

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 these 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.

bepress™ Santa Clara Law Review

Manage Submissions Upload Submission Usage Reports Configuration Mailing Lists My Account

Batch Upload

To batch upload content via an Excel Spreadsheet:

1. Download spreadsheet for Santa Clara Law Review
2. Complete spreadsheet - contact support@dc.bepress.com or call 510-665-1200 x2 for instructions.
3. Choose target issue:

lawreview/vol1/iss1
 lawreview/vol2/iss1
 lawreview/vol2/iss2
 lawreview/vol3/iss1
 lawreview/vol3/iss2
 lawreview/vol7/iss1
 lawreview/vol7/iss2
 lawreview/vol8/iss1
 lawreview/vol8/iss2
 lawreview/vol10/iss1
 lawreview/vol10/iss2
 lawreview/vol14/iss1
 lawreview/vol14/iss2
 lawreview/vol12/iss1
 lawreview/vol12/iss2

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.

Record Macro Use Relative References Add-Ins COM Add-Ins Insert Design Run Dialog Properties View Code Map Properties Expansion Packs Import Export Document Panel Modify

G101 **=MID(C101,FIND(" Vol. ",C101)+6,2)**

BigKahuna [Compatibility Mode]

Index	Title	Source	Subjects	Keywords	Abstract	Vol	Issue
90	50547	The Political Implicatio	Santa Clara Law Review;2010, Vol. 50 Issue 2, p547-568, Swift Boat			50	2
91	49565	The Private Securities F	Santa Clara Law Review;2009, Vol. 49 Issue 2, p565-603, Securities f			49	2
92	501043	The Right to Bear Arms	Santa Clara Law Review;2010, Vol. 50 Issue 4, p1043-10	United Stat		50	4
93	49997	The Roberts Court and	Santa Clara Law Re			49	4
94	491153	The Truth in Lending A	Santa Clara Law Re			49	4
95	49459	The War on Error—The	Santa Clara Law Re			49	2
96	49495	Third-Party Liability for	Santa Clara Law Re			49	2
97	50515	Thug Life: Hip-Hop's Cu	Santa Clara Law Re			50	2
98	491019	What Kind of Business-	Santa Clara Law Re			49	4
99	50303	When Self-Policing Doe	Santa Clara Law Re			50	2
00	50747	Who Is Responsible Wh	Santa Clara Law Re			50	3
01	501	Why Defense Attorneys	Santa Clara Law Re			50	1
02							
03							
04							
05							
06							
07							

Functions only extend to end of example metadata. You will need to copy/paste the functions down to the end of your metadata (assuming you have more lines of metadata than the examples).

COLLECTING THE DATA

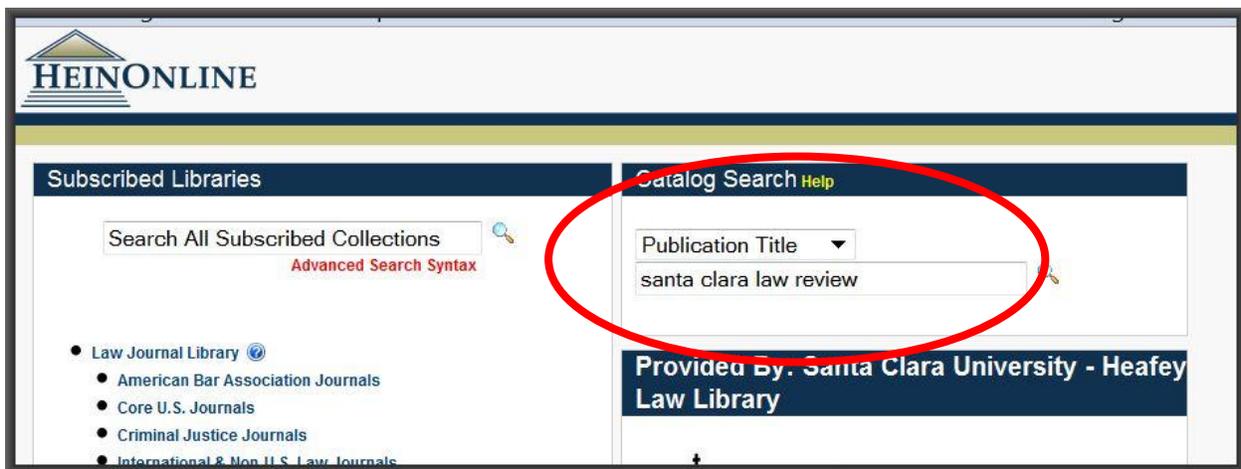
Outwit Hub

This process makes use of a screen scraper. I use [OutWit Hub](http://www.outwit.com/products/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 scrapers are available, but OutWit Hub was the only one used in this process.

