INSTRUCTIONS FOR CREATING DATA SETS IN EXCEL FOR JOURNALS IN BEPRESS DIGITAL COMMONS

Introduction

Moving law school journals to an open source publishing model using a digital commons (DC) is a no-brainer. However, entering all of the metadata for the backfile of these journals into the DC can be a daunting and very time consuming task. This usually involves typing or copy/pasting metadata about each article (title, authors, subjects, page number, etc.) into a template which is then uploaded.

This project is an extension of a presentation given at CALI 2012 in which I described a way to automate the process for gathering and uploading backfile metadata using spreadsheets. Santa Clara University Law School uses the Bepress digital commons, so theses instructions are written for that platform. However, it should be easy to modify the process for other platforms. Basically, the steps are:

- collect metadata from different sources into a set of spreadsheets
- parse the metadata into the appropriate fields
- collect and organize the parsed metadata into a final spreadsheet ready for uploading
- upload a spreadsheet for each article

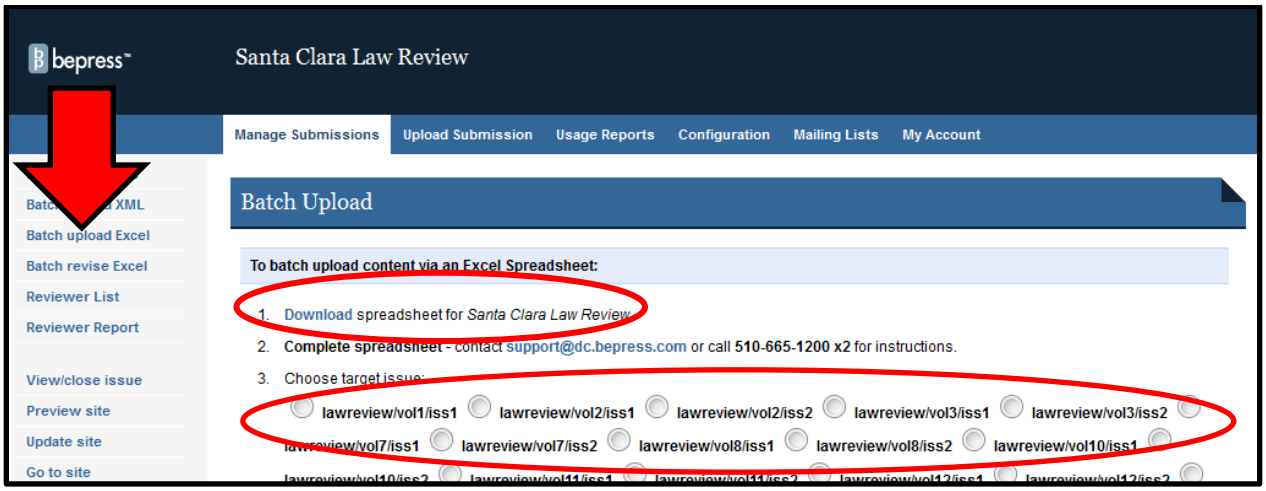
Using this process you should be able to collect all the metadata for a journal into a spreadsheet in one day and begin uploading the spreadsheets to the DC the next.

The Excel File

This process relies on Excel functions to parse metadata collected from various source. The BigKahuna workbook contains all of the functions used to create a spreadsheet ready for uploading to the digital commons for all three Santa Clara University Law School journals. All you need to do is collect the metadata for a desired journal (this process is described below) and paste that metadata into the corresponding columns in the BigKahuna spreadsheets. The included formulas will then parse the data and create a final spreadsheet.

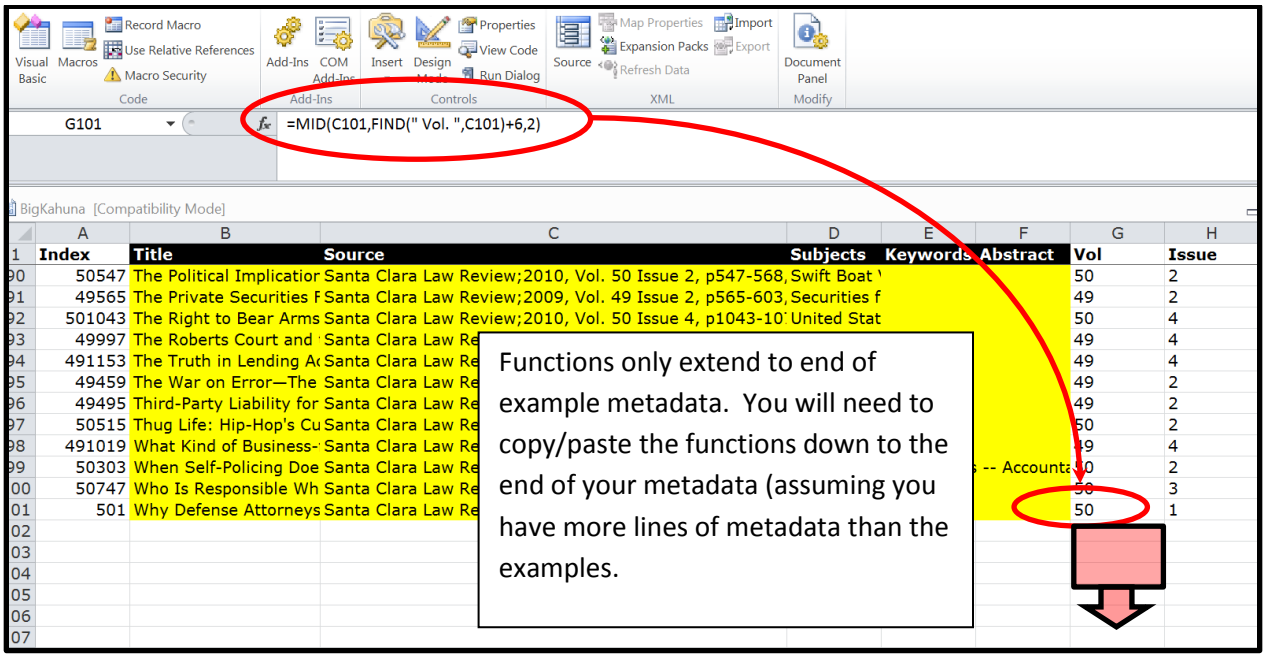
The BigKahuna Excel workbook contains three spreadsheets for metadata gathered from HeinOnline, Index to Legal Periodicals and Books, and Dropbox. The workbook also contains two additional spreadsheets – one to collate data and the other is the final spreadsheet. It is saved as an Excel 97-2003 workbook, which is the format that (currently) works with the Bepress DC.

