PHILADELPHIA AREA CONSORTIUM OF SPECIAL COLLECTIONS LIBRARIES

Spreadsheet to XML: Using the PACSCL Finding Aids Spreadsheet

Hidden Collections Processing Project

Holly Mengel

3/1/2012

Table of Contents

Introduction

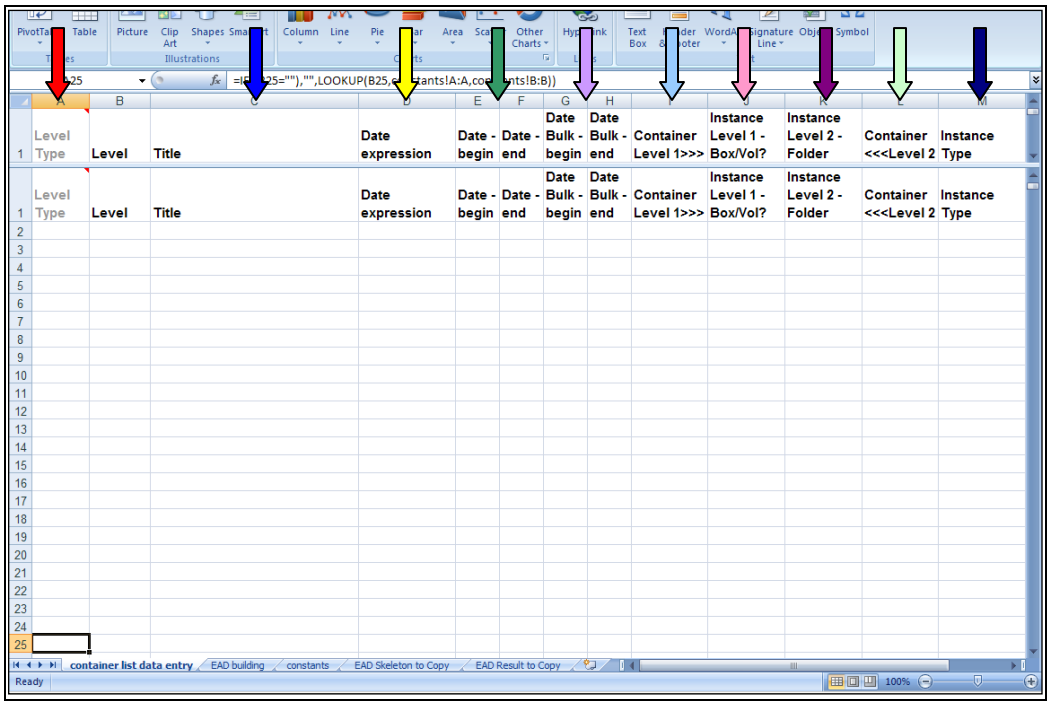
The PACSCL Finding Aids spreadsheet is a tool designed to import container lists into the Archivists' Toolkit. This spreadsheet can be used for original data entry or for converting electronic legacy finding aids to EAD finding aids. Regardless of the starting point, careful attention to data entry is important.

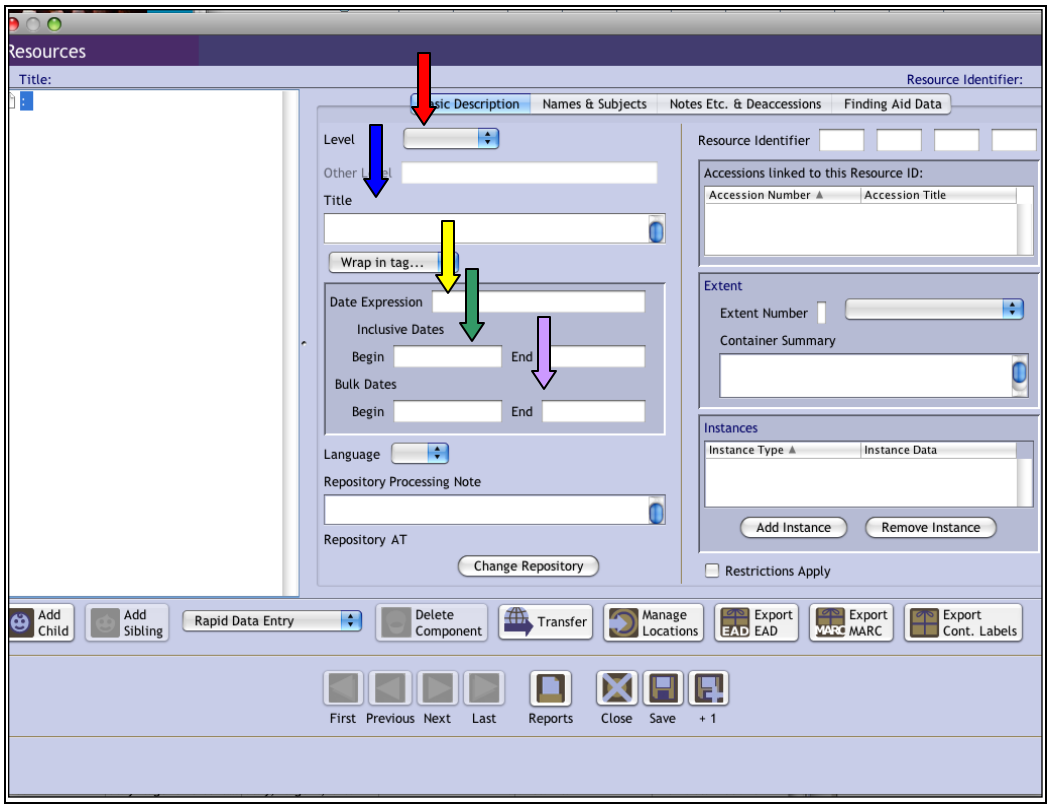
This spreadsheet was designed and shared by Matt Herbison, an archivist at the Drexel University College of Medicine Legacy Center. Technical expertise is not necessary to take advantage of this tool, however, working knowledge of MS Excel and EAD is helpful. Access to XML editing software, such as oXygen, XMetal or Dreamweaver, is also useful. XML editing software enables users to "validate" and edit XML code prior to import into the Archivists' Toolkit.

