10-8-2013

ETD_CON Utility and User Manual

Logan E. Jewett
Iowa State University, lejewett@iastate.edu

ETD Conversion Utility

Instruction Manual

Logan Jewett
October 2013

Digital Repository @ Iowa State University
http://lib.dr.iastate.edu

Acknowledgements

The author of this document would like to acknowledge Kelly Thompson for her assistance in bug testing, Wendy Robertson for sharing documentation and resources from the University of Iowa's own ETD conversion project, and Harrison W. Inefuku for his gracious leniency and connections in solving ETD conversions.

Table of Contents

About

Development

The "Electronic Theses & Dissertations Conversion Utility" or "ETD_CON" was developed in 2013 by Logan Jewett to be used by Iowa State University. The motivation for the project was to simplify and streamline processing those ETDs and preparing them for batch upload onto Iowa State's Digital Repository. This involved decompressing submissions, organizing them, and converting their metadata from a ProQuest format to the bepress format used by the Digital Repository.

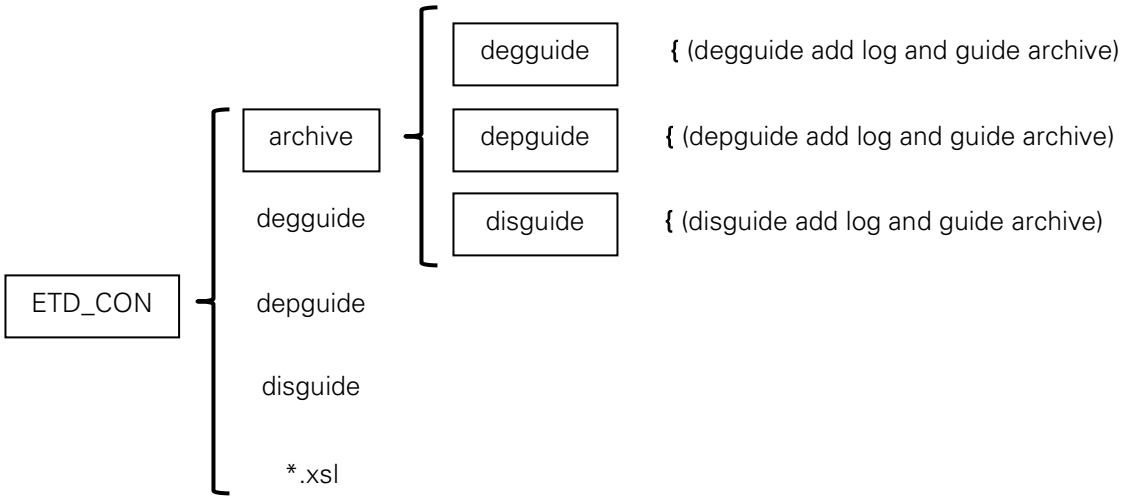
Technical Overview

ETD_CON was written in Java in the Eclipse IDE with WindowBuilder to aid in GUI development. It utilized the Saxon XSLT plugin to handle XSL processing. Decompressing and sorting ETD submissions was handled with the standard Java library. Converting the XML metadata involves utilizing the transformer factory from Saxon to evaluate and apply the XSL transform to the ProQuest metadata.[1] The degree, department, and discipline fields then are converted to their corresponding bepress values utilizing a series of plaintext crosswalks. To work properly, the directory containing ETD_CON requires the presence of a file with an 'xsl' extension to be utilized by the transformer factory, as well as three extensionless text files named 'degguide', 'depguide', and 'disguide' to be utilized as crosswalks.

Disclaimer

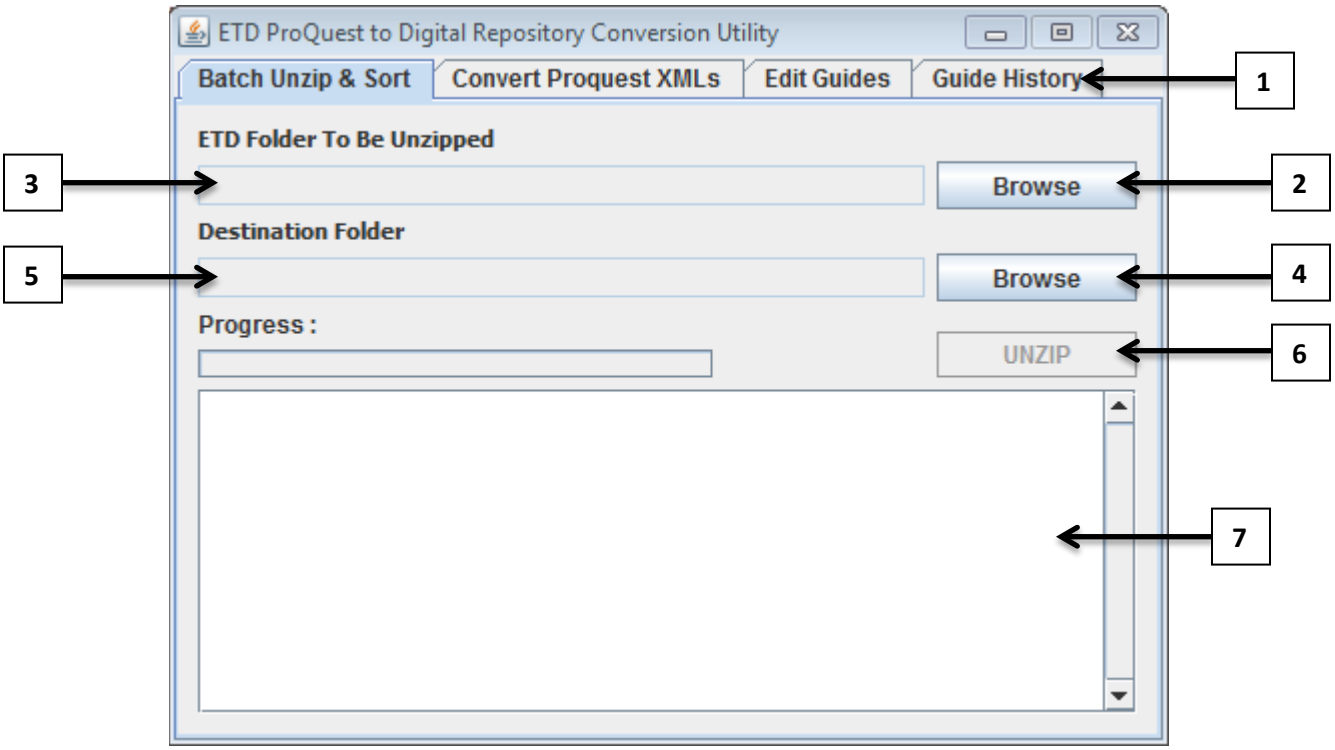
Although ETD_CON has been tailored to suit the needs of Iowa State University, the instruction manual, source code, coding documentation, as well as a working build of the program will all be made available free to use, modify, and distribute via the URLs provided below. The software is being distributed in the interest of being useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.

Contents of ETD_CON



[1] XSL transform was built on work done by the University of Iowa as described in Shawn Averkamp and Joanna Lee, "Repurposing ProQuest metadata for Batch Ingesting ETDs into an Institutional Repository," *code4lib Journal* 7 (June 26 2009): http://journal.code4lib.org/articles/1647

Unzipping Files



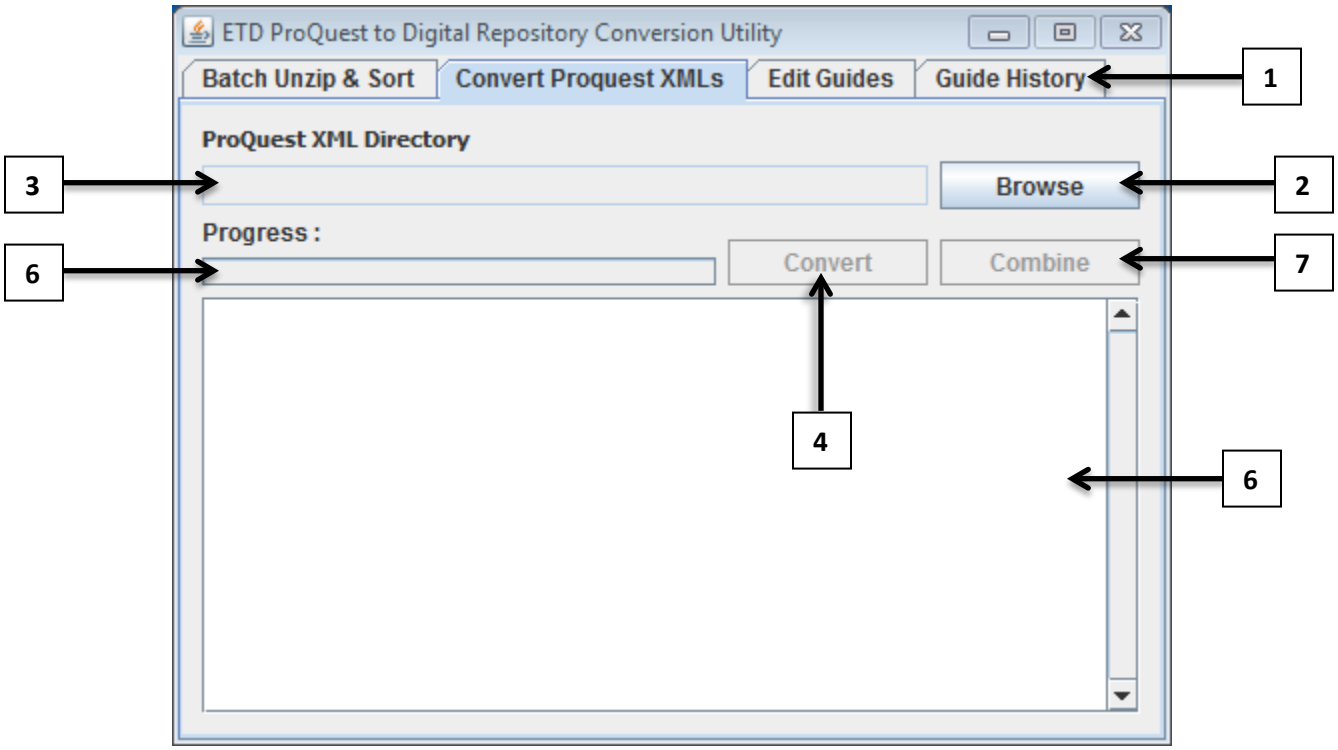
1. The different functions offered by the ETD conversion utility (Following info about bolded tab)
   Batch Unzip & Sort: unzip and sort directories of unprocessed ETDs.
   Convert Proquest XMLs: run the XSL transform to convert ETD metadata.
   Edit Guides: view and alter the crosswalks for the Degree/Department/Disciplines.
   Guide History: view (and restore from) the revision history of the various crosswalks.

2. Press to browse for and select the directory containing zipped ETDs that need unzipped and sorted.
3. Once the folder containing all of the zipped ETDs has been selected (utilizing the browse button indicated by [2]) this field will display the full pathname of the selected directory.
4. Press to browse for and select the directory that the zipped ETDs will be unzipped and sorted to.
5. Once the destination folder has been selected (using the browse button indicated by [4]) this field will display the full pathname of the selected directory.
6. Once both the relevant directories have been selected this button will become available. Press to unzip the files in the ETD directory and sort them to the destination folder.
7. Live-updated text field that indicates which file is currently being unzipped and where each of the individual contents is being sorted to.

Batch Unzip Instructions

1. Press the browse button indicated by [2] to select the directory containing the compressed/zipped ETDs that need unzipped and sorted. A new window will appear that allows navigation to the directory. The location of the folder once selected will appear in the text field indicated by [3].
2. Press the browse button indicated by [4] to select the directory that the ETDs will be decompressed/unzipped and sorted to. A new window will appear that allows navigation to the directory. The location of the folder once selected will appear in the text field indicated by [5].
3. The UNZIP button indicated by [6] should now be available once the two directories have been selected. Press to commence the batch unzip and sort.
4. The text area will update will update with information about what files are being decompressed and where the contents are being sent. After completion be sure to look through the contents of the text area and finished product for any errors.

Metadata Conversion



1. **The different functions offered by the ETD conversion utility (Following info about bolded tab)**
   Batch Unzip & Sort: unzip and sort directories of unprocessed ETDs.
   Convert Proquest XMLs: run the XSL transform to convert ETD metadata.
   Edit Guides: view and alter the crosswalks for the Degree/Department/Disciplines.
   Guide History: view (and restore from) the revision history of the various crosswalks.

2. Press to browse for and select the directory containing the ProQuest metadata XMLs that need to be converted to the bepress metadata schema. Directory is set to the sorted XMLs automatically when a Batch Unzip & Sort is run.
3. Once the folder containing all of the ProQuest metadata XMLs has been selected (utilizing the browse button indicated by [2]) this field will display the full pathname of the selected directory.
4. Once a directory has been selected this button will become available. Press to convert the ProQuest metadata XMLs to the bepress metadata schema utilizing the XSL transform and crosswalk guides in the program directory.
5. Live-updated text field that indicates which metadata file is currently being converted and what the Degree, Department, and Discipline fields are being mapped to.
6. Once a conversion has been performed this button will become available. Press to combine the converted metadata XMLs into a single batch-upload-ready XML.

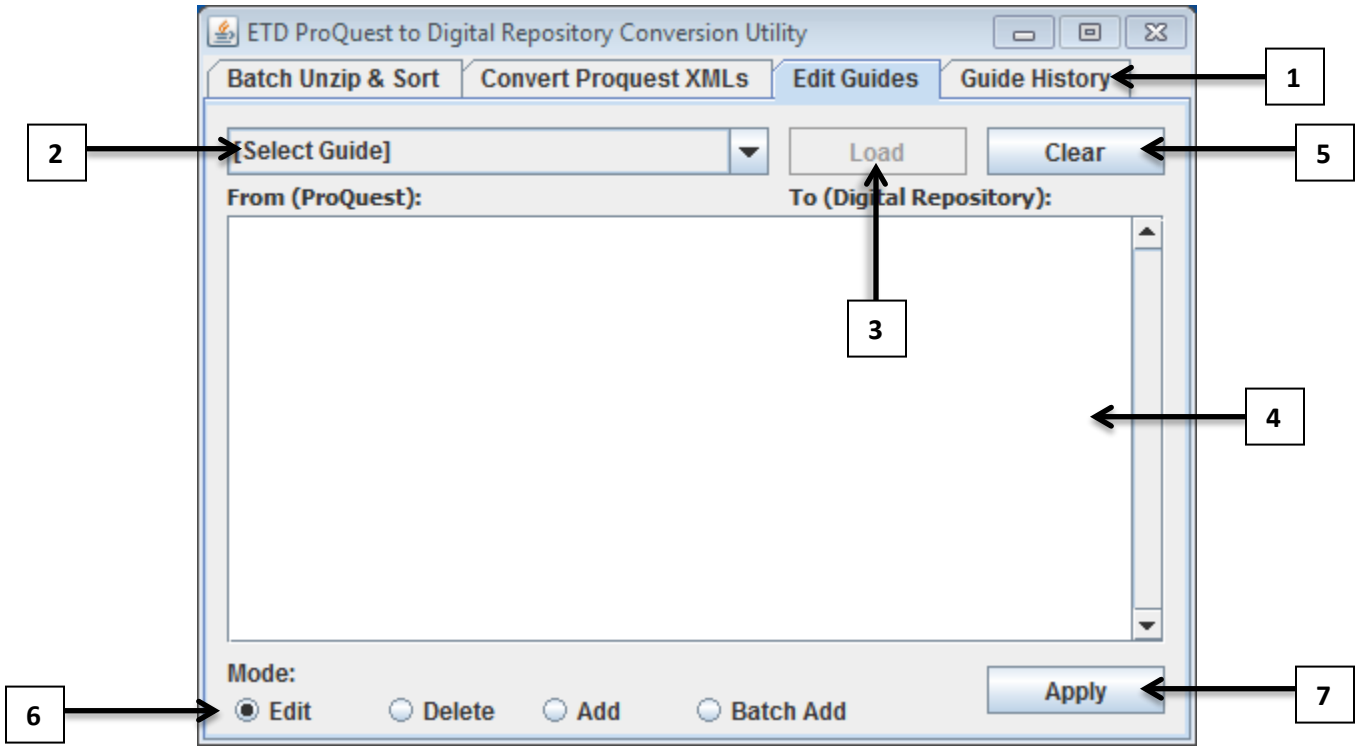
Convert Instructions

1. Press the browse button indicated by [2] to select the directory that contains the XML metadata that needs to be converted. If ETD_CON had already performed a batch unzip then the field indicated by [3] will be set to the ETDs being worked on automatically. Otherwise the location of the folder once selected will appear in the text field indicated by [3].
2. Once a directory is selected the convert button indicated by [4] will now be available. Press to convert the XML metadata utilizing the xsl and guides currently located in the directory containing ETD_CON.
3. As the conversion progresses, the text area will update with the metadata currently being worked on and what each of the degree, department, and discipline fields are being mapped to. Be sure to look through to see if something was mapped improperly.
4. At the end of a conversion, if a field did not have a crosswalk in its associated guide, the text area in the Edit Guides tab will populate with the offenders and a prompt will appear to indicate that some fields were not mapped. To remedy simply edit and add the corresponding crosswalks. If the guides are edited a conversion will need to be run again to re-map those fields.
5. Once the metadata has been converted, press the combine button indicated by [6] to concatenate the XMLs into a single batch upload ready document. The xml will be located at:

   [Directory of sorted/decompressed ETDs]>XML>Digital Repository XML

   The XML will be named *ETDmetadataYYYYMMDD* where the "YYYYMMDD" will be the date that the combine was run on.

Editing the Guides



1. The different functions offered by the ETD conversion utility (Following info about bolded tab)
   Batch Unzip & Sort: unzip and sort directories of unprocessed ETDs.
   Convert Proquest XMLs: run the XSL transform to convert ETD metadata.
   Edit Guides: view and alter the crosswalks for the Degree/Department/Disciplines.
   Guide History: view (and restore from) the revision history of the various crosswalks.

2. Dropdown menu containing all of the crosswalk guides used by the conversion utility. Click to select from this list which guide to view/work on.
3. Press to load the most current revision of the selected guide.
4. Editable text field used to display/edit a loaded guide, or to enter new crosswalks to add to the guide(s).
5. Press to clear the text field of the guide that was previously loaded. Only one guide can be loaded at time.
6. Modes (See following section for more information).
7. Click to apply any of the desired changes.

Modes

Edit: ETD_CON is always set to this mode by default. This mode is used for standard correcting of the crosswalk. ETD_CON will check the number of crosswalks currently entered against the most current revision to test for difference in length. A disparity will cause the appearance of an error message to prompt the user to switch to the appropriate mode to perform that particular action. ETD_CON will check to see if crosswalks are "properly formatted" and prompt the users which ones require correction.

Delete: Switching to this mode will remove the length check from edit mode and will allow for removal of crosswalks. ETD_CON will prompt the user before applying, with an additional notification if the user is attempting to remove more than half of the guide. ETD_CON will check to see if crosswalks are "properly formatted" and prompt the users which ones require correction. (Defunct as of release 1.0.0) ETD_CON does not allow the application of a blank guide.

Add: This mode will take the crosswalks entered into the text field and add them to the selected guide. ETD_CON will check to see if crosswalks are "properly formatted" and prompt the users which ones require correction. When an 'add' is performed the "add log" will be updated to include which crosswalks were added and when.

Batch Add: This mode allows for the addition of crosswalks to multiple guides in conjunction with a particular formatting. ETD_CON will switch to this mode and populate the text field if there are degrees, departments, or disciplines that it comes across that have no crosswalk while performing a conversion. If this occurs, ETD_CON will prompt the user after a conversion is complete.

Properly Formatting Crosswalks

ETD_CON utilizes crosswalk rules in the following format:

> ProQuest department/degree/discipline>bepress mapping field 1;field 2;etc.

The item that needs translated from the degree/department/discipline fields in a ProQuest XML appears first, followed by a '>' with no spaces on either side, followed by the item(s) in the bepress schema that they map to. Multiple fields are separated by a semicolon that also has no spaces on either side. Also, each crosswalk is separated onto its own line and must be contained on one line. As an example, this is the mapping utilized by Iowa State for Agronomy and Alternative dispute resolution:

> Agronomy>Agriculture;Agricultural Science;Agronomy and Crop Sciences
> Alternative dispute resolution>Dispute Resolution and Administration

Batch Adds

A "batch add" can be utilized when additions to multiple guides at one time is desired. ETD_CON utilizes the functionality of batch additions when it encounters fields of several different types that it doesn't know how to map when converting metadata. If this occurs while converting the metadata a prompt will appear to indicate that the guides may need additional crosswalks.

To format for a batch add, simply indicate with one of the following 3-letter tags what type the proceeding crosswalks are.

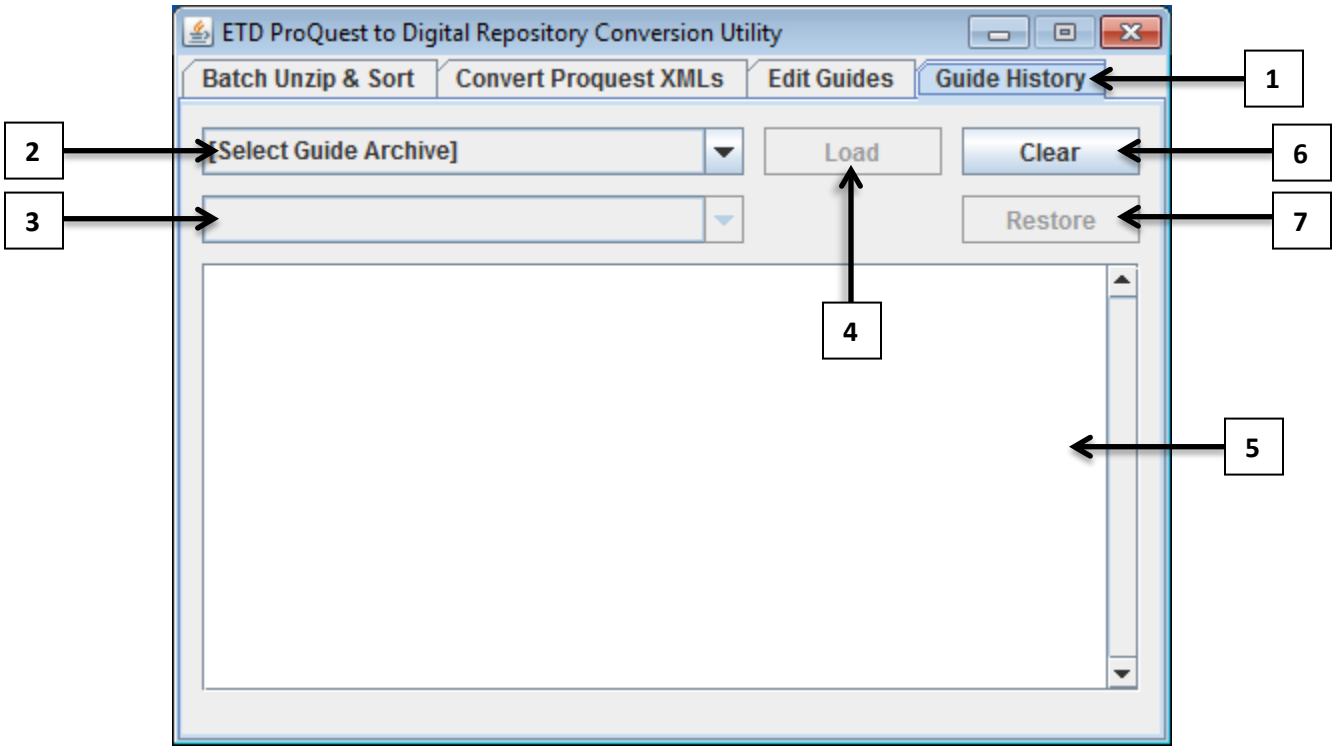
DEG = Degree
DEP = Department
DIS = Discipline

For example:

DEG
M.A.>Master of Arts
DEP
Plant Pathology>Plant Pathology and Microbiology
Engineering Mechanics>Aerospace Engineering
DIS
Wood sciences>Wood Science and Pulp, Paper Technology
DEG
M.Educ.>Master of Education

This example demonstrates several things about the formatting for batch adds. A tag can appear multiple times in a single batch add, so that making sure crosswalks are all grouped is unnecessary. In addition, multiple crosswalks can appear following a single tag, so that giving a tag to each crosswalk is also unnecessary. For both adds and batch adds crosswalks do not need to be listed in alphabetical order. Although the guides are alphabetized, they are routinely checked and ordered by ETD_CON.

Guide Revision History



1. **The different functions offered by the ETD conversion utility (Following info about bolded tab)**
   Batch Unzip & Sort: unzip and sort directories of unprocessed ETDs.
   Convert Proquest XMLs: run the XSL transform to convert ETD metadata.
   Edit Guides: view and alter the crosswalks for the Degree/Department/Disciplines.
   Guide History: view and restore the revision history for the various crosswalks.

2. Dropdown menu to select which of the crosswalk guide archives
3. Once a guide is selected with the dropdown menu in [2] this menu will populate with its revision history. Select a revision to load/restore. "Add Log" is a time-stamped record of all crosswalk additions.
4. Press to load the selected guide revision.
5. Text area that will display the loaded guide revision
6. Press to clear the text field of the guide revision that was previously loaded. Only one guide can be loaded at a time.
7. Press to restore a loaded guide revision to working status for ETD conversions. The most recent guide revision will be archived.

Troubleshooting

"XSL File Not Found"
A file with an 'xsl' extension isn't located in the same folder as ETD_CON

"Error Occurred While Constructing Transformer Factory With XSL"
An XSL that isn't properly formatted, isn't readable/accessible, or includes features that were added later than the XSL 2.0 conventions will throw this error.

"XSL File Not Found When Constructing Transformer Factory"
Generally this error would only be thrown if the file disappeared between assigning the pathname to a FileInputStream and constructing the Transformer Factory

"Not all Departments, Degrees, or Disciplines were converted. See Edit Guides tab for more options. Conversion will need to be run again if any guides are edited."
This error appears when items that did not have a crosswalk mapping appeared in the metadata while running a conversion. Switch to Edit Guides to see which fields they were and to add a mapping. If the guide(s) are edited it will be necessary to switch back to Convert ProQuest XML and run the conversion again.

"Error Occurred While Transforming With XSL"
May occur if an XML was not properly formatted. Check debug menu for more information.

"Guide(s) Not Found In Program Directory"
At least one of the guide files "degguide", "depguide", or "disguide" isn't located in the same folder as ETD_CON.

"Error occurred while parsing XML. Check fulltext URLs manually."
ETD_CON at the time of this publishing keeps track of submissions that are not to be published by leaving the full-text URL field blank. This error occurs when ETD_CON cannot parse a particular file's XML.

"Guide Not Found"
One of the guide files was altered after having been selected by ETD_CON. Check to see if the guide files are named properly and attempt to reselect/reload the guide.

"Please Select a Guide"
A guide from the dropdown menu is not currently selected.

"List of Rules in Text Area is shorter than in selected guide. To delete rules select Delete mode below."
This indicates that there is a discrepancy between the number of crosswalks between the working guide and guide currently loaded. If removing crosswalks was intentional switch to delete mode, otherwise try reloading the guide.

"Guide or Archive File Not Found"
This error will typically only appear if the selected file was removed after having selected the guide or archive file in ETD_CON.

"An Unknown Error Has Occurred"
This generally will only occur if ETD_CON is interrupted while attempting to read/write to files. For more information about what error occurred, check the stack trace in the debug menu.

"The Text Area is Empty. Please Enter Rules to be Added."
This error message will appear if an attempt to add a blank text area to one of the guides occurred. Add crosswalks to correct this error.

"Improperly Formatted Rules for a Batch Add"
This generally indicates crosswalks in the text area that do not have an associated 3-letter tag above them to denote which guide they are being sent to.

"Text Area Contains Improperly Formatted Rules"
Check the "Properly Formatted Crosswalks" section of the instruction manual for information about formatting.

"Error Occurred While Updating Department/Degree/Discipline Guide"
This error would typically occur if ETD_CON were prevented or stopped from copying files to the archive or rewriting one of the guides.

"Guide or Archive File Not Found."
User attempted an add or an edit but it could be that the crosswalk guides or archive directory are located in places ETD_CON doesn't recognize. See about section for more information about where files are supposed to be located

"This Action Would Clear the Guide and is Not Allowed."
This would indicate that the user attempted to remove all crosswalks from the selected guide in delete mode with an empty text area. See the section on delete mode for more information.

"Cannot Edit Guide"
This would typically indicate that ETD_CON wasn't able to alter one of the guides. Check to see if the guide is currently open, or if the current user has the ability to edit files on the drive.

"Archive File Not Found"
This would typically indicate that the selected archive file was altered sometime after being selected. Clear the text area and try to reload it.

"Error Occurred While Restoring Guide"
ETD_CON was unable to create a new archive file or rewrite one of the guide files. Check to make sure they aren't open and the current user has the ability to edit files on the drive.

Debug Window

All exceptions are handled and the stack trace is printed to a hidden "debug window". To view the debug window, press the hotkey:

ALT + SHIFT + D

This will cause the debug window to appear (see below). If no serious errors have been encountered this window will be blank, otherwise more information about any exceptions that occurred during operation will appear here. This can be a useful tool if attempting to debug the source code or to get more information about any unusual behavior during operation.