

@jerrywaller

LITA 2016 - JERRY WALLER
SYSTEMS LIBRARIAN, ELON UNIVERSITY

DATA MANIPULATION FOR ILS MIGRATION

Begin.

Greetings, Earthlings!

I'm Jerry Waller

Systems Librarian Elon University, and I'd like to...

Thank you...

thank everyone for expressing interest in what I have to say.

it's humbling :)

ILS Migration

This year our libraries migrated from iii Millennium to WMS
started in January
Finished in June.

Goals

My goals are to talk with you little bit about

Tools

A little bit about
the tools I used

What they are

A little bit about
what they are

Why I used them

A little bit about
why I used them

Strategies

Ideally, take away a better idea of strategies that you can use regardless of software.

PROLOGUE

WHEN I WAS A PROGRAMMER...

Phase I Clinical Trials

i was a biostats programmer analyzed data gathered when a new drug is administered to humans for the first time.

This is called Phase 1

Phase 1 doesn't study efficacy

Adverse events

Instead, it collects data on dosage—how it's administered; where it's administered and specifically looks for adverse events.

Side effects

Adverse events are basically side effects, except they're more serious.

FIRST Collected and Transcribed

all the data to be analyzed was collected,
then transcribed and somehow encoded

SECOND

Entered

Then someone in data entry had the job of entering that data.

THIRD

Cleaned

someone else, a programmer, cleaned the data so that it could be analyzed.

Collected
Entered
Cleaned

After I left Clinical Trials I realized this approach could be applied to libraries.

Why do we clean
data?

“DATA COLLECTED BY STATISTICAL AGENCIES MAY CONTAIN MISTAKES MADE DURING THE ACQUISITION, TRANSCRIPTION AND CODING PROCESS.”

**Riera-Ledesma, J., & Salazar-González, J.-J. (2007).
A branch-and-cut algorithm for the continuous
error localization problem in data cleaning*.
Computers & Operations Research, 34(9), 2790.**

Take 7 seconds to read this.

We are neither consistently accurate nor consistently correct when it comes to entering data.

In short, we are imperfect.

The data conform to a standard

BUT, we need to do the best we can to ensure the data conform to a standard.
Cleaning the data is one of those steps.

SYSTEMS LIBRARIAN...

Know thy data!

In order to establish a standard, you must be familiar with your data.

Know what “good”
data look(s) like

This does not mean memorizing every element

Malformed vs. Well-formed

what it means is knowing the difference between well-formed and malformed data

your use-case
your requirements

contingent upon **your** specific use-case and requirements.

CHAPTER 1

SAS

You've been waiting for this!

DISCLAIMER

I WAS A SAS INTERN

I am not employed, endorsed, or compensated by them.
so now that we've cleared the air...

SAS

Why SAS?

Why on earth would a systems librarian use a massively complex statistical analysis software suite?

SAS

I know how to use it.

I've been using SAS for over ten years, starting with that biostats programming I told you about.

SAS

SAS is primarily a
procedural language.

I actually love sas. It is primarily a procedural language, which I find very appealing
because

SAS

Very granular

I think it's easy to have granular control over how you manipulate your data.

SAS

Specify the order of code execution.

I can be particular about where and when I execute data.

Hard-core programmers will likely not find my way the most efficient.

But it does allow me to break apart datasets in ways that might provide insight later.

In other words, it helps me 'know my data'.

SAS

SQL Syntax.

Yay!

I LOVE SQL syntax!

SAS

Patron records

It is because of these reasons that I think SAS is great for cleaning up...

SAS

Circulation data

SAS is great for cleaning up...

SAS

Item records

SAS is great for cleaning up...

SAS

Massive batch
conversions of
conditional data.

SAS is great for...

SAS

What does that
mean?

SAS

Millennium → WMS

I can convert large amounts of Millennium output to WMS input
really, really fast

SAS

For example:

MILLENNIUM OUTPUT

Name	Fawkes, Guy J.
DOB	04-13-1570

This is a mockup of a Millennium export.
patron name is all one field
dob is m d y

WMS INPUT

First	Guy
Middle	J.
Last	Fawkes
DOB	1570-04-13

This is what WMS needed PAUSE
How did I get from one to the other?

SAS

[illegible]

As you can clearly see...