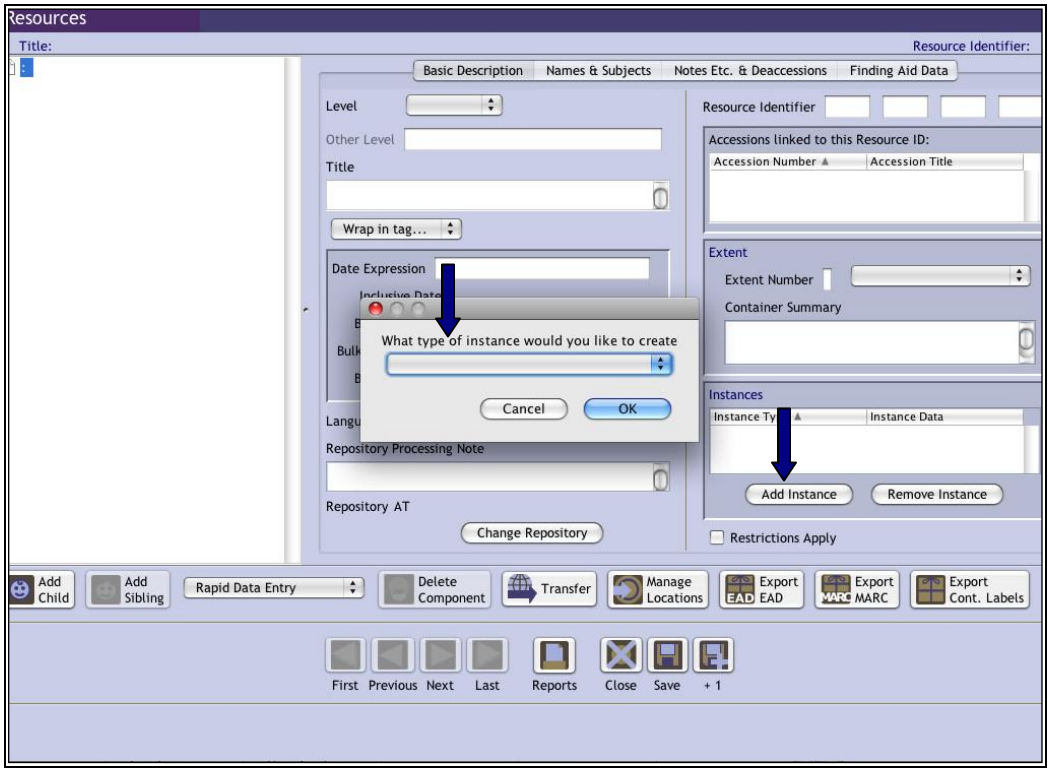
This guide demonstrates the mechanics of the PACSCL Finding Aids Spreadsheet and uses an extremely basic example. Those employing this guide for legacy conversions will find that few legacy finding aids are this simple or straightforward. For original data entry, follow instructions from Step 4 (page 6).

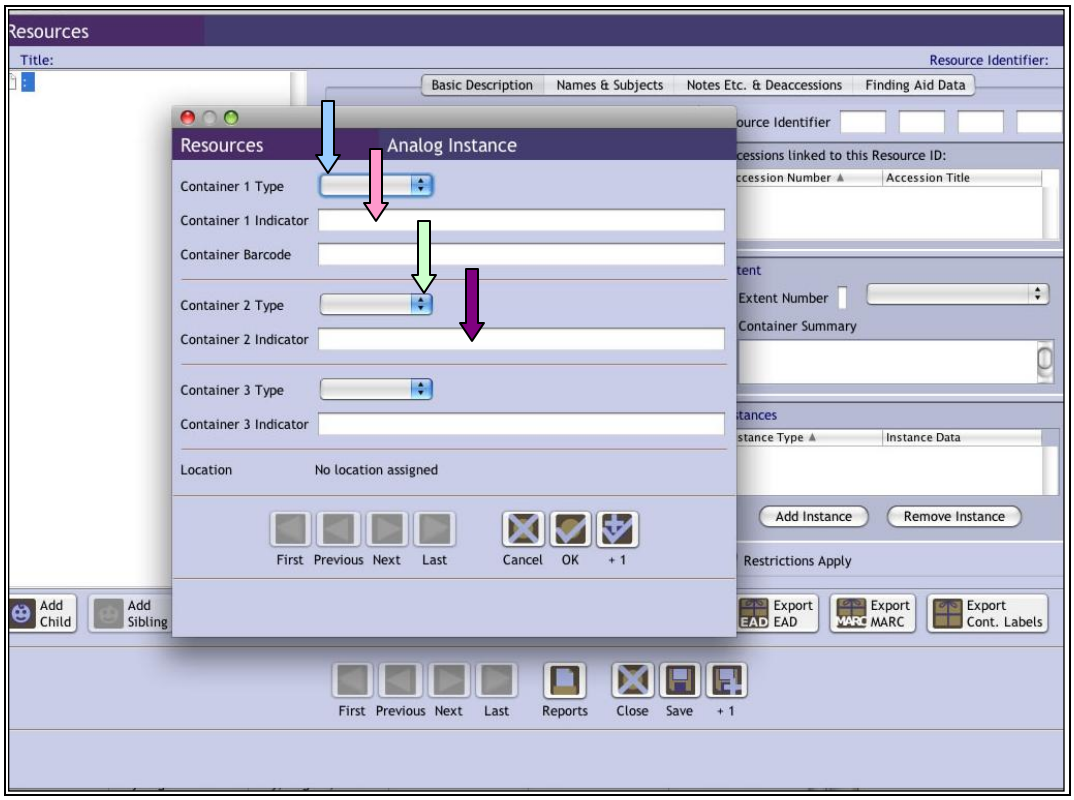
Familiarize Yourself with the Spreadsheet

Before beginning, open both the Archivists' Toolkit and a blank version of the spreadsheet (available at public.herbison.org/ead). The illustrations on the following pages show how the columns in the spreadsheet map directly to fields in the Archivists' Toolkit.




