@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.



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

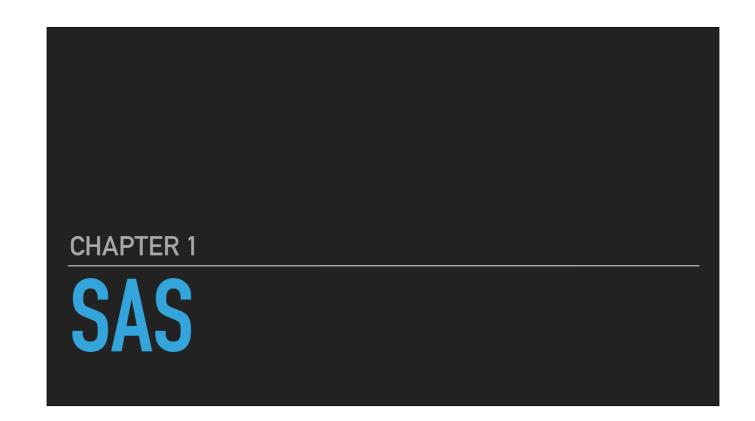
This does not mean memorizing every element

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.



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...

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

Very granular

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

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!

Patron records

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

Circulation data

SAS is great for cleaning up...

Item records

SAS is great for cleaning up...

Massive batch conversions of conditional data.

SAS is great for...

What does that mean?

SAS Millennium→WMS

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

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?

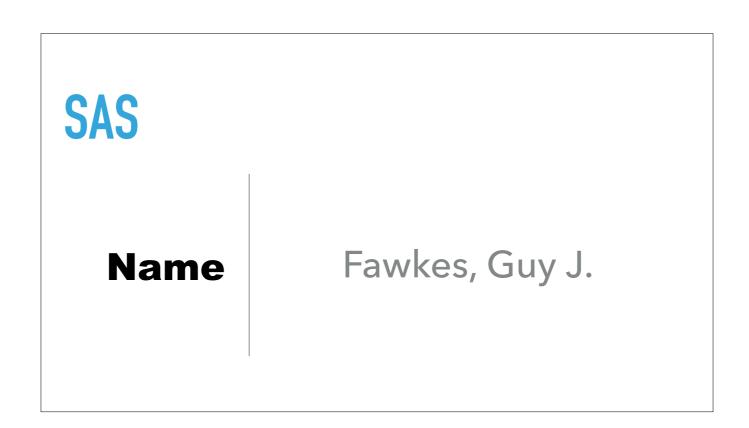
```
Section (Control (Con
```

As you can clearly see...

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 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.

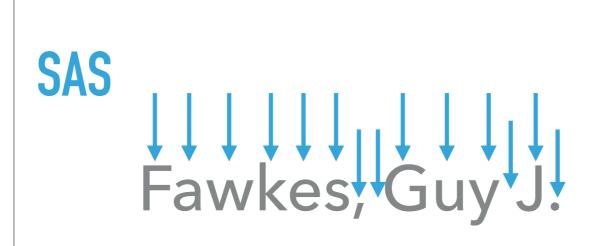
Fawkes, Guy J.

Let's break somethin'!

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.



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.

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

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.

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.

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 Iname Fawkes

and here's the result

However, not every name fits into these categories.

260 Outliers

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

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.

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.

Con:

Not great for MARC

not great for marc...

Yet:)

INTERMISSION BEGGING THE ???

open refine is really good for marc records.

DISCLAIMER

I am not a cataloger.

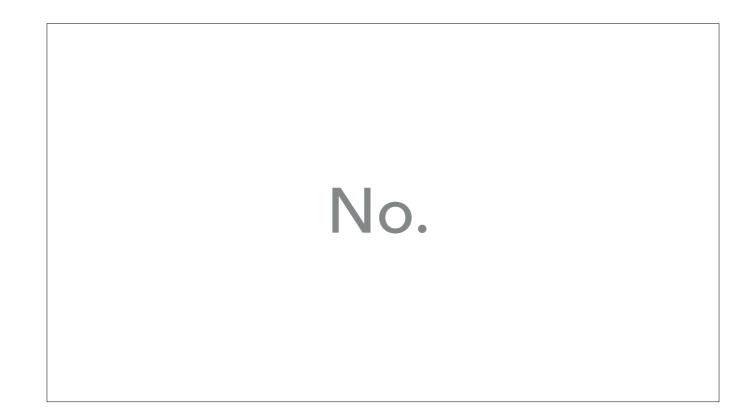
This is going on my tombstone. (past tense?)

