Smithsonian Libraries

Generating Batch DOI Upload Files for Crossref

Allyson Ota, Intern
8-9-2016

Generating DOI Batch Upload Files for Crossref

Contents

Contents....1
Initial Check of Data Received....3
Using OpenRefine to create DOIs and Clean Data....3
Weeding out records for further examination: Using blank facets in OpenRefine....4
Creating DOIs and formatting for XML generation....4
Monograph DOIs and Formatting....4
Create the **resource** URI....5
Serial DOIs....6
Generating XML and schema files in Excel....7
Required software and configuration:....7
Additional formatting in excel: Adding the timestamp column....7
Create a schema for the excel spreadsheet....8
Map a schema to the worksheet....8
Save XML schema (.xsd) for later use....10
Using oXygen to Transform XML into Required Format for Crossref Upload....11
Final Steps....12
Lessons Learned....13
Data issues in Excel regarding formatting for Barcodes and Catalog IDs, and other numerical fields....13
Removing titles that should not have DOIs assigned to them....13
Check against BHL holdings....13
SimpleText Query....13
Exclude Folklife Festival titles....13
Crossref Data Issues:....13
Year of Publication minimum....13
APPENDIX A: Working with Crossref & Additional Notes....15
Check if DOIs are registered in Crossref....15
Applying for DOIs....15
Crossref Schema....15
Test XML Batch File....15

Upload DOI Batch File 16
Review Crossref Submissions 16
Correct a DOI 16
APPENDIX B: Resources 17
Smithsonian Libraries Resources 17
OpenRefine 17
oXygen 17
Crossref 17
Microsoft Excel additional information 17

Generating DOI Batch Upload Files for Crossref

Initial Check of Data Received

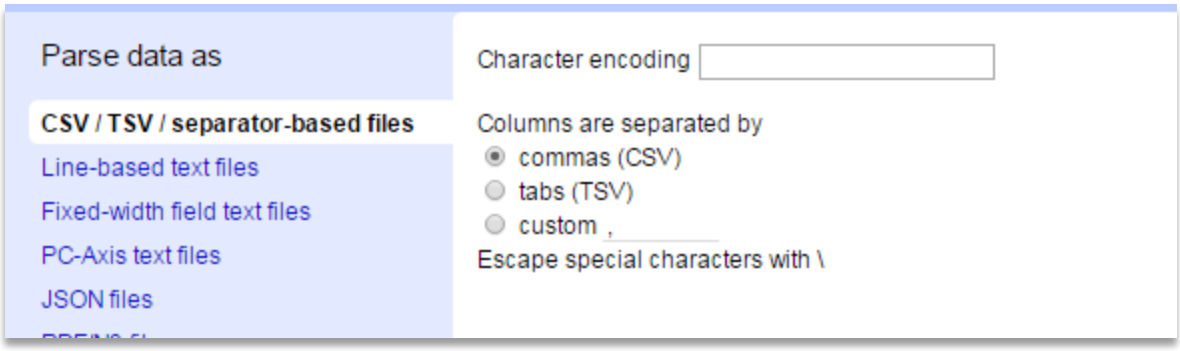
1. Check to ensure the export received has proper character encoding (recommend looking at the Title and Personal Name columns especially)
    a. Ensure diacritical marks are showing up properly
2. Check the Barcode/ISSN, and any other pertinent numbered columns to ensure their values are displaying properly and not in exponential notation before importing into OpenRefine
    a. Format Barcode and Catalog ID cells as text
3. Check to see if DOIs have already been registered for items being uploaded. See Appendix A for more information.

Using OpenRefine to create DOIs and Clean Data

Data cleanup can be done in Excel and/or OpenRefine. For a link to installation instructions for OpenRefine, see Appendix B.