Resources
Title:
Resource Identifier:
Basic Description
Names & Subjects
Notes Etc. & Deaccessions
Finding Aid Data
Level
Other Level
Title
Wrap in tag...
Date Expression
Inclusive Dates
Begin
End
Bulk Dates
Begin
End
Language
Repository Processing Note
Repository AT
Change Repository
Resource Identifier
Accessions linked to this Resource ID:
Accession Number
Accession Title
Extent
Extent Number
Container Summary
Instances
Instance Type
Instance Data
Add Instance
Remove Instance
Restrictions Apply
Add Child
Add Sibling
Rapid Data Entry
Delete Component
Transfer
Manage Locations
Export EAD
Export MARC
Export Cont. Labels
First
Previous
Next
Last
Reports
Close
Save
+ 1


Resources
Title:
Resource Identifier:
Basic Description
Names & Subjects
Notes Etc. & Deaccessions
Finding Aid Data
Level
Other Level
Title
Wrap in tag...
Date Expression
What type of instance would you like to create
Cancel
OK
Repository Processing Note
Repository AT
Change Repository
Resource Identifier
Accessions linked to this Resource ID:
Accession Number
Accession Title
Extent
Extent Number
Container Summary
Instances
Instance Data
Add Instance
Remove Instance
Restrictions Apply
Add Child
Add Sibling
Rapid Data Entry
Delete Component
Transfer
Manage Locations
Export EAD
Export MARC
Export Cont. Labels
First
Previous
Next
Last
Reports
Close
Save
+ 1



Notice: A spreadsheet with the data keyed into the wrong fields *may* import into the Archivists' Toolkit. However, in order for the data to import into the proper fields and display correctly, it is necessary for the data to be placed into the exact columns that map to the fields in the Archivists' Toolkit.

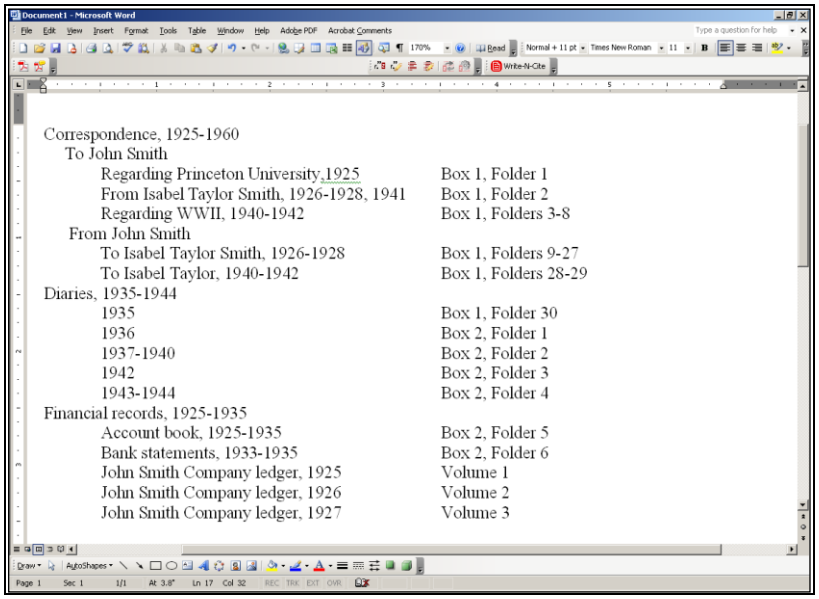
Notice: For instructions on proper data entry into Archivists' Toolkit fields, please see the PACSCL/CLIR Hidden Collections Processing Project, 2009-2012 Archivists' Toolkit Guide available at http://clir.pacscl.org/project-documentation/

Instructions for using the PACSCL Finding Aids Spreadsheet

Step 1: When starting with an electronic, MS Word document, copy and paste container list of an electronic finding aid into Notepad++. (If starting with an MS Excel document, follow instructions from Step 3).

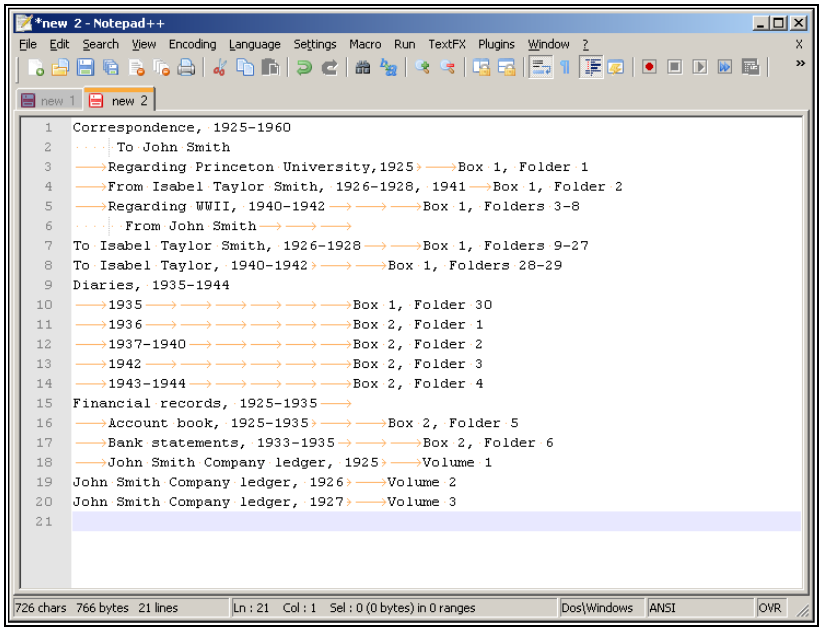
Notepad++ can be downloaded, free of charge, at http://notepad-plus-plus.org/download/

Original electronic finding aid:



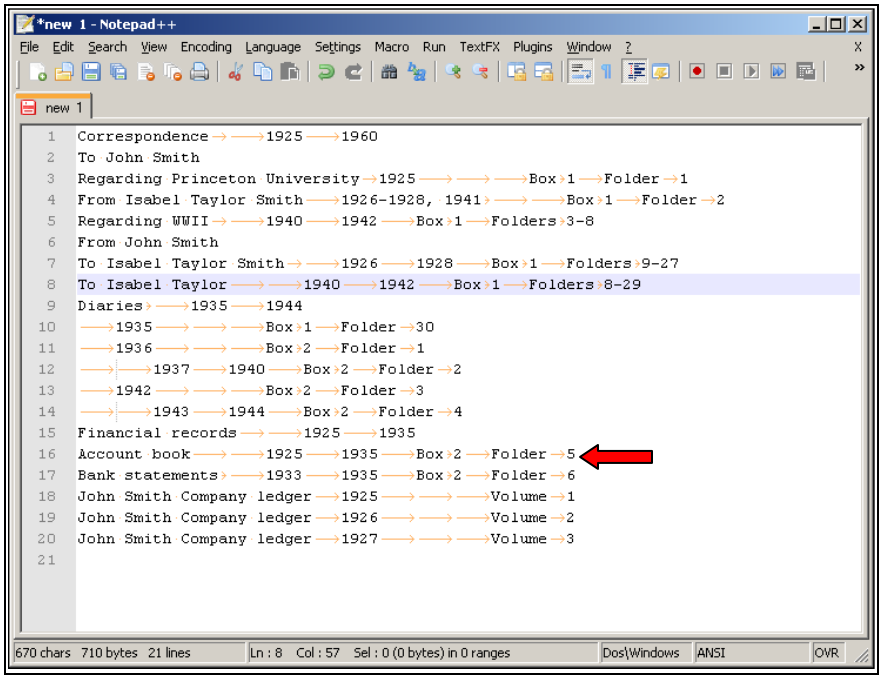
Correspondence, 1925-1960
To John Smith
Regarding Princeton University,1925 — Box 1, Folder 1
From Isabel Taylor Smith, 1926-1928, 1941 — Box 1, Folder 2
Regarding WWII, 1940-1942 — Box 1, Folders 3-8
From John Smith
To Isabel Taylor Smith, 1926-1928 — Box 1, Folders 9-27
To Isabel Taylor, 1940-1942 — Box 1, Folders 28-29
Diaries, 1935-1944
1935 — Box 1, Folder 30
1936 — Box 2, Folder 1
1937-1940 — Box 2, Folder 2
1942 — Box 2, Folder 3
1943-1944 — Box 2, Folder 4
Financial records, 1925-1935
Account book, 1925-1935 — Box 2, Folder 5
Bank statements, 1933-1935 — Box 2, Folder 6
John Smith Company ledger, 1925 — Volume 1
John Smith Company ledger, 1926 — Volume 2
John Smith Company ledger, 1927 — Volume 3

Original electronic finding aid, after being pasted into Notepad ++: (The arrows below represent tabs. The format of the original document pasted into Notepad++ will dictate where tabs exist. Each new document pasted into Notepad++ will look different.)



```
1  Correspondence, 1925-1960
2      To John Smith
3  →Regarding Princeton University,1925→→Box 1, Folder 1
4  →From Isabel Taylor Smith, 1926-1928, 1941→Box 1, Folder 2
5  →Regarding WWII, 1940-1942→→→Box 1, Folders 3-8
6      From John Smith→→→
7  To Isabel Taylor Smith, 1926-1928→→Box 1, Folders 9-27
8  To Isabel Taylor, 1940-1942→→→Box 1, Folders 28-29
9  Diaries, 1935-1944
10 →1935→→→→→→→Box 1, Folder 30
11 →1936→→→→→→→Box 2, Folder 1
12 →1937-1940→→→→→→Box 2, Folder 2
13 →1942→→→→→→→Box 2, Folder 3
14 →1943-1944→→→→→→Box 2, Folder 4
15 Financial records, 1925-1935→
16 →Account book, 1925-1935→→→Box 2, Folder 5
17 →Bank statements, 1933-1935→→→Box 2, Folder 6
18 →John Smith Company ledger, 1925→→Volume 1
19 John Smith Company ledger, 1926→→Volume 2
20 John Smith Company ledger, 1927→→Volume 3
21
```

Step 2: Separate folder titles, folder dates, box, box number, folder, and folder number with tabs. This will create a text delimited file that can be read and saved by most database and spreadsheet programs.

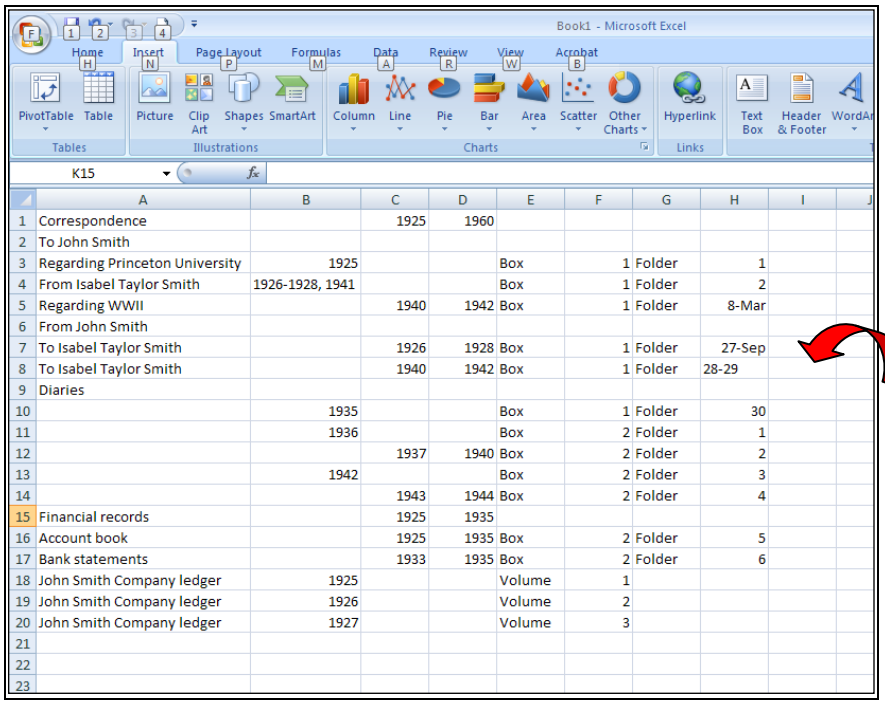


Notice: There are specific numbers of tabs between items, so that when pasting into the spreadsheet, fields fall into specific columns. In this example, data before the first tab will fall into the first column (title); data following one tab (shown with an arrow) will fall into the second column (date expression), etc. In this example, there were no bulk dates in the original finding aid, thus tabs for bulk dates are not included here, but will need to be compensated for in step 4.

Take note that in Box 2 Folder 5, there is a tab between 1925 and 1935. In the original finding aid, this is written 1925-1935, but because it is an "inclusive date" in the Archivists' Toolkit, it will fall into two separate columns "date begin" and "date end" in the PACSCL Finding Aids Spreadsheet.

Hint: Mistakes will be obvious after pasting into MS Excel. For example, if dates are in the title field, go back to the Notepad ++ document and add tabs as necessary.