The columns headers in the workbook represent fairly standard fields used to describe a journal article. However, it is possible that other schools will have more, less, or different metadata, depending on the design of their DC journal page. At some point you will need to download your spreadsheet from the DC to get the correct column headers for your journal.



Your column headers will need to be substituted for the column headers in the original Vlookup and Final spreadsheets of the BigKahuna workbook. The spreadsheets can then easily be modified to meet different metadata needs by copying/pasting formulas from the original columns into your matching columns. Column headers in red represent required fields. You will also need to create the issues for your journal within the DC.

When you first download the BigKahuna workbook you will see metadata in the yellow columns. This is provided as an example of data collected from HeinOnline, ILPB, and Dropbox. Your metadata should be pasted over the existing data (in the yellow fields). The functions are only copied down to the last line of the sample metadata. You will need to copy/paste the functions to the last line of your metadata.



COLLECTING THE DATA

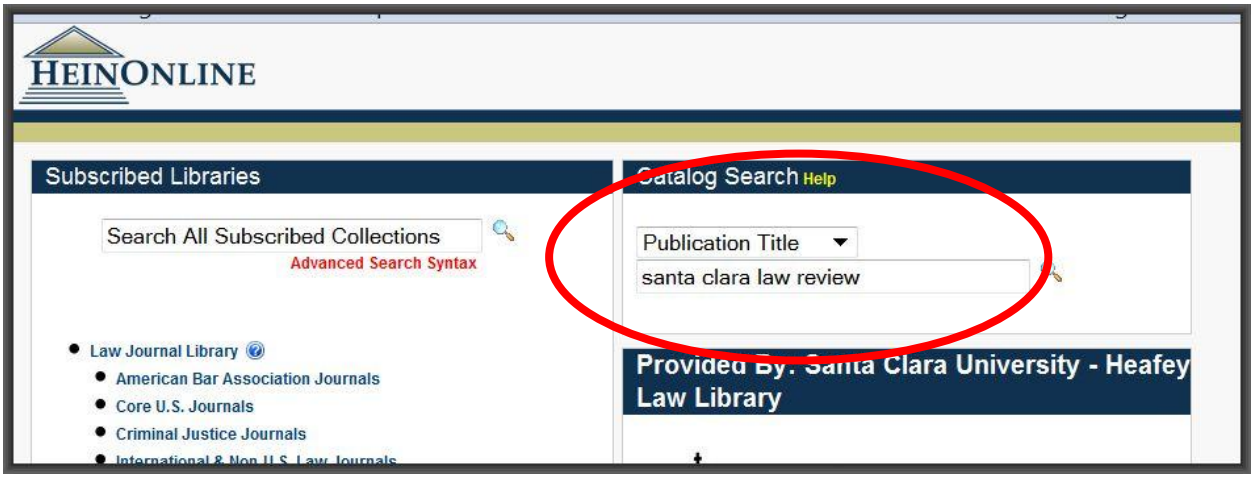
Outwit Hub

This process makes use of a screen scraper. I use OutWit Hub (http://www.outwit.com/products/hub/), which is an add-on for Firefox (a standalone version may also be available). You will be writing your own scrapers, so the professional edition is necessary (in 2011 a single license cost $60.00). Other low cost screen scapers are available, but OutWit Hub was the only one used in this process.