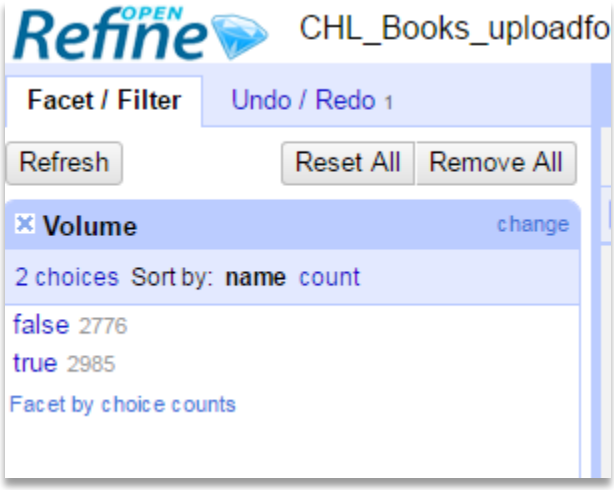
Note: You will probably receive a timeout error upon startup of OpenRefine for the Freebase reconciliation service, which can be ignored since the Standard Reconcile Service is no longer operational according to: https://github.com/OpenRefine/OpenRefine/wiki/Reconciliation-Service-API

1. Launch OpenRefine and a browser window should open to 127.0.0.1:3333; if the OpenRefine service is already running, you can input the URL above in a browser and it should bring you to the start page for OpenRefine.
2. Choose to **Create Project**
3. Click **Choose Files** and select your excel spreadsheet
4. Click **Next**
5. You can change the Project name in the upper-right (it defaults to the name of the file uploaded), click **Create Project**
6. When importing a .CSV, you can designate character encoding by clicking the empty field and selecting **UTF-8**



Weeding out records for further examination: Using blank facets in OpenRefine

7. If there are required fields, facets can help determine any blank/missing data, as well as any filled values where there should be none e.g. for **Volume** on monographs that are not a series, we would want to remove any rows for items that had volume information indicating a series.
    a. Select the arrow on the column you want to facet. E.g. if **Volume** should be blank, go to Volume, choose **Facet -> Customized facets -> facet by blank** and select **true** to see the volumes you want to work with
    b. Do the same for the column named **Serial_Volumes**



8. Facets allow you to export multiple views of the data by Choosing **Export -> Excel**
9. Multiple facets can be performed to then rule out other data e.g. missing values for **year** or other required elements
10. One-by-one, perform additional facets for required data in the fields: year, resource/path, title, etc.
11. Export to Excel

Creating DOIs and formatting for XML generation

This section outlines procedures used for single volume monographs and a series. Skip to the appropriate section below.

