Moving Digital Images

Michelle Sweetser

Marquette University

For over six years the Marquette University Archives managed patron-driven scanning requests using a desktop version of Extensis Portfolio while building thematically-based digital collections online using CONTENTdm. The purchase of a CONTENTdm license with an unlimited item limit allowed the department to move over 10,000 images previously cataloged in Portfolio into the online environment. While metadata in the Portfolio database could be exported to a text file and immediately imported into CONTENTdm's project client, we recognized that we had an opportunity to analyze and clean our metadata using OpenRefine as a part of the process. We also hoped to update our Portfolio database and the metadata embedded into the files themselves to reflect the results of this cleanup. This article will discuss the process we used to clean metadata in OpenRefine for ingest into CONTENTdm as well as the use of Portfolio and the VRA Panel Export-Import Tool for writing metadata changes back to the original image files.

Background – Portfolio Image Database

Beginning in 2008, the Marquette University Archives began to create, describe, and manage high resolution scans for patron-driven scanning requests. While a number of digital projects had been mounted online using CONTENTdm prior to that time, the department had treated patron-generated requests for scans as one-offs, discarding the files after a limited period of time. This resulted in a number of inefficiencies, most notably the repeated scanning of highly-requested images that did not fit the theme of our digital collection-building, scanned at whatever parameters were required by the patron, which may not have been sufficient for the needs of the next patron interested in them.

As the department sought a new approach to its patron-driven scanning, it made sense to turn to a workflow that embraced Extensis Portfolio. Across campus, the department had engaged in a number of conversations regarding the need for a digital asset management system. While resources were not available to support the purchase of a single networked system for use by multiple departments, several campus units were using single-license, desktop versions of Extensis Portfolio to manage their images. Among the offices using the product were the Office of Marketing and Communication and the university photographers in the Instructional Media Center. The University Archives had a copy of the software installed in the reading room to facilitate access to the images and metadata created by the university photographers. As a result, staff members were familiar with the product. Use of Portfolio would also benefit patrons who might seek resources from multiple departments in that they need only learn one interface to search for images (though they needed to go to each department separately to do so).

As the department embarked on the new workflow, images were scanned at specified resolutions based on a sliding scale determined by the size of the original image; title, description, keywords, and copyright information were embedded into the files using Photoshop; and the images were imported into Extensis Portfolio, where several other fields were recorded. As the campus users of Portfolio had met to discuss the need for a digital asset management system, the group developed a list of "Approved Photographic Keywords"[i] in late 2007, in anticipation of a day when our resources might be available and searchable in one repository. Partners agreed to make selections from this list when embedding metadata into digital images created by their units.

Initially conceptualized as a project to embed "quick and dirty" metadata that would largely rely on approved keywords for search and discovery, the cataloging workflow evolved over the ensuing years to encompass more complete descriptions of

the images in addition to the keywords, moving away from “quick and dirty” to reflect departmental practices for items catalogued in CONTENTdm collections.

Extensis discontinued Portfolio standalone sales at the end of 2013 and ceased technical support for the product on June 30, 2014. At this point in time, campus units continue to use the product but know that its life will be limited due to technological changes; there is a recognized need for a replacement but no specific plans have been put into place for adopting a new platform.

Background – CONTENTdm Collections

CONTENTdm has served as the department’s primary means of publishing images on the web; the desktop version of Portfolio does not include a web publishing component. All departmental CONTENTdm projects had been thematically based and the university archives had mounted several collections based on a variety of themes, including service activities, women’s athletics, Olympian Ralph Metcalfe, a complete run of student yearbooks, and two out-of-print university histories.[ii] When the department began cataloging and retaining patron-driven scans, there was concern among some that a collection built around patron requests would be too diverse and of little use or interest to the wider public. Furthermore, the department had a limited license to CONTENTdm and was approaching its item limit; other additions to the platform were seen as priorities.