SAS

`http://www.jerrywaller.org`

**SAS Programs for Academic and
Public Libraries**

No, I'm not going to do that to you. If you really want, you can look at my ugly code on github. I warn you, it is a hot mess

SAS

Name

Fawkes, Guy J.

sas can be tricky to explain Bear with me
since Millennium treats the patron name as one field, I needed to break it into elements.
And, for the sake of time, reformatting the DOB is pretty similar so I won't go over it here.

SAS

Fawkes, Guy J.

Let's break somethin'!

SAS


Fawkes, Guy J.

```
p=index(name,"");
```

```
result: p=7
```

the last name is the section of the name field that ends with the first occurrence of a comma.
the code creates a numeric variable based upon the position of the comma in the field.

SAS



Fawkes, Guy J.

```
n=length(name);
```

```
result: n=14
```

then the length of field—minus any trailing spaces—is calculated.

this example has a length of 14 because nested spaces count.

SAS

Fawkes, Guy J.

```
lname=substr(name,1,(p-1));
```

```
result: lname=Fawkes
```

the last name variable is created by taking the characters from
position 1 through p minus 1

SAS

Fawkes, Guy J.

```
fm_name=substr(name,(p+1),(n-p));
```

```
result: fm_name= Guy J.
```

because I like to iterate, I use that p variable to also
create a variable containing first and middle names
yes, there is a space at the beginning of that variable.

SAS

Guy J.

```
fname=scan(fm_name,1);
```

```
result: fname=Guy
```

the first name is determined by using the scan function
which identifies distinct words in a variable.
In this case it's the first distinct word.

SAS

Guy J.

```
mname=scan(fm_name,2);  
result: mname=J
```

the middle name or initial is the second distinct word

note: the scan function automatically ignores spaces and punctuation.

SAS

fname	Guy
mname	J
lname	Fawkes

and here's the result

However, not every name fits into these categories.

SAS

260 Outliers

In our migration we had
260 outliers out of 12,138 records

SAS

Pro tip:

As you iterate through your data, you get to Know Your Data.
Deal with the easier stuff first, then
you can modify your process for those outliers.

SAS

Con:

Learning curve

I've been working with SAS for over 10 years
and I had six weeks of intensive training when I started.
I'm always learning something new.

SAS

Con:

Not great for MARC

not great for marc...

SAS

Yet :)

INTERMISSION

BEGGING THE ???

open refine is really good for marc records.

DISCLAIMER

I am not a cataloger.

This is going on my tombstone.
(past tense?)

Why even mess with the marcs?

why do we even need to deal with marc records?

aren't they
universal?

aren't they a universal standard or something?

No.

The reasons why are beyond the scope of this presentation.

CHAPTER 2

OPENREFINE

but rest assured that open refine is really good for marc records. :)

CREDIT WHERE CREDIT IS DUE:

Terry Reese is smarter
than me.

first off: terry reese is smarter than me.
no one should accuse me of lacking humility :)

MARCEDIT

OpenRefine import/
export

Terry enabled openrefine import export functionality
Genius!

MARCEdit

Pro tip:

from experience, my suggestion is to [next]

MARCEDIT

Export as tab-delimited text

export mnemonic marc records from marcedit as tab-delimited text.

I could turn that sentence into a limerick.

MARCEdit

~~JSON~~

json export did not work as well for me.

MARCEdit

Windows

When it came down to it, I had better luck with marcedit on windows than on a mac.

Split large files

also, no way to get around it,
split .mrc files when migrating.

Split .mrc files into
manageable file sizes.

split those big 'ol marc containers into digestible chunks

Group .mrc files into like categories.

Group your marc container files.

One of the reasons for doing this is so that the data are more homogenous.

Know your data.

AN ASIDE:

1.237 GB .mrc file

this is all our records in a marc container
almost 1.25 billion characters in one continuous line of text
and we're a small university

OPENREFINE

The trick...

with openrefine, the trickiest trick may be...

OPENREFINE

is defining the
individual marc records

openrefine doesn't automatically know where a marc begins and ends.
we have to tell it.
Fortunately, openrefine is a good listener.

OPENREFINE

Steps

I'm going to walk through the steps needed to get openrefine to recognize marcs.
this may not make any sense at first,
but if I have time I'll give a quick two minute demo to reinforce it.

OPENREFINE

UTF-8

When importing, define your encoding as utf-8
Because diacritics!

OPENREFINE

bêçãüsé djǎčrìćš

Because diacritics!

OPENREFINE

コーヒー قهوة

Because non-Roman characters

OPENREFINE

~~Blank rows~~

First, when it comes to marc records and openrefine,
don't import blank rows .

OPENREFINE

No headers

We do not want to parse column headers.

OPENREFINE

No quotation marks

quotation marks are NOT used to enclose data

OPENREFINE

Make sure you are in the
"rows" view

This should be the default, but this is where you'll need to be in the beginning.

OPENREFINE

Take me to your LDR

(I'm so, so sorry)

Next, since the leader marks the beginning of an individual marc record...

The Leader marks the beginning of a record and consists of twenty-four characters, that may be numbers, letters, or blanks. These characters provide information about the record, such as its length.

OPENREFINE

Facet the first column for
"LDR"

we're going to facet the first column and isolate the rows that start with "LDR"

OPENREFINE

Star or Flag those rows

then star those rows.

OPENREFINE

Add a column based on
the LDR column

I'll add a new column based on the col containing the "LDR" value.

OPENREFINE

Oooh! What's its name?

I'm going to name the new column.
I like to call mine "Key".

OPENREFINE

rowIndex+1

use the formula aka grel syntax rowIndex+1 to create cell values

OPENREFINE

Close the facets

Once that's complete, close the facets dialog to show all the rows.

OPENREFINE

Move the new column to
the beginning

We'll make the new column "key" the first column in the dataset.

OPENREFINE

Click into “records” view

to verify that openrefine identifies individual marc records,
click into the records view.

Now that I’ve told you the steps, I’m going to show you the steps.

Refine

A power tool for working with messy data.

Create Project

Open Project

Import Project

Language Settings

Last modified	Item
today 12:42 PM	lta_test
today 11:40 AM	lta_test lev
a week ago	belkmatch
a week ago	cat_only.txt
a week ago	All_Checked_Out_Items_Report.txt
a week ago	lta_test
2 weeks ago	2111_customer_notify_audiobooks
2 weeks ago	ORL_modified_abc.clo.txt
a month ago	expmxy01_ip_user_count_201809
a month ago	expmxy02_user_count_201809
a month ago	expmxy01_user_count_201809
a month ago	NEO scan delete report 180318.txt
3 months ago	wms new titles 1
4 months ago	picking_out_archives lev
4 months ago	ArchivesA
5 months ago	belk_items_prefixes
5 months ago	everyitemrecord
5 months ago	wms_manual_checkins
5 months ago	postJune9checkouts.txt
5 months ago	patron_checkouts_from_June9th2018.txt
5 months ago	NEO_unknown_items
5 months ago	everybitrecord.txt
5 months ago	saxout_wms_loan_data.txt
5 months ago	neo_item_statistics_2018jun09
5 months ago	saxout_wms_item_stats.txt
5 months ago	2018 06-09 fines fees lev
5 months ago	clrc_load_data_file
5 months ago	lawEbooks
5 months ago	law_license_flagged
5 months ago	KB_databases

Version 2.9-v1.2 (7/16/18)

Help

About

Browse workspace directory

OPENREFINE

You've got MARC

TIPS 'N TRICKS

rowIndex+1

use rowIndex+1 to create cell values

OPENREFINE

Facets are awesome

OPENREFINE

Why can't I do this in
SAS?

why can't I do this in sas? well,

OPENREFINE

```
proc freq data=alpha;  
tables col1;  
run;
```

In SAS I'd have to write out a frequency statement to isolate the LDRs, but I wouldn't be able to interact with them like I can in OpenRefine.



THANK YOU

A migration is an opportunity to get to know your data better.
It is also an opportunity to clean it up and fix as many errors as you can.

Thank you.

thank you for expressing an interest in my talk this morning! :)

Questions?

SAS examples : <https://www.jerrywaller.org>

email: jwaller7@elon.edu

know your data :)

Any questions?



FINIS

A migration is an opportunity to get to know your data better.
It is also an opportunity to clean it up and fix as many errors as you can.