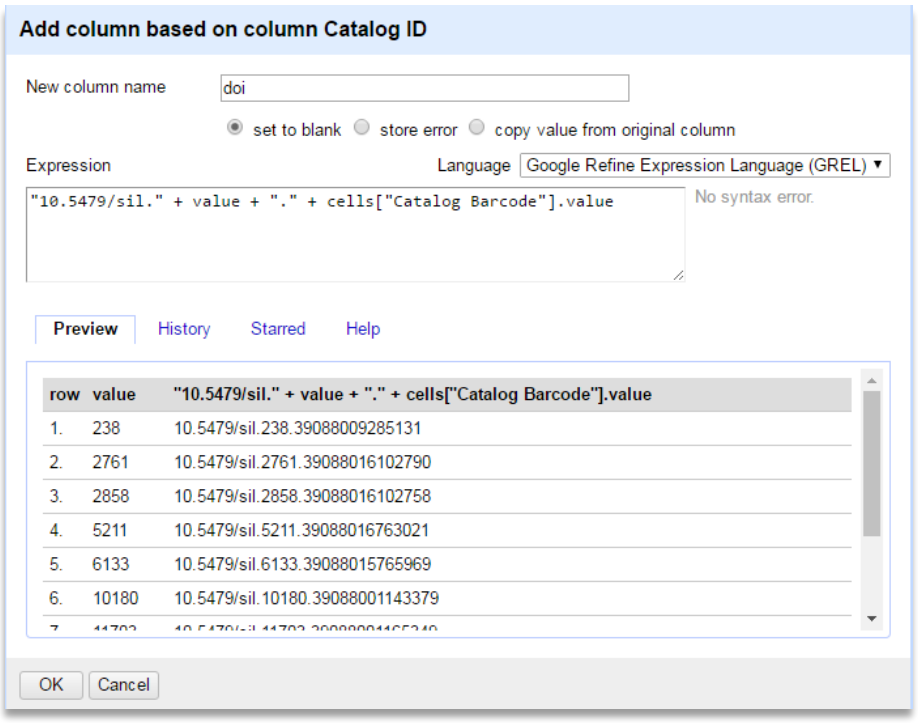
Monograph DOIs and Formatting

For monographs generate DOIs matching the format: **10.5479/sil.bibnumber.barcode** e.g. 10.5479/sil.307669.39088000225300

1. Click the arrow on the **Catalog ID** column, select **Edit column -> Add column based on this column**
2. Name the new column **doi**
3. Type the following into the Expression box:

   "10.5479/sil." + value + "." + cells["Catalog Barcode"].value

4. Click OK



Create the resource URI

1. Click the **Path** column, select **Edit column -> Add column based on this column**
2. Name the new column **resource**
3. Type the following into the Expression box:

   "https://library.si.edu" + value

4. Click **OK**
5. Table 1, indicates all required columns. Rename columns as appropriate (headings are case-sensitive):

Table 1. Monograph data field names

Crossref Required Name*	Column Name from Export
person_name	Personal Name
organization	Corporate Name
publisher_name	Publisher
year**	Year of Publication

resource**	Edit the **Path** column to create
title**	Title
doi**	Edit the **Catalog ID** column to create
language	Language

***required and cannot be blank for any record, recommend running facets to find blanks*

5. One by one, perform additional facets to search for blanks on all the following fields: **year, resource, title, doi, publisher_name** and select all **false** results in the filter pane before proceeding to the next column to facet by blank

1. Delete all additional columns from the spreadsheet, e.g: **Path, OCLC number**, **Catalog ID**, **Volume**, **Catalog Barcode, etc.**
2. Any item missing required information (e.g. year, path, etc.) cannot be included in the doi batch and should be made note of
3. Export to excel, by clicking **Export** in the upper right

Serial DOIs

For Serials use this format for DOI creation: **10.5479/si. + issn. + issue number. + starting page**. E.g. for Smithsonian Herpetological Information Service issue no. 143 that starts on page 1, you should have: 10.5479/si.23317515.143.1

1. Find the **sn** column with the ISSN in it, select **Edit Cells -> Transform**
2. In the **Expression** box, input the ISSN in quotes, without dashes e.g. “23317515”.
3. Click **OK**
4. Rename **sn** to **issn** (**Edit column -> Rename this column**)
5. Add a **doi** colulmn if one doesn’t already exist, if one exists skip to step 8
6. Select **issn -> Edit column -> Add column based on column**
7. In the **Expression** box, type:

"10.5479/si." + cells["*name of your issn column*"].value + "." + cells["*name of your issue number column*"].value + "." + cells["*name of first page number column*"].value

a. NOTE: For the Smithsonian Herpetological Information Service, there was no first page for some of the articles, so the last part was left off: **"10.5479/si." + cells["*name of your issn column*"].value + "." + cells["*name of your issue number column*"].value**

8. If **DOI** column exists, select DOI -> **Edit Cells -> Transform**
9. In the **Expression** box, input:

   "10.5479/si." + cells["*name of your issn column*"].value + "." + cells["*name of your issue no column*"].value + cells["*name of first page number column*"].value

10. Export to excel with filename of your choice, or if you'd like to do everything in OpenRefine, then export that is fine as well.
11. In either Excel or OpenRefine, delete all the columns that are not needed, and rename columns

Table 2. Serial data field names

Required column name	SRO export column name
year**	yr
issue**	is_no
title**	t1
subtitle	t2
person_name	author_primary_index
first_page	sp
last_page	op
doi**	DOI
resource**	ul