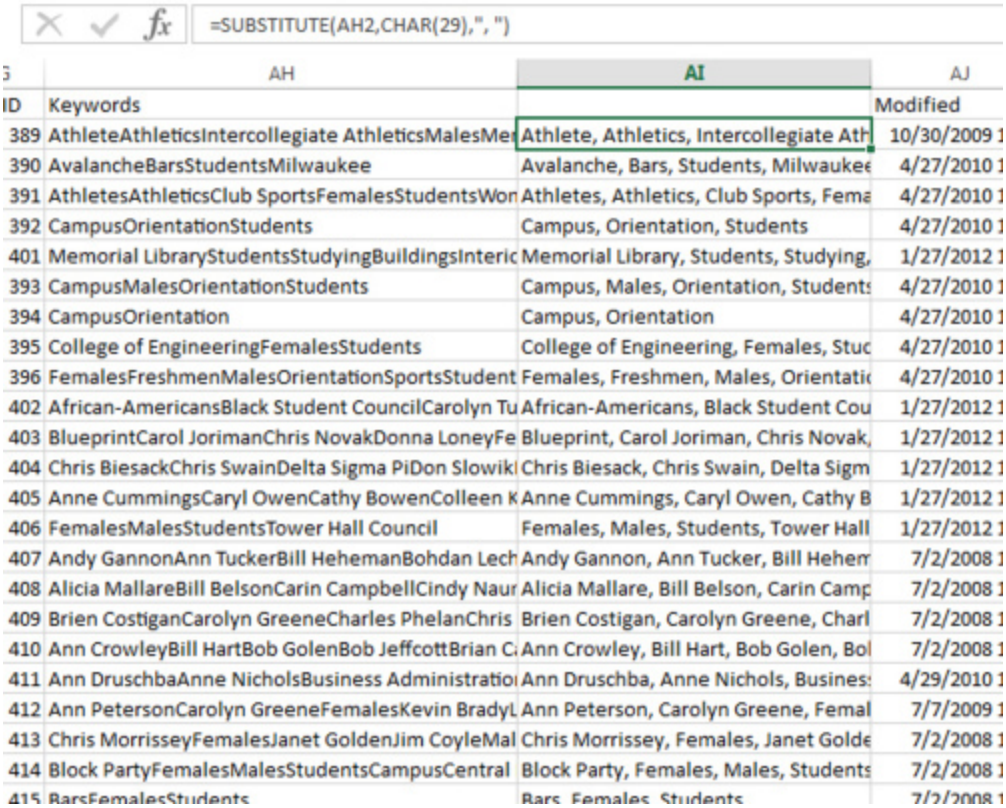
By early 2014, the department had scanned its 10,000th image for patron requests and the corpus had become a valuable resource for many requests of a general nature. The purchase of an unlimited license for CONTENTdm in late 2013 removed the remaining barrier to posting the images online in that venue. Knowing that some scanner-generated metadata would need to be massaged to conform to our standard practices for description in CONTENTdm, it seemed an opportune time to review and clean other metadata fields that had been generated by staff and student employees since the inception of the project.

Using OpenRefine to Clean Metadata

While the department worked from an approved keywords list, the master keyword list (controlled vocabulary) function was not employed within Portfolio because it was clunky and inconvenient to use. Instead, keywords were assigned within Adobe Photoshop, where it was possible for keystroking errors to result in misspellings. We had heard about OpenRefine (formerly GoogleRefine), a free open-source tool designed to assist in analyzing and cleaning data in bulk, but had not yet had an opportunity to use it.[iii] The program supports a variety of file formats, including CSV, Excel (.xls and .xlsx), XML, and RDF. Extensis Portfolio can export all field data as a text file, which paved the way for our use of OpenRefine.

Much basic massaging of data was accomplished via the use of string functions within Excel, for example concatenating the scanner-generated details on resolution with narrative text to generate our standard CONTENTdm entries related to file information. OpenRefine was primarily used to analyze, cluster, and edit keyword entries.

Initially, keywords appeared to have been concatenated during export, but upon further inspection, it was determined that they were separated by ASCII character 29, a control separator for separating a group. A simple substitution formula run in a new column quickly solved this issue.



	AH	AI	AJ
ID	Keywords		Modified
389	AthleteAthleticsIntercollegiate AthleticsMalesMe	Athlete, Athletics, Intercollegiate Ath	10/30/2009 1
390	AvalancheBarsStudentsMilwaukee	Avalanche, Bars, Students, Milwaukee	4/27/2010 1
391	AthletesAthleticsClub SportsFemalesStudentsWon	Athletes, Athletics, Club Sports, Fema	4/27/2010 1
392	CampusOrientationStudents	Campus, Orientation, Students	4/27/2010 1
401	Memorial LibraryStudentsStudyingBuildingsInteric	Memorial Library, Students, Studying,	1/27/2012 1
393	CampusMalesOrientationStudents	Campus, Males, Orientation, Students	4/27/2010 1
394	CampusOrientation	Campus, Orientation	4/27/2010 1
395	College of EngineeringFemalesStudents	College of Engineering, Females, Stuc	4/27/2010 1
396	FemalesFreshmenMalesOrientationSportsStudent	Females, Freshmen, Males, Orientatic	4/27/2010 1
402	African-AmericansBlack Student CouncilCarolyn Tu	African-Americans, Black Student Cou	1/27/2012 1
403	BlueprintCarol JorimanChris NovakDonna LoneyFe	Blueprint, Carol Joriman, Chris Novak,	1/27/2012 1
404	Chris BiesackChris SwainDelta Sigma PiDon Slowik	Chris Biesack, Chris Swain, Delta Sigm	1/27/2012 1
405	Anne CummingsCaryl OwenCathy BowenColleen K	Anne Cummings, Caryl Owen, Cathy B	1/27/2012 1
406	FemalesMalesStudentsTower Hall Council	Females, Males, Students, Tower Hall	1/27/2012 1
407	Andy GannonAnn TuckerBill HehemanBohdan Lech	Andy Gannon, Ann Tucker, Bill Hehem	7/2/2008 1
408	Alicia MallareBill BelsonCarin CampbellCindy Naur	Alicia Mallare, Bill Belson, Carin Camp	7/2/2008 1
409	Brien CostiganCarolyn GreeneCharles PhelanChris	Brien Costigan, Carolyn Greene, Charl	7/2/2008 1
410	Ann CrowleyBill HartBob GolenBob JeffcottBrian C	Ann Crowley, Bill Hart, Bob Golen, Bol	7/2/2008 1
411	Ann DruschbaAnne NicholsBusiness Administratio	Ann Druschba, Anne Nichols, Busines	4/29/2010 1
412	Ann PetersonCarolyn GreeneFemalesKevin BradyL	Ann Peterson, Carolyn Greene, Femal	7/7/2009 1
413	Chris MorrisseyFemalesJanet GoldenJim CoyleMal	Chris Morrissey, Females, Janet Golde	7/2/2008 1
414	Block PartyFemalesMalesStudentsCampusCentral	Block Party, Females, Males, Students	7/2/2008 1
415	BarsFemalesStudents	Bars, Females, Students	7/2/2008 1