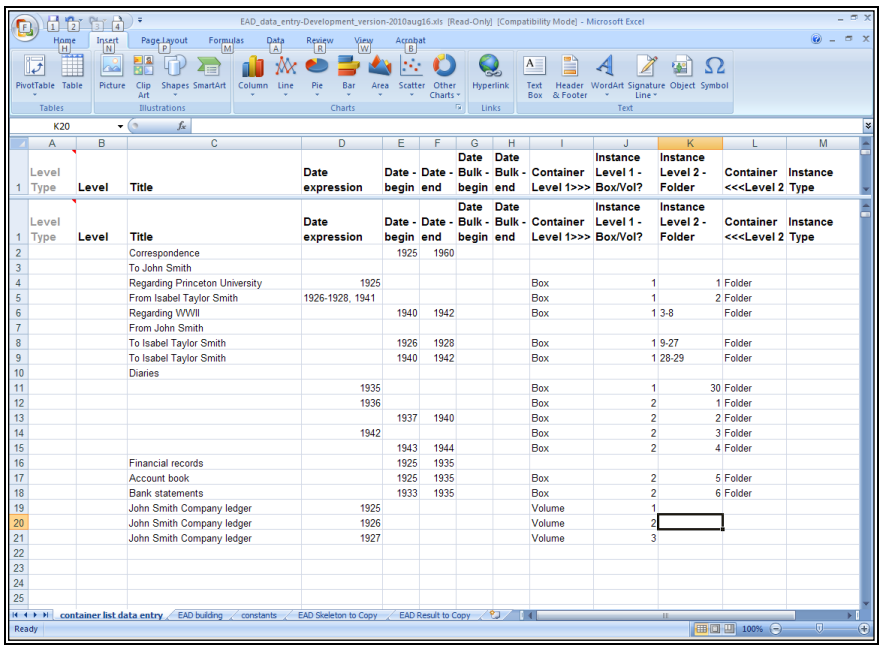
Step 3: Copy and paste text de-limited document into MS Excel, correct spelling and format titles and dates for DACS compliance (use regular expressions if you know how!). If original finding aid is an MS Excel document, start with this step.



	A	B	C	D	E	F	G	H	I	J
1	Correspondence		1925	1960						
2	To John Smith									
3	Regarding Princeton University	1925			Box	1	Folder	1		
4	From Isabel Taylor Smith	1926-1928, 1941			Box	1	Folder	2		
5	Regarding WWII		1940	1942	Box	1	Folder	8-Mar		
6	From John Smith									
7	To Isabel Taylor Smith		1926	1928	Box	1	Folder	27-Sep		
8	To Isabel Taylor Smith		1940	1942	Box	1	Folder	28-29		
9	Diaries									
10		1935			Box	1	Folder	30		
11		1936			Box	2	Folder	1		
12			1937	1940	Box	2	Folder	2		
13		1942			Box	2	Folder	3		
14			1943	1944	Box	2	Folder	4		
15	Financial records		1925	1935						
16	Account book		1925	1935	Box	2	Folder	5		
17	Bank statements		1933	1935	Box	2	Folder	6		
18	John Smith Company ledger	1925			Volume	1				
19	John Smith Company ledger	1926			Volume	2				
20	John Smith Company ledger	1927			Volume	3				
21										
22										
23										

Notice: The spreadsheet may not format all data correctly. Check dates and box and folder numbers.

Step 4: Copy columns into appropriate columns in the PACSCL Finding Aids Spreadsheet.



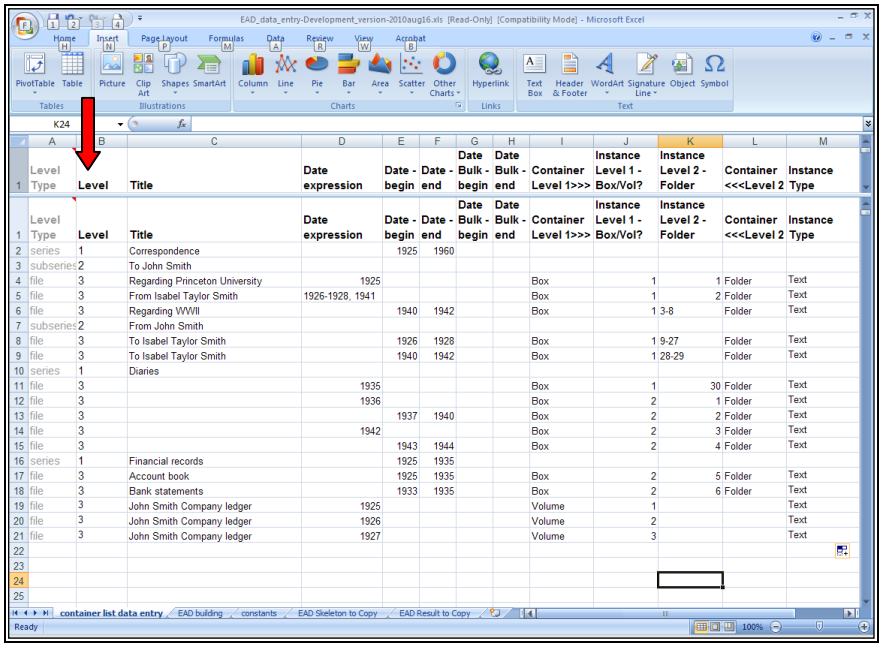
Level Type	Level	Title	Date expression	Date - begin	Date - end	Date Bulk - begin	Date Bulk - end	Container Level 1>>>	Instance Level 1 - Box/Vol?	Instance Level 2 - Folder	Container <<<Level 2	Instance Type
		Correspondence		1925	1960							
		To John Smith										
		Regarding Princeton University	1925					Box	1	1	Folder	
		From Isabel Taylor Smith	1926-1928, 1941					Box	1	2	Folder	
		Regarding WWII		1940	1942			Box	1	3-8	Folder	
		From John Smith										
		To Isabel Taylor Smith		1926	1928			Box	1	9-27	Folder	
		To Isabel Taylor Smith		1940	1942			Box	1	28-29	Folder	
		Diaries										
			1935					Box	1	30	Folder	
			1936					Box	2	1	Folder	
				1937	1940			Box	2	2	Folder	
			1942					Box	2	3	Folder	
				1943	1944			Box	2	4	Folder	
		Financial records		1925	1935							
		Account book		1925	1935			Box	2	5	Folder	
		Bank statements		1933	1935			Box	2	6	Folder	
		John Smith Company ledger	1925					Volume	1			
		John Smith Company ledger	1926					Volume	2			
		John Smith Company ledger	1927					Volume	3			

Hint: Make sure that data matches up: do several spot checks!