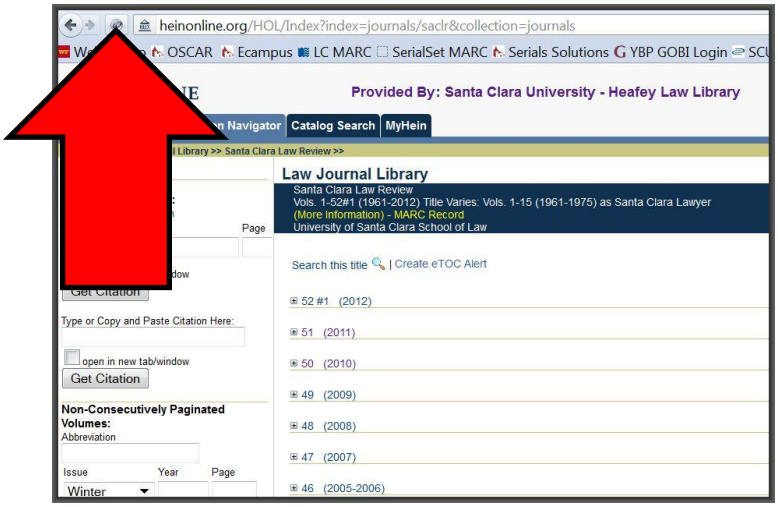
A screen scraper is used because of its ability to organize metadata on a web site(s) into a spreadsheet like format. The metadata can then be downloaded directly into an Excel file.

Scraping HeinOnline Metadata

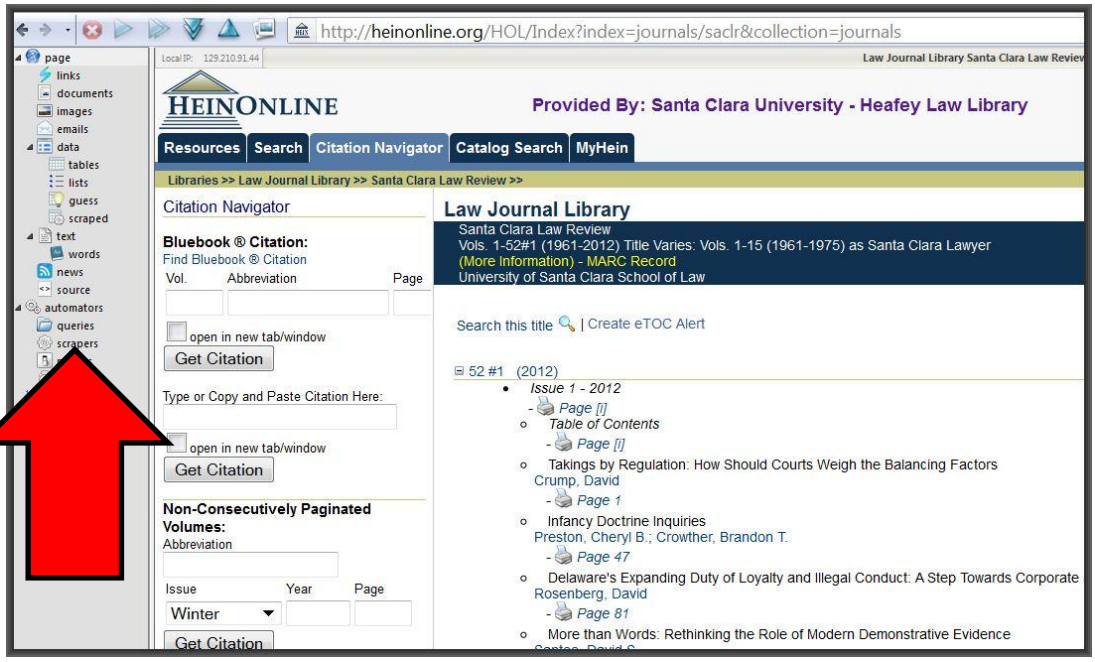
At the Heinonline site, look up your journal title:



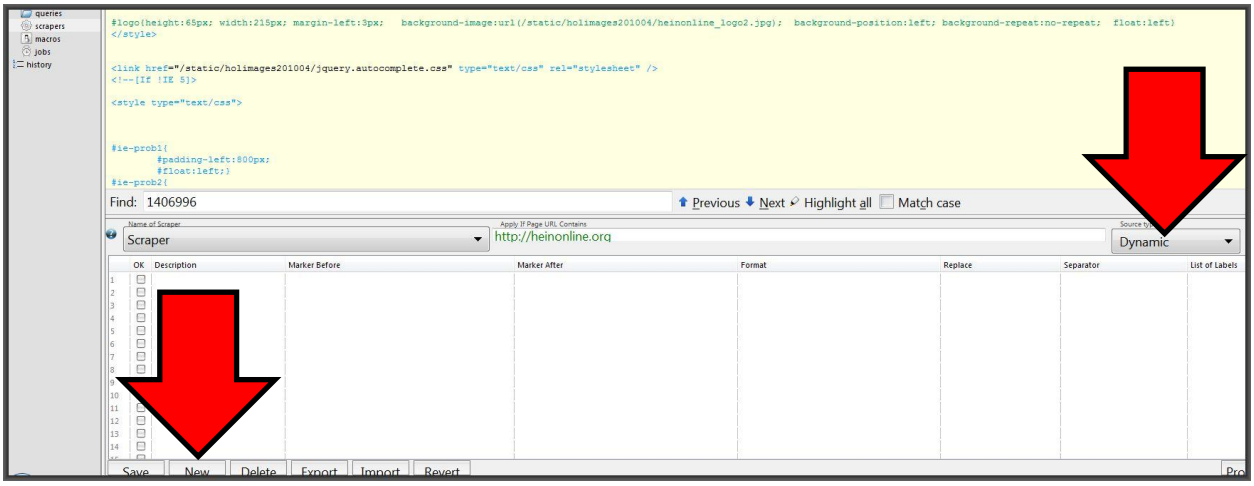
Next, open Outwit Hub:



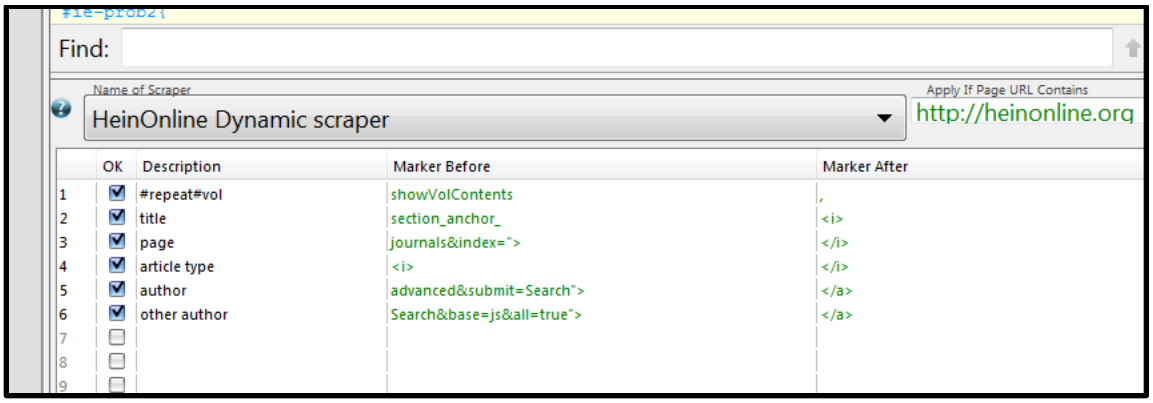
Outwit will open a new screen with the same web page displayed. On this new screen, click all of the plus marks ⊞ to expose the metadata for the entire journal:



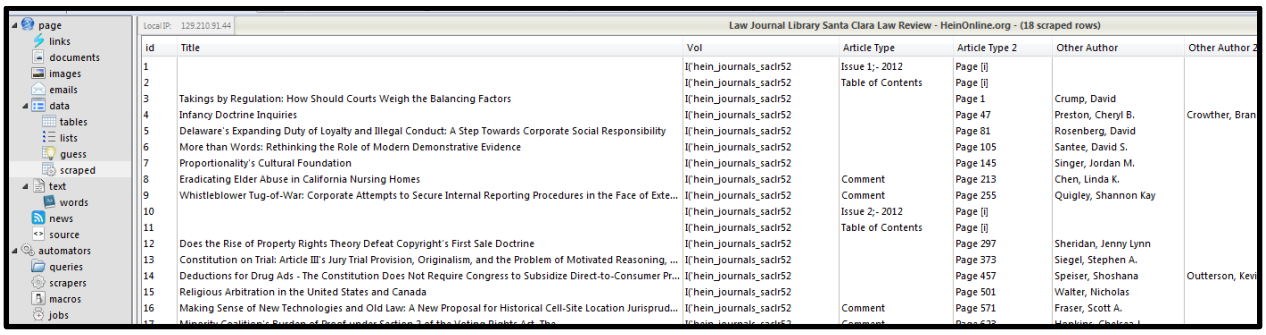
Now make a scraper to scrape the desired information from this page. First click in “scrapers” (see above). A screen like the one below will open. On that screen make sure that “dynamic” is selected, then select “new” to start a new scraper.



Name the scraper HeinOnline (or anything you want) and click OK. Enter the information below in your scraper and then select “save”.



Then run the scraper by selecting "execute". You will see a screen similar to:



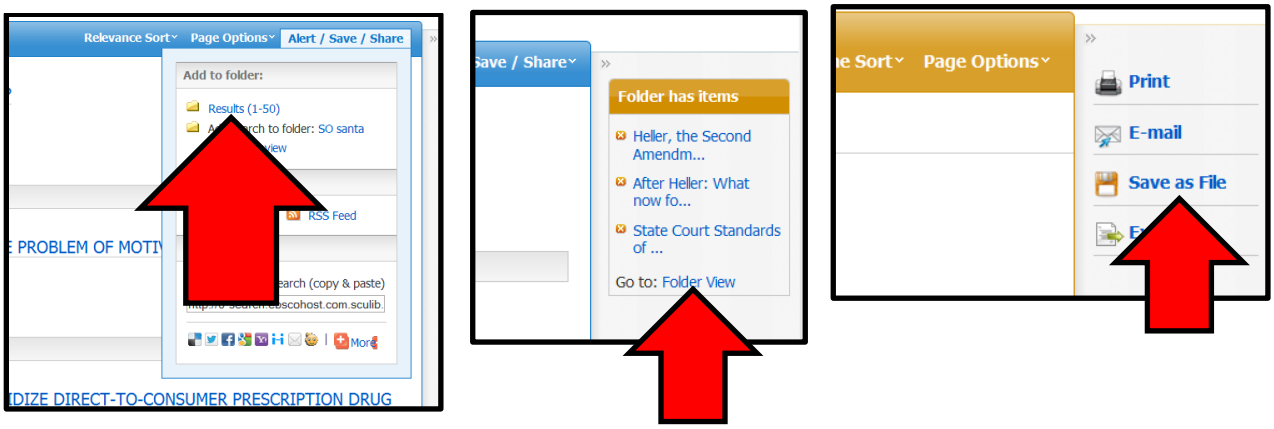
As you can see, the page number may display in the "Article Type 2" column. That is OK, you can rename the column at some point if you want. Also, there is trash information in the Article Type column (issue, table of contents, etc.). That is fine, the information will not be transferred to the final spreadsheet.