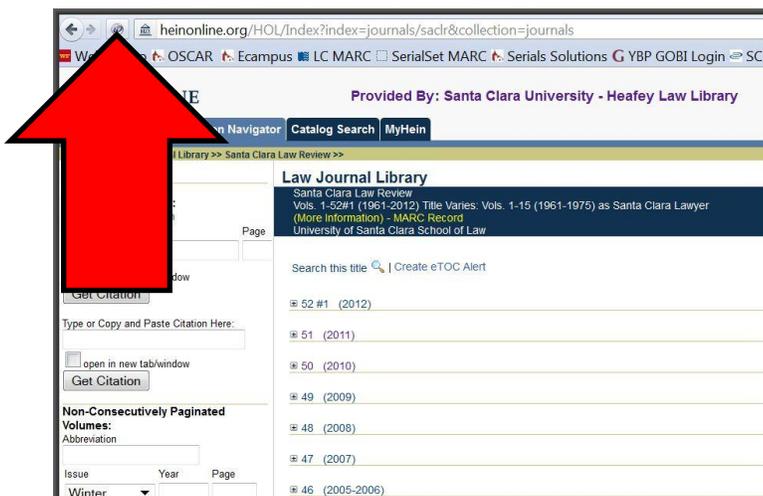
A screen scraper is used because of its ability to organize metadata on a web site(s) into a spreadsheet like format. The metadata can then be downloaded directly into an Excel file.

Scraping HeinOnline Metadata

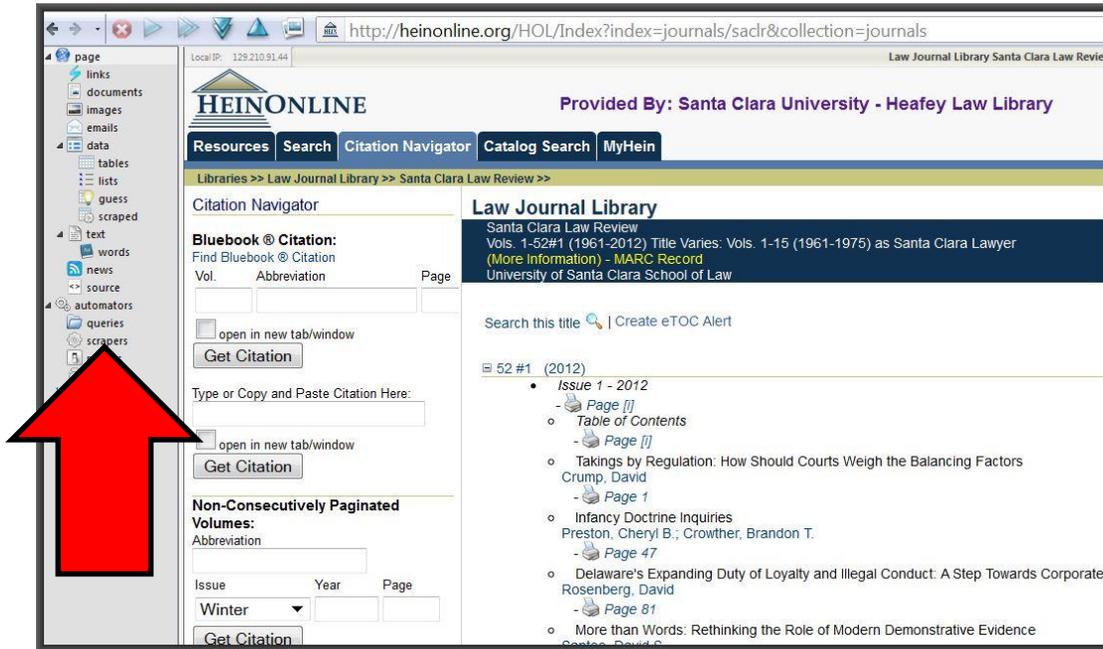
At the Heinonline site, look up your journal title:



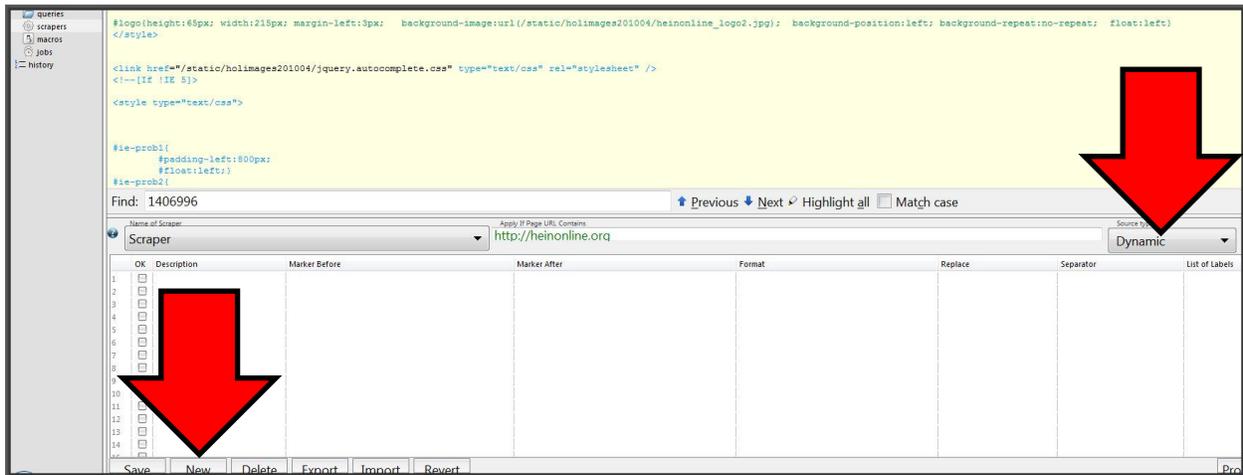
Next, open Outwit Hub:



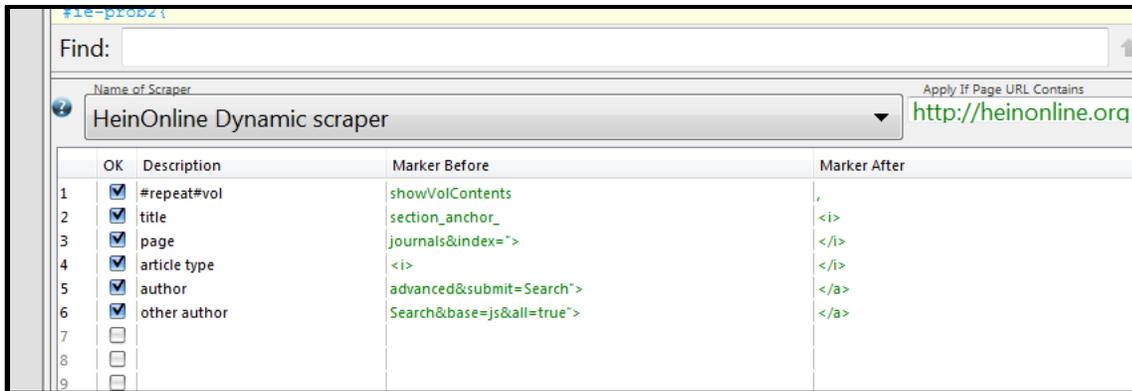
Outwit will open a new screen with the same web page displayed. On this new screen, click all of the plus marks  to expose the metadata for the entire journal:



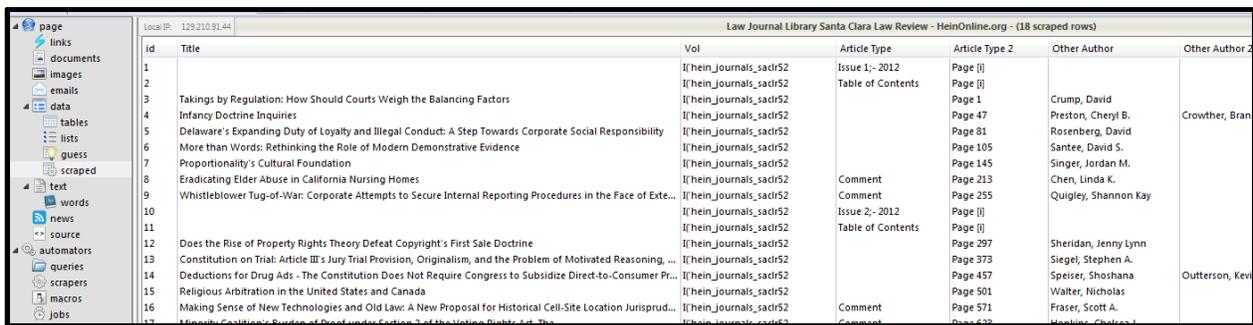
Now make a scraper to scrape the desired information from this page. First click in "scrapers" (see above). A blue like the one below will open. On that screen make sure that "dynamic" is selected, then select "new" to start a new scraper.



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:



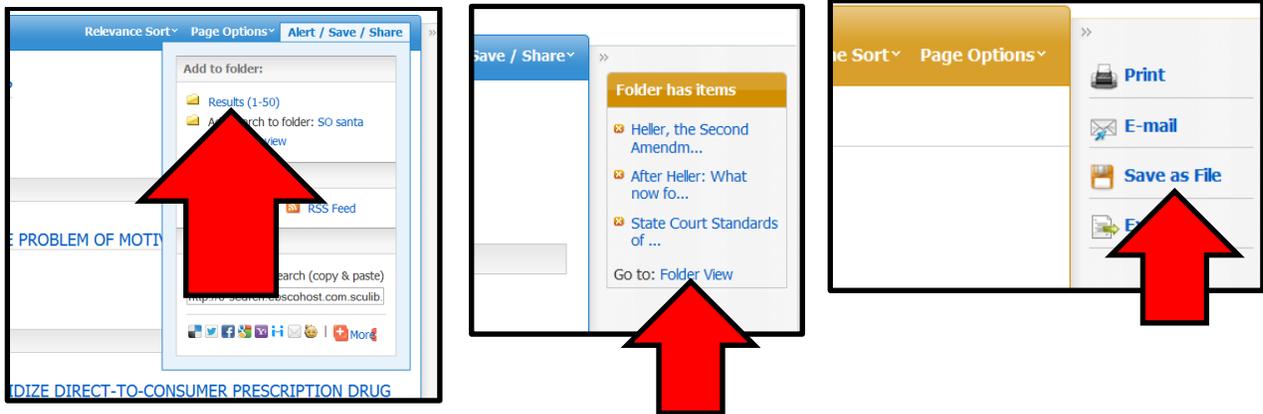
As you can see, the page number may display in the “Article Type 2” column. That is OK, you can rename the column at some point if you want. Also, there is trash information in the Article Type column (issue, table of contents, etc.). That is fine, the information will not be transferred to the final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

Articles

Number of items to be saved: 50

Remove these items from folder after saving

Save

For information on saving full text, see [online help](#). For information on using Citation Formats, see [online citation help](#)

Include when saving:

- HTML Full Text (when available)
- HTML link(s) to article(s)
- Standard Field Format**
Detailed Citation and Abstract
- Citation Format**
AMA (American Medical Assoc.)
- Customized Field Format**

Select Fields for Output

Fields in Common

<input checked="" type="checkbox"/> Abstract Information	<input type="checkbox"/> Author Information	<input type="checkbox"/> Authors
<input type="checkbox"/> Document Type	<input type="checkbox"/> Event Information	<input type="checkbox"/> Full Text Information
<input type="checkbox"/> Identifiers	<input type="checkbox"/> ISBN	<input type="checkbox"/> ISSN
<input checked="" type="checkbox"/> Keywords	<input type="checkbox"/> Language	<input type="checkbox"/> Legal Case
<input type="checkbox"/> Links	<input type="checkbox"/> Notes	<input type="checkbox"/> Other Title Information
<input type="checkbox"/> Physical Description	<input type="checkbox"/> Publication Information	<input type="checkbox"/> Region
<input type="checkbox"/> Review Info	<input checked="" type="checkbox"/> Source	<input checked="" type="checkbox"/> Subjects
<input checked="" type="checkbox"/> Title		