Figure 1: Using a substitute formula to remove ASCII character 29 and insert a comma.

When the file is imported into OpenRefine, data is presented as rows. In order to analyze and edit the keywords field, the data must be atomized, or split into records. In the sample set shown in Figure 2, 25 rows are atomized into 341 records using built-in functionality of OpenRefine.

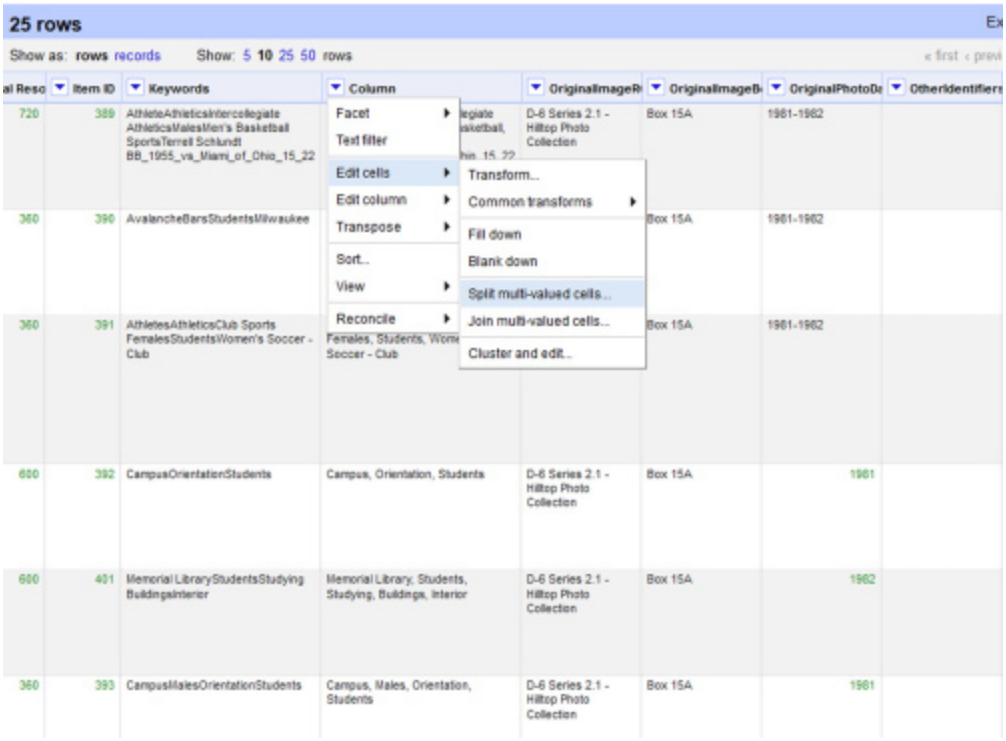


Figure 2: Using OpenRefine to atomize multi-valued cells in preparation for metadata analysis.

Once the keywords are atomized, the real power of OpenRefine can be brought to bear on metadata edit and cleanup through the use of facets, filters, and clusters. As this project focused on keywords, all efforts were targeted at one column of data, but multiple columns could easily be inspected using the same process. Using the built-in functionality illustrated in Figure 3, and by manually changing the number of unique values from the default limit of 2,000 unique terms, the metadata was successfully faceted into nearly 6,500 unique terms.