In the lower right portion of the screen you will see an "export" button. Export the file in an Excel format. Then, open the file in excel and copy/paste the metadata to the HeinOnline spreadsheet in the BigKahuna workbook.

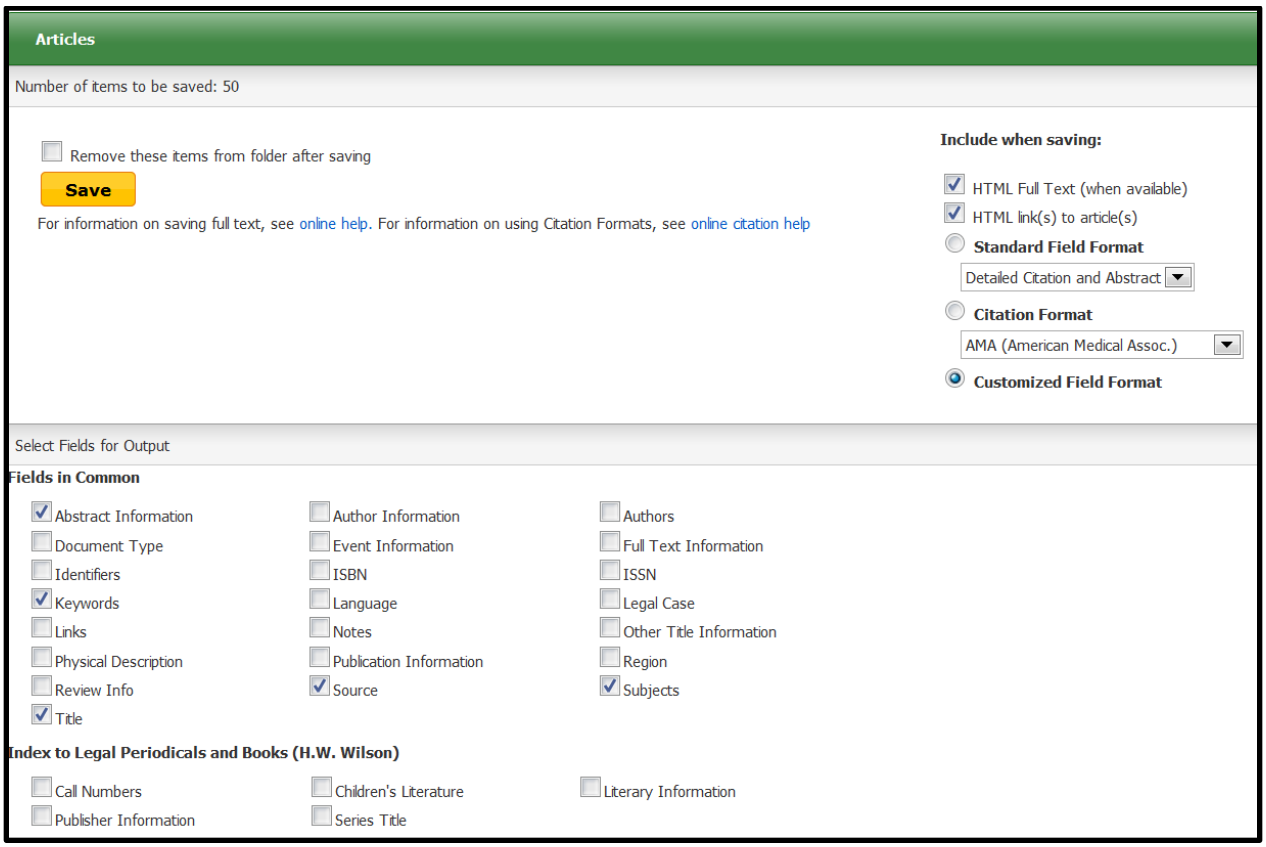
Scraping Index to Legal Periodicals and Books Metadata

Santa Clara University collected subject, keyword, and abstract metadata from ILPB. The university subscribes to ILPB via EbscoHost, so these directions will be specific to that platform. However, the scraper can be modified as necessary to scrape ILPB metadata from any platform.