***required and cannot be blank for any record*

12. The excel spreadsheet being used to generate an XML batch upload serials, should be populated with data for items with both DSpace and SRO ids, but lacking DOIs.
13. The only fields that can be blank for an item are: **subtitle**, **first_page** and **last_page**. All other fields must be populated with data.
14. For authors in the **person_name** field, use data from the **author_primary_index** column exported from SRO. The format should be: LastName, FirstName with individual authors separated by a semi-colon.
15. Export to excel, by clicking **Export** in the upper right

Generating XML and schema files in Excel

Required software and configuration:

Windows OS
MS Office Excel
Excel 2003 XML Tools Add-in
Show the Excel Developer Tab on the ribbon

Additional formatting in excel: Adding the timestamp column

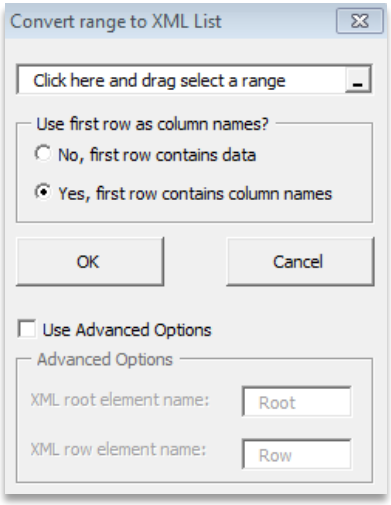
1. Open the exported excel file

2. Delete any unnecessary columns not listed in either **Table 1** or **Table 2**, above if any remain
3. Add a column named **timestamp**
4. Select the top cell in the column, choose to format it as text, and type the date/time in the following format: yyyymmddhhmmss
5. If there are more than 1,000 records create a New workbook with the same headings
    a. Create a new timestamp (steps 3 – 4 above) for each new workbook created. If you have less than 1,000 items, skip to the next section
    b. Ensure your timestamp(s) on additional sheets are unique, since this is used to generate a unique batch ID for Crossref

Create a schema for the excel spreadsheet

(For more information, see: [https://support.office.com/en-us/article/Create-an-XML-data-file-and-XML-schema-file-from-worksheet-data-e35400d4-0e10-4669-9a50-59a8c57d677e#feedbackText)](https://support.office.com/en-us/article/Create-an-XML-data-file-and-XML-schema-file-from-worksheet-data-e35400d4-0e10-4669-9a50-59a8c57d677e#feedbackText)

1. From the Menu, select **Add-Ins**-> **XML Tools** -> **Convert range to XML List**
2. Select your range
3. Under "**Use first row as column names**" select **yes** and click **OK**
    a. This creates an XML schema (.xsd file) and maps the cells to the schema, creating an XML table

