7-13-2018

A beginners guide to MarcEdit and beyond the editor: Advanced tools and techniques for working with metadata

Terry Reese
Ohio State University

Wendy C Robertson
University of Iowa

DOI: https://doi.org/10.1080/0361526X.2018.1439247

The Serials Librarian, 2018, Vol.74, Nos.1–4, 3–8, https://doi.org/10.1080/0361526X.2018.1439247

THE SERIALS LIBRARIAN
2018, VOL. 74, NOS. 1–4, 3–8
https://doi.org/10.1080/0361526X.2018.1439247

WORKSHOPS

A Beginners Guide to MarcEdit and Beyond the Editor: Advanced Tools and Techniques for Working with Metadata

Terry Reese[a] and Wendy Robertson[b]

[a]Presenter; [b]Recorder

ABSTRACT

Terry Reese taught two pre-conference workshops on MarcEdit at the NASIG 32nd Annual Conference: a full day session titled "A Beginner's Guide to MarcEdit" and a half day session titled "Beyond the Editor: Advanced Tools and Techniques for Working with Metadata." Attendees were given slides and sample files so that they could follow along or work through the exercises later at their own speed. The first day covered everything from installation and settings, to how to work with vendor records, and direct integration with library systems. The second day covered using OpenRefine and the features related to linked data.

KEYWORDS

MarcEdit; data manipulation; MARC; cataloging tools

MarcEdit is a freely available comprehensive library metadata editing application, developed and maintained by Terry Reese. Reese initially conceived of the software while a student at the University of Oregon, originally creating a set of programming libraries in order to learn how the machine-readable cataloging (MARC) format worked. Later, as the Map Cataloger at Oregon State University, Reese released the first public version of MarcEdit at the prompting of his long-time friend and mentor Kyle Banerjee. Since the software's release, Reese has continued to develop and support it, even as he's stepped away from actively working in technical services. Reese noted that today, he personally uses MarcEdit as an engine for his research related to linked data integration and use within libraries.

Reese is very responsive to requests and problem reports and even made tweaks the evening between the two sessions. The MarcEdit website has extensive information about the software, including videos tutorials.[1] There is also a MarcEdit discussion list where people can ask for assistance.[2] While the software is not open source, Reese is extremely open about his development process and provides a statement about why the software is not open source on his website.[3] Here he notes the various reasons why he currently keeps the software closed, including his concern that if the software were open sourced and forked, he would be unable to provide as much assistance as he currently provides to the community.

Reese describes MarcEdit as "MARC agnostic," meaning that it will work with any flavor of MARC, not only MARC21. It can work with UTF-8, MARC8, or any encoding standard supported by a user's operating system. Note that UTF-8 will match at the character level, not the byte level, so it works much better with CJK and Arabic characters. Reese recognizes the global use of the software and wants to ensure people the world over can continue to use it. Currently, one third of MarcEdit users do not use MARC21. Reese stated, "MarcEdit has been designed to work within the very heterogeneous metadata environment that we find ourselves in today." In order to work effectively with a wide range of library metadata standards, he considers integrations with OpenRefine, OCLC Online Computer Library Center, Inc. (OCLC), and other library cataloging systems to be important.

CONTACT Wendy Robertson ✉ wendy-robertson@uiowa.edu 📇 University of Iowa Libraries, Iowa City, Iowa, USA.

MarcEdit can work with millions of records; the real limitations are the size of the hard drive. Reese uses an external drive when working with large data sets, such as all the HathiTrust records or the May 2017 Library of Congress release of twenty-five million records. Many features in the software are customizable, including the font and size, making the software more accessible. Reese noted that accessibility will be one of the areas that will see the most improvements in the forthcoming MarcEdit 7, scheduled for release in the fall of 28 November, 2017.

During the past year, changes to MarcEdit include: linked data enhancements, BibFrame2 support, expanded command-line options, and integrated help. Of particular note is the addition of the Knowledge Bases And Related Tools (KBART) plugin, which was requested last year at NASIG. The software currently works with any version of Windows that supports .NET4.0. However, this requirement will change in the near future. MarcEdit 7 will have additional capabilities for linked data, such as native Resource Description Framework (RDF) and graph support, but these features require that the .NET framework moves to version 4.6. With this future release, Windows XP will no longer be supported. MarcEdit 7 will also have better tools to map eXtensible Markup Language (XML) to JavaScript Object Notation (JSON). Another major feature coming in version seven is that administrator security level will not be required to install the software. Reese is working closely with institutional system administrators to ensure that the installation will meet local requirements.

The individual user can make a variety of modifications to MarcEdit's settings, found by opening the Tools menu and selecting Preferences. These settings can be exported to ease the transition to a new computer by going to File, Share Configuration Settings, and then Export Settings. Windows 10 does not come with Arial Unicode MS, the MarcEdit default font. Reese uses the Google Noto Sans font and provides a knowledge-base article on its use.[4] He noted that the Noto Sans fonts are a significant improvement to the Arial Unicode MS font set, covering many more languages and characters.

A common workflow to revise a group of records using MarcEdit is to select a file of parsed MARC records (the extension may be .mrc, .dat, .bin, or something else) in MARC Tools, choose the location to save the .mrk file (the extension used for the MarcEdit mnemonic format), edit the records, save the file, compile the records back to MARC, and load the records into your cataloging system. Selecting files may be done by drag and drop, double clicking on a file with an extension associated with MarcEdit, or by navigating through the Open File menu option.

Another standard procedure is to convert MARC records into or out of MARCXML. The software can also change the encoding of records, for example out of MARC8. Additional common tasks include appending individual or groups of records using the MARCJoin utility, dividing a large set into smaller groups using the MARCSplit utility, and using the batch process. The batch process can now run tasks, allowing many procedures to be easily automated. Previously this feature could only be used through the command line, which is faster, but less user friendly. These commands are all available on the Tools menus, and if they are used frequently may be set as one of the eight default programs on the opening screen.

The MarcEditor is a notepad application designed to work with MARC records. There are editable templates to create new records in all formats. In the MarcEditor section of the Preferences, there is a choice to use *Resource Description and Access* (RDA) templates, to set the number of records to display per page, and other options. MarcEdit also includes a Preview Mode, which opens a snippet of the file into the Editor. Reese noted that the Preview Mode works best for large file sets as it removes the need for the application to load large amounts of data into memory. Reese recommended that users consider trying the preview mode when working with files of 150 MB or larger, and demonstrated this process using a 350 MB file that loaded in less than a second.

An entire set of records or a subset of records from a file may be edited. Editing can be as simple as a standard Find and Replace, to adding and changing fields based on conditions in other fields. Data in existing fields can be moved to a new field with the Swap function, or the data can be copied to a new field and reused with the Build function. In the Replace function, the program will retain the last ten replacements in the drop-down list. The Find All option allows movement directly to a record when there are multiple pages.

Much of the training was built around specific tasks catalogers need to do and questions that Reese has received. For example, a specific use for exporting tab delimited records is to create a list of titles and Uniform Resource Locators (URLs) for subject selectors from a file of vendor records. Using the "normalize field data" option will remove the subfield codes so the data can be easier to read (e.g., \0$aRevegetation$zNew Mexico$vMaps. versus Revegetation–New Mexico–Maps.). Another example was using OCLC's Classify service to add call numbers to a vendor-supplied file of e-journal and e-book records. If there are multiple options for a call number, MarcEdit gives preference to the one from the record with the most holdings. Globally adding a field is useful for identifying the source of batch loaded records in a local field.

People commonly select a subset of records to edit; it is only possible to select a subset based on one criteria. However, a second subset could be extracted from an existing subset to create a sub-subset, adding an additional criteria to identify the needed records. It is easy to identify extra records from a set by using the record deduplication function, using any field selected. The duplicates can be deleted (retaining the first, last, or most recently edited record) or output to a new file for further investigation.

Another commonly used feature is merging additional content into existing records. Some examples of scenarios that might necessitate this approach include a set of very brief records that need many changes, records that need a specific field added, such as contents notes, or duplicate records. If there are no identifiers, MARC21 may be used for matching, with a customizable list of fields used as match points. Reese indicated that the default 70% confidence rating usually results in good matches. The existing data may be overlaid or only the unique content may be added.

Converting spreadsheet data to MARC records is another use for MarcEdit. In this case, there are a few common technical issues. If the computer has a 64-bit version of Microsoft Office, this condition must be specified in preferences, in the Other section. Other problems tend to be related to Excel converting International Standard Book Numbers (ISBNs) to scientific notation or only providing the first 255 characters of a long field, such as an abstract. The MarcEdit site has guidance on these and other common issues.[5] When joining fields, in order to have multiple subfields, select the option to make a field repeatable. For example, if the data has column 1: Geology, column 2: Oregon, column 3: Corvallis, marking the field repeatable will yield = 650 \0$aGeology$zOregon $zCorvallis.

MarcEdit includes reports to analyze groups of MARC records. The field count report displays a count of all the fields in the file, and when downloaded includes subfield counts. It also provides a count of all the records in the file by looking at occurrences of a mandatory, non-repeatable field like an 001. The material type report identifies the number of records of each format and allows the user to jump to specific records. Other reports validate the record in a variety of ways, including the MARC itself, the ISBNs, the International Standard Serial Numbers (ISSNs), or the headings. The MARC validator includes a count of likely duplicates. Validate Headings includes an option to check to see if the services are online before checking the headings.

MarcEdit's functionality grows considerably through the use of regular expressions. Currently they can be used as part of several functions: Delete, Edit, Copy, Swap, Build New Field, Validation, and Extract/Delete selected records. Since the software is written using .NET, it makes use of the .NET Regular Expression Library. Reese noted that the .NET Regular Expression Language is very Perl-like, and suggested referring to Microsoft's documentation.[6] The standard command \W will find any non-word character in any language, not only in Latin scripts. When people ask questions on the MarcEdit list, Reese and other people frequently provide an answer using a regular expression. Examples of changes that are possible with regular expressions include: adding a period to the end of a 500 if it is missing, changing a 300 from 32 pages to 1 online resource (32 pages), and splitting an 856 with multiple u subfields into two separate 856 fields. It is important to note that by default, MarcEdit handles one field at a time, but if you need to check against multiple files you should add "/m" to the end of the command (e.g., to change videodisc to blu-ray in the 300 when a 538 says blu-ray). The edit menu has a "special undo" function for your last global update, including those using regular expressions. Several of the commonly used regular expressions are available as "edit shortcuts" including changing uppercase characters to sentence-case and removing smart characters. By turning on logging in Preferences, one may view a report to see all changes, which is useful if a user has made many global changes.

Reese also shared lesser known functionality, including the ability to process a file of remote records from within MARCTools. The file may be in MARC or XML and may even be zipped. He suggested that if there is a problem loading vendor records into a local system, using the loose algorithm to break the records into MarcEdit may resolve structural errors, missing fields, or record marker problems. When viewing records in the MarcEditor, select "Validate MARC records" from the "Tools" menu to identify the problems. Another tip is to use the compare records feature when training a cataloger to see what changes the individual made.

Task automation allows people to create defined task lists that can be run automatically with no difference from running the task manually. Anything that can be done in MarcEditor may be automated as a task. Tasks are particularly useful when processing a set of vendor records since most of the checks and edits are done on every set of records that arrives. Task libraries may also be shared and they can be used within another task list. If tasks are saved to a network folder, MarcEdit will make a copy locally so that the tasks can still be used even when not connected to the network. If the preference to enable logs has been selected, the changed records can easily be extracted into a new file in advanced log management so that only the changed records are loaded into the library system. The command line tool allows tasks to be scheduled.

The RDA helper includes options to quickly add 3XX tags, remove the General Material Designation (GMD), expand abbreviations, and make other modifications as desired to change a record from *Anglo-American Cataloguing Rules*, Second Edition (AACR2) to RDA. The many options give you great control over which fields should have their abbreviations expanded. Fields can be specified for inclusion in the process. Reese has not included RDA fields in crosswalks because the RDA helper can be used.

MarcEdit will transform XML from one system to another using either XSLT or XQuery documents. Within MarcEdit, Reese noted that for translations that he creates himself, he utilizes an indirect translation process, mapping metadata schemas through a control schema. In the case of MarcEdit, MarcEdit makes use of MARCXML as its control schema, as this format results in minimal data loss. Reese utilizes a control schema to enable reusable crosswalks. By mapping data through an intermediary schema, MarcEdit can automatically offer access to all the other format translations. Of course, this model does not preclude direct one-to-one format crosswalking. By using crosswalks, Reese allows more flexibility in the software. These XML functions can be modified or added to as needed from the "Tools" menu of MARC Tools to extend what MarcEdit can do and meet local needs. The "Search" option provides access to additional, existing crosswalks, such as ONline Information eXchange (ONIX) to MARCXML. Crosswalks may be edited and additional crosswalks may be created. The simplest navigation to the existing crosswalks is with the "Application Shortcuts, XSLT Data Path" option on the "Help" menu. Reese suggests keeping one's own XSL crosswalks in a separate folder to ensure they do not get lost during an update.

MarcEdit can harvest Open Archives Initiative (OAI)–formatted XML. In the server field, include everything before the "?" in a standard OAI command. For the set name, include everything after the "&set=" portion of the OAI string. The metadata option Dublin Core will use the metadata prefix "oai_dc." By typing a different prefix into the metadata box, additional metadata formats may be pulled into MarcEdit. OAI-harvested Dublin Core metadata can be transformed into MARC and then transformed into another format, extending the use of MarcEdit to metadata librarians.

Plugins may be activated for several other functions, including converting MARC to a KBART file, as mentioned previously, converting RIS formatted records to MARC, and harvesting data from the Internet Archive and packaging for HathiTrust. Plugins are not part of the standard functionality because relatively few people use these options. Each plugin needs to be enabled in the Add-ins menu. It is then used either from the home screen or from within MarcEditor, depending on its specific functions. Plugins can be written in any .NET language. The existing plugins are available on Reese's GitHub account.[7]

One plugin that Reese demonstrated was the OCLC Connexion plugin. This plugin enables MarcEdit to read a local OCLC Connexion local save file, edit the records in MarcEdit, and save the records back into the local save file. One of the challenges that Reese had to overcome when creating this plugin was the Connexion application itself. MarcEdit has been created to run as a 64-bit application on 64-bit systems and as a 32-bit application on 32-bit systems. This is a problem, because Connexion always runs as a 32-bit application, even when run on 64-bit versions of Windows. Because a 64-bit Application cannot "talk" to a 32-bit process, MarcEdit includes a special option that enables MarcEdit to run in a virtual 32-bit environment. Reese noted that this option should only be used when MarcEdit has to interact with 32-bit processes on 64-bit versions of Windows.

MarcEdit can directly integrate with library systems if they have an application programming interface (API) that allows searching, creating, and updating of bibliographic records. At this time, it supports direct integration with Koha and Alma, and supports search and discovery via Search/Retrieve via URL, Z39.50, or local API. In addition to library management system integration, MarcEdit also provides a special set of integration tools when working with OCLC. Using the metadata API, MarcEdit can work directly with WorldCat data. This may be especially useful for Mac users who otherwise are unable to edit WorldCat records. The most common use Reese's library makes of the OCLC integration is to remove holdings as part of their withdrawal process. More information about the API can be found on the OCLC site and on Reese's blog.[8]

An important subset of functions in MarcEdit can be found in the MARCNext category. MARCNext represents tools and features that reflect Reese's current research interests and advanced integration work. Presently, this toolset includes tools related to OpenRefine integration, a BibFrame testbed, linked data tools, as well as a SPARQL Protocol and RDF Query Language and JSON viewer.

The OpenRefine tool was created to help fill a gap for catalogers wanting to work with the software. At first glance, OpenRefine looks similar to spreadsheet software like Excel, but it is so much more than that. OpenRefine is a powerful tool for working with messy, unstructured data. Within the catalog community, OpenRefine has become popular when doing reconciliation work, or large editing projects. However, moving MARC data into OpenRefine and back into MARC has been a barrier. The MarcEdit OpenRefine tool simplifies this process, providing a tool that can read and write MARC data in formats that OpenRefine can understand. OpenRefine versions have different import processes, so it is important to note the version Reese is using (currently 2.7). After a project is created in OpenRefine, a JSON or tab delimited file exported from MarcEdit may be imported. The JSON file is limited to 5,000 records, so a tab delimited file should be used for very large sets. When working with JSON, users must select a node for a whole record, create a project, and then move the indicator column immediately after the field name. When working with a text file, they must change the import settings to not parse the first column as headings or cell data into numbers and dates, and do not use quotation marks to enclose cells. After edits have been made in OpenRefine, users can export the results as tab delimited and import to MarcEdit using the OpenRefine transfer. Optionally, a .mrc file may be imported to OpenRefine as long as there is no need to get the data back into MARC format.

Reese's interests have moved to linked data, and MarcEdit now has many tools built in to support this. He is active in the linked data community and works closely with a Program for Cooperative Cataloging's working group looking at embedding subfield 0's for the authority record control number into MARC data. The Link Identifiers function will add uniform resource identifiers (URIs) to records, including into 880 fields. Reese works hard to abide by standards and best practices so as to not burden servers. MarcEdit is one of the few tools the Library of Congress will allow to run in batch against their live data. It is possible to process extremely large quantities of records; Reese was able to add URIs to one million records in three days. MarcEdit can also add links to local controlled vocabularies. Reese believes such local linked data services may be more important in the future, especially for special collections with extensive, specialized vocabularies. The Getty Vocabularies are a particularly good model of how to create a local vocabulary (preferably in JSON). If an identifier exists in a subfield 0 but is not structured as a URI, the rules files can define a pattern so that the existing data is replaced with a link [e.g., German National Library headings can have $0(DE-588)4134080-2 converted to http://d-nb.info/gnd/4134080-2]. The linked data profile configuration may be accessed through the "Edit Linked Data Rules" option on the Addins menu. There is even a convenient link to the "How do I profile new endpoints?" section of the MarcEdit knowledgebase. As changes are made to MARC to incorporate linked date, Reese modifies the rules. A copy of the XML rules is available through GitHub so that everyone in the MARC community may use it, even outside of MarcEdit. Options may also be set on the Linked Data tool to limit resolution to only a specific vocabulary. Reese has included a button to check the status of the services before running the function.

The Bibliographic Framework Transition Initiative (BIBFRAME) testbed will convert records to either BIBFRAME 1.0 or 2.0. It is especially valuable for catalogers who are not yet involved in the conversations around BIBFRAME, as it enables them to see their data transformed and potentially feel confident about taking part in this work. Reese encourages more catalogers to become involved to ensure that their perspectives are incorporated in the end result. Several institutions, including Stanford University, have been working on a Metadata Object Description Schema (MODS) to BIBFRAME 2.0 crosswalk. After they have completed their work, Reese will compare their results with taking MODS record to MARC and then to BIBFRAME using MarcEdit. He hopes MarcEdit will work just as effectively.

A brief article such as this can be no more than a cursory overview of a day and a half of training. However, the software is widely used by a supportive community. Reese is also tremendously generous with his time, both in answering questions and in developing new features and solving problems. He has made many tutorials on his blog and YouTube channel to teach people how to make the best use of the software. In the last year, Reese has made many changes to MarcEdit, and with version seven due this fall, the next year will see even more new features. The software is powerful and it is well worth taking the time to learn its many functions.

Notes

1. MarcEdit Development, http://marcedit.reeset.net/ (accessed July 25, 2017). Video tutorials are available on Reese's YouTube channel, https://www.youtube.com/channel/UC7OLudoObYgiN_EmyDtZ_DQ (accessed July 25, 2017).
2. MarcEdit ListServ, https://listserv.gmu.edu/cgi-bin/wa?A0=marcedit-l (accessed July 25, 2017).
3. "Is MarcEdit Open Source?," http://marcedit.reeset.net/is-marcedit-open-source (accessed July 25, 2017).
4. "Replacement Unicode Fonts," http://marcedit.reeset.net/replacement-unicode-fonts (accessed July 25, 2017).
5. "Correct ISBNs Converted to Scientific Notation in Excel," http://marcedit.reeset.net/correct-isbns-converted-to-scientific-notation-in-excel and "Troubleshooting the Delimited Text Translator," http://marcedit.reeset.net/troubleshooting-the-delimited-text-translator (accessed July 25, 2017).
6. "Regular Expression Language—Quick Reference," https://msdn.microsoft.com/en-us/library/az24scfc(v=vs.110).aspx (accessed July 25, 2017).
7. https://github.com/reeset (accessed July 25, 2017).
8. "WorldCat Metadata API," http://oclc.org/developer/services/worldcat-metadata-api (accessed July 25, 2017); Terry Reese, "MarcEdit and the OCLC Metadata API: Introduction," Terry's Worklog. November 5, 2013, http://blog.reeset.net/archives/1245 (accessed July 25, 2017).

Notes on contributors

Terry Reese is Head, Digital Initiatives, The Ohio State University, Columbus, Ohio.

Wendy Robertson is Institutional Repository Librarian at the University of Iowa, Iowa City, Iowa