Hint: This is where, in supplied example, the compensation for bulk dates occurs, Notice that columns G and H are empty.

Hint: Make sure that instance levels 1 and 2 are next to each other.

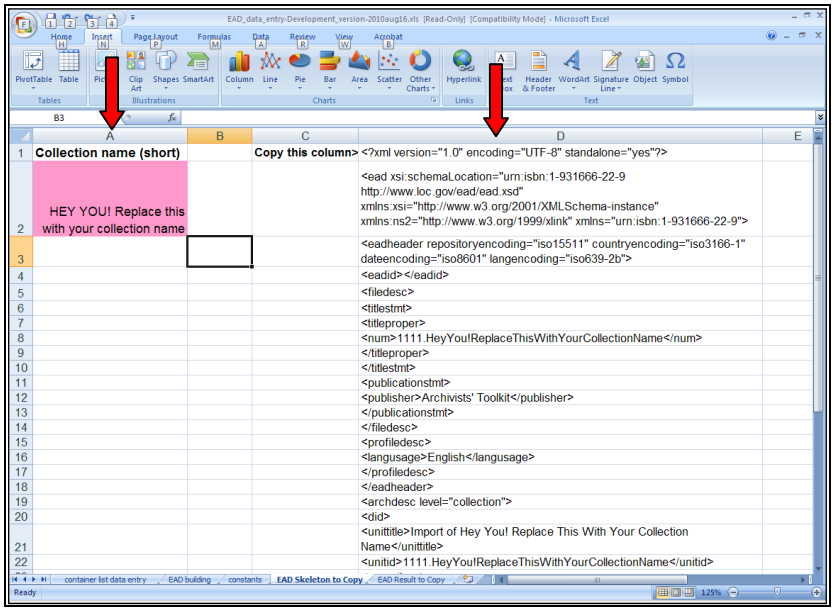
Step 5: Add level (series, subseries, file) to create hierarchy, and instance type (text, graphic material)



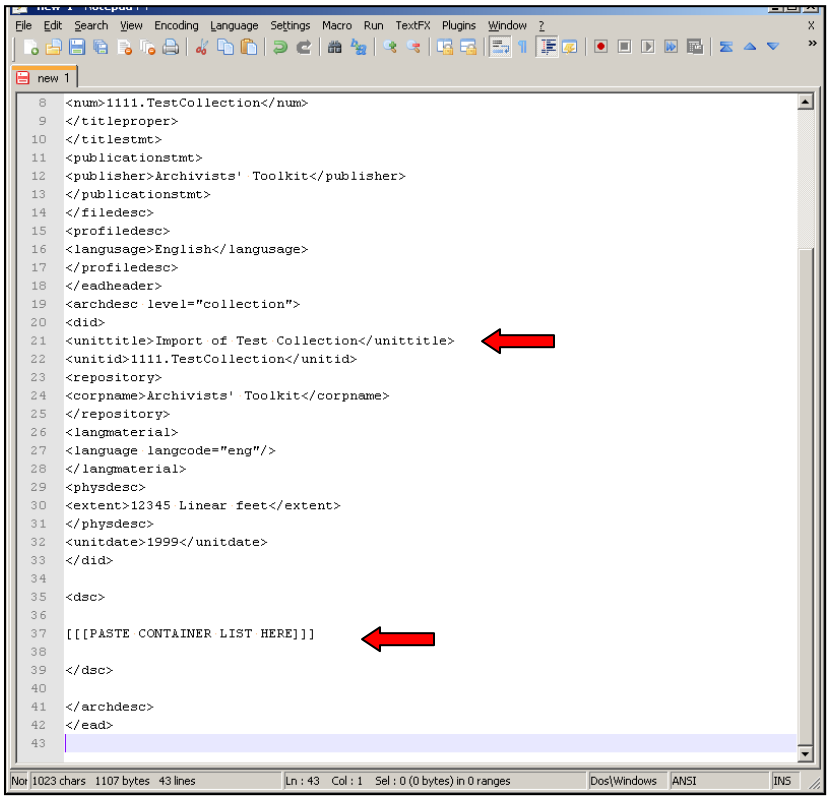
Level Type	Level	Title	Date expression	Date - begin	Date - end	Date Bulk - begin	Date Bulk - end	Container Level 1>>>	Instance Level 1 - Box/Vol?	Instance Level 2 - Folder	Container <<<Level 2	Instance Type
series	1	Correspondence		1925	1960							
subseries	2	To John Smith										
file	3	Regarding Princeton University	1925					Box	1	1	Folder	Text
file	3	From Isabel Taylor Smith	1926-1928, 1941					Box	1	2	Folder	Text
file	3	Regarding WWII		1940	1942			Box	1	3-8	Folder	Text
subseries	2	From John Smith										
file	3	To Isabel Taylor Smith		1926	1928			Box	1	9-27	Folder	Text
file	3	To Isabel Taylor Smith		1940	1942			Box	1	28-29	Folder	Text
series	1	Diaries										
file	3		1935					Box	1	30	Folder	Text
file	3		1936					Box	2	1	Folder	Text
file	3			1937	1940			Box	2	2	Folder	Text
file	3		1942					Box	2	3	Folder	Text
file	3			1943	1944			Box	2	4	Folder	Text
series	1	Financial records		1925	1935							
file	3	Account book		1925	1935			Box	2	5	Folder	Text
file	3	Bank statements		1933	1935			Box	2	6	Folder	Text
file	3	John Smith Company ledger	1925					Volume	1			Text
file	3	John Smith Company ledger	1926					Volume	2			Text
file	3	John Smith Company ledger	1927					Volume	3			Text

Notice: There are 3 levels of hierarchy to work with (see red arrow): 1 is series, 2 is subseries, 3 is file. All three levels are not required, but hierarchy is needed for a successful import. Type the number in column B and the appropriate level type automatically populates column A.

