

Capitalizing on Automated Workflow

Expediting Metadata Creation for Electronic Theses and Dissertations at Oklahoma State University

MADISON CHARTIER

Metadata Librarian, Oklahoma State University

Oklahoma Library Association

Technical Services Round Table Workshop

30 October 2020

Outline

I. Establishing Context

A. OSU's Metadata Team

B. OSU's Electronic Theses & Dissertations (ETDs) Collection

C. UMI ProQuest

II. Workflow Tools

A. eXtensible Stylesheet Language Transformations (XSLTs)

B. OpenRefine

III. Live Demo: OSU's ETD Processing Workflow

IV. Summary

Context

OSU's Metadata Team
OSU's ETD Collection
UMI ProQuest

OSU's Metadata Team

EDMON LOW LIBRARY
& Branch Libraries

Today's Hours
24 Hours
View all hours

Renew / Login Directory Calendar Quicklinks

Search the library website

Research Tools & Collections Help & Services Study Spaces & Computers About The Library Ask Us Chat

Popular Links

- [Request Forms](#)
- [Library Services](#)
- [Endnote & Zotero](#)
- [About the OSU Libraries](#)
- [Give to the Library](#)
- [Friends of the Library](#)
- [Branches & Departments](#)
- [Anywhere Library Access](#)
- [Library Jobs](#)
- [Give Us Feedback](#)

Metadata

The Metadata Unit supports the mission and vision of the Library in the planning, development, creation, enhancement, and implementation of non-MARC metadata standards that facilitate discoverability and access to scholarly and cultural heritage resources in digital collections.

Department Staff

[Chartier, Madison](#)

Contact Information

Metadata
215 Library
Edmon Low Library
Stillwater, OK 74078
P: 405-744-9161

Sparse, lean, & mean.

OSU's ETD Collection

OSU's largest online collection

- 25,400+ items
- 2019:
 - 554 new entries
 - 36 separate uploads
 - Max. of 96; min. of 1

