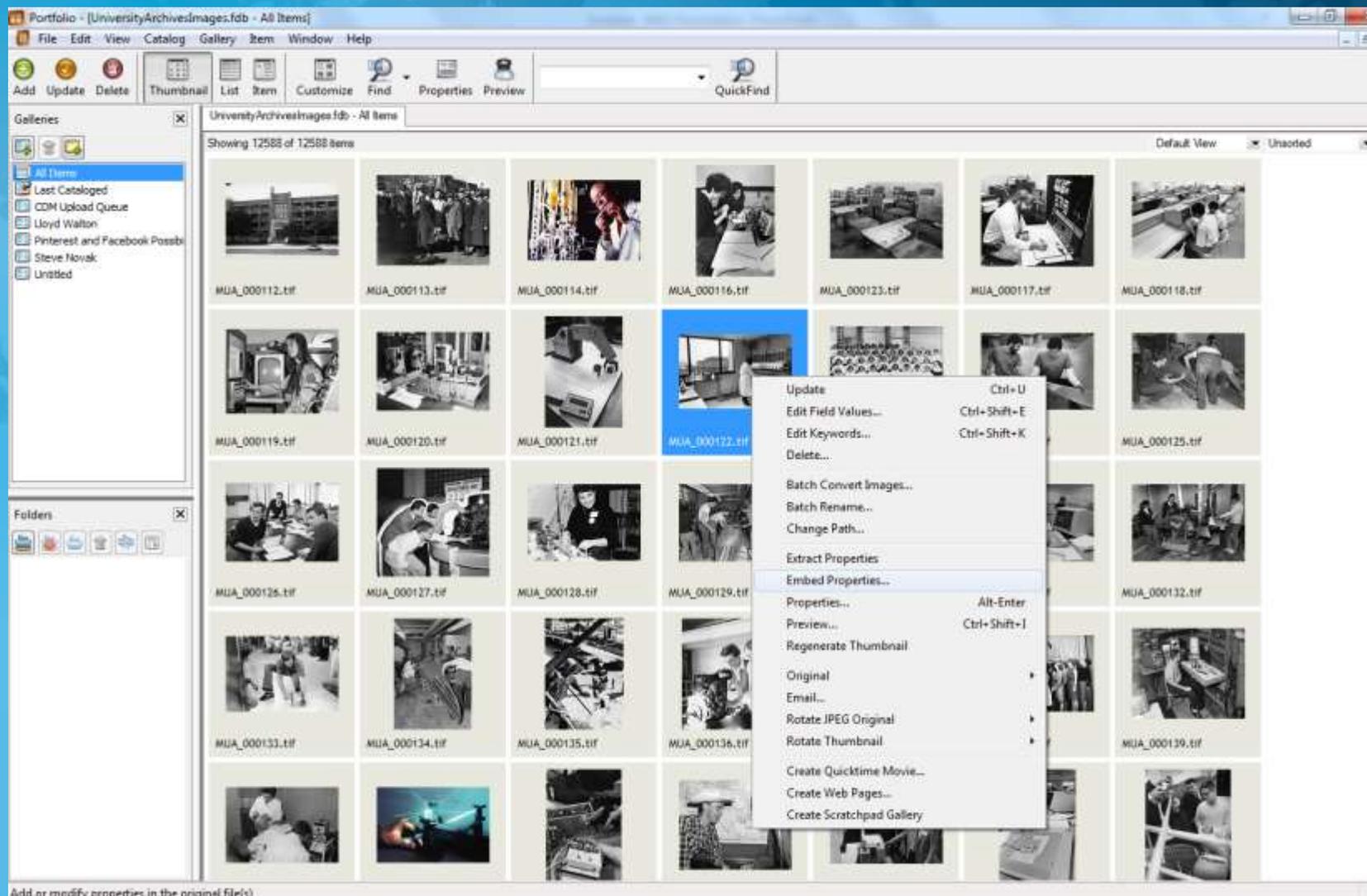
- Two strategies were explored
  - A combination of Portfolio's built in "Import Field Values" and "Embed Properties" functionalities
  - VRA Export-Import Plugin
- Tested both after making backup copy of the database and a sample set of image files

# Portfolio as Tool for Writing Metadata Changes

- Imports data from a plain text file
- Map fields in .txt file to those in Portfolio
- Select the check box to Replace Multi-valued Field Data
- Used Portfolio's Embed Properties feature to write the new data to linked files



