

# Importing finding aid container lists into Archivists' Toolkit using Excel and Notepad++

July 16, 2012

## General

Use these guidelines with Excel file, *AT-EAD\_containerList.xlsx*. The steps will show you how to generate EAD code for finding aid container lists only. Once imported into Archivists' Toolkit, the container list can be merged with an existing collection-level finding aid, or can be used as a starting point for a new finding aid.

You will need:

- Excel
- Notepad++ (open-source xml editor): <http://notepad-plus-plus.org/download/v6.1.5.html>
- Archivists' Toolkit

and possibly an xml validator, such as oXygen to verify the EAD code. Invalid code will not import into AT.

*Note:* xml developer programs, like oXygen and XMetaL require a licensing fee. oXygen offers a free 30-day trial.

A basic understanding of EAD and XML is helpful. The spreadsheet is built to expose the container list code, but it is not foolproof. Careful review of tags and xml syntax is needed for a successful import.

*Acknowledgements:* This methodology was adapted from PACSCL. The spreadsheet was modified from Matt Herbison's Excel file, *Copy of EAD\_data\_entry-Development\_version-2010aug16*. More information can be found on the PACSCL blog at:

<http://clir.pacscl.org/2012/03/19/excel-to-xml-the-spreadsheet-from-heaven>

## Documents to use

Copy Excel spreadsheet **AT-EAD\_containerList.xlsx**. You may want to rename the file as a working copy.

It contains 3 worksheets:

- container list code
- copy&paste from finding aid
- PACSCL EAD Skeleton2Copy (written by Matt Herbison)



Locate and open the **finding aid document** you wish to import.

*Note:* The following steps offer guidelines for finding aids written in the AMNH finding aid Word template. If it is using the current template, you should be able to copy and paste the container list with little manipulation. If not, a more hands-on approach to repurposing the data may be needed.

## REPURPOSING CONTAINER LISTS IN EXCEL

Using the sheet titled *copy&paste from finding aid*, copy and paste the entire container list from your finding aid document into the cells. Depending on how the author set up the Word document, the data may not paste consistently – compare and review the data carefully before moving information around (see the item circled in red below). In addition, long titles may be broken up using several tabs or a hard return. Always review and adjust the data to make sure that title information for each folder is represented in a single cell.

Box	Folder	Title	date attribute	date expression	notes
<b>Box 1 Summary</b> Contains 13 notebooks from the 1890s. There are five notebooks with Hovey's class notes from when he studied at the university in Heidelberg, Germany in 1891. The majority of this box is comprised of field notebooks from localities around					
1	1	Connecticut Field Note Book No. 1, undated.			
1	2	Chemistry notes I. 1891.			
1	3	Trotchet Mountains and Hudson River Valley, N.Y., undated. (field notebook)			
1	4	Connecticut Field Note Book No. II. 1893			
1	5	South Dakota and Wyoming. 1897			
1	6	Italy. 1897. (field notebook)			
1	7	Virginia. 1898. (field notebook)			
1	8	Topographic Map of the Yellowstone National Park. 1898.			
1	9	Map of Connecticut, undated. (field notebook)			
1	10	Petrography and Optical Mineralogy. 1892.			
1	11	Special Mineralogy. 1891.			
1	12	Petrography. 1890.			
1	13				
<b>Box 2 Summary</b> Contains nine field notebooks from Hovey's studies of the Hudson River Valley in 1900, the Black Hills of South Dakota in 1901, St. Vincent and Martinique from 1902-1903 and Mexico in 1905. There is also a volume of the American Mus					
14	14	Hudson River Valley. 1900. (field notebook)			
14	15	Black Hills. S.D. 1901. (field notebook)			
14	16	St. Vincent and Martinique. 1902. (field notebook)			
14	17	St. Vincent and Kingstons. 1902. (field notebook)			
14	18	Martinique. 1903. (field notebook)			
14	19	Mexico. 1905. (field notebook)			
14	20	Cauca Valley. undated. (field notebook)			
14	21	American Museum Journal Cash Accounts. 1902-1905.			