Step 6: Once hierarchy, data entry, etc. are correct, click on the worksheet titled "EAD Skeleton to Copy." Fill in the pink box (in column A) with the name of the collection. Click on Column D, copy, and then paste into a new Notepad++ document.



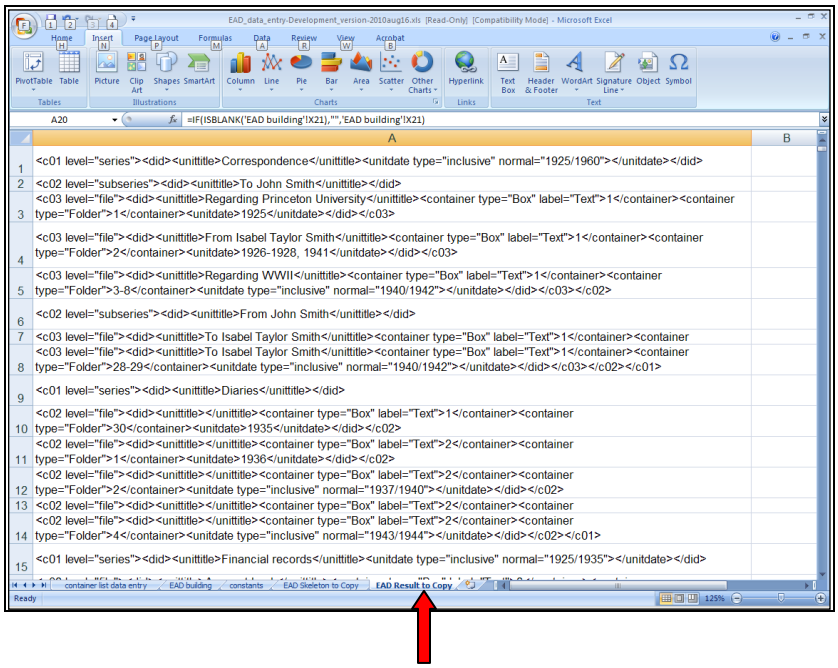
In Notepad ++, data will look like this:



Notice: The collection title is included.

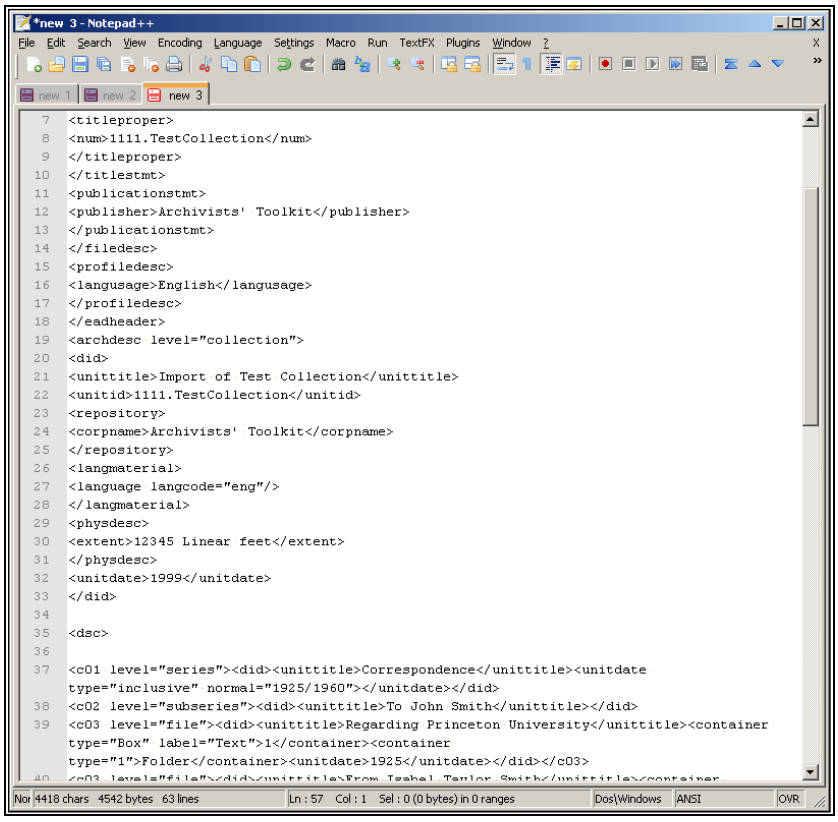
Notice: There is a place to paste the container list.

Next, returning to the PACSCL Finding Aids Spreadsheet, click on the worksheet entitled "EAD Result to Copy" (pictured on the following page). Click on column A and copy. In Notepad++, delete [[[PASTE CONTAINER LIST HERE]]], and then paste information from column A into that space.



Hint: If the word "REF" appears in column A, something has gone wrong. Frequently, this can be fixed by saving the data, and opening a clean version of the PACSCL Finding Aids Spreadsheet. Once the clean version is opened, paste the data back into the spreadsheet. If the problem persists, check data entry!

Notepad++ should look like this:



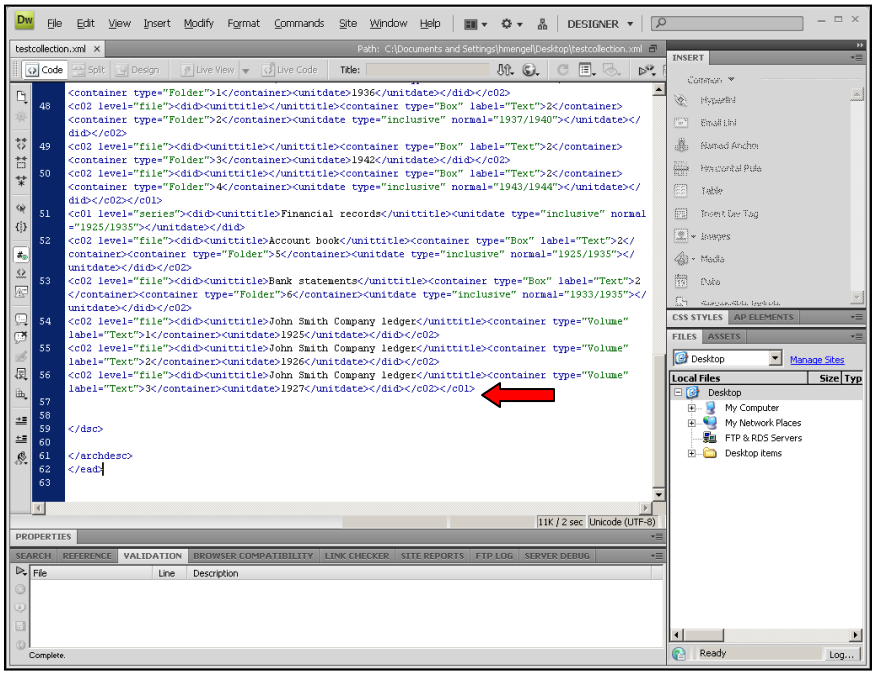
Notice: The container list is here.