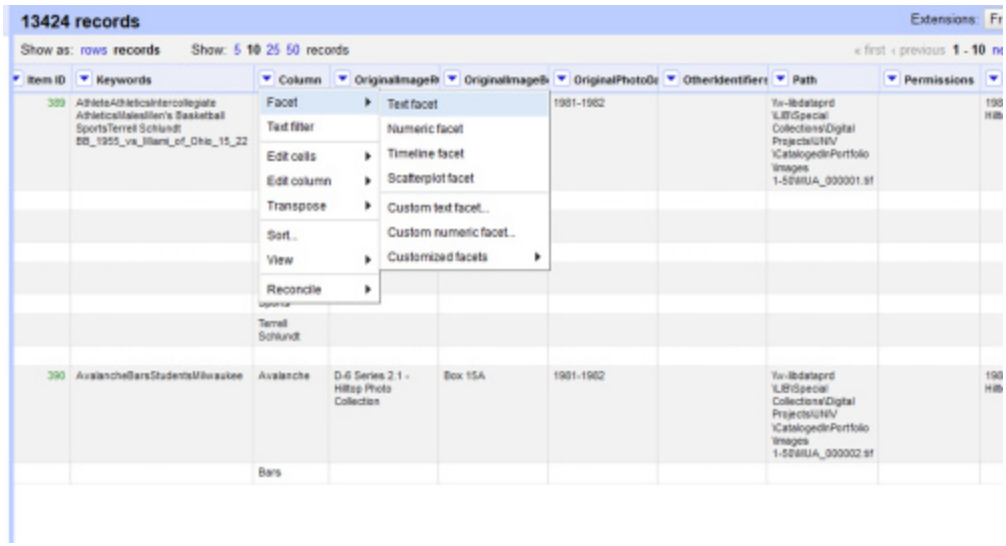


Figure 3: Using OpenRefine to facet a column for further analysis.

Facet results show up at a panel to the left on the screen and can be sorted by name (alphabetical) or by count. Scrolling through the resulting list of faceted terms, it was easy to identify some errors due to keystroking. Several file names had been accidentally embedded as metadata, and occasionally time stamp information crept in as well. Sometimes clusters of similar terms popped out as problem areas. As seen in Figure 4, there were no fewer than six different spellings of the term “men’s basketball” in our metadata.

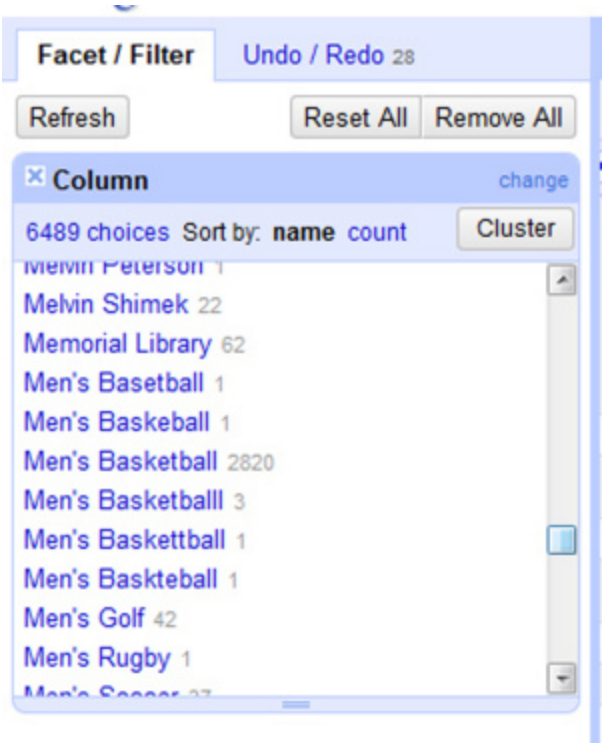


Figure 4: Facet Results upon Initial import.

Editing and correcting the errors is quite straightforward: by holding the cursor over the row in the panel, an option to edit it appears as a text link. Once that link has been selected, a text box appears as in Figure 5 to allow for editing of the term and the application of that correction to the appropriate field.

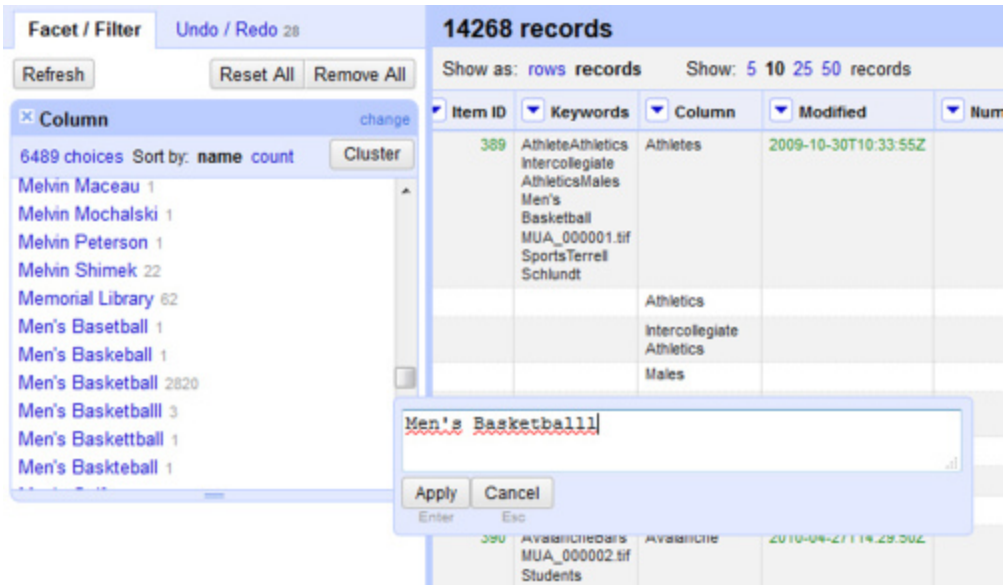


Figure 5: Editing terms in the facet view.