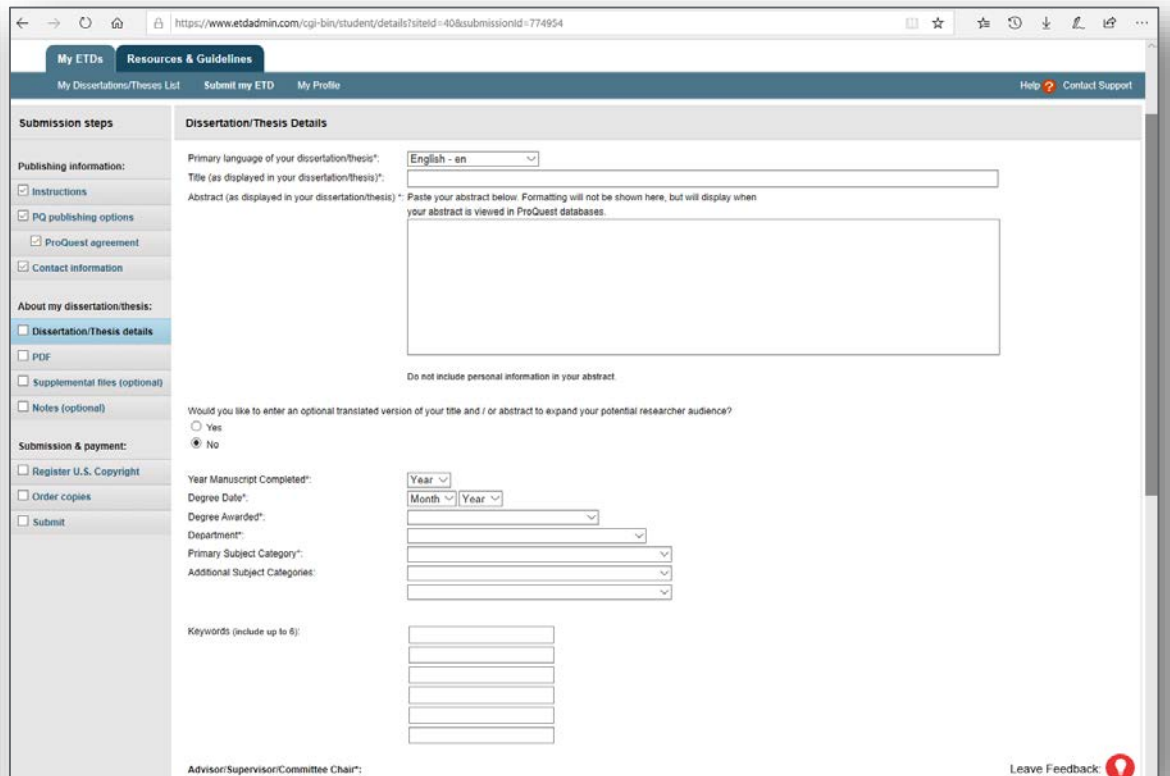
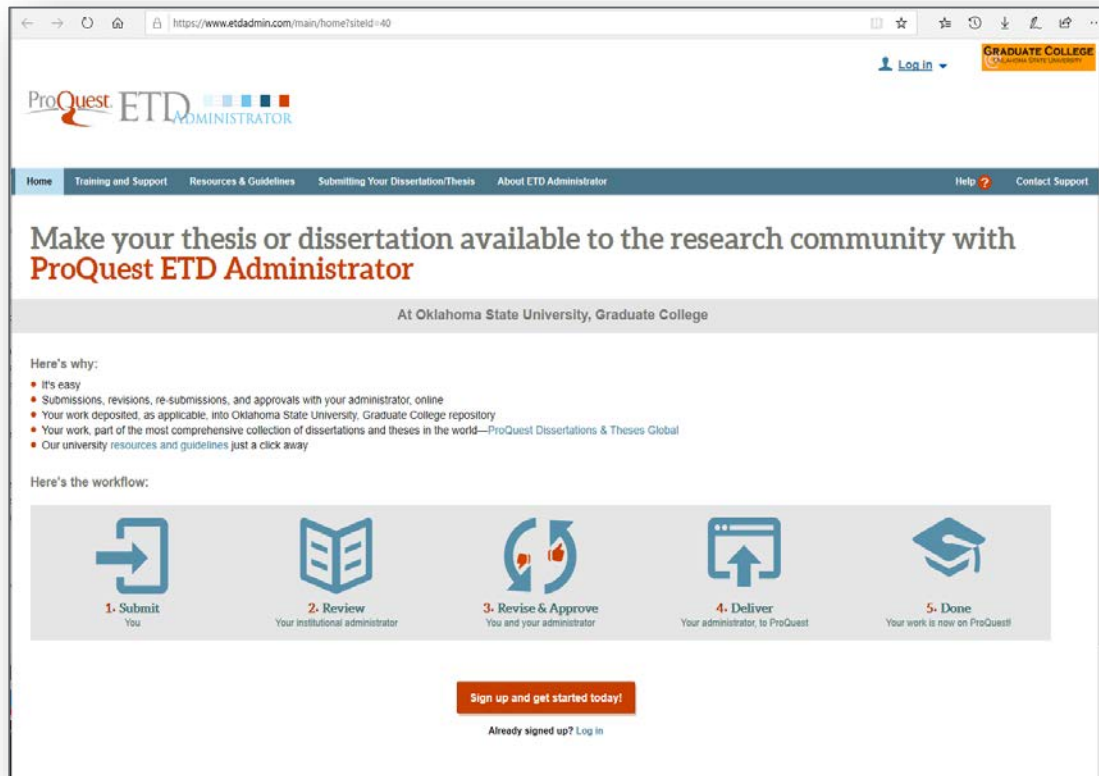
High-priority collection

- Student-faculty work is the pride of OSU!
- We process & publish ASAP upon ingest!

The screenshot displays the SHAREOK website interface. At the top, the logo 'SHAREOK' is followed by the tagline 'advancing Oklahoma scholarship, research and institutional memory'. Below this is a breadcrumb trail: 'SHAREOK Home / Oklahoma State University / OSU - Electronic Theses and Dissertations'. The main content area is titled 'OSU - Electronic Theses and Dissertations'. On the left side, there is a search bar and a navigation menu with options: 'This Community' (selected), 'Search SHAREOK', and a 'BROWSE' section with links for 'This Community', 'By Issue Date', 'Authors', 'Titles', 'Subjects', 'By Series', 'All of SHAREOK', and 'Communities & Collections'. The main content area features a large banner with a graduation cap icon and the text 'Electronic Theses & Dissertations' and 'DIGITAL COLLECTIONS @ OSU LIBRARY'. Below the banner, a section titled 'Collections in this community' lists: 'OSU Dissertations [9597]', 'OSU Master's Report [734]', and 'OSU Theses [15078]'.

A tall order for a single metadata librarian...

UMI ProQuest



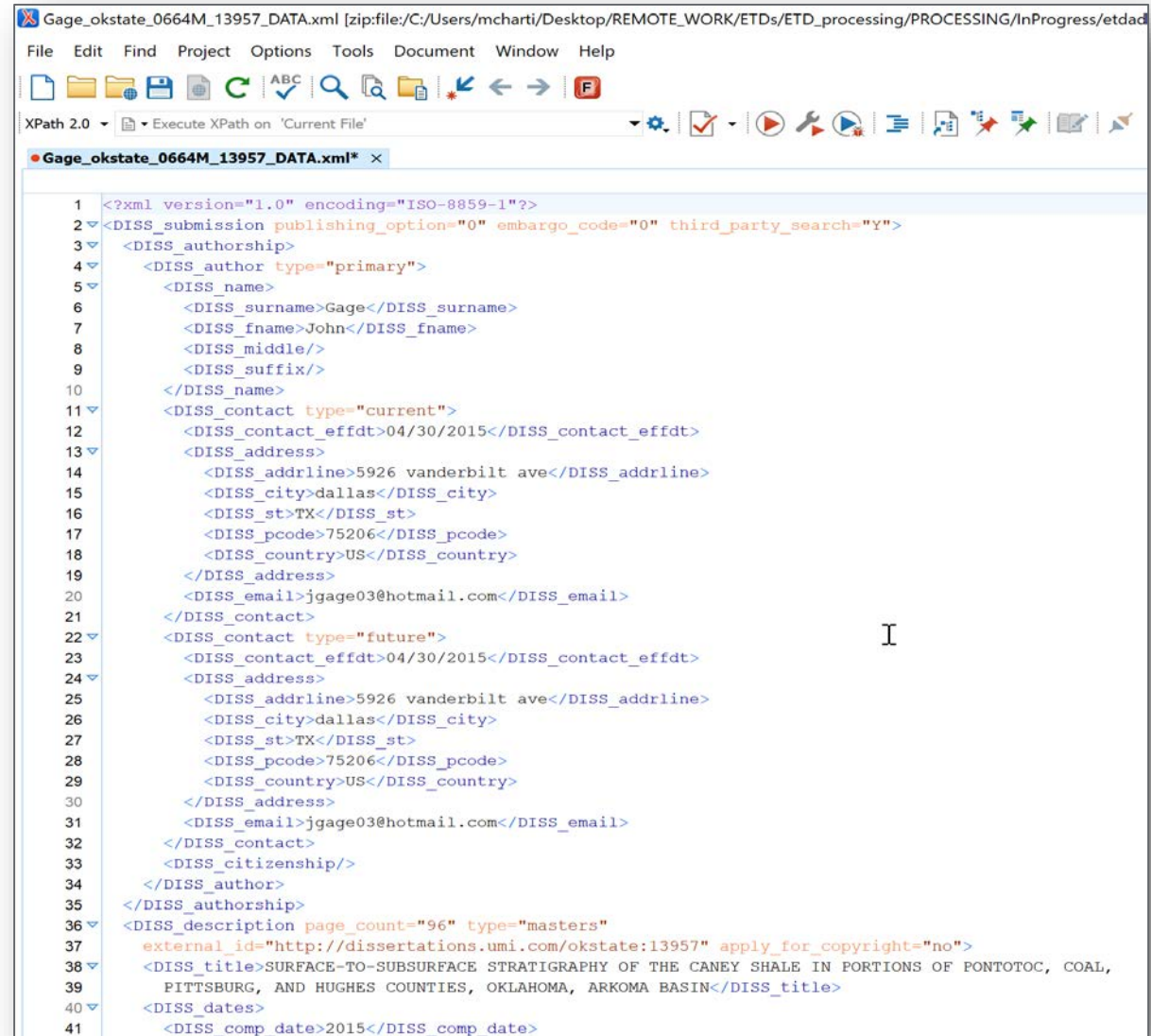
Images provided by Dr. Jean Van Delinder, Senior Associate Dean of the Graduate College.

Consistent, uniform metadata → **AUTOMATED BATCH PROCESSES!**

Consistent XMLs

Consistency enables everything!

- Same fields/tags
- Same value formats
- Individual values for each ETD
- Every item's metadata is set up the same way
- Permits batch processes & automation



```
1 <?xml version="1.0" encoding="ISO-8859-1"?>
2 <DISS_submission publishing_option="0" embargo_code="0" third_party_search="Y">
3   <DISS_authorship>
4     <DISS_author type="primary">
5       <DISS_name>
6         <DISS_surname>Gage</DISS_surname>
7         <DISS_fname>John</DISS_fname>
8         <DISS_middle/>
9         <DISS_suffix/>
10      </DISS_name>
11     <DISS_contact type="current">
12       <DISS_contact_effdt>04/30/2015</DISS_contact_effdt>
13       <DISS_address>
14         <DISS_addrline>5926 vanderbilt ave</DISS_addrline>
15         <DISS_city>dallas</DISS_city>
16         <DISS_st>TX</DISS_st>
17         <DISS_pcode>75206</DISS_pcode>
18         <DISS_country>US</DISS_country>
19       </DISS_address>
20       <DISS_email>jgage03@hotmail.com</DISS_email>
21     </DISS_contact>
22     <DISS_contact type="future">
23       <DISS_contact_effdt>04/30/2015</DISS_contact_effdt>
24       <DISS_address>
25         <DISS_addrline>5926 vanderbilt ave</DISS_addrline>
26         <DISS_city>dallas</DISS_city>
27         <DISS_st>TX</DISS_st>
28         <DISS_pcode>75206</DISS_pcode>
29         <DISS_country>US</DISS_country>
30       </DISS_address>
31       <DISS_email>jgage03@hotmail.com</DISS_email>
32     </DISS_contact>
33     <DISS_citizenship/>
34   </DISS_authorship>
35 </DISS_authorship>
36 <DISS_description page_count="96" type="masters"
37   external_id="http://dissertations.umi.com/okstate:13957" apply_for_copyright="no">
38 <DISS_title>SURFACE-TO-SUBSURFACE STRATIGRAPHY OF THE CANEY SHALE IN PORTIONS OF PONTOTOC, COAL,
39 PITTSBURG, AND HUGHES COUNTIES, OKLAHOMA, ARKOMA BASIN</DISS_title>
40 <DISS_dates>
41 <DISS_comp date>2015</DISS_comp date>
```

Consistency → Batch Processing → Automation

Workflow Tools

eXtensible Stylesheet Language Transformations
OpenRefine

XSLT #1: Metadata Extraction & Prep

- 1) Pulls and formats metadata from XML files
- 2) Assigns Qualified Dublin Core field labels
- 3) Sorts out embargoed and restricted entries
- 4) Generates 2nd XML file with prepped metadata

```
<xsl:template match="/">
  <xml>
    <table>
      <xsl:choose>
        <xsl:when test="DISS_submission[@third_party_search = 'Y'] and DISS_submission[@embargo_code = '0']">
          <xsl:apply-templates select="*/DISS_submission[@third_party_search = 'Y']">
            <xsl:sort select="DISS_authorship/DISS_author/DISS_name/DISS_surname"/>
          </xsl:apply-templates>
        </xsl:when>
        <xsl:otherwise/>
      </xsl:choose>
    </table>
  </xml>
</xsl:template>

<xsl:template match="DISS_submission">
  <!-- Changed from tr to th to better format XML transformation. Previous way read each data entry as a separate row vs. a column cell. Save time on OpenRefine cleanup MVC -->
  <thead>
    <xsl:text>filename</xsl:text>
  </thead>
  <th>
    <td>
      <xsl:apply-templates select="DISS_content/DISS_binary"/>
    </td>
  </th>

  <thead>
    <xsl:text>dc.contributor.advisor</xsl:text>
  </thead>
  <th>
    <td>
      <xsl:apply-templates select="DISS_description/DISS_advisor"/>
    </td>
  </th>

  <thead>
    <xsl:text>dc.contributor.author</xsl:text>
  </thead>
  <th>
```

Preps metadata to meet standards established in the collection's Metadata Application Profile.

XSLT #2: Metadata File Merger

```
<?xml version="1.0" encoding="UTF-8"?>
<xsl:stylesheet xmlns:xsl="http://www.w3.org/1999/XSL/Transform"
  xmlns:xs="http://www.w3.org/2001/XMLSchema" exclude-result-prefixes="xs" version="2.0" xmlns:fn="http://www.w3.org/2005/xpath-functions">

  <!-- Below is a recommended insert from Oxygen XML Editor online: https://www.oxygenxml.com/forum/topic11375.html.
  This is meant to help with merging the multiple XML files into one.

  This XSLT serves as STEP 2 in the ETD workflow. The following code should take all of the XML files in a batch (make sure they're separated

  <xsl:output method="xml"/>
  <xsl:template match="/">
    <xsl:copy>
      <xsl:apply-templates mode="rootcopy"/>
    </xsl:copy>
  </xsl:template>

  <!-- PROBLEM: in this section. Makes multiple copies. Copies all files 30 times. ??? -->
  <xsl:template match="node()" mode="rootcopy">
    <xsl:copy>
      <xsl:variable name="folderURI" select="resolve-uri('.', base-uri())"/>
      <xsl:for-each
        select="collection(concat($folderURI, '?select=*.xml;recurse=yes'))/*/*node()">
        <xsl:apply-templates mode="copy" select="."/>
      </xsl:for-each>
    </xsl:copy>
  </xsl:template>

  <!-- Deep copy template -->
  <xsl:template match="node() | @*" mode="copy">
    <xsl:copy>
      <xsl:apply-templates mode="copy" select="@*" />
      <xsl:apply-templates mode="copy" />
    </xsl:copy>
  </xsl:template>

  <!-- Handle default matching -->
  <xsl:template match="*" />

</xsl:stylesheet>
```

I

Merges individually prepped files into one mass file.

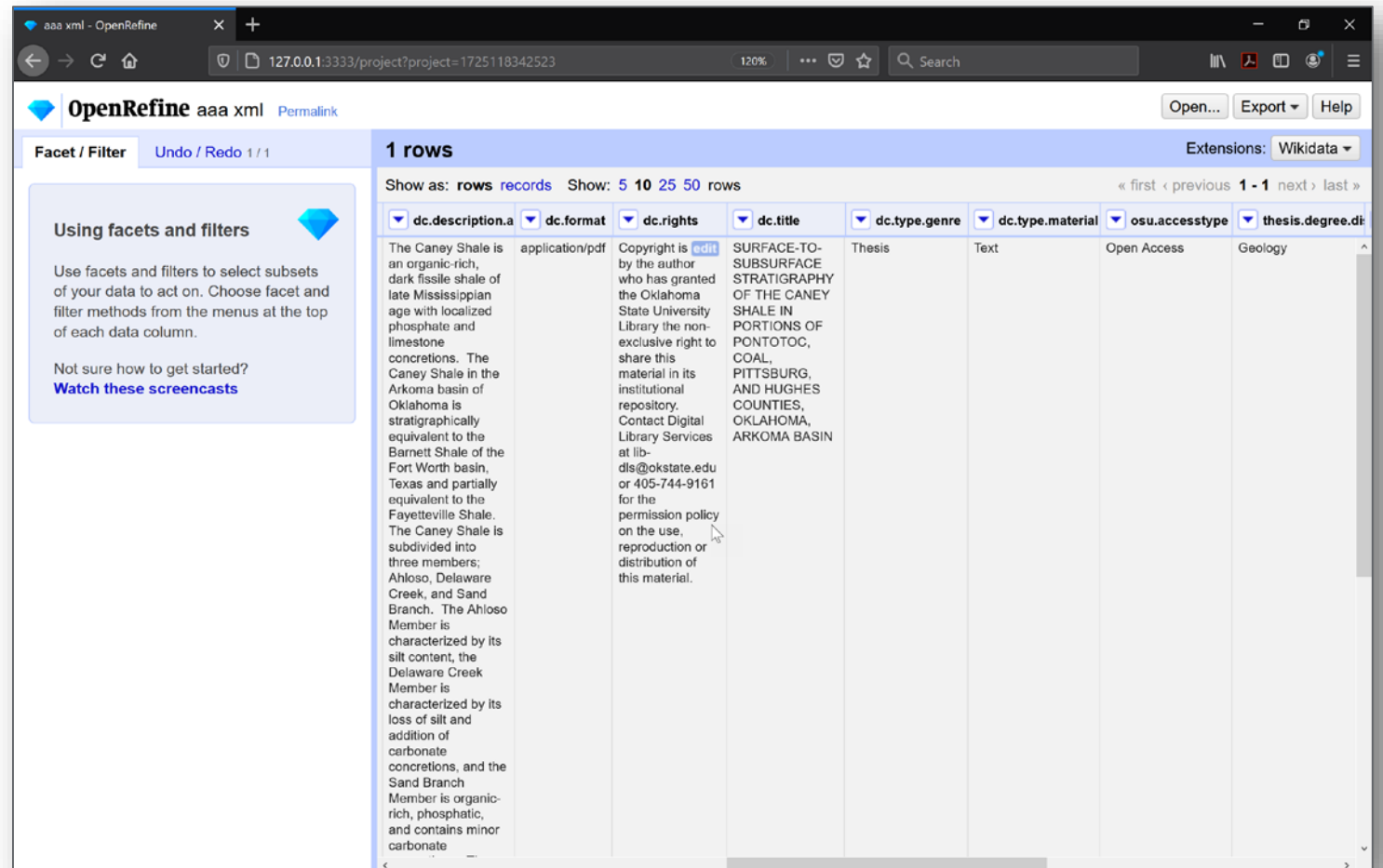
OpenRefine

Open-source tool in popular use at OSU Library

- Cleans messy data!
- Use for metadata cleaning

Enables our file conversion!

- Can upload compiled XML file...
- ...clean metadata...
- ...and export as a comma-separated values (CSV) file!



All that remains = accuracy checks against PDF files!

Live Demo

OSU's ETD Processing Workflow

Summary

Madison Chartier

Digital Resources & Discovery Services

O | 405.744.9536

E | madison.chartier@okstate.edu

215A Edmon Low Library

<https://info.library.okstate.edu/c.php?g=916329>

THANK YOU!

