Librarian as Data Migrator: A Functional Pathway from Millennium to Koha

Christopher R. Todd

Cataloging/Systems Librarian, Barco Law Library, University of Pittsburgh School of Law, Pittsburgh, Pennsylvania, USA

Abstract

Purpose - This case study demonstrates that an in-house ILS migration can be accomplished by a dedicated team of librarians without advanced tools or prior experience with data migration or systems integration.

Design/methodology/approach – This migration was accomplished by academic librarians using freely-available tools: OpenOffice Calc, MarcEdit, and the Koha Integrated Library System.

Findings – The data migration pathway presented here was developed and successfully used to transfer over 48,000 records in less than two months.

Practical implications – This case study presents an original process that is particularly effective for smaller libraries.

Originality – While similar case studies exist, most employ expensive third-party contractors for data migration or rely heavily on institutional IT departments.

Keywords Integrated Library System, Koha, Millennium, data migration, library cataloging, systems integration

Paper type Case Study

Introduction

While Integrated Library Systems (ILS) are a crucial element of modern library services, many institutions are hampered by expensive proprietary systems with limited access to training or customization. A great deal of professional literature explores the benefits of migrating to an open-source ILS, however many libraries outsource the more technically challenging migration processes. This case study records the experiences of faculty librarians at a small college library who completed a Millennium to Koha migration without reliance on IT personnel or third party assistance.

Our migration team performed an extensive literature review in preparation for the transition to Koha. The majority of case studies published on this topic fell into two categories: ILS migrations that relied heavily on an IT department and those that were performed by a third party contractor. Migrations at the New York University Health Science Library and the Arcadia University library depended on substantial involvement of a robust IT department (Walls 2010, Kohn and McCloy 2010). Whereas the Paine College Library and the New York Academy of Medicine utilized outside companies for data migration services (Dennison 2011, Genoese and Keith 2011).

Restricted by a timeline of less than two months, our team at the Northern Marianas College Library resolved to complete the migration from Millennium to Koha without any outside support. While this decision was driven by necessity, a secondary goal of the in-house migration was to develop local capacity to service and troubleshoot the system. Throughout this process our faculty librarians were involved with every step of the data migration, developing advanced databasing skills and increased familiarity with the new system.

Background

The Northern Marianas College Library was first automated in the late 1990s, using grant funding to implement SirsiDynix Horizon as part of a state-wide consortia agreement. After exiting the consortium in 2010, the library migrated to Innovative Enterprises Millennium ILS. In both cases, the data migration was handled by third party contractors with little input from library personnel.

By 2014, the Millennium system was becoming increasingly problematic. Left in an “out of the box” state for years, many of the features of this system were never fully brought online. These problems were compounded by the departure of key employees with extensive training in Millennium, greatly reducing the institutional knowledge of the system. With no funding available to train new staff, the library began exploring open-source Integrated Library Systems as a possible alternative.

Ongoing ILS support costs were also a major factor in the decision to migrate. Rising service fees and dwindling library funding resulted in roughly 50% of the library’s annual budget being allocated to ILS maintenance fees. With no low-cost training options available, migrating to a stable open-source platform with ample online training materials became a top priority for the department.

Library staff aggressively evaluated open-source Integrated Library Systems for ease of implementation, robust community support, and patron-friendly features. Of the products reviewed, our team concluded that Koha best met the particular needs of our institution. The benefits of the Koha ILS and Koha Community are well-documented throughout the professional literature. Koha has been continuously updated since its release in 2000 and has been implemented in thousands of libraries across the globe (Koha Community, 2017).

Compared to the majority of both open-source and proprietary Online Patron Access Catalogs, the Koha ILS included a polished user interface which we hoped would be more appealing to early college students. Additionally, the OPAC was easily customizable by library staff with minimal HTML/CSS knowledge. The wide availability of training materials and a competitive market for third-party support also factored heavily in our decision. Finally, since the Koha ILS was freely available, the library was quickly able to implement the new system without initiating a lengthy Request for Proposal process.

Project Planning

With a short deadline and no supplementary funding for the project, our only option was to complete the ILS installation and data migration using only library employees. Our in-house migration team consisted of two faculty librarians with no systems experience and only limited ILS training. The college's IT department identified a staff member to handle the server side installation and updating of the Koha modules. Despite the limited personnel and overall lack of experience, our team developed a functional process that allowed us to complete the migration of over 48,000 MARC records in less than two months.

As we triaged tasks and created an aggressive timeline to completion, it was clear that our primary concern would be finding a simple method for exporting 20 years of catalog information from the Millennium ILS. At the time this project was undertaken in late 2014, few detailed accounts of Millennium to Koha data migrations were available. Our major source of inspiration was a 2011 article documenting the first recorded Millennium to Koha migration at the New York University Health Sciences Library (Walls, 2011). While NYU completed the same task within a similarly tight timeline, our data handling processes were significantly different.

While the NYUHSL IT department coded complex "helper scripts" to automate record retrieval and transfer, our team developed a simple yet functional alternative method of extracting catalog information from Millennium (Walls, 2010). Our process did not export complete MARC records; rather, it exported the crucial components of each bibliographic record using the native reporting features of Millennium. Once compiled outside of the Millennium environment, this metadata could be reassembled into Koha-compliant MARC records through a multistage process.

After deciding on a minimal set of MARC tags to export, we created a reporting template to isolate this data. To reduce server lag and create manageable files, this process was completed in five batches based on location codes (Archives, Main Library, Education Library, Special Collections, and Branch Campus). While the NYU migration team briefly explored this option, it was ultimately rejected, noting that "exporting as a list was not feasible due to the complexity of completely capturing all the MARC fields in a comma-delimited format" (Walls, 2011). While it is true that this technique does not capture complete MARC records, it can be used to export a set of descriptive elements that can be reassembled into MARC format. By carefully removing the MARC formatting, we were able to extract and use the data contained within.

Working with a much smaller collection than the NYUHSL, we were able to utilize this technique with great success. The Millennium reporting module was clearly not designed to function as a catalog-level export tool, since it only allowed us to extract field information without the accompanying MARC formatting. A more appropriate tool would have been the

Millennium Global Exchange Module (Crago, 2015). However, this feature was not brought online during installation in 2010.

Our team developed a process using the Create List tool in Millennium to deconstruct MARC records into exportable plaintext to be reassembled into functional records downstream in the migration. This process consisted on four distinct stages: Migration Preparation, Data Export, Data Mapping, and Upload (Figure 1). Our two-person team learned the rudiments of database schema and established a "crosswalk" pathway for each exported field. Through this process we converted our Millennium export information into a common database format (Comma Separated Value, or CSV) and then back into an OPAC viewable MARC file for ingest into the Koha system.

Migration Preparation

To ensure that our export process transferred all data fields uniformly, we began the migration preparation by surveying our catalog records for internal consistency. This included the standardization of location codes and elimination of cataloging inconsistencies that had accumulated through successive ILS platforms and catalogers. The collection was also heavily weeded to reduce the amount of records that would need to be removed after the migration.

During this phase we uncovered several major cataloging issues that presented data migration challenges. While developing the export process, it quickly became clear that many of the bibliographic records in our system did not adhere to accepted cataloging standards. Numerous records were integrated from regional library systems or cataloged locally without adherence to AACR2 or RDA guidelines. Additionally, previous data migrations had introduced systematic errors in holdings records, particularly though the assignment of system-generated "placeholder" numbers for items without barcodes.

As we corrected these inconsistencies, our IT department created a partition on the current ILS server, allowing us to run a test installation of Koha alongside our active Millennium system. This enabled us to test our data mapping process and to verify that information had migrated successfully by crosschecking both systems.

Data export

Prior to export, our team decided on a minimum set of critical MARC tags to retain from each record. Based on the Library of Congress National Level Minimal Requirements, we created an export template including: Author, Title, Edition Statement, ISBN, Call Number, Barcode, LCCN, Description, Publication information, Series Title, General Note, Additional Statement, Location, and multiple Subject Headings (Library of Congress, 2010). Since our process exported this information as plaintext without the context of a full MARC-encoded record, it was crucial to record which MARC tag corresponded with each sequentially-numbered export field (Figure 2). This information was necessary for compiling the exported data back into a MARC file later in the migration process.

Since this report would generate a .txt file with basic CSV functionality, several steps were needed to ensure the data would move seamlessly through several downstream systems without introducing errors. While Millennium uses a comma as the standard delimiter character, we changed the program defaults to instead use a percent character (%) as a unique delimiter. This enabled us to create a CSV file that would not confuse plaintext commas, such as those in the title and author fields, with the delimiting character (Figure 3).

This process was repeated for each of the five location codes used by the library. At the conclusion of this phase, all of our MARC records had been deconstructed into columns of information stored in a CSV file. Several steps were needed to translate this plaintext information into Koha-ready MARC records.

Data Mapping

This deconstructed MARC data was checked for errors and lightly formatted during the Data Mapping phase of the migration. The delimited text was reformatted into a tabular data file that could be easily shared with downstream applications. We used the Calc spreadsheet application in the Apache OpenOffice suite to translate the comma-separated text into tab-delimited text. Calc allows the user to make several important customizations to the raw CSV export file during ingest. The upload dialog box includes an option to identify a non-standard delimiter character (Figure 4). By changing the delimiter to “%”, our information was cleanly restructured as a tabular data file.

Once imported into Calc, the exported data displays in an easily-readable spreadsheet format. Each catalog item is represented by a single row of bibliographic information demarcated by columns that correspond to the original MARC tags from Millennium (Figure 5). In this form, the exported data can be checked for formatting errors and processed into Koha-compliant MARC records. We relied heavily on regular expressions (regex), advanced “wildcard” characters, to modify text and prepend information to specific fields (such as adding a sub-collection code before a range of call numbers). After all cleanup operations, the spreadsheet was converted back into CSV format.

Up to this point, the data extracted from Millennium existed as raw text in several relational database formats. The information itself was preserved, but the MARC formatting was abandoned during the extraction process. To reconstitute these records, we utilized MarcEdit, a powerful suite of freely-available metadata editing tools, to “map” the spreadsheet fields back into complete MARC records.

The “Delimited Text Translator” feature of MarcEdit greatly simplified this process. Using the CSV file from Calc, again setting “%” as the delimiter character, MarcEdit quickly rendered this information back into an easily-viewed tabular data file. Similar to Calc, MarcEdit displays a “Data Snapshot” window which allows the user to determine if the delimiting character is properly recognized (Figure 6).

The major benefit of using MarcEdit is its ability to map data into MARC formatting during the ingest process. The Delimited Text Translator quickly allows the user to associate each CSV column with the appropriate MARC tag used by the Koha ILS. Using the dropdown menu, each column can be mapped to a three digit MARC tag, including subfields and indicator values (Figure 7).

Since our data was exported as five separate files, we iteratively tested a single batch of data to develop a MarcEdit template that would uniformly translate each successive file. While the MARC tags used in Millennium did not match those in Koha one-for-one, we developed a simple crosswalk to translate our exported MARC elements into the necessary Koha format. Once this mapping information was finalized, we used the aforementioned MarcEdit dialog box to save our mapping arguments, creating a local file to be applied to later data sets.

The Delimited Text Translator produced an editable document in the MarcEdit mnemonic file format (extension ".mrk"). At this stage, final alterations to the data were performed in MarcEdit using its intuitive editing features, which also accept regular expressions. The .mrk file was easily converted to a Koha-ready MARC (.mrc) file for import. To prevent duplicate uploads and to refine our data mapping pathways, our team tested the viability of our reconstructed MARC files by uploading them first to a Koha sandbox demo provided by Bywater Solutions (Bywater, 2017). This extra step allowed us to isolate our process, catching and correcting errors before any data entered our clean installation of Koha.

Upload

Records uploaded to Koha are first held in a "Staging" phase, followed by a separate "Import" phase. This allows the system to reject any records with MARC formatting errors. We found this to be a straightforward process that is covered extensively in Chapter 9 of the *Koha 3.2 Manual* (Engard, 2011).

After our first dataset proved viable in Koha, we repeated the process for the four remaining files. Using this multistage process, we were able to export, translate, and upload the remainder of the collection in just a few days. We utilized a similar technique to migrate patron records between the two systems. The only major difference was the application of a concatenate formula in Calc to join patron first name and last name fields to create a simple username for each enrolled student.

The success of this project depended on a positive working relationship with the college's IT department. The only two responsibilities of our IT contact were partitioning the existing ILS server and installing Koha on a virtual machine. After this installation was complete, library staff handled all aspects of the data migration. The willingness of our librarians to do the heavy lifting of data migration greatly eased our interactions with IT and created an atmosphere of collaboration between the two departments.

If repeating this process today, OpenRefine could be an effective substitute for the Calc program. OpenRefine, formerly Google Refine and Freebase Gridworks, is a data cleanup and transformation application with many similarities to popular spreadsheet programs (Magdinier, 2016). Recognizing both JavaScript Object Notation (JSON) and regex, OpenRefine could easily replace Calc while also completing some of the operations performed in MarcEdit.

OpenRefine also allows users to save all data transformations as JSON commands in a text file that could be shared with other libraries performing similar in-house migrations.

Despite our success, the technique used in this case study was far from a straightforward transfer of MARC records from one ILS to another. Deconstructing, exporting, transforming, and reconstructing MARC records through numerous computer applications presented several risks. Each program that opens the data file has the potential to alter either the source information (such as spreadsheet programs routinely removing leading zeros) or formatting (such as changing the CSV delimiting character or altering spreadsheet tabs). Vigilance is crucial to catch these errors.

A major downside to the technique presented here was its inability to export item status. Instead of overtaxing our rudimentary process by finding yet another workaround, we readily relied on manual record correction for issues we could not solve with our limited databasing skills. To address this issue, we again utilized the Create List feature in Millennium. We simply ran a stock report for items with an outstanding status (lost, claims returned, overdue, withdrawn) and manually updated these records in Koha after the migration was complete.

While our actual data migration was fairly quick, an extensive quality control review was necessary. Once Koha was fully operational, staff and student workers were brought on board to manually update item statuses and merge duplicate records. Additionally, our team used basic sampling techniques to select and scan several hundred circulating items to ensure that all information had fully and correctly transferred to Koha.

As noted by other publications on the topic of ILS migrations, patron fines and circulation information rarely export cleanly to a new ILS. After noticing this trend in the pre-migration literature review, we decided to migrate patron fines manually only for lost books and issue a general amnesty on all fines for returned books. This significantly decreased the amount of clean-up work for the library staff and created a degree of excitement around the new ILS for students.

Despite these downsides to our process of daisy-chaining applications, it also yielded several benefits. Due to the rigid structure of this migration pathway, we were forced to identify and address unexplored or ignored legacy data issues. We discovered that a great deal of barcode information was incorrect or missing; we were able flag these records and fix this information during the post migration cleanup phase. Ultimately, the quality of our records was greatly improved by this hands-on migration process.

Conclusion

This case study presents a functional model for a fully in-house data migration from Millennium to Koha. While this project presents several challenges, it can be accomplished by dedicated librarians with no prior training and only minimal support from IT. By handling all aspects of data extraction, clean up, and formatting, our library gained an in-depth knowledge of our collection and the underlying principles of digital information management.

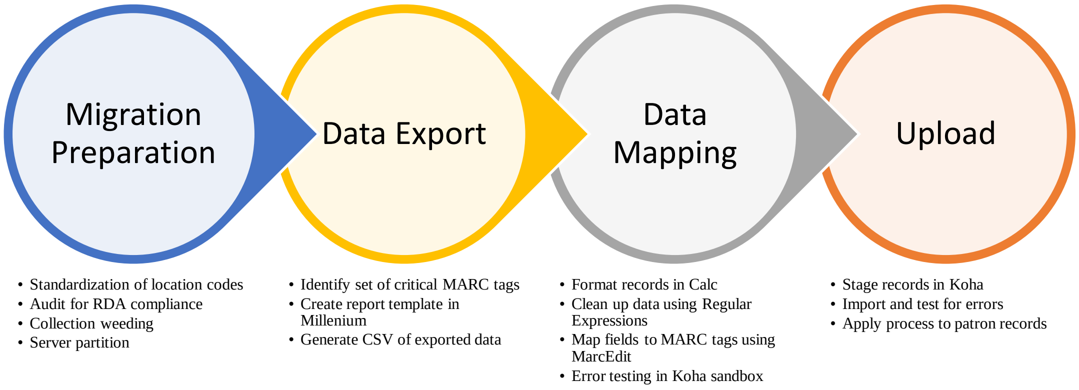
While this data migration was a no-cost operation, there are long-term costs associated with maintaining an open-source ILS. Our library was extremely fortunate that our extant ILS server could be repurposed to run Koha without any significant investment in additional hardware. Regardless, local hosting will always require funds for server upgrades, maintenance, incremental backups, and increased power and connectivity costs.

Time proved to be the most valuable resource in this process. The migration team needed several weeks to learn the basics of database operations, systems integration, data mapping, use of regular expressions, and to acquire enough specialized language to communicate effectively with our IT liaison. Even after the successful migration, several months of part-time cleanup work were needed to ensure all records had migrated accurately. While this project was nominally free to complete, the amount of staff hours devoted to the migration were high.

This case study demonstrates that small libraries with little training and minimal IT support can successfully complete an in-house ILS migration. There was a general fear that we could not accomplish this migration with such limited resources, but our small size and agility proved to be a major benefit. While some more elegant solutions were out of reach, this case study shows that a data migration can be accomplished using resources that are available to any librarian frustrated with their ILS.

Figures

Figure 1 – Project Phases.



- Standardization of location codes
- Audit for RDA compliance
- Collection weeding
- Server partition

- Identify set of critical MARC tags
- Create report template in Millenium
- Generate CSV of exported data

- Format records in Calc
- Clean up data using Regular Expressions
- Map fields to MARC tags using MarcEdit
- Error testing in Koha sandbox

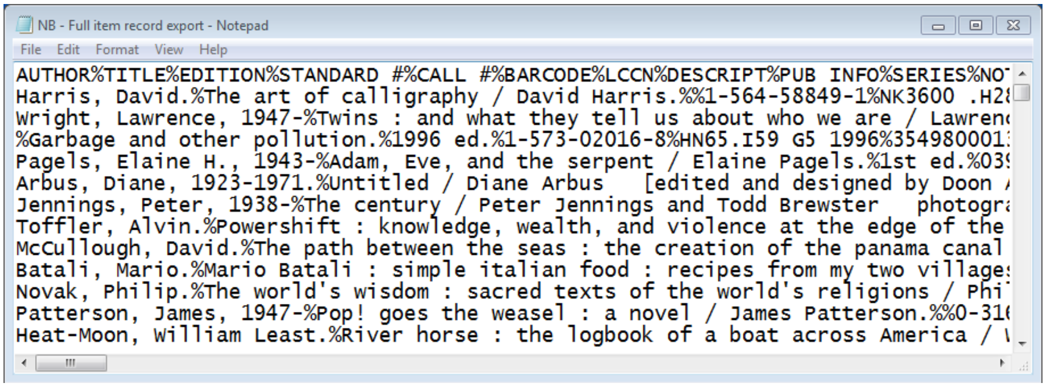
- Stage records in Koha
- Import and test for errors
- Apply process to patron records

Figure 2 - Overview of data mapping phase.

Millennium Plaintext	Mapping	Koha MARC Tag
LCCN	Library of Congress Control Number	010 ‡ a
ISBN	ISBN	020 ‡ a
Call Number (split on cutter period during Calc phase)	Call Number Remainder of Call Number	050 ‡ a 050 ‡ b
Author	Author	100 ‡ a
Title (split into three subfields using Calc)	Title Remainder of Title Statement of Responsibility	245 ‡ a 245 ‡ b 245 ‡ c
Edition	Edition Statement	250 ‡ a
Publisher	Publisher	260 ‡ a
Series	Series Statement	490 ‡ a
Note	General Note	500 ‡ a
MFHD (see Koha guidelines	Call Number	952 ‡ o

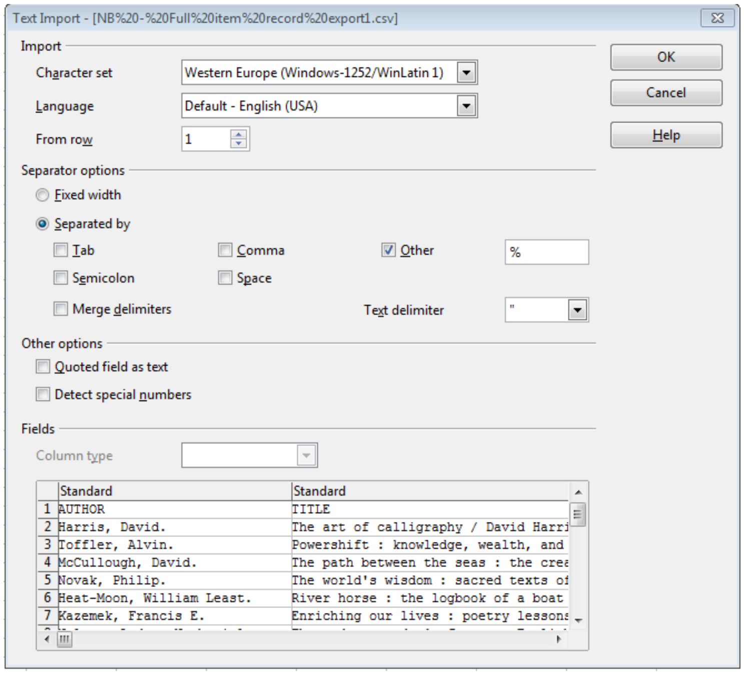
for 952 field)	Item Type	952 ‡ y
	Barcode	952 ‡ p
Subject Heading (repeated several times to accommodate records with large number of subject headings)	Subject Heading 1	650 ‡ a
	Subject Heading 2	650 ‡ a
	Subject Heading 3	650 ‡ a
	Subject Heading 4	650 ‡ a
	Subject Heading 5	650 ‡ a

Figure 3 – CSV export file. Note that empty fields will display as %% or two sequential empty fields as %%% in the text file, these will be reconstructed later as empty tabs.



```
AUTHOR%TITLE%EDITION%STANDARD #%CALL #%BARCODE%LCCN%DESCRIPT%PUB INFO%SERIES%NO
Harris, David.%The art of calligraphy / David Harris.%%1-564-58849-1%NK3600 .H2
Wright, Lawrence, 1947-%Twins : and what they tell us about who we are / Lawren
%Garbage and other pollution.%1996 ed.%1-573-02016-8%HN65.I59 G5 1996%354980001
Pagels, Elaine H., 1943-%Adam, Eve, and the serpent / Elaine Pagels.%1st ed.%03
Arbus, Diane, 1923-1971.%Untitled / Diane Arbus   [edited and designed by Doon
Jennings, Peter, 1938-%The century / Peter Jennings and Todd Brewster   photogr
Toffler, Alvin.%Powershift : knowledge, wealth, and violence at the edge of the
McCullough, David.%The path between the seas : the creation of the panama canal
Batali, Mario.%Mario Batali : simple italian food : recipes from my two village
Novak, Philip.%The world's wisdom : sacred texts of the world's religions / Phi
Patterson, James, 1947-%Pop! goes the weasel : a novel / James Patterson.%%0-31
Heat-Moon, William Least.%River horse : the logbook of a boat across America /
```

Figure 4 - The preview field in the bottom of the window will automatically adjust the defined delimiting character.



Text Import - [NB%20-%20Full%20item%20record%20export1.csv]

Import

Character set: Western Europe (Windows-1252/WinLatin 1)

Language: Default - English (USA)

From row: 1

Separator options

○ Fixed width

● Separated by

□ Tab □ Comma ☑ Other %

□ Semicolon □ Space

□ Merge delimiters Text delimiter "

Other options

□ Quoted field as text

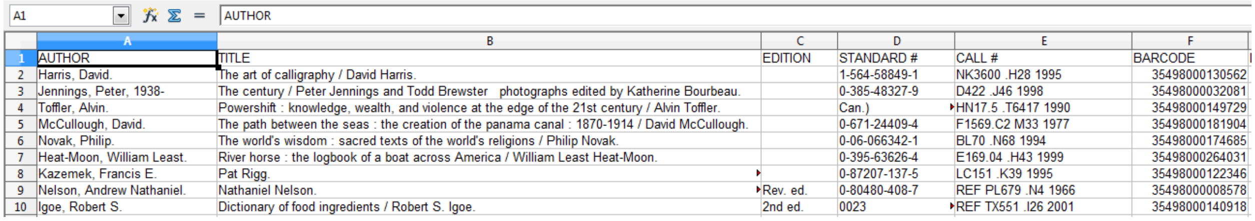
□ Detect special numbers

Fields

Column type

	Standard	Standard
1	AUTHOR	TITLE
2	Harris, David.	The art of calligraphy / David Harri
3	Toffler, Alvin.	Powershift : knowledge, wealth, and
4	McCullough, David.	The path between the seas : the crea
5	Novak, Philip.	The world's wisdom : sacred texts of
6	Heat-Moon, William Least.	River horse : the logbook of a boat
7	Kazemek, Francis E.	Enriching our lives : poetry lessons

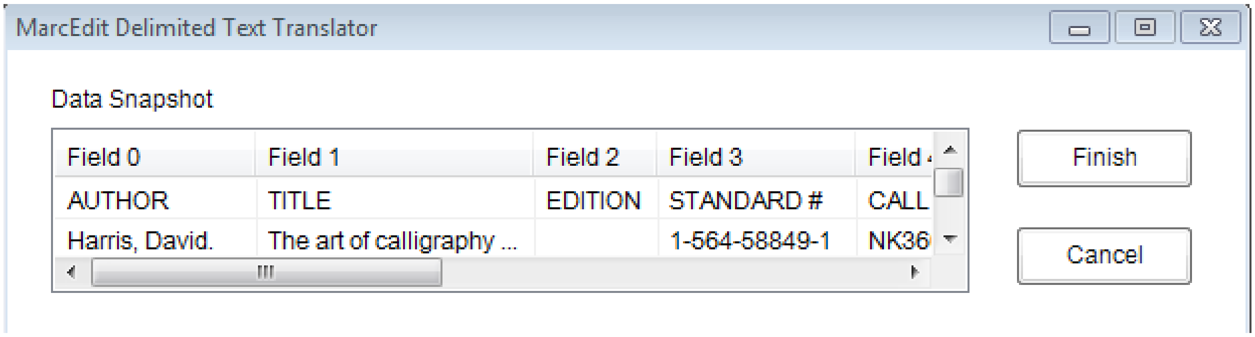
Figure 5 – Export data rendered as a spreadsheet in Calc.



A1 AUTHOR

	A	B	C	D	E	F
1	AUTHOR	TITLE	EDITION	STANDARD #	CALL #	BARCODE
2	Harris, David.	The art of calligraphy / David Harris.		1-564-58849-1	NK3600 .H28 1995	35498000130562
3	Jennings, Peter, 1938-	The century / Peter Jennings and Todd Brewster photographs edited by Katherine Bourbeau.		0-385-48327-9	D422 .J46 1998	35498000032081
4	Toffler, Alvin.	Powershift : knowledge, wealth, and violence at the edge of the 21st century / Alvin Toffler.		Can.)	HN17.5 .T6417 1990	35498000149729
5	McCullough, David.	The path between the seas : the creation of the panama canal : 1870-1914 / David McCullough.		0-671-24409-4	F1569.C2 M33 1977	35498000181904
6	Novak, Philip.	The world's wisdom : sacred texts of the world's religions / Philip Novak.		0-06-066342-1	BL70 .N68 1994	35498000174685
7	Heat-Moon, William Least.	River horse : the logbook of a boat across America / William Least Heat-Moon.		0-395-63626-4	E169.04 .H43 1999	35498000264031
8	Kazemek, Francis E.	Pat Rigg.		0-87207-137-5	LC151 .K39 1995	35498000122346
9	Nelson, Andrew Nathaniel.	Nathaniel Nelson.	Rev. ed.	0-80480-408-7	REF PL679 .N4 1966	35498000008578
10	Igoe, Robert S.	Dictionary of food ingredients / Robert S. Igoe.	2nd ed.	0023	REF TX551 .I26 2001	35498000140918

Figure 6 – MarcEdit ingest dialog box.

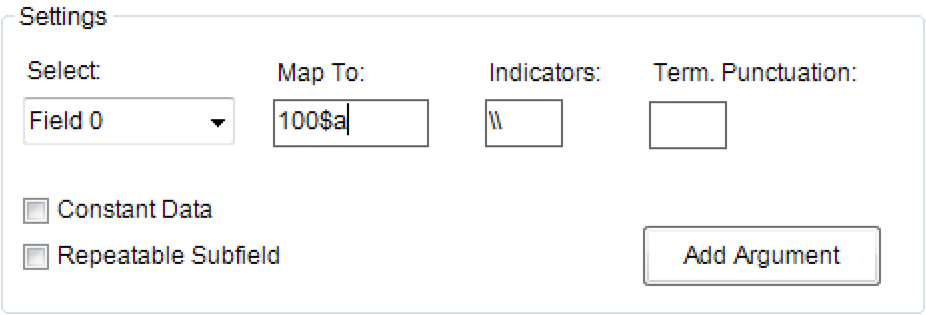


MarcEdit Delimited Text Translator

Data Snapshot

Field 0	Field 1	Field 2	Field 3	Field 4
AUTHOR	TITLE	EDITION	STANDARD #	CALL
Harris, David.	The art of calligraphy ...		1-564-58849-1	NK36

Figure 7 - Note that MarcEdit format utilizes the "$" as the subfield character in place of the "‡" symbol used by many ILS.



Settings

Select: Field 0

Map To: 100$a

Indicators: \\

Term. Punctuation:

□ Constant Data

□ Repeatable Subfield

References

Bywater Solutions (2017), "Koha Demo", available at: http://bywatersolutions.com/demos/ (accessed 12 August 2017).

Crago, R. (2015), "Millennium iii Bibs Export for Koha", available at: https://wiki.koha-community.org/wiki/Millennium_iii_Bibs_Export_for_Koha (accessed 8 September 2017).

Dennison, L. (2011), "Small and open-source: Decisions and implementation of an open-source integrated library system in a small private college", *Georgia Library Quarterly*, Vol. 48 No. 2, pp. 6-8.

Engard N.C. (2011), "Stage MARC Records for Import", in *Koha 3.2 Manual*, available at: http://manual.koha-community.org/3.2/en/stagemarc.html (accessed 9 July 2017).

Genoese, L. and Keith, L. (2011), "Jumping ship: One health science library's voyage from a proprietary ILS to open-source", *Journal of Electronic Resources in Medical Libraries*, Vol. 8 No. 2, pp.126-133.

Koha Library Software (2017), "Koha Users Worldwide", available at: https://wiki.koha-community.org/wiki/Koha_Users_Worldwide (accessed 24 August 2017).

Kohn, K. and McCloy, E. (2010), "Phased Migration to Koha: Our Library's Experience", *Journal of Web Librarianship* Vol. 4 No.4, pp. 427–434.

Library of Congress (2010), “MARC 21 Format for Bibliographic Data National Level Full and Minimal Requirements”, available at: https://www.loc.gov/marc/bibliographic/nlr/ (accessed 6 September 2017).

Magdinier, M. (2013), “OpenRefine History”, available at: http://openrefine.org/2013/10/12/openrefine-history.html (accessed 24 August 2017).

Walls, I. (2010), “Becoming Truly Innovative: Migrating from Millennium to Koha”, paper presented at Code4Lib, 24 February, Asheville, North Carolina, available at: https://code4lib.org/conference/2010/walls (accessed 24 August 2017).

Walls, I. (2011), “Migrating from Innovative Interfaces' Millennium to Koha: The NYU Health Sciences Libraries' experiences”, *OCLC Systems & Services: International digital library perspectives*, Vol. 27 No. 1, pp. 51-56.