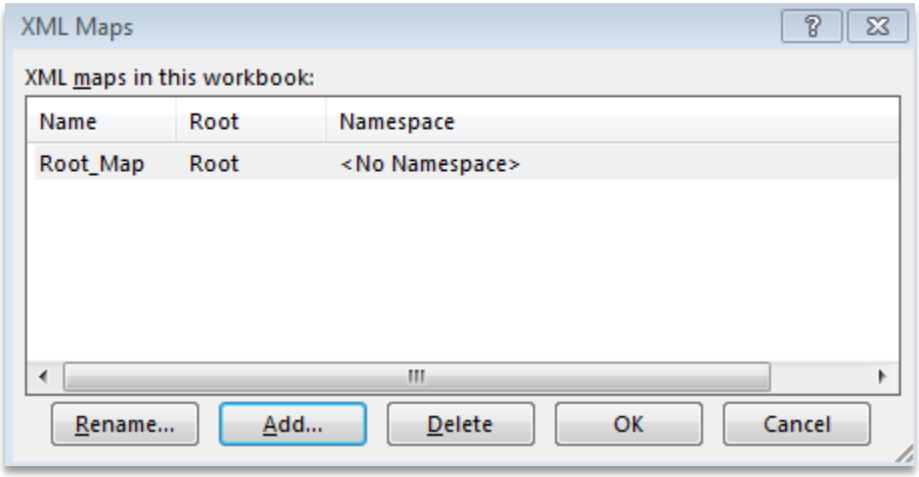


4. If the Visual Basic Editor appears and displays a Visual Basic for Applications (VBA) compile error, do the following:
    a. Click OK
    b. In the highlighted line in the VBA code module, delete "50" from the line: e.g. change **XMLDoc As msxml2.DOMDocument50** to **XMLDoc As msxml2.DOMDocument**
    c. Press F5 to advance to the next line, and delete "50"; continue until no errors are found
    d. Close Visual Basic editor
5. Choose to save as an .xsd file; this schema file can be used for additional files if your batch contains more than 1,000 items.

Map a schema to the worksheet

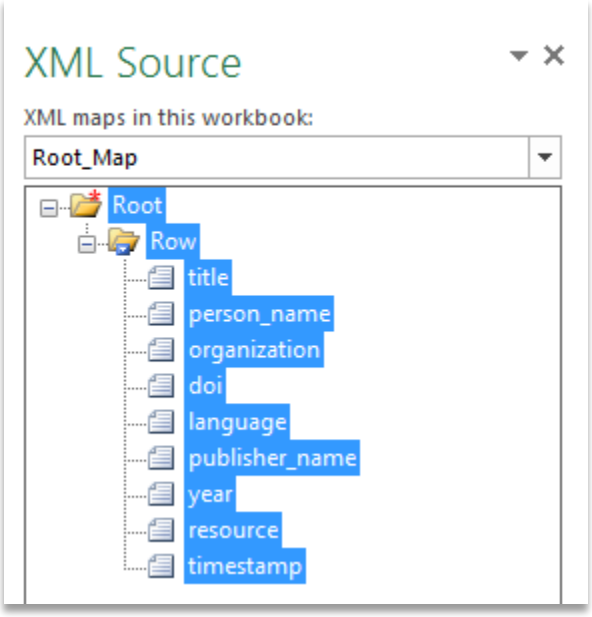
6. From the excel menu's **Developer** tab select **Source**
7. An **XML Source** side pane should open
8. Select XML Maps, an XML Maps window should open
    a. Click **Add**

b. Select the .XSD (schema file) file



c. Click **OK**

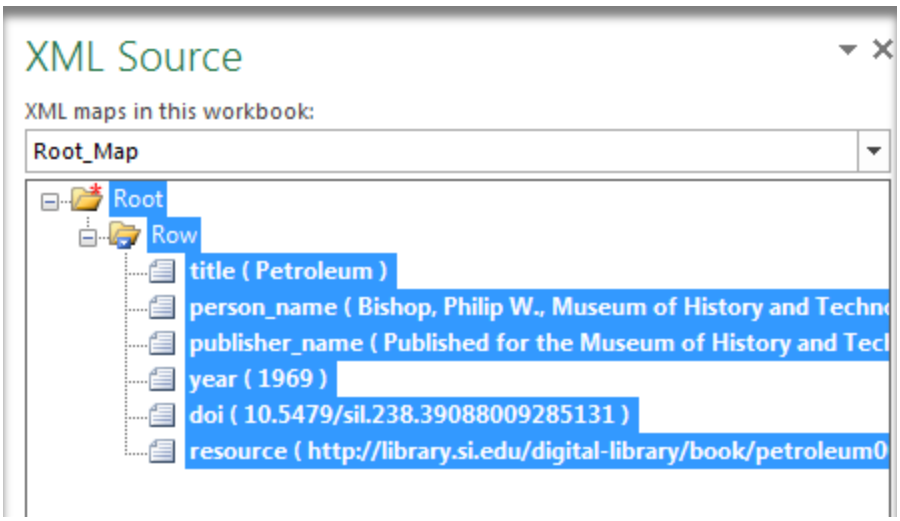
d. The Root map should appear in the **XML Source** Pane



e. In the XML Source pane, select the root folder icon and the entire directory tree should highlight

f. Click the Root directory in the Root_Map, then drag and drop it to the top-leftmost cell to map it to your worksheet