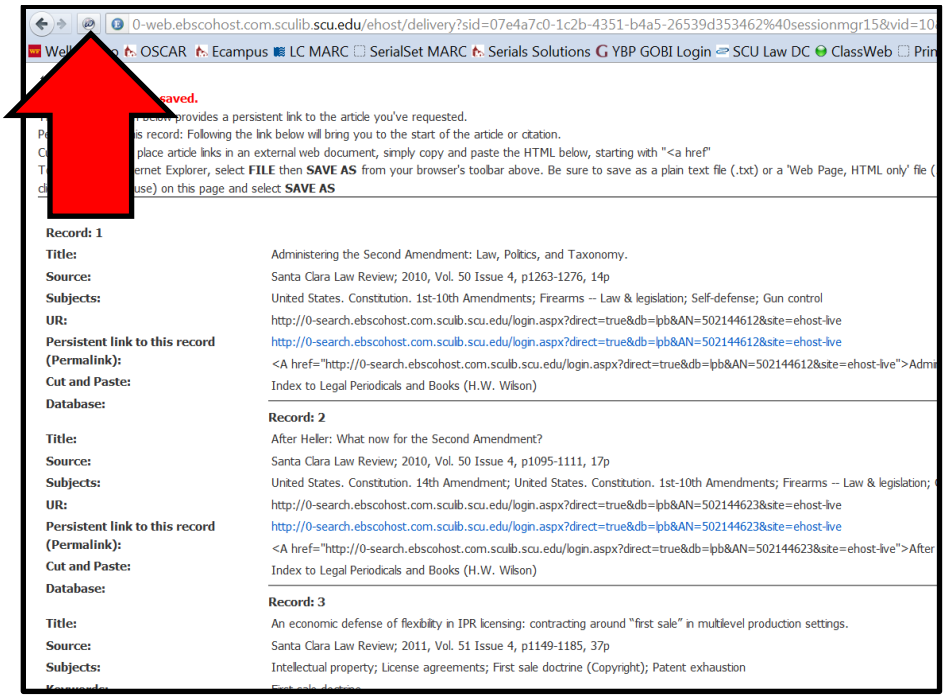
Open the ILPB website and search for your journal title as "source". Then, under "Page Options", choose to display the maximum results per page (for SCU that was 50). Next, under "Alert/Save/Share", choose to add results to a folder and then view the folder:



Next, choose “Customized field format”, select the five fields below, and “save”:

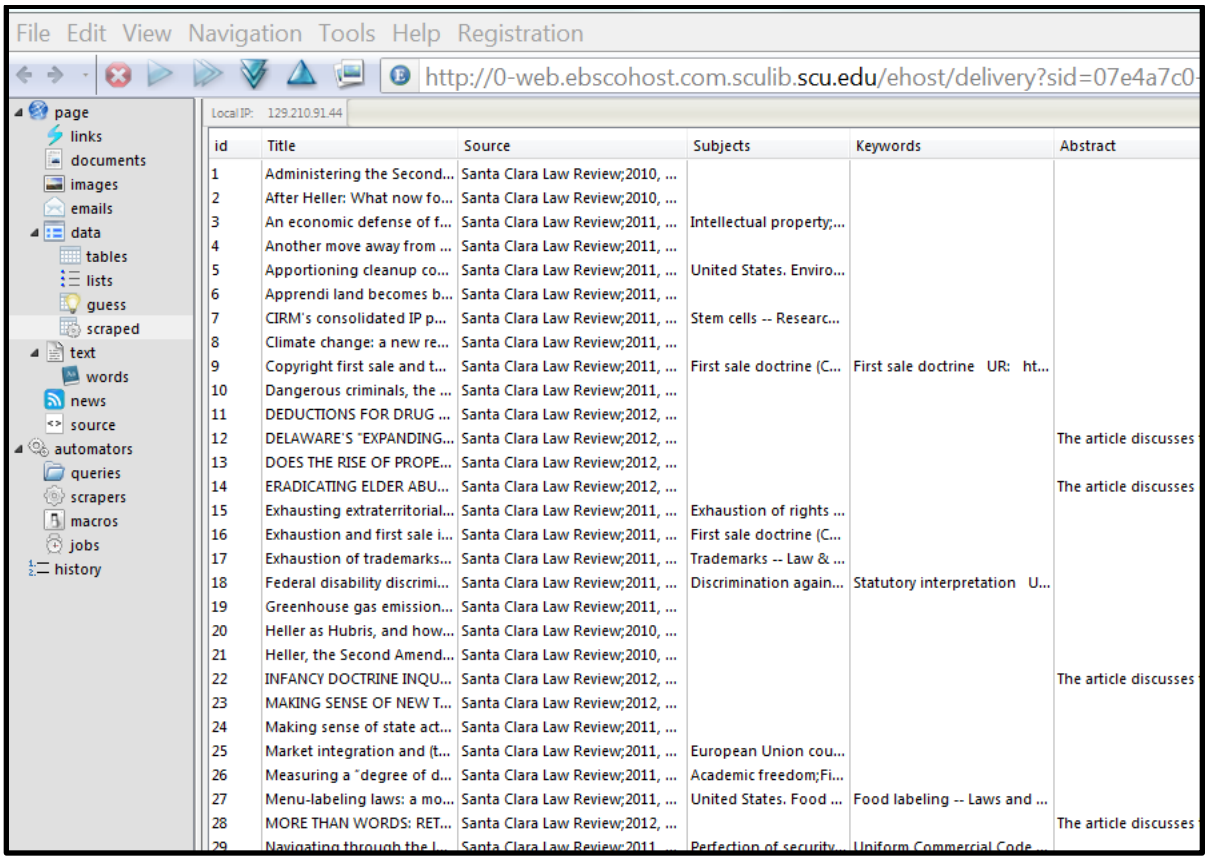


You should see a screen similar to the one below. From here, launch Outwit Hub:



In the Outwit Hub screen, select "scraper", then "new" to create a new scraper. Populate the scraper with the following fields and "save" then "execute".

Once you have executed the scraper you should see a screen similar to this:



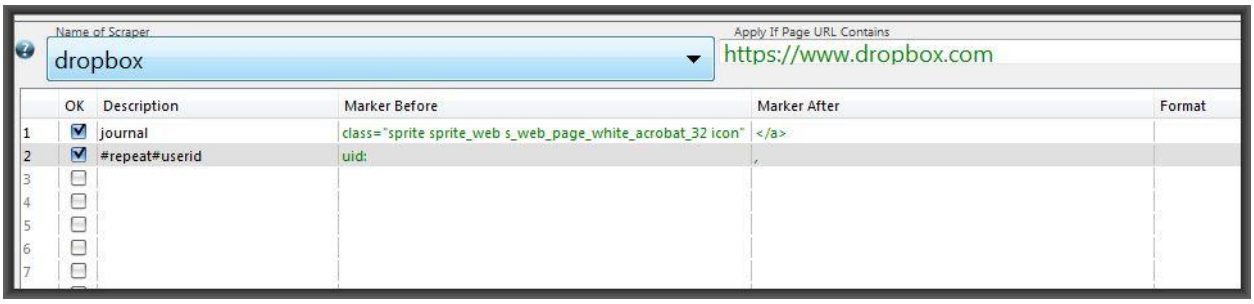
Export the data as an Excel file. Open the Excel file and copy/paste the 5 columns of data into the matching columns of the ILPB worksheet in the BigKahuna workbook.

You will need to repeat this process until all metadata is collected from the ILPB database.

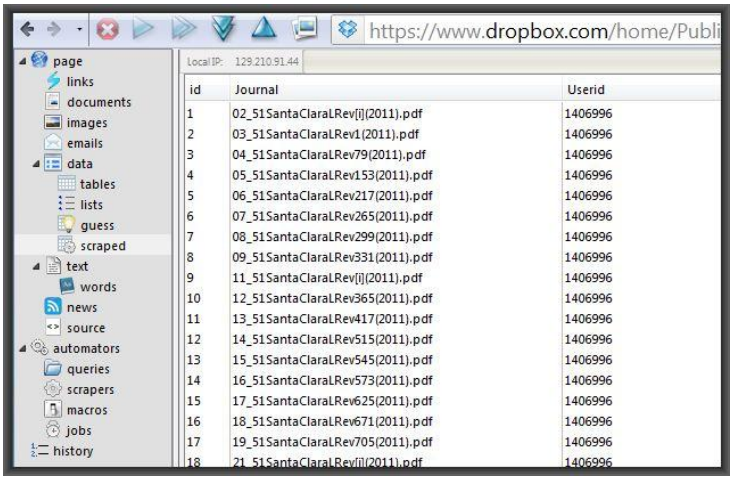
Scraping data from a Dropbox Public File

Uploading metadata to the DC via spreadsheets requires a publicly available URL for each article. We uploaded the PDF files purchased from HeinOnline to a public folder in Dropbox. There are many file hosting services available. If you choose to use a file hosting service other than Dropbox you will need to adjust the spreadsheet functions accordingly.

Once files are uploaded the metadata can be scraped using this small scraper:



The scraped metadata should look something like this:

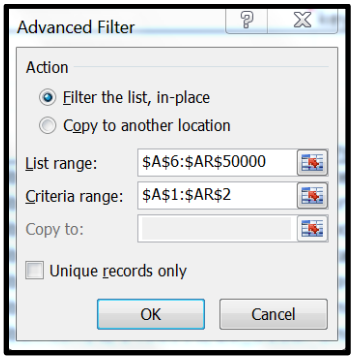


Save the metadata as an excel file and copy/paste the *Journal* and *Userid* columns into the Dropbox spreadsheet of the BigKahuna workbook.

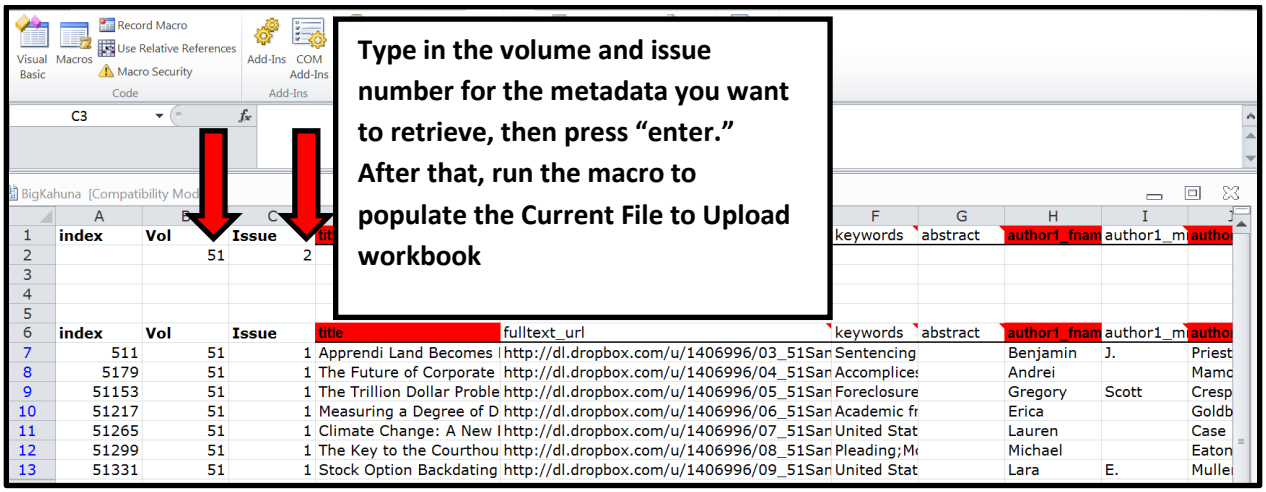
Creating the Final Spreadsheet

The vlookup spreadsheet of the BigKahuna workbook should self-populate with the parsed data from the other three spreadsheets. The Final spreadsheet is created simply be copying the vlookup spreadsheet and pasting the ***values*** into the Final spreadsheet beginning in cell A5. Make sure you paste the values, not the formulas.

