MarcEdit: A simplified metadata processing tool

Terry Reese
Gray Family Chair for Innovative Library Services
Oregon State University
Email: terry.reese@oregonstate.edu
Before we start

I’m going to talk about MarcEdit but…

- Open Source development options
  - Python libraries
  - Perl Libraries
  - Ruby Libraries
  - PHP Libraries
  - Etc.
Roadmap

- What is MarcEdit?
- What can MarcEdit do?
  - MARC Tools
  - Editing MARC records
  - Lite-weight management/validation functionality
  - Supported conversion functions
    - Conversion to MARC
    - Conversion to XML-based markups
- Building your own solutions
- Miscellaneous functions
  - MarcEdit Script Maker
What is MarcEdit?

- Started development in 1999
  - Originally coded in 3 programming languages: Assembler (libraries), Visual Basic (UI) and Delphi (COM).
  - Initially designed as a replacement for LC’s DOS-based MARCBreakr/MARCMake software
What is MarcEdit?

Today:
- Written in C#
- Continues to be freely available
- Supports both UTF/MARC8 character sets
- MARC Neutral
- XML aware
Important notes

- Installation notes
  - As a C# application, it requires the installation of the .NET 2+ framework and MDAC 2.8 components.
  - If Using a previous version (prior to January 2009, you should *uninstall* then reinstall MarcEdit

- System Requirements
  - Any version of Windows that supports .NET
  - Fully supported on Linux
  - Partially supported on MAC (using MONO)

- Upgrade/Support
  - Upgrade cycle is approximately 4-6 months, with bug fixes released as they are reported.
  - I answer every question I get about MarcEdit.
  - Will be starting a listserv for users to ask and answer their own questions.
Getting Help

- MarcEdit Help File
- MarcEdit Tutorials
  - Online & YouTube
- MarcEdit ListServ

- Contacting the author
  (terry.reese@oregonstate.edu)
Edit MARC records in MarcEdit

- Two things to know about editing MARC records in MarcEdit
  1. MarcEdit is MARC agnostic
     - Does not enforce MARC21 conventions
     - Does not enforce character set homogeneity
  2. MarcEdit’s MarcEditor translates MARC records into a mnemonic format for editing – so you need to remember to convert editing mnemonic records back to MARC before loading.
Editing Records – Getting Started

• Two Workflows

1. *Most Common*: Break your record in the MarcBreaker; Edit the records in the MarcEditor; Compile records back into MARC using the MarcMaker

2. *Fewest Steps*: Preview your MARC record in the MarcEditor (does automatic MARC=>Mnemonic conversion); Edit records; Compile to MARC from within the MarcEditor
MARC Tools
Special Notes about MARC Tools

- MARC Tools represents the part of the application for converting files from one type to another.
- Access to the MARC functions
- Access to the XML Functions
- Access to Character conversion functions
About Character Conversions

- Today, ILS systems are fragmented regarding the type of character set that they will support
  - Two primary character sets:
    - MARC8 (ANSEL) – legacy
    - UTF8
  - Most vendors send records in one format or the other, meaning that character conversions are sometimes necessary.
MARC Engine Settings

- **Of Note:**
  - Use Diacritics turns mnemonics on and off
  - MARCXML XSLT determines how data moves between MarcEdit’s mnemonic format and MARCXML
  - XSLT Engine
    - Saxon.net supports XSLT 2.0
    - MSXML supports XSLT 1.0, but is orders of magnitude faster
  - Unicode Normalization
    - New feature designed to allow international users to break away from MARC21’s preferred KD normalization
Character set conversion in MarcEdit

- Two types:
  - Direct character set conversion on the MARC Tools window (when dealing only with UTF8 and MARC8)
  - Character conversion tool for translating data from any known character set to either UTF8 or MARC8

- *Important* -- when dealing with character sets, MarcEdit can correct the bytes, but you need to have a font that can render the data (applies mostly to Linux users)
MARC Character Conversions

- Supports moving between any known system character set and MARC8.
- Can be run from the Breaker/Maker – or as its own standalone utility.
MarcEdit’s MARCEngine

- MARCEngine is the heart of the application
  - Two important facts:
    - MarcEdit’s MARCEngine can correct a number of structural errors within MARC records. IE., if the leader is in-correct, the record directory is wrong, etc. MarcEdit can likely fix it.
    - Because of this, MarcEdit uses two MARC breaking algorithms. There is MARC-strict and MARC-loose. MarcEdit always utilizes MARC-strict, but when a processing error occurs, it falls back to MARC-loose before generating a parsing error.
Invalid Records

When MarcEdit’s MARC-loose processing algorithm is used, the results bar returns data in *red*
Isolating Invalid Records: MarcValidator

- MARCValidator
  - Originally developed for use at Oregon State to manage vendor records
  - Validator has two settings:
    - Field validation: Users can create a profile to test for the presence of field/field data.
    - Structure validation: Allows users to clean files with structurally invalid MARC records.
XML 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 transformations will ever take place. Example: MODS $\rightarrow$ MARCXML $\rightarrow$ EAD
MarcEdit Crosswalking model
MarcEdit: Crosswalks for everyone
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.
Batch Record Processor

- Allows MarcEdit to process “lots” of files.
- Can utilize any built-in or derived XML Function transformation.
MARCJoin/MARCSplit

- **MARCJoin**
  - “Join” lots of MARC files back into one large file.

- **MARCSplit**
  - “Split” MARC Records into a bunch of smaller bits
Little Known Functionality

- MARC Tools can process remote data
  - In the Input area – if you enter a full URL, MarcEdit will go get it and process the data.

- MarcEdit’s MARC Tools supports multiple XML engines, settings.

- Character conversion isn’t limited to known – pre-populated items. You can define your own character-sets for process.
Editing Records in the MarcEditor

- MarcEditor
  - Specialized Textpad designed specifically for MARC records.
  - Is UTF8 aware – can be used to generate records in MARC8 (though mnemonics) or UTF8 charactersets.
Editing MARC

• MarcEditor
  ◦ Supports a number of global editing functions:
    • Find/Replace functionality
    • Globally Add/Delete MARC fields
    • Globally Edit Subfield data
      • Conditionally add/remove field data
    • Globally Edit Indicator data
    • Globally Swap field data
    • Record Deduplication
    • Record Sorting
    • Macros
    • Z39.50 Cataloging
Editing MARC – Find/Replace

- Works like a normal Find/Replace in most Textpad utilities.
- Unlike most Textpads, Replace supports UTF-8 (when working with UTF-8 files) and regular expressions.
Find all function was designed for use with the Paging mode

Allows users to find any text across all pages

Generates a jump list that can be used to find individual records for edit
Jump List

- Find All
Jump List

- Jump List Example
Jump List

- When using the jump list:
  - Will jump to the page and record within the set
  - Will save (temporarily) any items modified or pages automatically (though to set saved items, you need to actually save the page)
Editing MARC – Global Add/Delete Field

- Globally add fields to all MARC records
  - Allows users to set insertion position.
- Globally delete fields
  - Allows global delete
  - Allows conditional delete
- Supports Regular Expressions
Editing MARC – Modifying subfield data

- Allows for the modification of variable MARC field subfield data (MARC fields >10)
- Allows for the modification of control field data by position or range of positions
- Allows users to prepend and append data to subfields.
- Allows users to change subfield tagging.
Editing MARC – Modifying subfield data

- Allows users to insert new subfields and define subfield placement.
- Allows users to move field data from one field to another.
- Supports:
  - UTF-8 with UTF-8 files
  - Regular Expressions
  - Adding new subfields.
Editing MARC – Modifying subfield data
Editing MARC – Swapping Fields

- Swap parts of MARC Fields or entire MARC fields
  - Define field, indicator and subfields to move.
  - Can move field data and delete the original field or clone the field data and move the clone to the new location.
  - Can add data to an existing field.
Fixing Boo-boos

- MarcEdit’s Special Undo
  - Allows you to step back one global change.
MarcEdit provides multiple sorting types:

- **Control Number**
  - Sorts record position within the file
- **Title**
  - Sorts record position within the file
- **Author**
  - Sorts record position within the file
- **Call Number**
  - Sorts record position within the file
- **0xx Fields**
  - Sorts the 0xx fields within individual records (does not change record position within a file)
- **All Fields**
  - Sorts all fields within individual records (does not change record position within a file)
- **Custom Sort**
  - Sorts all defined fields within individual records (does not change record position within a file)
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)
Field Counts

- Field Count
  - Provides a quick count of fields
  - Report of subfields used within a particular field
  - Detailed reports of all fields/subfields used within a fileset.
In-Line Validation

- MarcValidator-lite
  - Can access MarcValidator for quick validation of data elements found in the file set
  - Validation can use any defined rules set.
Delimited text translator

- Delimited Text Translator
  - Translates Tab, comma, pipe, Excel (Office 2000-2007), Access (Office 2000-2007) files into MARC
  - Can save translation maps
  - Can create constant data
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  

Building your own solutions

Why?
- Allows you to extend MarcEdit’s editing functionality
  - i.e., conditionally add field data based on data in other fields.
  - Dynamically building field data from other field data.
  - Creating automated data processing solutions for repurposing.
Building your own solutions

- MarcEdit functionality is exposed via the Windows COM architecture.
- MarcEdit exposes:
  - All MARCEngine functions
  - All TCP/IP (Exporting) functions
  - Z39.50 functions
  - [http://wiki.library.oregonstate.edu/confluence/display/ME/For+Programmers](http://wiki.library.oregonstate.edu/confluence/display/ME/For+Programmers)
MarcEdit Script Wizard

- What is the script wizard?
  - A general tool that can be used to automate simple record edits

- What is it best for?
  - Being used as a template generator for more complicated scripts.
    - Includes templates for working with MARC Engine
    - Sorting datafields
    - Fixing III specific export data
    - Able to generate both VBScript and PERL examples.
MarcEdit Script Wizard
Getting Help

- Youtube videos (just search for marcedit)
- You can ask me: 
  terry.reese@oregonstate.edu
- Questions