It is easiest to identify errors in the facet view when errors sort with alphabetic proximity. The cluster tool facilitates in identifying these clusters and locating other entries that are similar in some way; "this feature helps you find groups of different cell values that might be alternative representations of the same thing."[iv] A variety of similarity methods can be applied to the data; in this data set, each method applied discovered clusterings not previously revealed. As the data set included a large number of personal names, many clusterings were difficult to assess without external verification of the individuals depicted in the image and had to be ignored. Overall, however, our data was improved by the correction of spacing, capitalization, pluralization, and spelling differences.

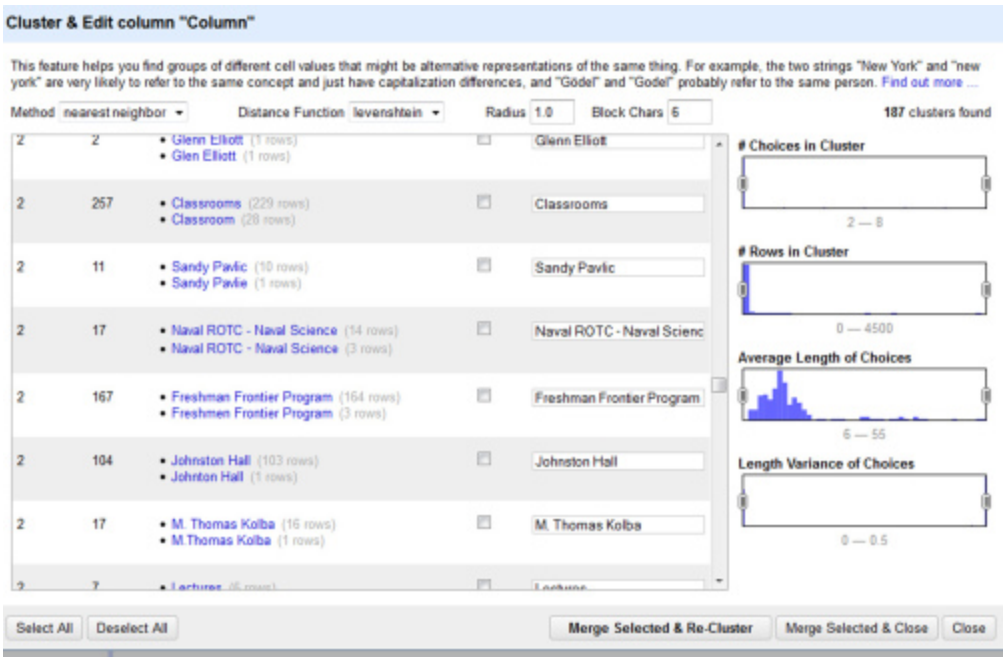


Figure 6: The clustering and editing feature shows issues with plurals, spacing, and general spelling.

Once satisfied with the results of the editing via the facet and clustering features, the data can be collapsed by joining the multi-valued rows using steps similar to those illustrated in Figure 2. We experienced some issues with this process, perhaps related to the fact that some of our records did not have values in the first few fields. Finally, the cleaned metadata can be exported to a local machine in a number of formats, including comma- and tab-separated values, Excel, and more, as seen in Figure 7.