Index to Legal Periodicals and Books (H.W. Wilson)

<input type="checkbox"/> Call Numbers	<input type="checkbox"/> Children's Literature	<input type="checkbox"/> Literary Information
<input type="checkbox"/> Publisher Information	<input type="checkbox"/> Series Title	

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

0-web.ebscohost.com.sculib.scu.edu/ehost/delivery?sid=07e4a7c0-1c2b-4351-b4a5-26539d353462%40sessionmgr15&vid=10

saved.

provides a persistent link to the article you've requested.

is record: Following the link below will bring you to the start of the article or citation.

place article links in an external web document, simply copy and paste the HTML below, starting with "<a href"

Internet Explorer, select **FILE** then **SAVE AS** from your browser's toolbar above. Be sure to save as a plain text file (.txt) or a 'Web Page, HTML only' file (use) on this page and select **SAVE AS**

Record: 1

Title: Administering the Second Amendment: Law, Politics, and Taxonomy.

Source: Santa Clara Law Review; 2010, Vol. 50 Issue 4, p1263-1276, 14p

Subjects: United States. Constitution. 1st-10th Amendments; Firearms -- Law & legislation; Self-defense; Gun control

UR: <http://0-search.ebscohost.com.sculb.scu.edu/login.aspx?direct=true&db=lp&AN=502144612&site=ehost-live>

Persistent link to this record (Permalink): <http://0-search.ebscohost.com.sculb.scu.edu/login.aspx?direct=true&db=lp&AN=502144612&site=ehost-live>

Cut and Paste: Admin

Database: Index to Legal Periodicals and Books (H.W. Wilson)

Record: 2

Title: After Heller: What now for the Second Amendment?

Source: Santa Clara Law Review; 2010, Vol. 50 Issue 4, p1095-1111, 17p

Subjects: United States. Constitution. 14th Amendment; United States. Constitution. 1st-10th Amendments; Firearms -- Law & legislation;

UR: <http://0-search.ebscohost.com.sculb.scu.edu/login.aspx?direct=true&db=lp&AN=502144623&site=ehost-live>

Persistent link to this record (Permalink): <http://0-search.ebscohost.com.sculb.scu.edu/login.aspx?direct=true&db=lp&AN=502144623&site=ehost-live>

Cut and Paste: After

Database: Index to Legal Periodicals and Books (H.W. Wilson)

Record: 3

Title: An economic defense of flexibility in IPR licensing: contracting around "first sale" in multilevel production settings.

Source: Santa Clara Law Review; 2011, Vol. 51 Issue 4, p1149-1185, 37p

Subjects: Intellectual property; License agreements; First sale doctrine (Copyright); Patent exhaustion

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".

Name of Scraper: ILPB

Apply If Page URL Contains: <http://0-web.ebscohost.com.sculib.scu.edu>

	OK	Description	Marker Before	Marker After	Format
1	<input checked="" type="checkbox"/>	title	Title:	 </dd>	
2	<input checked="" type="checkbox"/>	Source:	Source:	 </dd>	
3	<input checked="" type="checkbox"/>	Subjects	Subjects:	keywords	
4	<input checked="" type="checkbox"/>	Keywords	Keywords:	Abstract	
5	<input checked="" type="checkbox"/>	Abstract	Abstract:	UR:	
6	<input type="checkbox"/>				
7	<input type="checkbox"/>				
8	<input type="checkbox"/>				
9	<input type="checkbox"/>				
10	<input type="checkbox"/>				