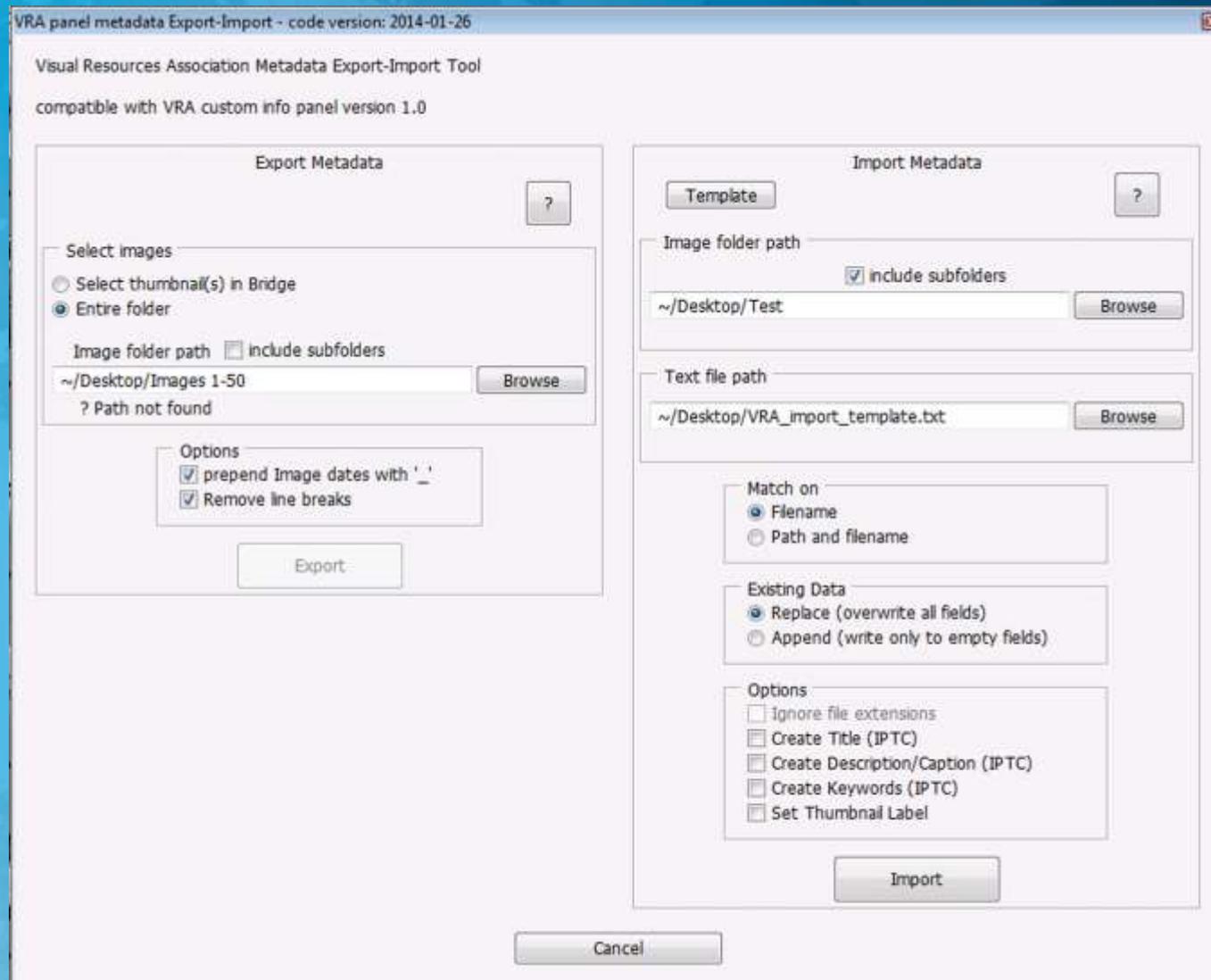
# Portfolio Embed Properties Feature



# VRA Export-Import Plugin

- Free JavaScript plugin for Adobe Bridge CS3 and higher
- Allows for the export of embedded metadata to .txt files and the import of metadata from .txt files onto a group of images
- Import requires
  - .txt file uses established column headers
  - replace and overwrite all fields or append data only to empty fields
- Easier to generate pre-formatted export and paste updated keyword information into appropriate field

# VRA Export-Import Plugin



# Takeaways

- Overall, process went smoothly, but bumps should be expected
  - Many answers to problems can be found online
  - Test your process with copies of files and in smaller numbers until you know it works
  - Know limits put in place by products you are using (e.g. 10,000 records in CONTENTdm at once)

# Takeaways

- If using Open Refine, only upload the fields you absolutely need and consider doing your analysis in batches
  - Trade-off between crashes and ability to facet, cluster, analyze all data at once
  - It can be difficult to interpret the results of clustering if you have a lot of names in your data

# Takeaways

- Maintaining Portfolio and CONTENTdm side by side has not been onerous; benefits outweigh costs
  - Workflow for mounting online embraces Portfolio first
  - Use CONTENTdm to catch vocabulary errors

# Thanks!



**MARQUETTE**  
UNIVERSITY

---

**Be The Difference.**

**Michelle Sweetser**  
**University Archivist**

[Michelle.sweetser@marquette.edu](mailto:Michelle.sweetser@marquette.edu)