9. Your worksheet should also appear linked. (see images below)



	A	B	C	D	E	F
1	title	person_name	organization	doi	language	publisher_n
2	Incipit Arit	Boethius		10.5479/si	lat	
3	Auctoritate	Aristotle		10.5479/si	lat	
4	Theophras	Theophrastus		10.5479/si	lat	Per Bartholo
5	T. Lucreti (	Lucretius Carus, Titus		10.5479/si	lat	Paulus Fride
6	Epytoma I	Regiomontanus, Joannes		10.5479/si	lat	Johannes Ha
7	Ortus sanitatis.			10.5479/si	lat	Johann Prüs
8	Kleines Di	Brunschwig, Hieronymus		10.5479/si	ger	Johann (Rein
9	Cosmogra	Apian, Peter		10.5479/si	lat	impensis P.
10	De la pirot	Biringucci, Vannoccio		10.5479/si	ita	C. Navò
11	De historia	Fuchs, Leonhart		10.5479/si	lat	In officina Isi
12	Nicolai Co	Copernicus, Nicolaus		10.5479/si	lat	Apud Ioh. Pe

Worksheet should appear linked with XML Root Map selected

1. Click **Export** and **Save as type** should be XML
2. This file will be transformed into another XML file for upload to Crossref

Save XML schema (.xsd) for later use

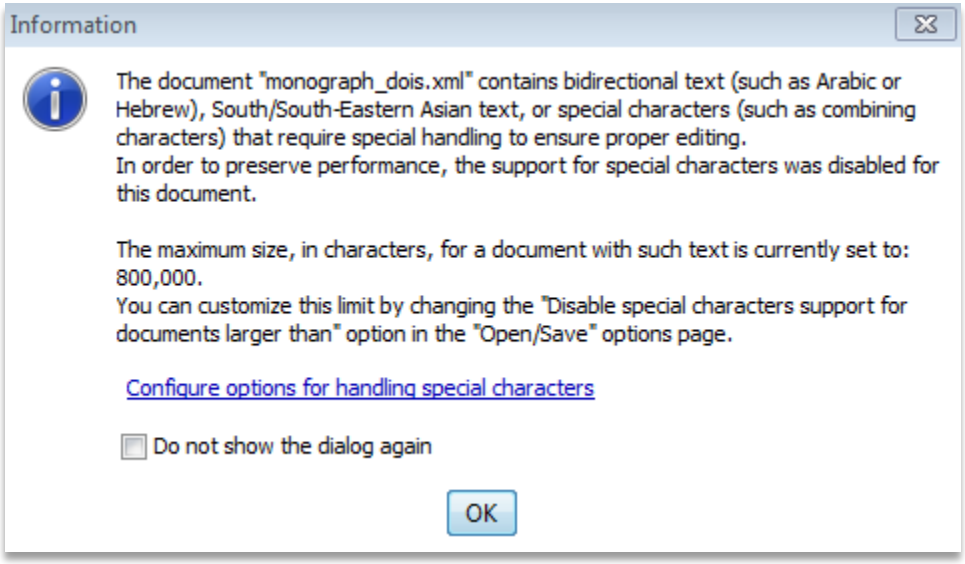
If you will be doing more uploads for the same data you can save the generated schema for later use

1. Select any cell in the open, mapped, excel table
2. From menu, go to the **Add-ins** -> **XML Tools** -> **Create XSD Files for the XML Schema at the active cell**
3. If there are Visual Basic Compile errors, in the highlighted lines of the VBA code module, delete "50" from the line: e.g. change **XMLDoc As msxml2.DOMDocument50** to **XMLDoc As msxml2.DOMDocument,** Press F5 to advance to the next error, and delete '50' until there are no more instances
4. Close Visual Basic window
5. A notepad document will automatically generate the .XSD file
    a. Save the file: Save As type should be "All files" and the extension should be .XSD