Finally, save your document as an xml.

Step 7: Prior to import into the Archivists' Toolkit, open the XML document in an XML editor (if there are issues with the XML code, the document will not import into Archivists' Toolkit). In the editing software, "validate" the XML code in order to pinpoint issues in coding and make necessary edits.

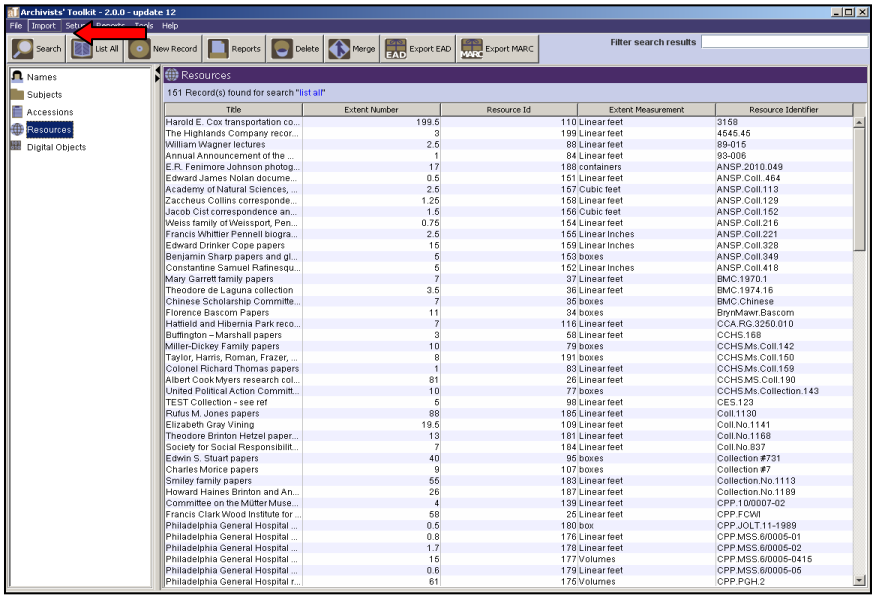
Common problems to look for:

- Ampersands need to be coded as &
- If a "date expression" was placed within the first column of "inclusive dates," the finding aid will not validate.
- Diacritics may be an issue
- If the container list is very long, sometime the spreadsheet has difficulty closing all the open tags. Check out the end of the finding aid and </c01> should be present almost at the bottom. If it is not, close the tags (see below):



Notice: The </c01> tag is present.

Step 8: After fixing any mistakes, save the file. Then open the Archivists' Toolkit, click on Resources, then click on Import on the menu bar along the top of the page, and then, from the drop-down menu, click on Import EAD. See illustration on the following page:

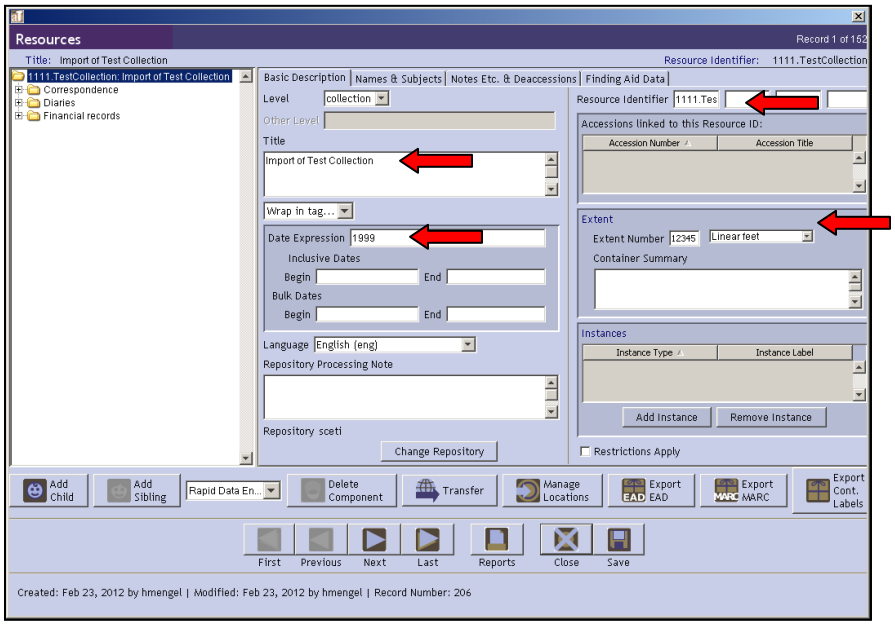


Depending upon the size of the finding aid, the import could take a few seconds or several minutes. After the import is complete, double click on Resources so that the new file loads. It will always be titled "Import of [name of the collection]."

Step 9: Add collection level information.

Change the following fields (remember to be DACs compliant):

Title
Date
Resource Identifier
Extent
Language of the collection (only if it is not English)



For a DACs compliant finding aid, add:

Abstract
Biographical/historical note
Scope and content note
Creator of the collection

If the original finding aid has these components, simply copy and paste them into the appropriate fields in AT.

Step 10: To get the collection into the PACSCL Finding Aids Site, export the EAD.

First save the xml file to the test web folder, and the next day, check for accuracy on the test site (which is not publicly available to those without the url).

Don't forget to:

Spell check

Make certain that hierarchy is correct (it is more difficult to see hierarchy mistakes in the spreadsheet than in the Archivists' Toolkit)

If OCR was used to make the finding aid a searchable, electronic document, plan to read the finding aid, word for word.

Make all corrections to the finding aid in the Archivists' Toolkit. Re-export the final version of the finding aid and save the xml file to the production web folder. The next day, it should appear on the PACSCL Findings Aids Site (findingaids.pacscl.org)

The container list of the finding should look like this: