Making Collection Management Decisions with Data

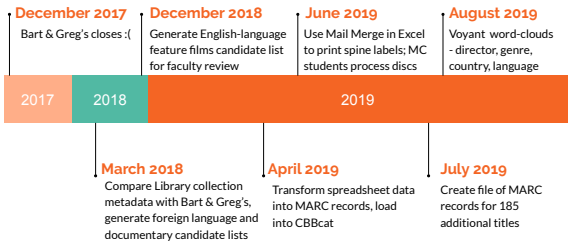
A Library Project Overview by Karl Fattig
presented at
Innovative Users Group 2020 - Minneapolis

Doing more with delimited (meta)data

When *Bart & Greg's DVD Explosion* store closed shop in December of 2017, the Library was asked to consider acquiring part or all of the store's collection of DVD and BluRay titles. In order to make a data-driven decision, the College Librarian tasked me with comparing our own Library video collection with that of *Bart & Greg's*.

Bowdoin

Bart & Greg's Explosion Project Milestones



Project Participants

- Marjorie (College Librarian)
- Joan (Collections Librarian)
- Bart (Bowdoin Annex)
- Karl (Sytems)
- Kat (Associate Librarian - TS)
- Erin (Associate Librarian - PS)
- Carmen (Media Librarian)
- Elizabeth (Media Supervisor)
- Media Commons students
- Mary (Catalog Librarian)
- Rachel (AV Cataloger)
- Film Studies faculty
- Annex Team

It Takes a Village...

The first project goal is to create lists of potential **ACQUISITION CANDIDATES** that would complement Bowdoin's own video collections

Tip

CBBcat records generally have one title field. Records in Bart & Greg's database have two (Explosion title and IMDB title). Upon examination The IMDB title more often matches CBBcat's title.

1. Compare

Select a tool to compare Library catalog metadata with Bart & Greg's

- **Library catalog metadata**
  Export delimited text file from MARC records in library services platform (Innovative Interfaces Sierra)
- **Bart & Greg's metadata**
  Copied from MS Access tables into MS Excel spreadsheet.
- **Problems**
  No common numeric identifier shared in both data sets; inconsistent title data

Which of the 26,000 *Bart & Greg's* titles are unique, not held in Bowdoin Library's 7,600-title video collections?

Tip

Since the Library and Explosion metadata share no common numeric identifiers (ISBN, ISSN, IMDB-ID) we'll have to use **natural language processing** to compare title data.

Tool - OpenRefine

- Already used by electronic resources librarians to create knowledge bases across the scholarly publications supply chain (GOKb - Global Open Knowledgebase)
- Open-source
- Was Google Refine, Freebase Gridworks
- Transformation expressions using GREL (General Refine Expression Language) sort of like Excel formulae

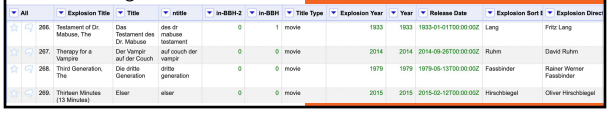
Tip

Not a web-hosted service, but locally installed, creates a server on your workstation, run via **web browser interface.**

We use **OpenRefine** to compare the title fields in Bowdoin's metadata set

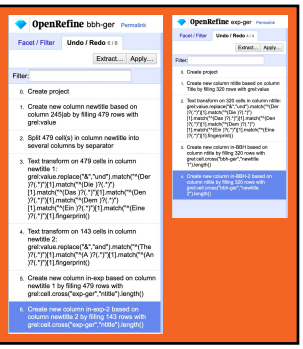
All	TITLE	245jab	newtitle 1	in-exp	newtitle 2	in-exp-2	newtitle 3	008 Date One	COUNTRY	LANG	REC
8.	Das testament des Dr. Mabuse = Testament of Dr. Mabuse / Nero-Film AG : a joint venture between Janus Films and Home Vision Cinema : producer, Fritz Lang writers, Fritz Lang, Thea von Harbou ; directed by Fritz Lang	Das testament des Dr. Mabuse =Testament of Dr. Mabuse /	des dr mabuse testament	1	dr mabuse of testament	0		2004	xxu	ger	b137168

with the IMDB title field in Bart & Greg's metadata set



All	Explosion Title	Title	ntitle	in-BBH-2	in-BBH	Title Type	Explosion Year	Year	Release Date	Explosion Sort T	Explosion Direct
266.	Testament of Dr. Mabuse, The	Das Testament des Dr. Mabuse	des dr mabuse testament	0	1	movie	1933	1933	1933-01-01T00:00:00Z	Lang	Fritz Lang
267.	Therapy for a Vampire	Der Vampir auf der Couch	auf couch der vampir	0	0	movie	2014	2014	2014-09-26T00:00:00Z	Ruhm	David Ruhm
268.	Third Generation, The	Die dritte Generation	dritte generation	0	0	movie	1979	1979	1979-05-13T00:00:00Z	Fassbinder	Rainer Werner Fassbinder
269.	Thirteen Minutes (13 Minutes)	Elser	elser	0	0	movie	2015	2015	2015-02-12T00:00:00Z	Hirschbiegel	Oliver Hirschbiegel

Before performing the comparison we use a stack of **GREL commands** to normalise the title data in each metadata set; this is called a transform, and it creates a new data element, a quasi linguistic fingerprint, which will serve as the comparison point.



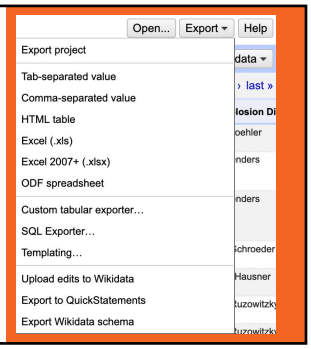
The title comparison command stack populates a new column with a **numeric value** representing the number of matching lines/rows

- Click 'OK'. This will populate a new column with a number representing the number of matching lines found in ListB
  - If the "ListB Comparison" column contains a zero (0) then no match has been found
  - If the "ListB Comparison" column contains a one (1) then a single match has been found
  - If the "ListB Comparison" column contains a two (2) then two matches have been found
  - etc.
- Facet on the new 'ListB Comparison' column to find those lines in ListA that do not appear in ListB (a zero in the column)
- To identify journals that are in ListB but not in ListA, the same process is carried out starting with the 'title.identifier.issn' column in the 'ListB' project

We can then perform a **text facet** on the data in the comparison column

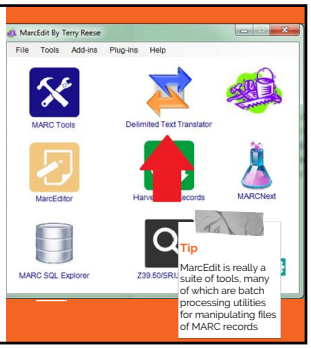
- Click 'OK'. This will populate a new column with a number representing the number of matching lines found in ListB
  - If the "ListB Comparison" column contains a zero (0) then no match has been found
  - If the "ListB Comparison" column contains a one (1) then a single match has been found
  - If the "ListB Comparison" column contains a two (2) then two matches have been found
  - etc.
- Facet on the new 'ListB Comparison' column to find those lines in ListA that do not appear in ListB (a zero in the column)
- To identify journals that are in ListB but not in ListA, the same process is carried out starting with the 'title.identifier.issn' column in the 'ListB' project

Once faceted to the value zero, we can **export** the relevant metadata for our non-matching titles, which are the acquisitions candidates for review by library collection managers



With the numeric data we can create a **chart** to visualize the collection management scenario, and present the data to administration

The second project goal is to transform the spreadsheet data into **MARC RECORDS** that can be batch-loaded into CBBcat



2. Create MARC

Once collection management decisions have been made, we end up with a spreadsheet with metadata for the titles we will be adding to the library. Now we can use this metadata to create *brief* MARC records for CBBcat - n.b. these are not full catalog records - **for that we'd need Rachel for at least a year**

- **Tab-delimited data**
  Tab (control character 9) is not present in the actual metadata content
- **MARC specifications**
  Stakeholders - video cataloger, catalog librarian, collections librarian, technical services manager, College librarian.

How long would it take one full-time cataloger to create records for the 6,000 titles selected for acquisition?

Tip

On average it takes a cataloger 2.5 hours to create a full-featured catalog record for a **DVD.**

375 days

Tip

Another option would be **outsourcing** the project to a library cataloging vendor. But this would be costly.

How much would it cost to outsource the project to a third-party vendor?

Tip

Depending on degree of difficulty, language, availability of "copy" cataloging, per title charges range from $17-$60.

$190,000 - with shelf ready option

LANGUAGE FAMILIES AND GROUPS	UNEDITED COPY	FULL EDITING COPY	ORIGINAL CATALOGING	AV/MEDIA SURCHARGE
English	$5	$15	$23	$5
African: Niger-Congo and various other African languages	9	30	50	10
Afro-Asiatic: Arabic, Hebrew, Amharic, Tigrinya	7	25	45	5
Altaic: Mongolian, Turkish, Ottoman Turkish, Turkic	7	25	45	5
Austronesian: Tagalog, Malaysian (Malay), Indonesian, Hawaiian	7	25	45	5
Baltic: Latvian, Lithuanian	7	25	45	5
Basque	7	22	34	5
Celtic: Welsh, Scots, Irish Gaelic, Breton	7	25	45	5
Chinese (Mandarin, Cantonese), Japanese, Korean	7	22	34	5
Dravidian: Tamil, Telugu, Malayalam	9	30	50	10
Finno-Ugric: Estonian, Finnish, Hungarian	7	25	45	5
Greek, Albanian, Armenian, Caucasian, Georgian	7	25	45	5
Indic: Bengali, Gujarati, Hindi, Marathi, Punjabi, Sanskrit, Urdu	7	25	45	5
Indo-Iranian: Nepalese, Oriya, Persian (Farsi)	7	25	45	5
Mayan languages	7	25	45	5
Romance: French, Italian, Portuguese, Romanian, Spanish (and Blindex headings), Catalan, Galician, Provençal, Latin	7	22	34	5
Scandinavian: Danish, Swedish, Norwegian, Icelandic, Faroese	7	22	34	5
Sino-Tibetan: Burmese, Tibetan	9	30	50	10
Slavic: Belarusian, Bosnian, Bulgarian, Church Slavic, Croatian, Czech, Macedonian, Polish, Russian, Serbian, Serbo-Croatian, Slovak, Slovenian, Ukrainian	7	22	34	5
Tai-Kadai: Thai	7	25	45	5
Vietnamese	7	25	45	5
Western Germanic: Dutch, German	7	22	34	5
Yiddish, Ladino	7	25	45	5

How can we transform the spreadsheet metadata into brief MARC records?

Tip

The actual transformation takes less than 10 seconds. Data preparation and clean-up required a couple of days of my attention.

–

Tool - MarcEdit

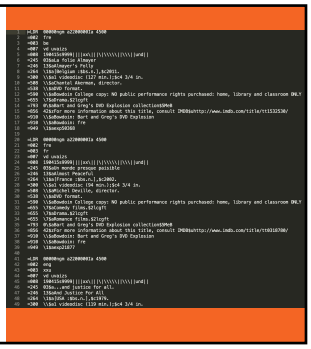
- Freeware
- Developed in 1999 by Terry Reese (Oregon State Univ.)
- Supports use of XSLT to map metadata from one schema to another
- Leverages grep for sophisticated/nuanced discovery
- Allows for human-readable MARC record editing

Tip

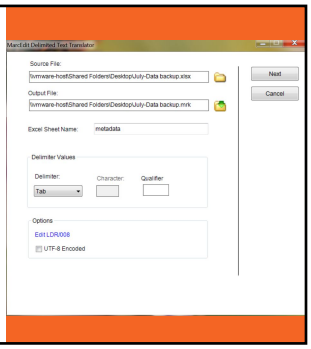
We already use **MarcEdit** to edit records, perform database maintenance, and enhance catalog records.

This is what a file of **MARC records** looks like to our library services platform (Sierra) - note the tags and indicators are at the beginning of the record, with the delimited field data following.

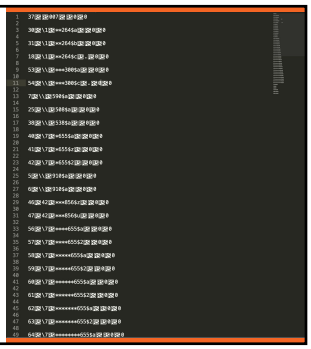
This is what a file of **MARC records** looks like to the human eye in MarcEdit's Marc Editor.



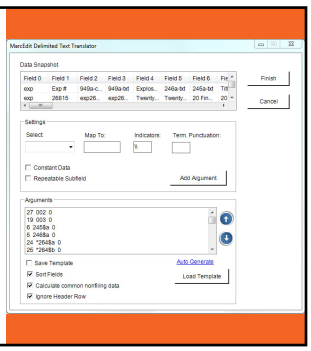
We will use MarcEdit's **delimited text translator** to reconfigure the spreadsheet data into a file of MARC records.



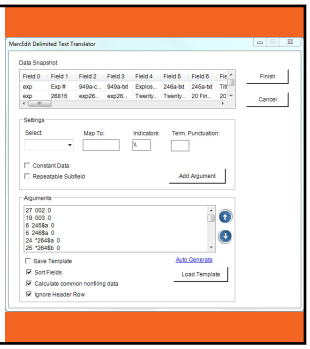
The delimited text translator employs a **template file** to map the appropriate columnar data into the correct MARC field/subfield.



MarcEdit's delimited text translator **shows a preview** of the spreadsheet data; subfields can be joined to create MARC fields

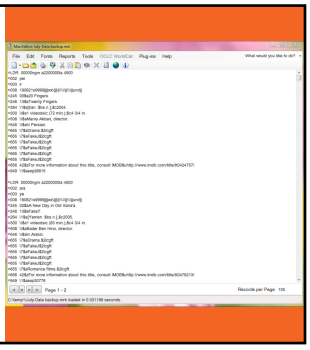


MarcEdit's delimited text translator claims to be able to **calculate non-filing indicators (skip characters),** this feature works for English-language articles only, and I was unable to get a customized non-filing indicator matrix to work.

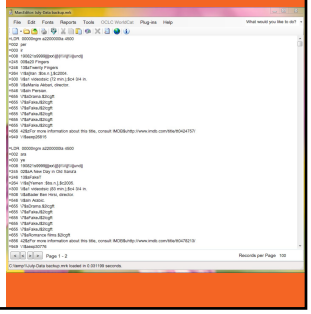


My solution was to sort the master list by title and **create individual sheets** based on the non-filing indicator values

MarcEdit's delimited text translator **creates an editable file** of MARC records; some data cleanup is required in the Marc Editor

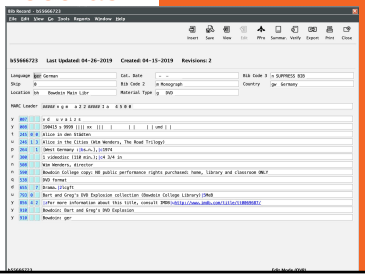


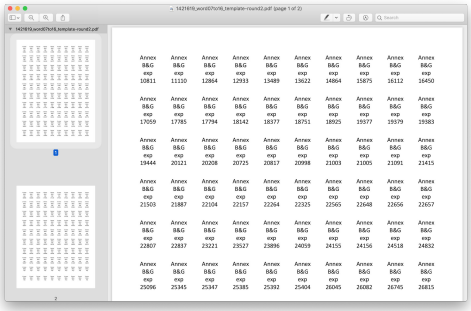
In April, I used spreadsheet data to **create all the MARC fields in the records;** in July (for a smaller set of additional titles we added), I used spreadsheet data only to create unique MARC fields, and used MarcEdit to add common fields (538, 590, 910, etc.) globally.

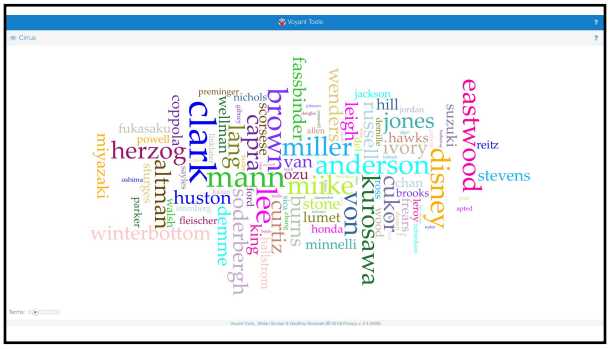


In order to load the records into CBBcat, I created a **custom load profile for Sierra Data Exchange -** the load profile instructs the system how to index and display MARC fields

It took about 7 minutes **to load the records** into CBBcat







Contact information

Karl Maria Fattig
kfattig@bowdoin.edu
(207) 725-3027
Bowdoin College Library
3000 College Sta
Brunswick, ME 04011-8421

presentation available at:
http://bit.ly/3aFEidh