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

File Edit View Navigation Tools Help Registration

http://0-web.ebscohost.com.sculib.scu.edu/ehost/delivery?sid=07e4a7c0

Local IP: 129.210.91.44

id	Title	Source	Subjects	Keywords	Abstract
1	Administering the Second...	Santa Clara Law Review;2010, ...			
2	After Heller: What now fo...	Santa Clara Law Review;2010, ...			
3	An economic defense of f...	Santa Clara Law Review;2011, ...	Intellectual property;...		
4	Another move away from ...	Santa Clara Law Review;2011, ...			
5	Apportioning cleanup co...	Santa Clara Law Review;2011, ...	United States. Enviro...		
6	Apprendi land becomes b...	Santa Clara Law Review;2011, ...			
7	CIRM's consolidated IP p...	Santa Clara Law Review;2011, ...	Stem cells -- Researc...		
8	Climate change: a new re...	Santa Clara Law Review;2011, ...			
9	Copyright first sale and t...	Santa Clara Law Review;2011, ...	First sale doctrine (C...	First sale doctrine UR: ht...	
10	Dangerous criminals, the ...	Santa Clara Law Review;2011, ...			
11	DEDUCTIONS FOR DRUG ...	Santa Clara Law Review;2012, ...			
12	DELAWARE'S "EXPANDING...	Santa Clara Law Review;2012, ...			The article discusses
13	DOES THE RISE OF PROPE...	Santa Clara Law Review;2012, ...			
14	ERADICATING ELDER ABU...	Santa Clara Law Review;2012, ...			The article discusses
15	Exhausting extraterritorial...	Santa Clara Law Review;2011, ...	Exhaustion of rights ...		
16	Exhaustion and first sale i...	Santa Clara Law Review;2011, ...	First sale doctrine (C...		
17	Exhaustion of trademarks...	Santa Clara Law Review;2011, ...	Trademarks -- Law & ...		
18	Federal disability discrimi...	Santa Clara Law Review;2011, ...	Discrimination again...	Statutory interpretation U...	
19	Greenhouse gas emission...	Santa Clara Law Review;2011, ...			
20	Heller as Hubris, and how...	Santa Clara Law Review;2010, ...			
21	Heller, the Second Amend...	Santa Clara Law Review;2010, ...			
22	INFANCY DOCTRINE INQU...	Santa Clara Law Review;2012, ...			The article discusses
23	MAKING SENSE OF NEW T...	Santa Clara Law Review;2012, ...			
24	Making sense of state act...	Santa Clara Law Review;2011, ...			
25	Market integration and (t...	Santa Clara Law Review;2011, ...	European Union cou...		
26	Measuring a "degree of d...	Santa Clara Law Review;2011, ...	Academic freedom;Fi...		
27	Menu-labeling laws: a mo...	Santa Clara Law Review;2011, ...	United States. Food ...	Food labeling -- Laws and ...	
28	MORE THAN WORDS: RET...	Santa Clara Law Review;2012, ...			The article discusses
29	Navigating through the L...	Santa Clara Law Review;2011, ...	Perfection of security...	Uniform Commercial Code	

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.

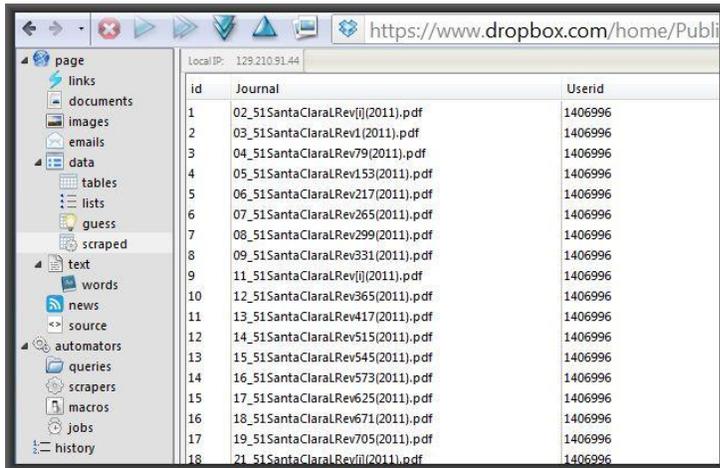
Once files are uploaded the metadata can be scraped using this small scraper:

Name of Scraper:

Apply if Page URL Contains:

	OK	Description	Marker Before	Marker After	Format
1	<input checked="" type="checkbox"/>	journal	class="sprite sprite_web_s_web_page_white_acrobat_32 icon"		
2	<input checked="" type="checkbox"/>	#repeat#userid	uid:		
3	<input type="checkbox"/>				
4	<input type="checkbox"/>				
5	<input type="checkbox"/>				
6	<input type="checkbox"/>				
7	<input type="checkbox"/>				

The scraped metadata should look something like this:



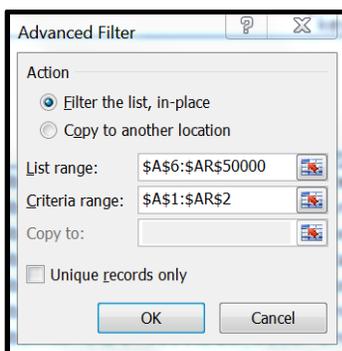
id	Journal	Userid
1	02_51 Santa Clara Rev[i](2011).pdf	1406996
2	03_51 Santa Clara Rev1(2011).pdf	1406996
3	04_51 Santa Clara Rev79(2011).pdf	1406996
4	05_51 Santa Clara Rev153(2011).pdf	1406996
5	06_51 Santa Clara Rev217(2011).pdf	1406996
6	07_51 Santa Clara Rev265(2011).pdf	1406996
7	08_51 Santa Clara Rev299(2011).pdf	1406996
8	09_51 Santa Clara Rev331(2011).pdf	1406996
9	11_51 Santa Clara Rev[i](2011).pdf	1406996
10	12_51 Santa Clara Rev365(2011).pdf	1406996
11	13_51 Santa Clara Rev417(2011).pdf	1406996
12	14_51 Santa Clara Rev515(2011).pdf	1406996
13	15_51 Santa Clara Rev545(2011).pdf	1406996
14	16_51 Santa Clara Rev573(2011).pdf	1406996
15	17_51 Santa Clara Rev625(2011).pdf	1406996
16	18_51 Santa Clara Rev671(2011).pdf	1406996
17	19_51 Santa Clara Rev705(2011).pdf	1406996
18	21_51 Santa Clara Rev[i](2011).pdf	1406996

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

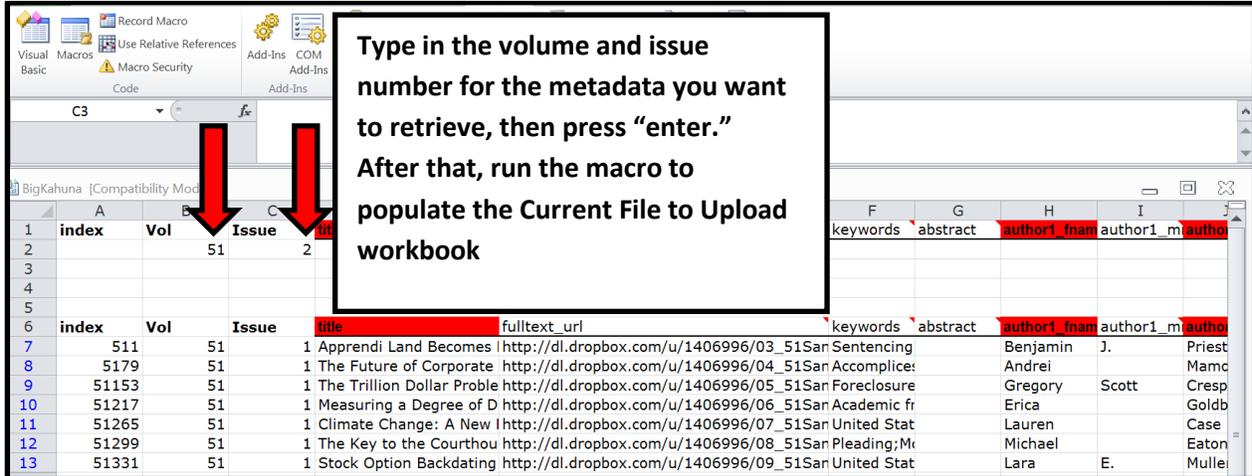
Creating the Final Spreadsheet

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 by 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.



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 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