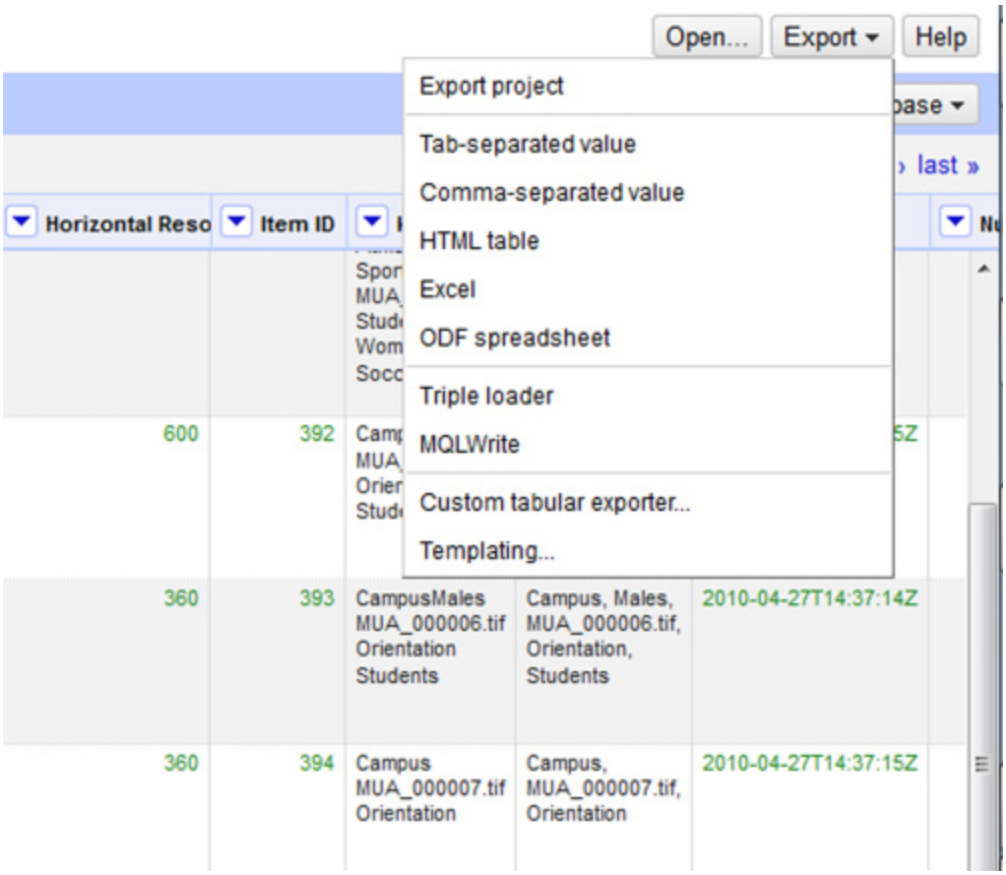


Figure 7: Options available for data export after cleaning.

Using Cleaned Metadata – Ingesting into CONTENTdm

At this point, the cleaned metadata spreadsheet was ingested into the Project Client for CONTENTdm using the spreadsheet method in the Project Client. After encountering a few minor challenges, including the discovery of a 10,000 record limit and several images in unusual formats (.psd and .eps files received from donors and added to the Portfolio database as a means by which to manage small accessions of digital files absent an electronic records management program), the process worked smoothly. The images are now discoverable and available for download from the Marquette University History Online CONTENTdm site[v], opening them up to a much wider audience than before.

Using Cleaned Metadata – Writing to Image Files

While CONTENTdm works well in making images discoverable online, it lacks at this time many basic features that facilitate the collection and delivery of images to patrons en masse. Images of interest must be downloaded one at a time. To request higher resolution versions, patrons must either record the reference URL or note the unique identifier for each desired image; staff members must then locate and collect files one at a time for delivery using outside mechanisms. All image metadata other than the title is lost upon download from CONTENTdm, so patrons must make note of the relationship between the image and the descriptive record or come up with their own way of capturing important information about their selections. Many regular internal users from the Office of Marketing and Communication and University Advancement do not look to CONTENTdm as their primary resource for image needs; instead, they make individual requests of archivists and ask for the delivery of a group of potentially acceptable images. This means that staff members experience these burdens on behalf of patrons when limited to CONTENTdm.

Compared to CONTENTdm, Extensis Portfolio and other digital asset management programs offer significant enhancements in the delivery of images to patrons. Patrons can flag multiple images of interest at one time and the program automatically tracks their selections for them. Built-in features allow for the automatic collection of selected image files from multiple locations on the LAN due to the network paths recorded at import. Metadata remains embedded in the files during transfer, and while we recognize that there are many opportunities for this metadata to be stripped from the file as it is used and changed by patrons, we have found that embedded metadata has at times made our jobs easier when patrons have changed file names and returned to us later for assistance in identifying the image (at times, years have lapsed). For users who are unable to access embedded metadata or don't know how, field information for groups of selected images can be exported from Portfolio as a .txt file and can be provided to patrons as a separate file. As archivists who deal with internal

clients who frequently request upwards of 50 images at one time, these features are a boon; it is for these reasons that we continue to use the Portfolio database as a part of our workflow in delivering images to internal users who make high-volume requests.

In order to continue to use Portfolio as a part of our workflow, the metadata in the database as well as that embedded in the files stored on our LAN needed to be updated to reflect the changes made in OpenRefine. Two strategies were explored for the process of writing the metadata to the files and importing it to the database: a combination of Portfolio's built in "Import Field Values" and "Embed Properties" functionalities, as well as the VRA Export-Import Plugin. As a precaution, a backup copy of the database was made prior to testing so that any errors or lessons learned during the process could be corrected without loss of data. Similarly, a group of sample image files were copied to a test area on the computer to examine the process of writing metadata to files.

Writing Metadata to Image Files Using Portfolio

Portfolio allows for the import of data from a plain text file, a clean version of which was created in Excel after an export from OpenRefine. As the keyword field was the only one modified in OpenRefine, the import was a simple process of keying on the identifier field (a unique field in our naming system) and importing only the data for the keyword field. Portfolio provides an interface for mapping field import values, as illustrated in Figure 8. It was necessary to select the check box to Replace Multi-valued Field Data so that the new values overwrote the values in the database instead of being appended to the list, lest we perpetuate the keywording issues we sought to fix.