Once the values are pasted into the Final spreadsheet it is helpful to create an advanced filter to make a spreadsheet for a single issue. Instructions for creating an advanced filter are in the Excel help file. The one for this project is:



A macro titled Sub UploadFile() should be included as a workbook module. The macro is designed to copy the metadata about a single issue to a stand-alone spreadsheet titled “Current File to Upload”. You define the single issue by typing in the volume and issue number in the advanced filter portion of the Final spreadsheet.



	index	Vol	Issue	title	fulltext_url	keywords	abstract	author1_fnam	author1_m	author
1	index	Vol	Issue	tit		keywords	abstract	author1_fnam	author1_m	author
2		51	2							
3										
4										
5										
6	index	Vol	Issue	title	fulltext_url	keywords	abstract	author1_fnam	author1_m	author
7	511	51	1	Apprendi Land Becomes	http://dl.dropbox.com/u/1406996/03_51San	Sentencing		Benjamin	J.	Priest
8	5179	51	1	The Future of Corporate	http://dl.dropbox.com/u/1406996/04_51San	Accomplices		Andrei		Mamc
9	51153	51	1	The Trillion Dollar Proble	http://dl.dropbox.com/u/1406996/05_51San	Foreclosure		Gregory	Scott	Cresp
10	51217	51	1	Measuring a Degree of D	http://dl.dropbox.com/u/1406996/06_51San	Academic fr		Erica		Goldb
11	51265	51	1	Climate Change: A New I	http://dl.dropbox.com/u/1406996/07_51San	United Stat		Lauren		Case
12	51299	51	1	The Key to the Courthou	http://dl.dropbox.com/u/1406996/08_51San	Pleading;Mc		Michael		Eaton
13	51331	51	1	Stock Option Backdating	http://dl.dropbox.com/u/1406996/09_51San	United Stat		Lara	E.	Mulle

Before using the macro, you will first need to create a blank spreadsheet and save it as an Excel 97-2003 workbook with the title “Current File to Upload”; make sure you save it in the same folder as the BigKahuna workbook.

Once the Current File to Upload workbook has been populated by the macro it is ready to be uploaded to the DC.

If the macro is not in the BigKahuna workbook, paste the following text into a macro in the Visual Basic editor. You will need to make sure the “Developer” tab is accessible in Excel to create a macro.

```
Sub UploadFile()


'

' This macro is used in combination with the advanced filter.  Just change the

'    either the volume or issue in the advanced filter, then run this macro.  The

'    filtered information will be copied from the main spreadsheet (BigKahuna) to

'    another worksheet titled Current File to Upload and the file will be saved.

'    The Current File to Upload should then be uploaded to the Digital Commons.

' Its is most helpful if you assign this macro to a key so that all you have to do is
```

```
'    press the assigned key to create a new Current File to Upload spreadsheet.



' Run the advanced filter

    Range("A6:AR50000").AdvancedFilter Action:=xlFilterInPlace, CriteriaRange:= _
       Range("A1:AR2"), Unique:=False

' Lets get rid of the existing metadata in Current File to Upload
    Workbooks.Open Filename:= _
       ThisWorkbook.Path & "\Current File to Upload.xls"
    Cells.Select
    Selection.ClearContents
    Range("A1").Select

' Back to the main spreadsheet to copy the issue metadata

    Windows("BigKahuna.xls").Activate
    ActiveWindow.WindowState = xlNormal
    ActiveWindow.WindowState = xlNormal
    Rows("6:6").Select
    Range(Selection, Selection.End(xlDown)).Select
    Selection.Copy

' Now paste the issue metadata into Current File to Upload

    Windows("Current File to Upload.xls").Activate
    ActiveWindow.WindowState = xlNormal
    ActiveWindow.WindowState = xlNormal
    Range("A1").Select
```

```
    ActiveSheet.Paste


' Delete the unnecessary columns of metadata and save the file
    Columns("A:C").Select
    Selection.Delete Shift:=xlToLeft
    Range("A1").Select

    Application.CutCopyMode = False
    ActiveWorkbook.Save

    'Back to main file to enter another vol/number and begin again

    Windows("BigKahuna.xls").Activate
    ActiveWindow.WindowState = xlNormal
    ActiveWindow.WindowState = xlNormal
    Range("B2").Select
End Sub
```

That's it. I hope you find the process helpful and the steps outlined in this instruction document easy to follow. I also hope that the functions in the BigKahuna spreadsheet work for your metadata as well as they work for the metadata collected for Santa Clara University's law journals. If you are collecting the metadata from the same sources, the functions should work fine.

If you need help with any part of the process described above, or if you should need help tweaking the functions for some reason, please do not hesitate to contact me at:

Whitney Alexander
Director of Technical Services
Heafey Law Library
Santa Clara University
Santa Clara, CA 95056

Phone: 408-554-2733

Fax: 408-554-5318
walexander@scu.edu