Headers are provided in the spreadsheet, however you may want to create new columns or move columns to best work with the data you have. The screenshot above shows descriptive summaries for the box contents. This information will not be used in the conversion, but you can manually add it to the AT finding aid after the container list import.

### Break out the dates from the titles

A quick & easy way to automate separating dates into new cells

1. Select the column with the folder titles
2. Use Find+Replace to add a unique delimiter to the lines
3. Choose Data > Text to Columns. With Delimiter selected, check Other: [unique delimiter]

*Example:*

Chemistry notes I. 1891.

Find = “. 1”

Replace = “. \$1”

Use “\$” as the delimiter to break the dates out of the columns

Result: Chemistry notes I. \$1891.

*Note:* Using the period at the end of the title in the Find+Replace command keep from converting every instance of the number “1” into “\$1”. For instance finding “. 1” converted the line to read “Chemistry notes I. \$1891” and not “Chemistry notes I. \$189\$1”. Carefully review data to make sure other title information did not get converted by mistake.

Example using undated folders:

Connecticut Field Note Book No. 1. undated.

Find = "undated."

Replace = "\$undated"

Result: Connecticut Field Note Book No. 1. \$undated.

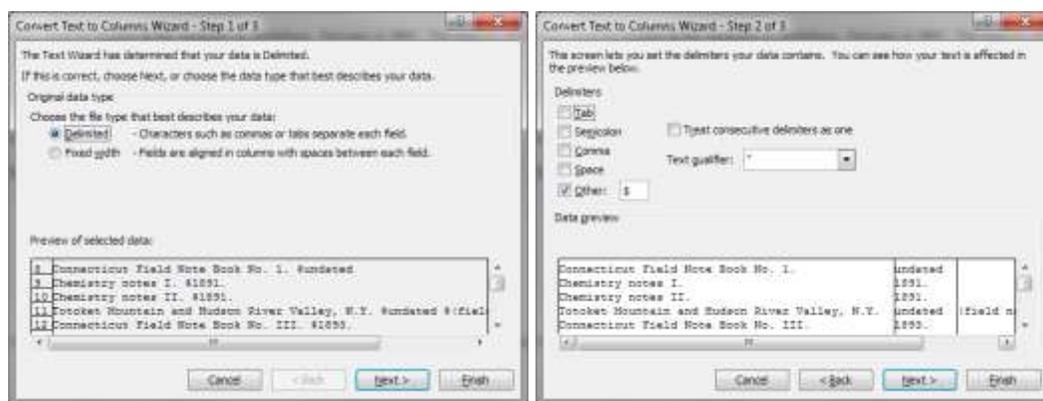
At this point, you may also want to add delimiters to parenthetical data or manually add delimiters where the Find+Replace did not apply. If there are any parenthetical notes or descriptions that follow the folder titles, they can be added into Archivists' Toolkit manually after the import.

### Data > Text to Columns

With the title column selected, go to Data > Text to Columns

Select Delimited

Uncheck Tab and check Other: \$



Make sure dates are in one column and notes are in another. You may have to manually move data around to align in their correct columns.

### Create an additional column for dates

You will need 2 columns for dates: date attribute is inserted into the code as EAD data, while date expression is how the date is displayed in a finding aid. In the EAD code, the inclusive dates are written with a beginning and an end date with a backslash in between. e.g., "1891/1891".

1. Copy the entire column for dates.
2. Insert copied column into the spreadsheet.
3. With one column of dates selected, run a Find+Replace for hyphens (-) and replace with slashes (/). This is your date attribute column.
4. Clean up the other date column to make DACS-compliant and remove terminal periods.

*Notes:* Single dates in the date attribute column should still be expressed by a beginning and end date. Undated titles will have no data for this cell. Be sure to include the quotation marks for the date attribute column, otherwise the xml will not validate. You can add them quickly by inserting a couple of columns of quotation marks on either side then combining the cells; go back and remove quotation marks for undated titles.

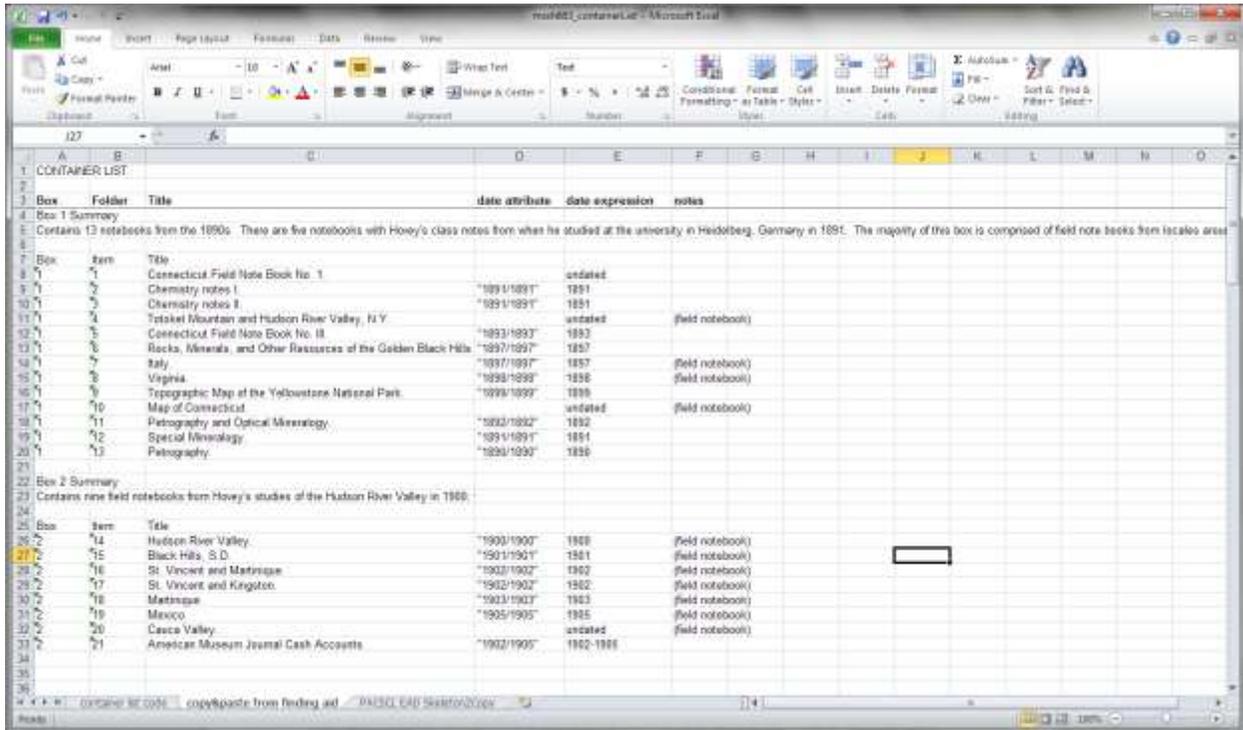
### Replacing invalid characters in data

Folder titles may include diacritics and characters that will affect the xml validation of your EAD code. You will need to replace them with UTF-8 characters. Below are a couple of examples of characters to look out for. If special characters are used, you may have to manually adjust them in AT.

- & – spell out "and" or use "&amp;" instead
- slanted quotations and apostrophes – use straight quotes and apostrophes:

single opening quote	'	replace with	'
single closing quote	'	replace with	'
double opening quote	"	replace with	"
double closing quote	"	replace with	"

Cleaned up data ready to paste into code should look like this:



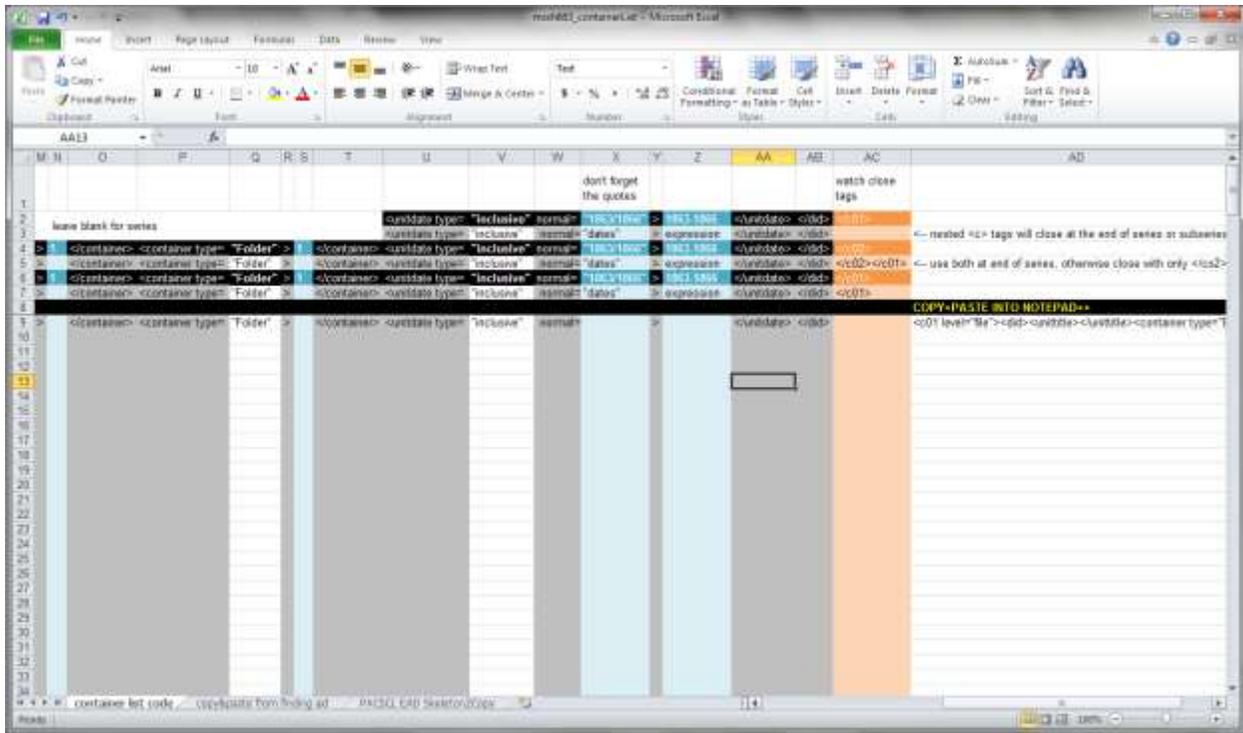
### Copying data into EAD code

Copy and paste clean columns of data into the code template laid out in the worksheet titled *container list code*. Watch for series and folder levels – mind your <c01>s and <c02>s! If you have a container list with series and subseries, pay careful attention to closing tags within nested containers. A series should hold information for its constituent subseries and folders. Its closing tags should wrap up all the elements included: </c03></c02><c01>. Most likely you will be working with a simple folder list which will only consist of one container element: <c01>.

The columns are color-coded to give an at-a-glance view of the data that needs to be unique vs. data that can be repeated.

