***Step-by-step instructions on getting Islandora metadata into Open Refine***

Prepared October 2018 by Wilhelmina Randtke, Florida Academic Library Services Cooperative

**Software to install.**

Before starting, install MarcEdit and OpenRefine.

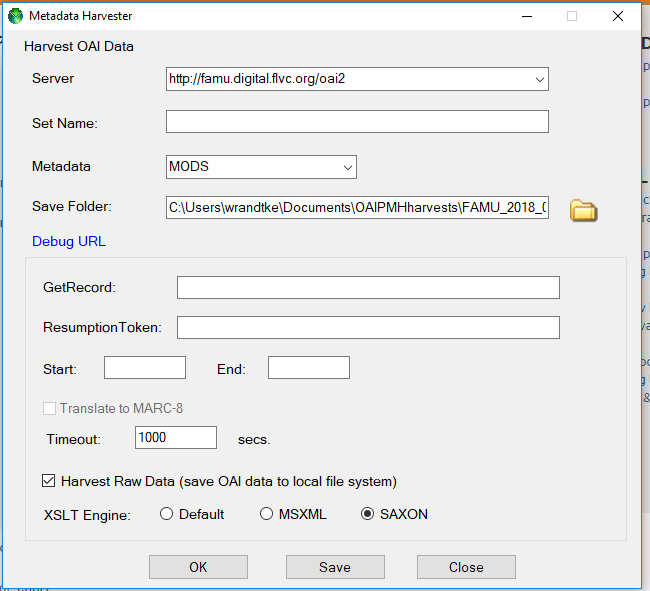
MarcEdit: https://marcedit.reeset.net/downloads

OpenRefine: <http://openrefine.org/download.html>

**Step 1: Get records with MarcEdit.**

OAI-PMH can import from an XML feed, but won't follow the resumptionTokens, so can't be used to get a big set of records by OAI-PMH. You have to use some tool other than Open Refine to get the MODS XML out of Islandora.

* Install MarcEdit.
* Choose the option to “Harvest OAI Records”
* Fill out info (sample screenshot below):
  + Click to "Harvest Raw Data". Checking this box changes some of the other boxes, so that's why you do it first.
  + First, check the box for "Harvest Raw Data"
  + Server = Your site's OAI-PMH repository. For example, <http://spc.digital.flvc.org/oai2> .
  + Set Name
    - You have to use this to get one Islandora collection at a time. If you harvest, the whole site, there's no way to later narrow it to one single collection; getting data for the collection vs site has to be done at harvest.
    - To see the sets available for your site, run the ListSets query. Go to something like <http://spc.digital.flvc.org/oai2?verb=ListSets> in your web browser, and look at what sets are available. If you have many collections, use the browser to get more with the resumptionToken . Go to something like <http://fsu.digital.flvc.org/oai2?verb=ListSets&resumptionToken=601740735> in your web browser.
  + Metadata = MODS
  + GetRecord: leave blank
  + Resumption Token: leave blank
  + Start: leave blank
  + End: leave blank
  + Timeout: By default, this is 100 seconds. Consider changing it to a bigger number.
  + Harvest Raw Data: box should be checked
  + XSLT Engine: It doesn't matter. SAXON may be preferable.
* Click "Save"
* This will save the harvest as a series of XML files in whatever folder on your computer you chose.



**Step 2: Import the MODS XML into Open Refine.**

* Install Open Refine.
* Open Refine is to where you run it on your desktop.  When you double click to run it, it opens a black window with white text and a command line interface.  You don’t actually do anything with that command line interface.  Instead, you visit [http://127.0.0.1:3333](http://127.0.0.1:3333/) in a web browser and you use the web browser to use Open Refine. If you haven't used it before, then try working through this tutorial <https://programminghistorian.org/en/lessons/cleaning-data-with-openrefine> . That tutorial is a way to click through something quick and get past major roadblocks to running the software.
* From the home page of Open Refine at [http://127.0.0.1:3333](http://127.0.0.1:3333/) , look in the top left corner and click "Create Project".
* Get data from "This Computer", click to "Choose Files", and select the files that you downloaded with MarcEdit.
* Click "Next".
* Open Refine will pop up a list of files. You don't need to make any changes and can click to "Configure Parsing Options".
* On the "Configure Parsing Options" page:
  + At the bottom of the screen, make sure that it is selected to Parse data as "XML files".
  + At the top of the screen under the heading "Click on the first XML element corresponding to the first record to load.", scroll down until you see the <metadata> and <mods:mods> elements. Click on the <mods:mods> element (this element has lots of attributes in it and is really long; look at the screenshot to get the idea).
  + Clicking this will take you to a spreadsheet style preview of your data. You can experiment with checking and unchecking boxes in the settings at the bottom of the screen and then click "Update Preview" to try and get a better first pass importing the XML to the Open Refine spreadsheet.
  + In the top right corner, type a "Project Name".
  + Click to "Create Project".
* Now your MODS XML records have been imported into Open Refine.

**Step 3: Understand the limitations of Islandora OAI-PMH configuration and of Open Refine.**

* Islandora has configuration options for OAI-PMH. As of June 2018, those options are set to *NOT* send out info about:
  + Each issue of each Serial. Instead, Islandora will send out a record about the Serial parent object.
  + Each issue of each Newspaper. Instead, Islandora will send out a record about the Newspaper parent object.
  + Each child of each Compound Object. Instead, Islandora will send out a record about the Compound Object parent, but will not send out info about each child object.
* XML allows repeated elements. For example, most elements in MODS can repeat. You can have multiple subjects assigned, multiple creators, etc. Most Open Refine tutorials focus on CSV files and things where there isn't a repeated element with different values. In your newly opened MODS, scroll left and right. At first, it will look as if there are many blank rows, but really what has happened is that when an element is repeated, for example when a MODS record has several subject headings in it, then Open Refine has made one column for that element but started each occurrence on a new row.
* You cannot go from OpenRefine back to XML. In order to do so, you would have to export from OpenRefine to Excel, then use an XSLT stylesheet with Excel to export XML records. Setting up the XSLT stylesheet is complicated and tedious and requires XSLT programming skills. So, the value of getting your records into OpenRefine is that you can verify what is and isn’t a problem, and then you can know whether or not you have a clean up project and how big of a clean up project. You can find and understand problems if you have them. You can also verify that you don’t have problems, if you don’t.