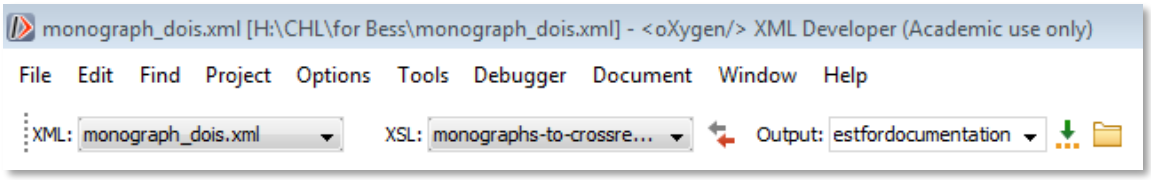
6. An XSLT transform in oXygen will be run against this .xml to generate an additional XML file with the correct formatting and structure, allowing upload into Crossref.org

Using oXygen to Transform XML into Required Format for Crossref Upload

1. Create a folder on your computer (name it anything you like). Place the .xsl file you'll be using, along with **languages.xml,** and finally place the **exslt.org** directory in there as well.
2. In oXygen, go to the menu **Project -> New Project**
3. Name your project and select the output directory
4. Choose **File -> Open** and select your XML file which is populated with rows of data about each title
5. Choose **File -> Open** and select the .XSL file which will transform it to produce an XML in the format required by Crossref
6. An Information window might open about bidirectional text. Click the link to **Configure options for handling special characters** at the bottom



7. A **Preferences** window should open. Choose to **Enable support for special characters**
8. Next to the Output field, click the folder icon to select the save location and file name that you want to give the transformed .XML file
9. Both the XML and XSL files should be displayed in oXygen. In the ribbon area, select the **Output** directory location



10. Click the blue arrow to run

11. Your output should appear in the 3rd pane on the right
12. This file is ready to be tested/uploaded against Crossref. See Appendix A.

Final Steps

1. Provide a list of all newly registered DOIs that need to be added to SRO and DSpace to the appropriate parties
2. All data that required further examination and did not get assigned a DOI should be forwarded onwards so any discrepancies can be addressed. E.g. duplicate DSpace ids for different articles, items missing required information, e.g. ISSN and Journal title, year, etc.

Lessons Learned

Data issues in Excel regarding formatting for Barcodes and Catalog IDs, and other numerical fields

This is mentioned in the "Initial Data Check of Data Received" section, but to give more detail, there were some issues with received .csv files where numerical values e.g. barcodes were displaying in scientific notation, so the complete barcode was not displayed. Catalog ID also did the same for some longer values. Before importing into OpenRefine, these fields should be formatted as text and displaying correctly.

Removing titles that should not have DOIs assigned to them

Check against BHL holdings

A check was done to see if any of the titles were also in the Biodiversity Heritage Library (BHL) since those publications would already have a DOI. Keri Thompson was able to run a report and return listings of Catalog IDs already in BHL. I was able to use a text editor to get rid of extra commas and single quotes, and create a column of numbers which I then pasted to the bottom of the **Catalog ID** column and then used conditional formatting to highlight duplicates in Excel and deleted matching rows.

SimpleText Query

A SimpleTextQuery was done against Crossref to ensure items with Year of Publication listed as 2000 onwards were not already assigned DOIs. As a note (and it says this on their page), entries should be submitted in alphabetical order, or as a numbered list. See Appendix A for more information on this process.

Exclude Folklife Festival titles

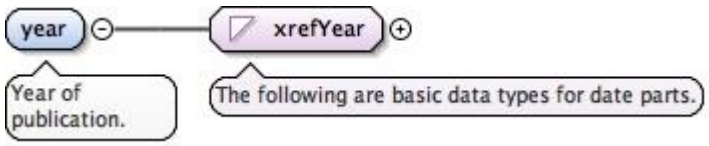
Keri Thompson asked to exclude items with titles related to the Folklife Festival. A search was done to remove "Festival of American Folklife" or "Folklife Festival" in their title.

Crossref Data Issues:

Year of Publication minimum

Crossref rejected a title in testing that had a **year** value of 1300. When checking their schema documentation, it noted a minimum date of 1400. (See MinInclusive value below from Crossref).

Element year

Namespace	http://www.crossref.org/schema/4.3.7
Annotations	Year of publication.
Diagram	
Type	**xrefYear**

Properties	Content:	**simple**
Facets	TotalDigits	**4**
	MaxInclusive	**2200**
	MinInclusive	**1400**
Used by	Complex Type	**date_t**
	Elements	**approval_date**, **creation_date**, **publication_date**, **update_date**
Source		

```
<xsd:element name="year" type="xrefYear">
  <xsd:annotation>
    <xsd:documentation>Year of publication.</xsd:documentation>
  </xsd:annotation>
</xsd:element>
```

APPENDIX A: Working with Crossref & Additional Notes

The following was taken from notes and documentation provided by Bess Missell.

Check if DOIs are registered in Crossref

This was performed for the Herpetological Information Service series but will not apply to every batch being worked on. Skip or refine this section if not applicable.

1. Search for journal title in SRO e.g. "Smithsonian Herpetological Information Service" (http://research.si.edu/)
2. Copy and paste the citations from SRO into MS Word and number them
3. Then copy and paste them into Notepad to remove any extra notations/diacritics
4. Then paste 50 citations at a time into the SimpleTextQuery here: http://www.crossref.org/SimpleTextQuery/
    a. You need a registered email account to use this service – try either naplesr@si.edu or hutchinsona@si.edu
5. Any citations with a black DOI and no matching red DOI is not registered
6. Pull out all citations needing DOIs (blank DOI) or needing their DOI to be registered
7. Add these titles to the excel spreadsheet with all pertinent information for DOI upload

NOTE: In the instance where DOIs are being checked against publications, if they end with a "." e.g. 10/5479/sil.1234567890.x. Crossref's SimpleText Query does not appear to recognize it as an end character. It ignores the period. While it is permissible to end a DOI with a period, it would probably be best to ensure DOIs do not end with periods since this gets overlooked when inputting citations and checking for duplicates with Crossref's Simple Text Query.

Applying for DOIs

- Use templates in s:\ISD\CrossRef\templates
- Contributions Template – use me.xml
- Atoll Template – use me.xml
- Database Template – use me.xml

Crossref Schema

For new formats, follow the Crossref schema at

http://www.crossref.org/help/schema_doc/4.3.7/4.3.7.html

http://www.crossref.org/help/schema_doc/4.3.3/4.3.3.html

Test XML Batch File

Test your completed xml with the Crossref Metadata Quality Check at

http://www.crossref.org/02publishers/parser.html

Upload DOI Batch File

1. Upload your DOI xml submission at

   https://doi.crossref.org/servlet/submissionAdmin?sf=showUpload

   Login: smit

   Password: smit1129

2. Put completed submissions in s:\ISD\CrossRef\in_process\Uploaded with Success

Review Crossref Submissions

You can review your submissions using the Submission Administration function –

http://help.crossref.org/viewing_the_submissions_administration_report

Correct a DOI

1. Go to https://doi.crossref.org/servlet/submissionAdmin?sf=showUpload and login
2. Go to Queries, DOI Query and enter the DOI and choose format = UNIXREF
3. Click submit and the xml will be returned to you. Cut and paste into xml editor – add header info, updating doi_batch_id, timestamp and corrected metadata. Upload.

APPENDIX B: Resources

Smithsonian Libraries Resources

- SRO – Smithsonian Research Online: research.si.edu
- DSpace Repository : repository.si.edu
- Example of XML to use for DOI registration: S:\ISD\CrossRef\Batch loads

OpenRefine

- http://openrefine.org/
- Installing OpenRefine: https://github.com/OpenRefine/OpenRefine/wiki/Installation-Instructions

oXygen

- https://www.oxygenxml.com/

Crossref

- www.crossref.org
- http://help.crossref.org/using_best_practices_depositing
    - Best Practices for Depositing
        - Includes optimum file size for uploads

Microsoft Excel additional information

- Excel 2003 XML Tools Add-in
- Show the Excel Developer Tab on the ribbon