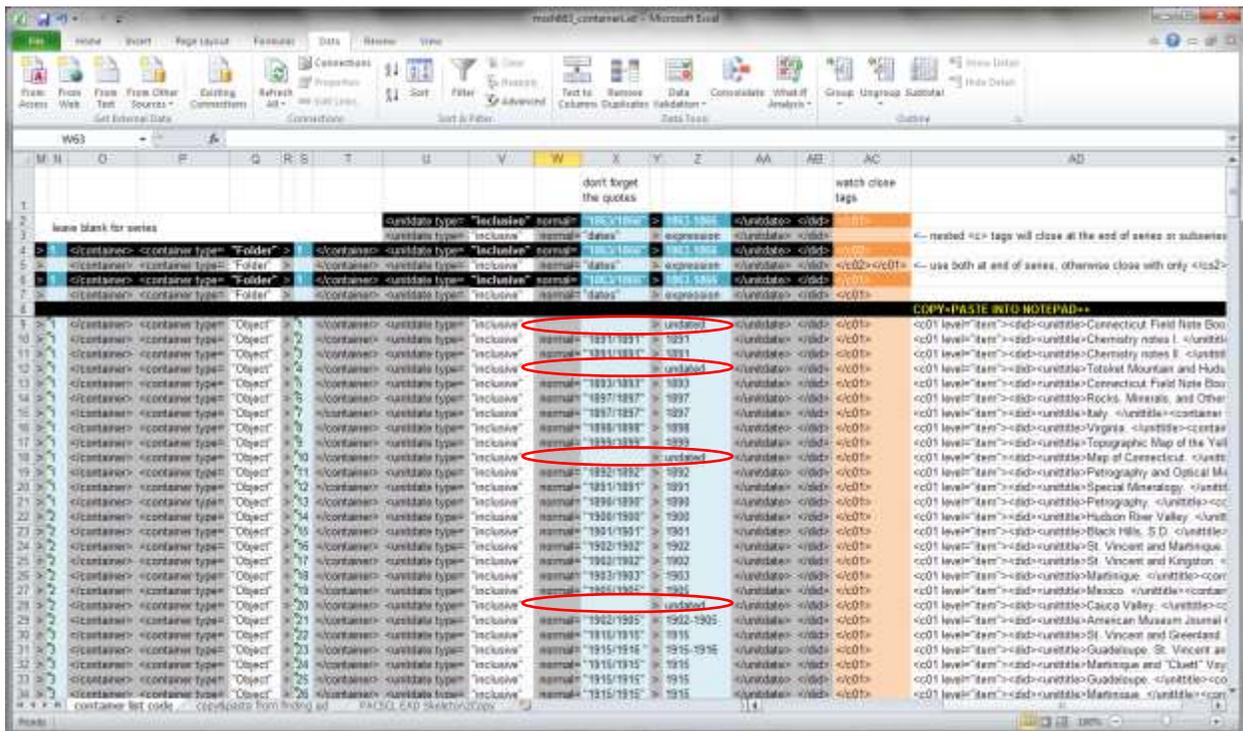
grey	fill down UNLESS you have a series, then match control
white	variable info, but most likely a fill down
light blue	new data from container list – use cleaned up columns from working sheet
light orange	closing tags – if you only have folder level (only <c01>s), then fill down, otherwise watch the end tags preceding a new series!

The very last last column is a combination (excel equation) for each line item. This is the data you will be copying and pasting into Notepad++



### Last bits of clean up

- Make sure you have copied and/or filled down the correct code. Check columns that contain only ">".
- Delete " normal=" from columns for titles with no dates, or marked "undated". See below.
- Run a Find+Replace for quotations and apostrophes.
- Do a thorough check for diacritics.



## USING NOTEPAD++

Open a new xml file in Notepad++. You will copy&paste two sets of information into this file:

- The code from worksheet *PACSCL EAD Skeleton2Copy*
- The column COPY+PASTE INTO NOTEPAD++ from the sheet *container list code*

For container lists you want to merge with existing collection-level resources already in AT, you do not have to fill out any top-level information such as <title>, <unitid>, <unitdate>, etc. Simply copy&paste the PACSCL skeleton code, then paste the new container list into the skeleton code.

Save the file as an *eXtensible Markup Language File* (.xml) and import the EAD into Archivists' Toolkit.

## IMPORTING INTO ARCHIVISTS' TOOLKIT

In the top menu options, select Import > Import EAD. Select your new xml file and import. The new resource should appear in the list (remember to select *List All* to view all the finding aid resources).

If the file does not import, review the error message and fix the code. Use a validator to help identify issues with the EAD. Watch out for those diacritics!

## Merge container list with a collection-level finding aid

Select both the collection-level finding aid and the container list and click *Merge*. Select the collection-level resource to merge into. This will retain all the top-level information and add the new "children" data to it.

Resulting resource should look like this:

