# MarcEdit: Working with Data

TERRY REESE

HEAD OF DIGITAL INITIATIVES

THE OHIO STATE UNIVERSITY

REESET@GMAIL.COM

# Roadmap

Look at the types of data MarcEdit can process

Explore the MARC Tools tooling

Look at:

| Splitting | Joining | Merging |
|---|---|---|
| Character Conversions | Batch Processing | Comparison Tools |
| Exporting Tab Data | Importing Tab Data | |

XML Processing

- Working MarcEdit's XML Processing Toolkit

"This is gobbledygook. I asked for mumbo-jumbo."

# MarcEdit and Data

# File Types

MarcEdit registers and understands 3 file types natively:

- .mrc – this is a binary MARC file
  This is the file that you would load into your ILS

- .mrk – MarcEdit's mnemonic file format
  This is the file type that you would edit in the MarcEditor

- .mrk8 – Legacy file type; this is MarcEdit's mnemonic file format, but defined as saved as UTF8.
  Generally, this extension is only supported for legacy purposes. The application handles both .mrk and .mrk8 files identically at this point.

# Working with other file types

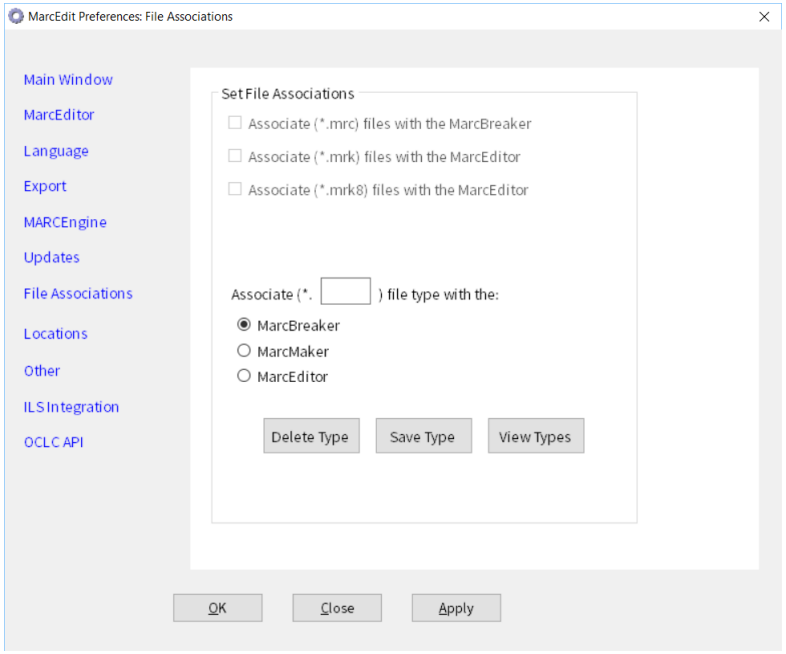
So what is a file type?

- Generally, these are shortcuts that helps the operating system understand what application can read a particular file type.

- However, this doesn't mean that MarcEdit can only read .mrc, .mrk, and .mrk8 files. Any binary MARC file or mnemonic encoded file can be read in MarcEdit. This includes innovative interfaces .bin files, OCLC's .dat files, vendor .001, .marc, .bin files. As long as the files are a format MarcEdit understands, the extension is generally meaningless.

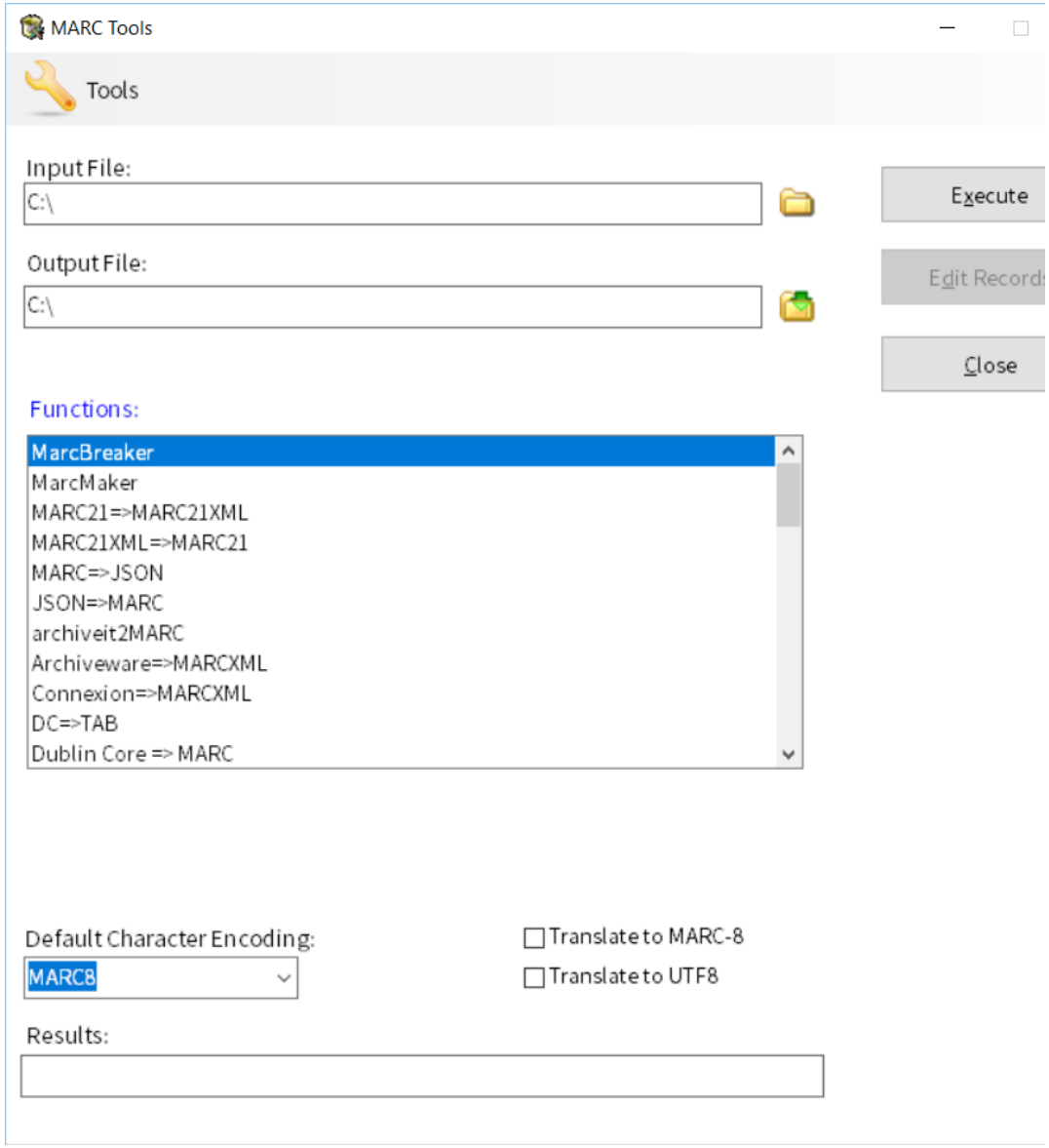
# Working with other file types

And remember, you can always associate different file types with MarcEdit by updating the file associations in the preferences



# MARC Tools

MarcEdit's MARC Tools Hub

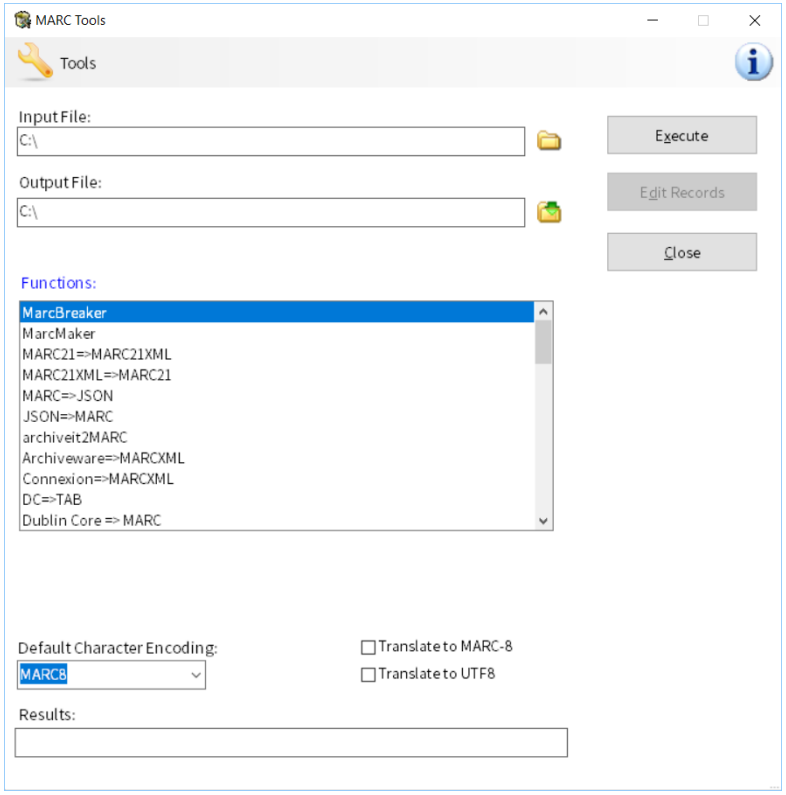


# MARC Tools

I created MARC Tools to be a data hub, or the graphical representation of the MARCEngine component in MarcEdit.

As such, this is one of the places where you can gain access to a number of tools that can be utilized to process MARC, XML, JSON, or delimited data

- Most of these tools can also be accessed via the main window menus.



# I have MARC Data – now what?

Common Workflows 1:

1. Open MarcEdit
2. Open MARC Tools
3. Select the MARC File to Break
4. Select a Save Path
5. Execute
6. Edit in the MarcEditor
7. Make your Changes
8. Save the File
9. Compile back to MARC
10. Load into your ILS

# I have MARC Data – now what?

Common Workflows 2:

1. Open MarcEdit
2. Open the MarcEditor
3. Select File/Open (or control+O)
4. Navigate to your file – open
5. Edit records
6. Save File
7. Compile back to MARC

# Marc Tools

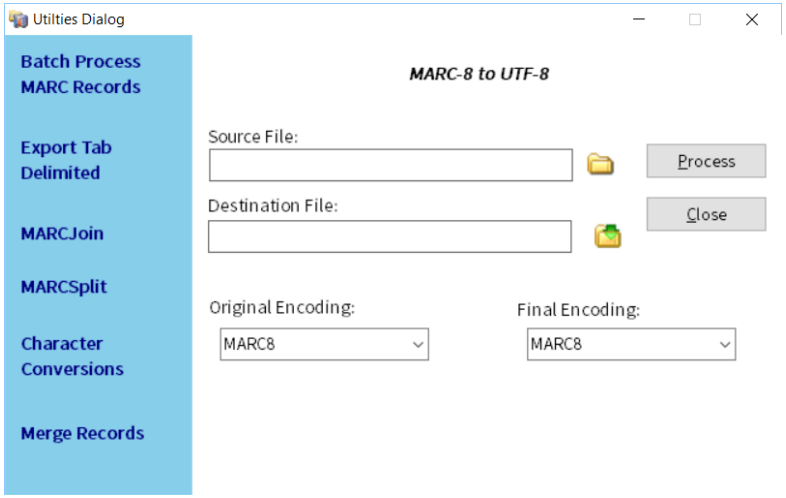
Built-in functions

- MarcBreaker – Tool used to convert MARC records to the MarcEdit mnemonic format
- MarcMaker – Tool used to convert MarcEdit mnemonic format to MARC
- MARC=>MARC21XML – converts MARC to MARC21XML
  - Automatically converts data from MARC-8 to UTF8
- MARC21XML=>MARC – converts MARC21XML to MARC
  - Doesn't automatically convert data from UTF8 to MARC8 – will leave data in UTF8
- MARC => JSON
  - Converts data using MARC to MARC JSON
- JSON => MARC
  - Converts MARC JSON to MARC

# MARC Character Conversions

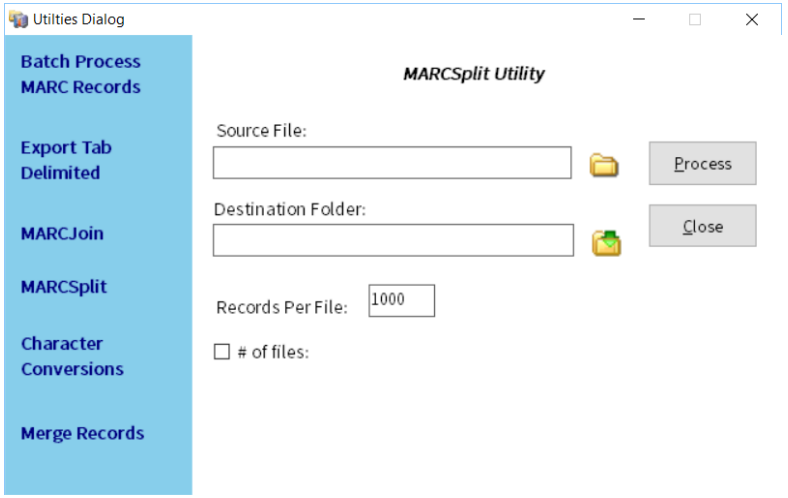
Supports moving between any known Windows Characterset and MARC8.

Can be run from the Breaker/Maker – or as its own standalone utility



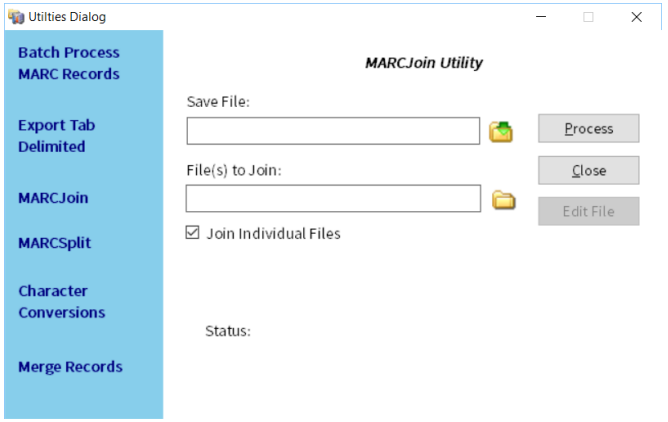
# MARCSplit

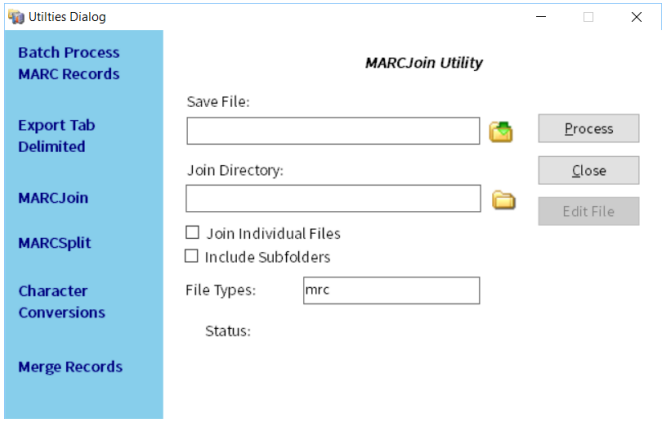
Utility used for splitting large MARC record sets into smaller files



# MARCJoin

Utility used for joining large sets of MARC data to a single file



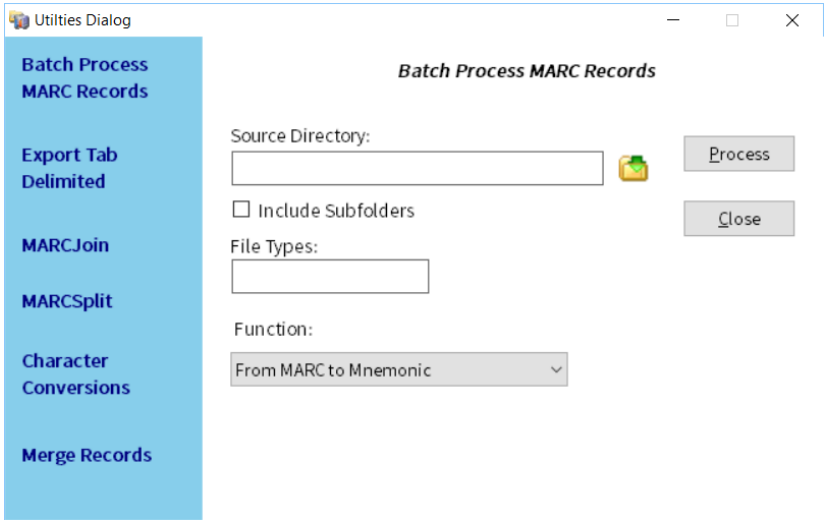



# Batch Record Processor

Allows MarcEdit to process “lots” of files.

Files can be processed against an entire folder’s contents or by file type

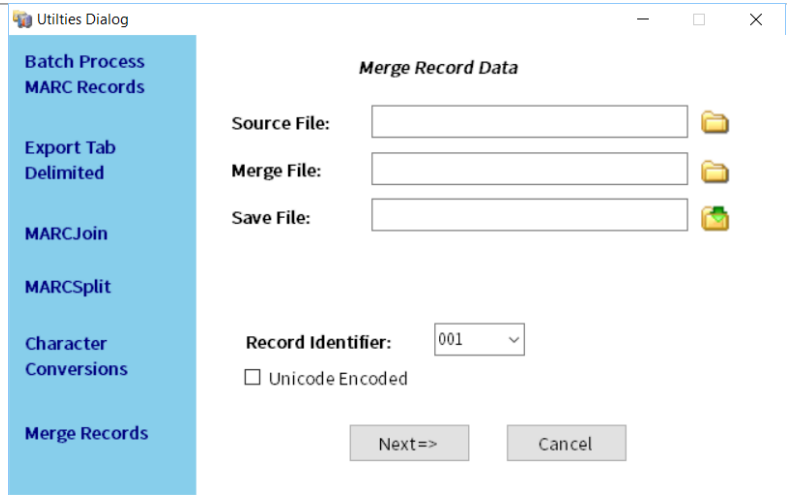
Can utilize any built-in or derived XML Function transformation



# Merge Records Tool

Allows users to merge MARC data from two files

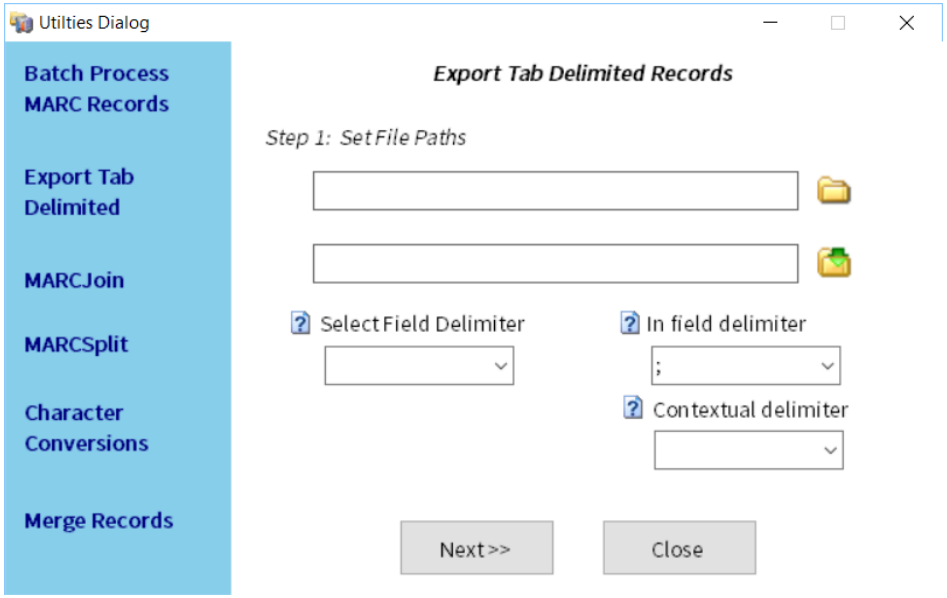
Allows users to merge unique data, selected data and all data.



# Export Tab Delimited Records

Developed as a simple report generating tool

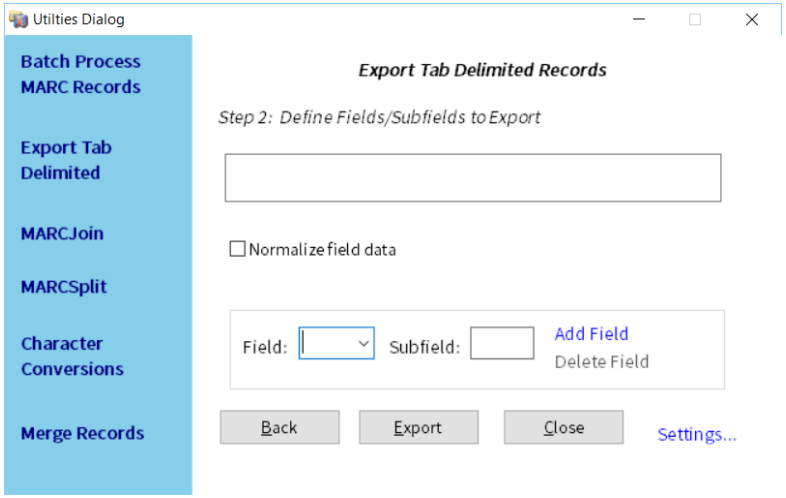
Allows users to create spreadsheets utilizing data found within the MARC records.



# Export Tab Delimited Records

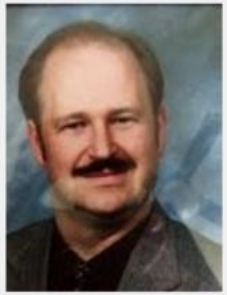
Options:

- Export by position (when working with field < 10
- Export field or field and subfield
- Export multiple subfields in a single string



# MARC Compare

**About RobertCompare**



RobertCompare was originally conceived out of the mind of Dr. Robert Ellett, though at the time, he'd yet to earn the title doctor. Very rarely do I create programs to meet very specific user needs. I've always taken the approach with MarcEdit that tools should be generalizable, and largely, not tied to a specific project. RobertCompare was different. The tool was created to support Mr. Ellett's research on his Ph.D. theses, and only after completion, generalized for wider use.

When I moved MarcEdit from the 4.x to 5.x codebase, I dropped this application because it had seemed to have run its course. This was something Bob would periodically poke at me about occasionally -- I think that he liked the idea of RobertCompare kicking around. Of course, the program was terribly complicated, and without folks asking for it, converting the code from assembly to C# was a pretty steep undertaken.

Well, that changed last year when Bob passed away. I liked Bob a lot -- he was immeasurably kind and easy to get along with. After his passing, I decided I wanted to bring RobertCompare back...I wanted to do something to remember my friend. It's taken a lot more time than I'd hoped, in part due to a move, a job change, and the complexity of the code. However, after an extended absence, RobertCompare has been reintroduced into MarcEdit. I hope folks can find it as useful as my friend.

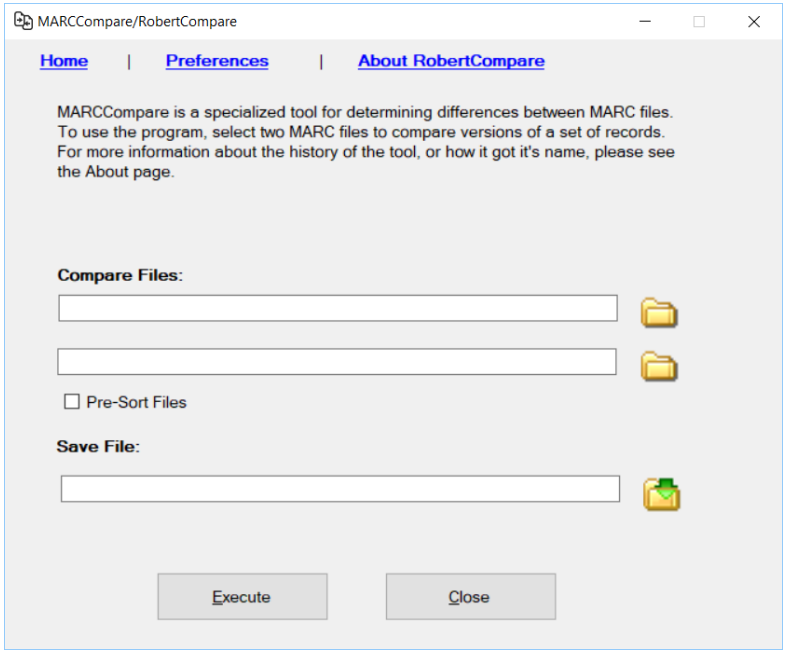
Thanks Bob.

--TR

# MARCCompare

Allows users to compare two MARC files together

- Returns an HTML page that notes differences between the two records.
- Colors noting adds and deletions is configurable
- Note: the process uses standard diffing techniques, so changes between fields can sometimes look like both adds and deletions.
  - (ie, MARC data is sometimes confusing for diff tools)



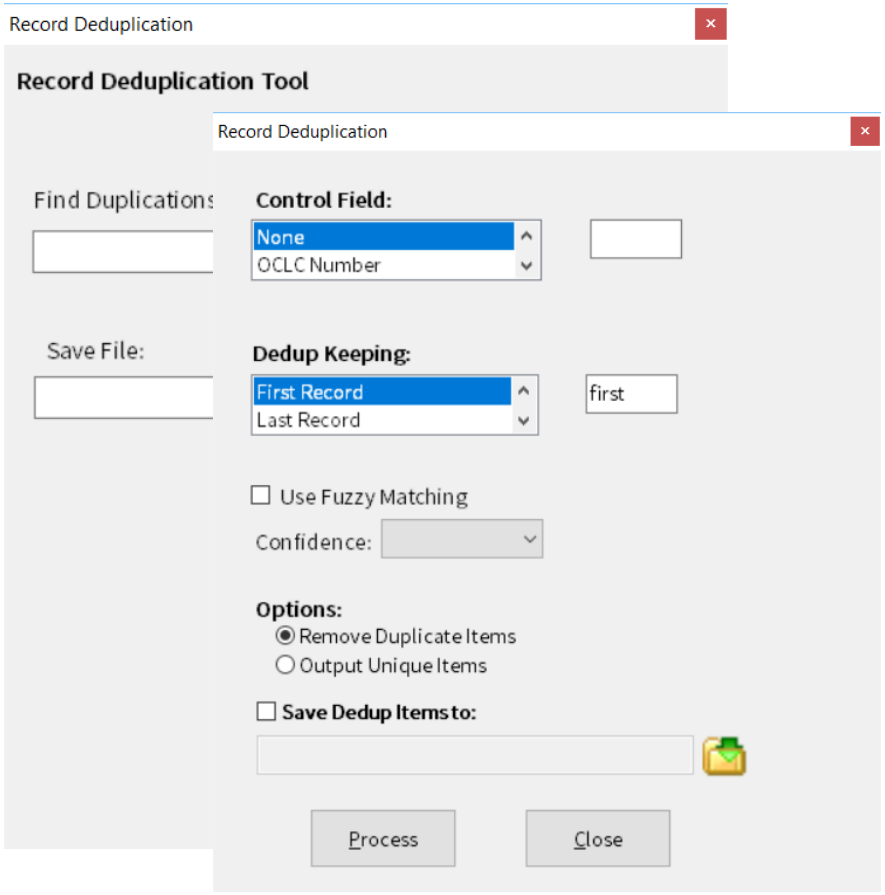
# Record Deduplication

MarcEdit provides a simple dedup tool that can:

- Dedup on a defined control field (any field)
- Dedup on a transaction field (or using an additional transaction field)

Output

- Removes all duplications and saves the duplications to a file
- Prints just unique items within the file (i.e., those without a duplicate pair)

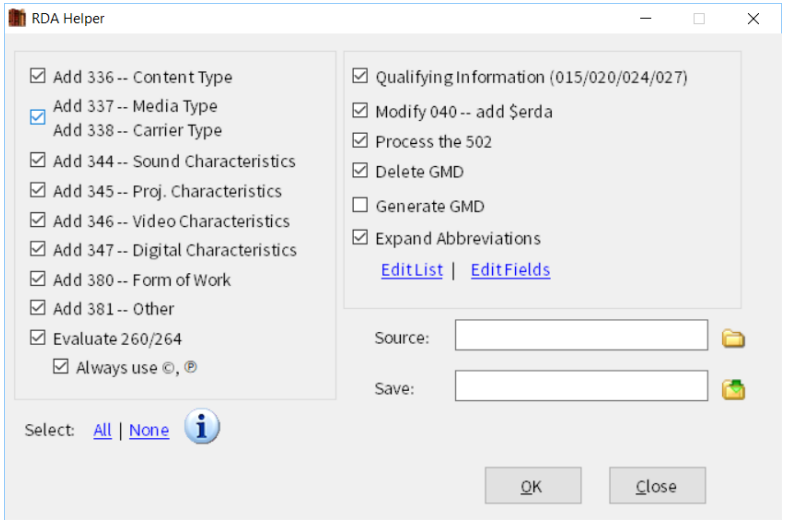


# RDA Helper

I created the RDA Helper with the help of PCC members 6 months prior to LC officially transitioning to RDA

Purpose of the tool was to help facilitate translation between Pre and AACR2 records to RDA

Tools has been gradually expanded to handle not only conversions, but conversions of 880 data pairs as well.



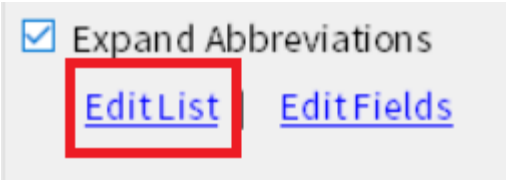
# RDA Helper

RDA Helpers has a number of Options

- You get to determine which fields the Helper processes when looking for abbreviation Expansion


- By default, fields checked are: =245$c, =260, 264, =300, =5

- Likewise, Abbreviations to expand can be found in the Edit List



# RDA Helper

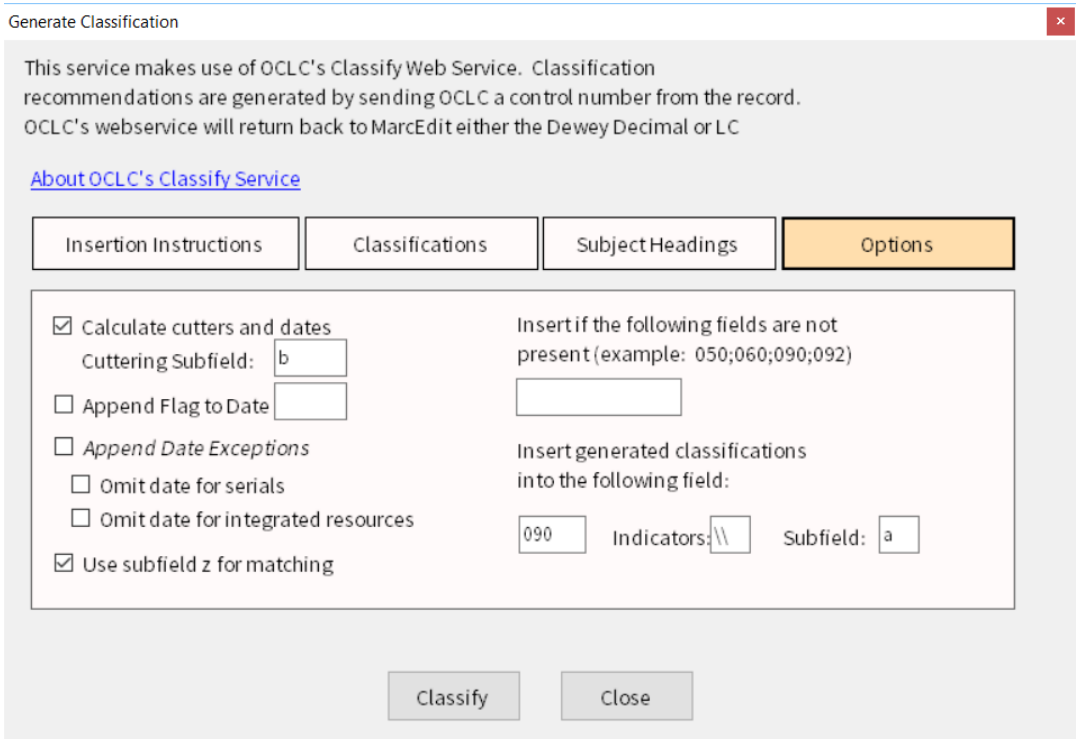
Special Instructions:

- 380 – Because this isn't a controlled field, MarcEdit makes use a genre list at the Library of Congress. This means that these values can be more general than if done by a cataloger.
- 260/264 – Handles many different forms of the field. When the tool is always set to generate a symbol, the tool will utilize MARC8 or UTF8 encoding based on the data.
- Qualifying information – moved qualifiers into a $q. Example: 020 $a02312123 (electronic) to 020 $a02312123$qelectronic
- Process the 502 – converts a dissertation note into a delimited format. Example: 502 $aThesis (M.A.)--University College, London, 1969. to: 502 502 $gThesis$bM.A.$cUniversity College, London$d1969.
- Generate GMD (works on AACR2 encoded data or RDA Encoded data)
- Abbreviation expansion can be customized (using regular expressions) and fields where abbreviations are run can be customized.

# OCLC Classification Services

Developed to allow for automatic call number generation

- Developed utilizing WorldCat's API Services
- Great for adding call numbers to E-books
- Supports LC Classification and Dewey
- Open to all users (not just OCLC Members)



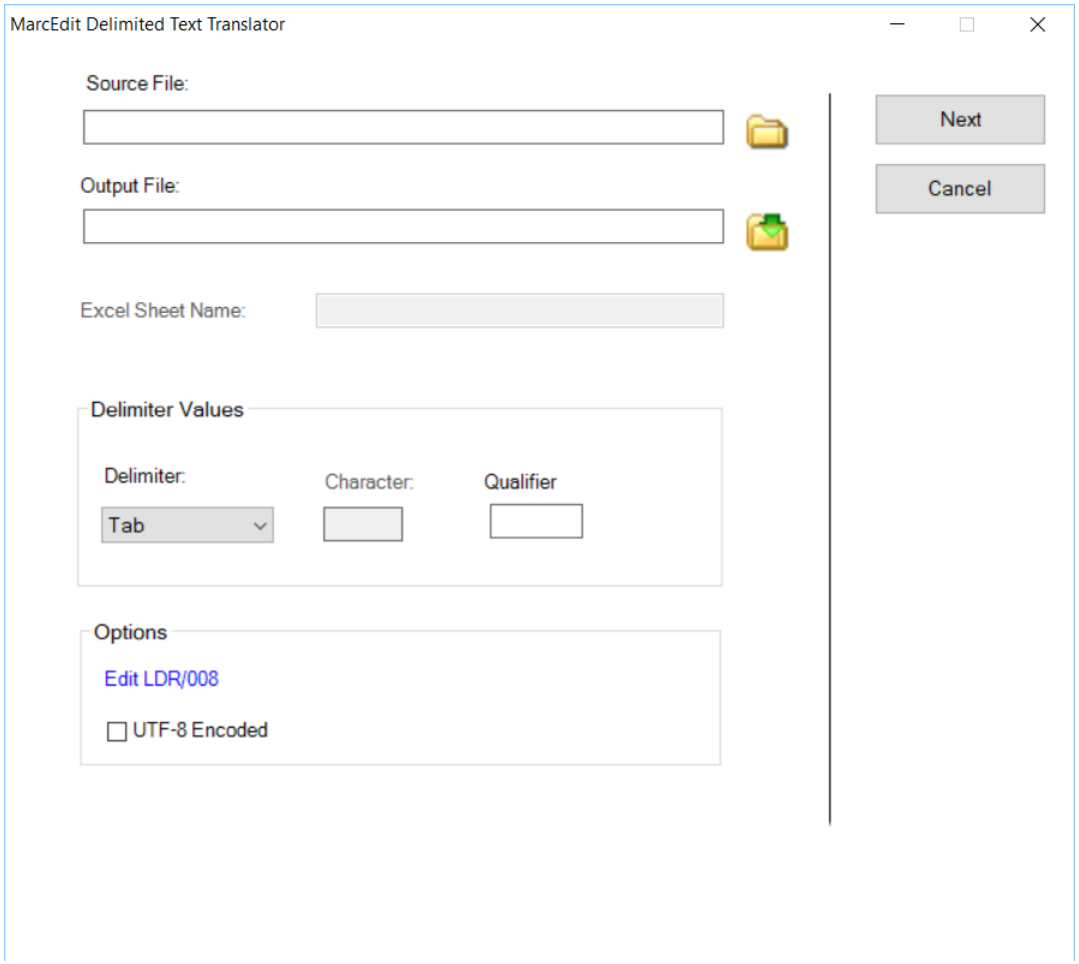
# OCLC Classify Service

MarcEdit can leverage OCLC WorldCat to generate call numbers automatically for files

- Fields used:
  - 001
  - 010$a$z
  - 020$a$z
  - 022$a$z
  - 024$a$z
  - 1xx$a
  - 776$w$z

# MarcEdit Delimited Text Translator

Developed to ease the transition of Worksheet data to MARC

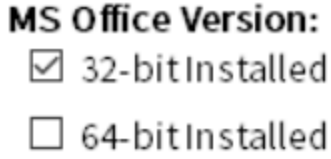


# Delimited text translator

Translator supports all Excel and Access formats, as well as most delimited formats.

Delimiter works with both 32 and 64 bit versions of Office

- However, if you use a 64-bit version of office, you ***must*** designate that in the MarcEdit Preferences.



# Troubleshooting

Most common error that occurs are the following:

- Not all my data is extracted (it only 255 characters are showing up)?
- My ISBN data is all wonky, what happened?
- I'm getting an error when I try to process Excel data (either .xsl or xslx files)?

# Troubleshooting

Error:

- Not all my data is extracted (it only 255 characters are showing up)?
- My ISBN data is all wonky, what happened?

Excel does predictive data typing, so if the first 6 fields are less than 255 characters, Excel will assume that your data is only 255 characters (regardless of it's actual size). This means, the API only sees 255 characters. Likewise, if your ISBN data looks like scientific notation (or another number format), it will convert all data into that format when working with the API. How do you fix this – see this knowledge base article: http://marcedit.reeset.net/correct-isbns-converted-to-scientific-notation-in-excel

# Troubleshooting

Error:

- I'm getting an error when I try to process Excel data (either .xsl or xslx files)?

This error occurs because MarcEdit isn't able to access the Office OBDC component that allows the tool to work with Excel/Access Data. This problem occurs for one of two reasons

1. You are using a 64 bit version of Office. Check the preferences and make sure the correct version of Office is selected.
2. You are using Office 365. Office 365 installs office into a hypervisor, which means that the programming components are not exposed for 3rd party use. To enable this support, you have to install a component from Microsoft to make the application visible.

Information related to solving these issues can be found at the following Knowledge-base article: http://marcedit.reeset.net/troubleshooting-the-delimited-text-translator

# Delimited text translator Options

Wizard-like interface

Supports Unicode data (in excel or delimited file)

Joining (relating) fields

Editing global 008/LDR

# Delimited Text Translator: Mapping format

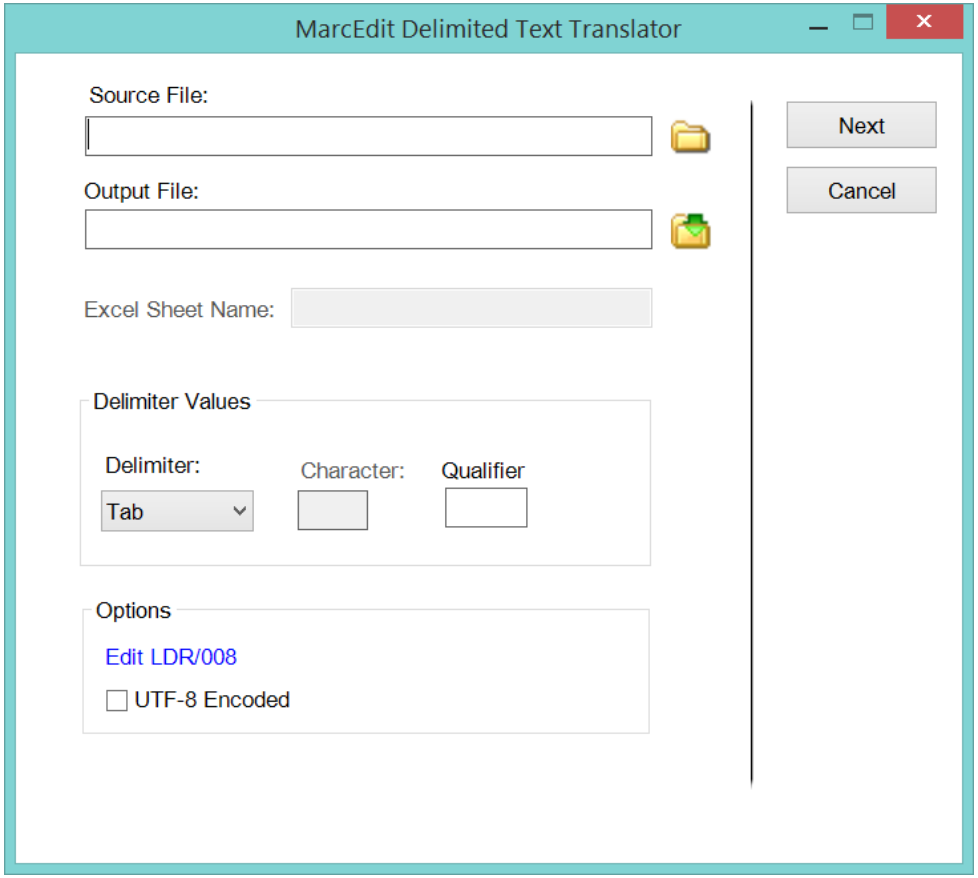
Map to: Field + subfield

Indicators:  Indicator values

Term Punct.:  Trailing punctuation

Arguments – Joining defined items (select and right click on items)

Ability to save templates



# Common Joining techniques

When would I mark a field as repeatable?

- By default, when the Delimited Text translator encounters two like subfields on the same field, it creates a new field. For example:
  column 1: This is a note
  column 2: This is a note 2
  if I mapped column 1 500$a and column 2 to 500$a, by default, MarcEdit would generate the following output:
  =500 \\$aThis is a note
  =500 \\$aThis is a note 2
- However….

# Common Joining techniques

When would I mark a field as repeatable?

- If I need to have multiple, like subfields on the same field, for example, like a subject field – we would mark the field as repeatable:
  column 1: Geology
  column 2: Oregon
  column 3: Corvallis
  If these fields were not marked as repeatable, the output would look like:
  =650  \0$aGeology$zOregon
  =650  \0$zCorvallis

  However, if these fields were marked as repeatable, the output would look like:
  =650  \0$aGeology$zOregon$zCorvallis

# Little Known Functionality

## Ability to process remote records from within the MARC Tools area

- Can be either MARC or XML data
- Can be zipped (if zipped, the file will be preprocessed)

# Little Known Functionality

Character conversion isn't limited to known – pre-populated items. You can define your own character-sets for process.

- MarcEdit can utilize any known Windows Characterset when doing characterset conversion.
  - Defined MS charactersets: http://msdn.microsoft.com/en-us/library/dd317756(VS.85).aspx

# MarcEdit and bad records

Two MARC breaking algorithms

- Strict MARC algorithm
- Loose breaking algorithm

Loose algorithm can heal MARC records (sometimes)

- Structural errors
- Missing field or record markers

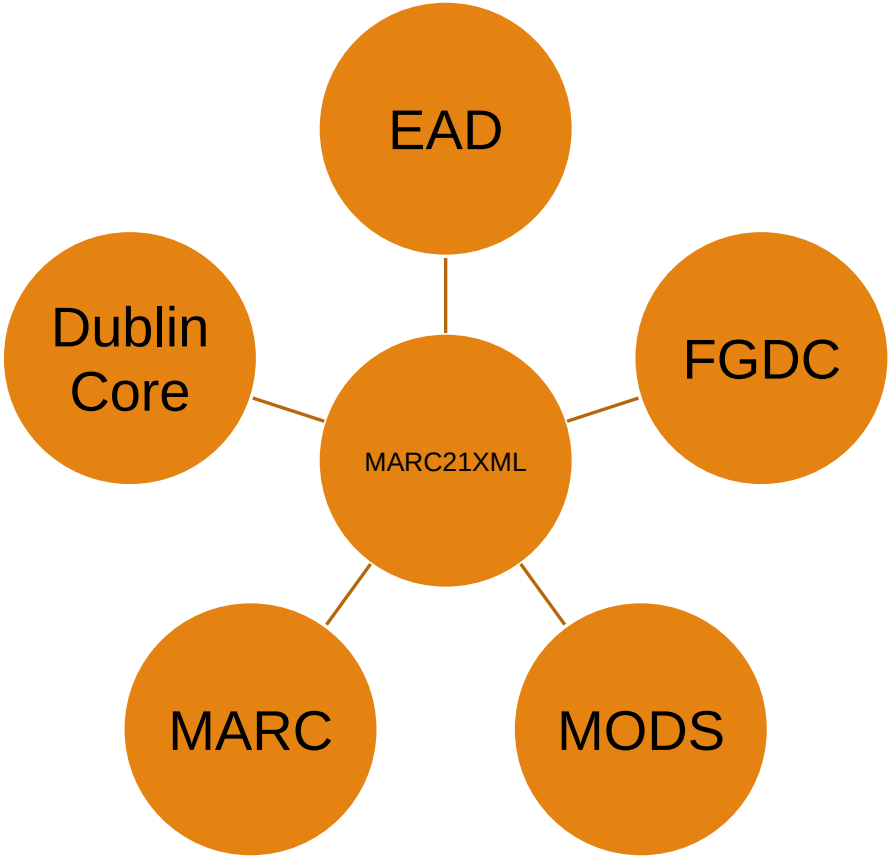
# MARC Conversions

# MarcEdit: crosswalking design

MarcEdit model:

- So long as a schema has been mapped to MARCXML, any metadata combination could be utilized. This means that no more than two tranformations will ever take place. Example: MODS → MARCXML → EAD

# MarcEdit Crosswalking model

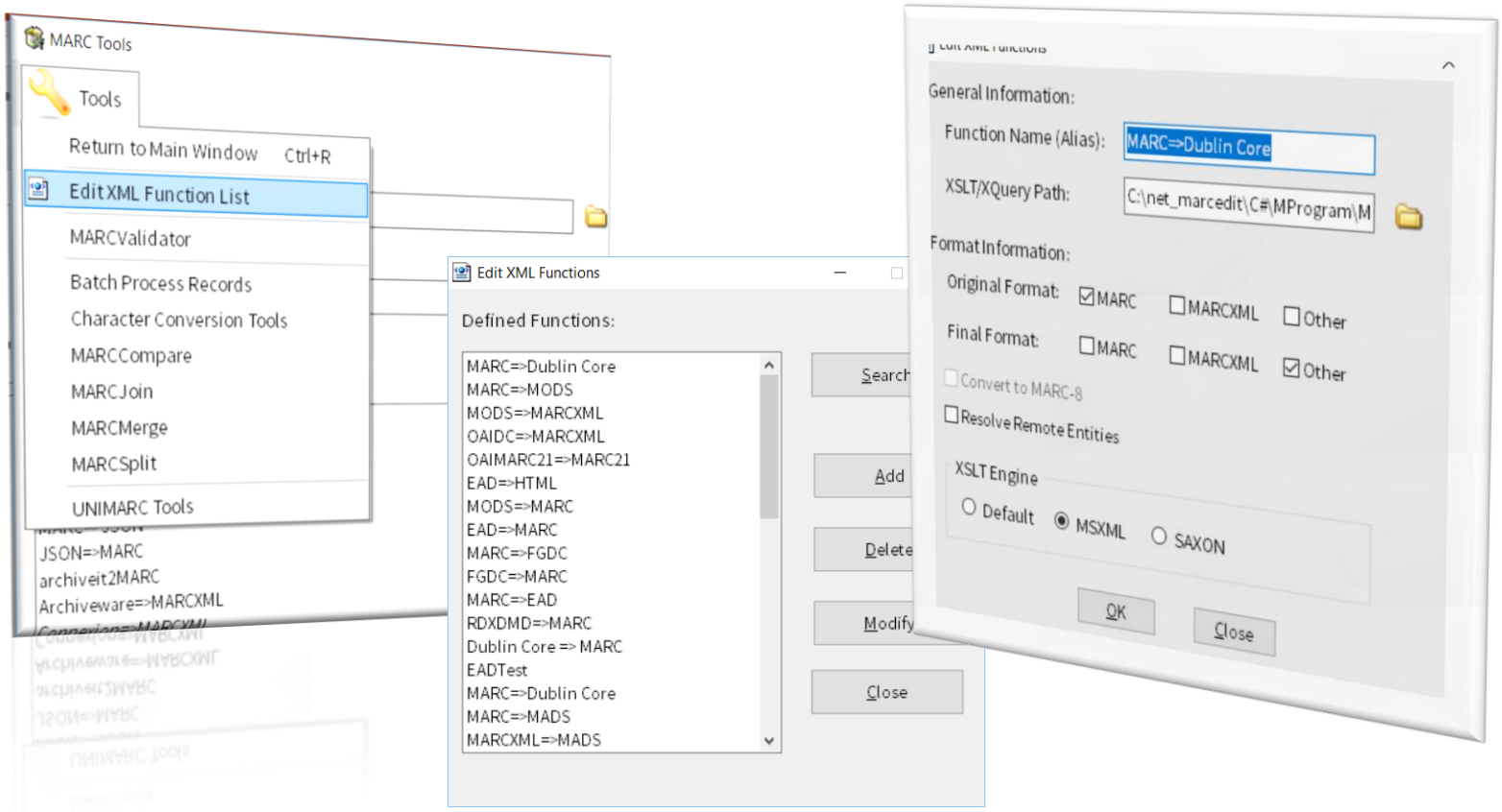


# MarcEdit: Crosswalks for everyone

What's MarcEdit doing?

- Facilitates the crosswalk by:
  1. Performing character translations (MARC8-UTF8)
  2. Facilitates interaction between binary and XML formats.

# Setting up Crosswalks




# Questions