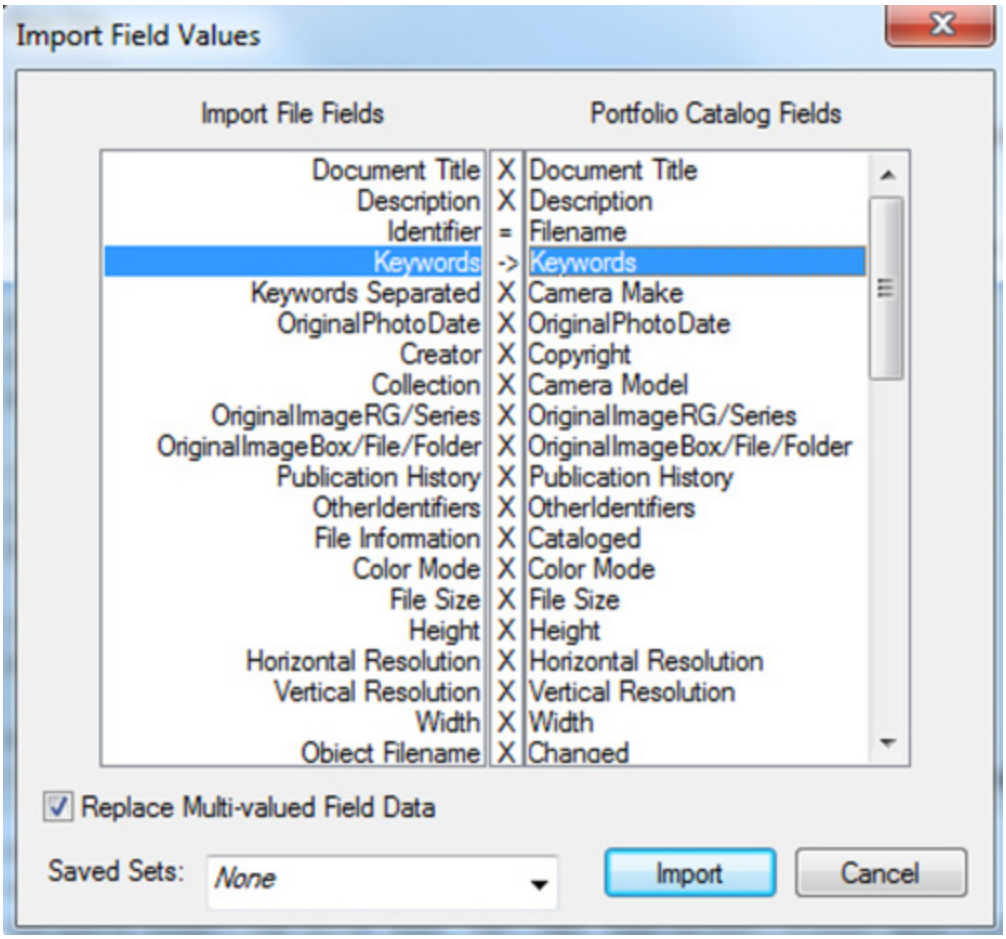


Figure 8: Mapping field values to be imported in Portfolio.

At this point, the process of writing the changes from the database was quite simple as well: by selecting all of the images in the database and clicking on Portfolio's embed properties feature. Given the number of files to be written to, this process took a bit of time, but worked without incident.

Writing Metadata to Image Files Using the VRA Panel Export-Import Tool

While the Portfolio tools worked well for the process, it was only at the eleventh hour that we realized they were available to us, after having experimented with the VRA Panel Export-Import Tool as a solution as well. The tool is a free JavaScript plugin for Adobe Bridge CS3 and higher that allows for the export of embedded metadata to tab-delimited text files and the import of metadata from tab-delimited files into a group of

images[vi] (see Figure 9). For import, the tool requires the use of a tab delimited text file with established column headers. The only options for handling the data are to replace and overwrite all fields (meaning every column must to be complete or data will be lost) or to append the data only to empty fields. Neither option fit our needs exactly, where just one field needed to be updated. While a downloadable metadata template with required column headers is available from within the plug-in, the data to complete it had to be cut and pasted from the file generated by our Portfolio export and OpenRefine cleanup, which used different column names and an alternate column order. This added step seemed unnecessarily cumbersome. Instead, it was easier to generate a pre-formatted export for all of the images using the export functionality of the tool and to paste the updated keyword information into the one appropriate field before writing importing the metadata back onto the files.

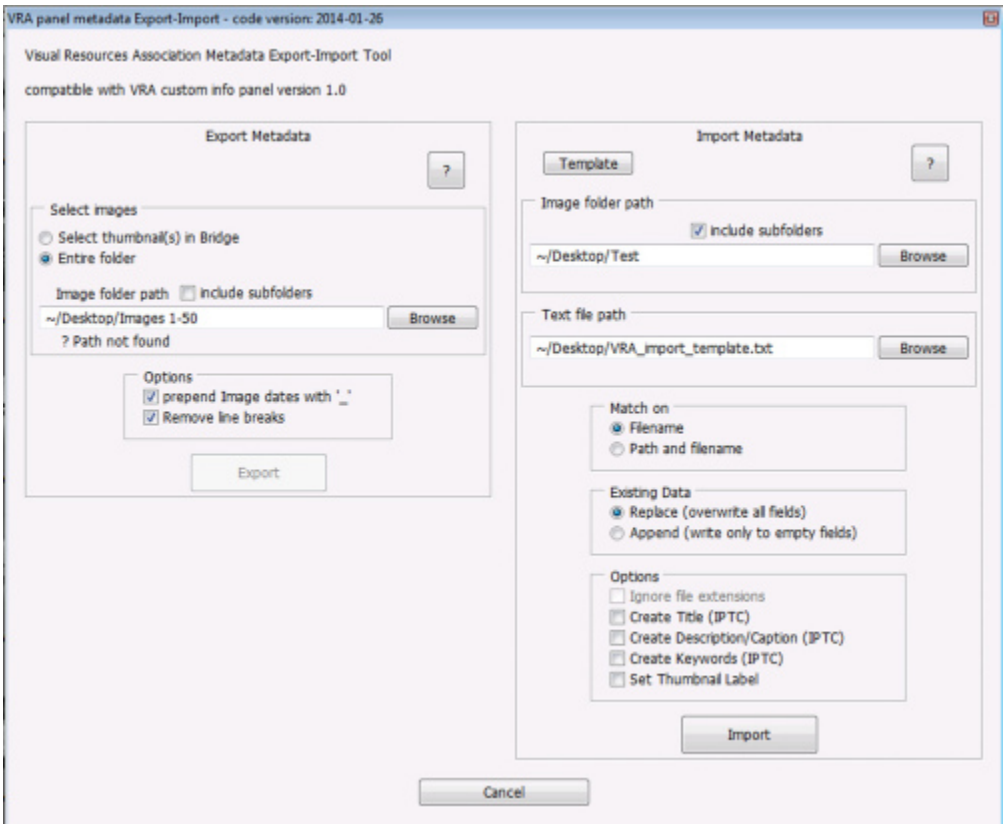


Figure 9: The VRA panel metadata Export-Import tool for Adobe Bridge.

As we do not yet know how long the Portfolio database will continue to run on our computers, nor do we know whether a campus-wide digital asset management software will become a reality, we are pleased that there is a second workable option for writing batch metadata edits back to the metadata embedded in our files and that it works with relative ease. Use of the VRA panel to embed metadata to the DC Title, Summary Description/Caption, Summary Keywords, and Headline fields and does not require our patrons to download the VRA panel to view. These fields can viewed when opening the image in Photoshop and are also visible in the properties panel in Windows Explorer.

Most of the images scanned for our early CONTENTdm collections do not have embedded metadata and have not been imported into the Portfolio system. In the future, it is possible that we will use the Export-Import tool to write metadata batch exported from CONTENTdm back to our archival masters to facilitate delivery of caption-related information to internal users who make large requests.

Conclusion

As with any project that moves data from one system to another, there are challenges along the way and the process inevitably consumes more time than one might expect. However, we found OpenRefine to be simple to install and to integrate into our workflow for CONTENTdm projects, given that we frequently employ the spreadsheet method for import. Lacking the presence of a controlled vocabulary in CONTENTdm and the existence of thousands of keywords, OpenRefine gave us a means to correct keystroking errors before establishing the controlled vocabulary we now use for our additions to the collection. Other organizations with data in uncontrolled fields will find the tool powerful and scalable in meeting their needs for data analysis and cleaning. Similarly, organizations with access to Adobe Bridge and a desire to write metadata to files are likely to find the VRA Export-Import tool a satisfying solution.

Finally, while Portfolio Standalone is no longer supported by Extensis, our smooth experience in reading and writing metadata from within the product means we will look for any future replacement to have similar functionality. At this point, a team has not yet been brought together to search for a solution and a timeline has not been set for doing so. While CONTENTdm lags in many enhancements that would aid in the delivery of high-resolution files to patrons, it remains the primary means by which we can provide online access to our collections at this time.

About the Author

Michelle Sweetser is the University Archivist in the Department of Special Collections and University Archives at Marquette University. Her work encompasses the acquisition, appraisal, description, digitization, and preservation of materials documenting the history of the university. She holds an AB in anthropology from Dartmouth College and an MSI from the University of Michigan.

Notes:

[i] See the most current version of the keyword list at http://www.marquette.edu/library/archives/documents/UAKeywords10-22-12.pdf.

[ii] See all of the department's digital collections at http://digitalmarquette.cdmhost.com/.

[iii] To download the software and to read documentation, visit http://openrefine.org/.

[iv] "Clustering in Depth." Accessed November 14, 2014. https://github.com/OpenRefine/OpenRefine/wiki/Clustering-In-Depth.

[v] The site is accessible at http://cdm16280.contentdm.oclc.org/cdm/landingpage/collection/p16280coll1.

[vi] The tool is available for download at http://metadatadeluxe.pbworks.com/w/page/48025141/VRA%20Panel%20Export-Import%20Tool. A number of short videos posted to the site provide useful instruction in the basics of using the tool.