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?



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!

Pro tip:

from experience, my suggestion is to [next]

Export as tabdelimited text

export mnemonic marc records from marcedit as tab-delimited text. I could turn that sentence into a limerick.

JSON

json export did not work as well for me.

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

The trick...

with openrefine, the trickiest trick may be...

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.

Steps

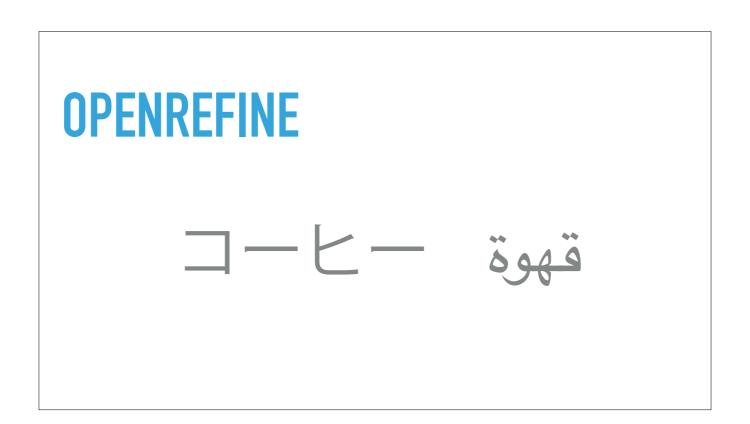
I'm going to walk through the steps needed to get openrefine to recognize marcs. this may not make any sense at first, but if I have time I'll give a quick two minute demo to reinforce it.

UTF-8

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

bêçãüśe dįåčritićš

Because diacritics!



Because non-Roman characters

Blank rows

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

No headers

We do not want to parse column headers.

No quotation marks

quotation marks are NOT used to enclose data

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.

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.

Facet the first column for "LDR"

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

Star or Flag those rows

then star those rows.

Add a column based on the LDR column

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

Oooh! What's its name?

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

rowIndex+1

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

Close the facets

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

Move the new column to the beginning

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

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	Lower tool for working with me	ory Mile.
	Last modified	Name
Create Project	today 12:42 PM	No. (not
Open Project	today 11.40.488	Na_led to:
Import Project	a week ago	bektomatch
Language Settings	a week ago	call_only tot
	a week ago	Al_Checked_Out_Rems_Report tel
	a week ago	No. (see
	2 weeks ago	2111_outbner_nolve_audiobooks
	2 weeks ago	OR_modified_abc clio bit
	a month ago	egray(1_b_ser_cost, 201608
		egray(2_ser_sourt_201609
		egray(1_ser_sourt_201609
	a month ago	NEO soan delete report 190316 txl
	3 months ago	wins new titles 1
	4 months ago	picking_out_entitives tev
	X rename 4 months ago	ArchivesA
	5 months ago	bek_tems_prefixes
		everytemrecord
	Emortis apr	wms_manual_checkins
	5 months ago	post/unell/deckouts bid
	5 months ago	patror_checkouts_from_june99/2015 tid
		MEQ_unknown_items
		everybitmoord bit
		secul_ame_loan_date tol
		nec_tem_statistics1_2016pu09
		sesout_erre_item_stats.txt 2016 06 09 fines fees.txv
		circ joud, data_file
	5-months ago	on one one one
	Emorite ago	law lowne faged
Yellow 2 E-c 2 (TRUNK)		NS. databases
	Emoritia ago	The statement of the st
Nep		
About	Drowse workspace directs	

You've got MARC

TIPS 'N TRICKS

rowIndex+1

use rowIndex+1 to create cell values

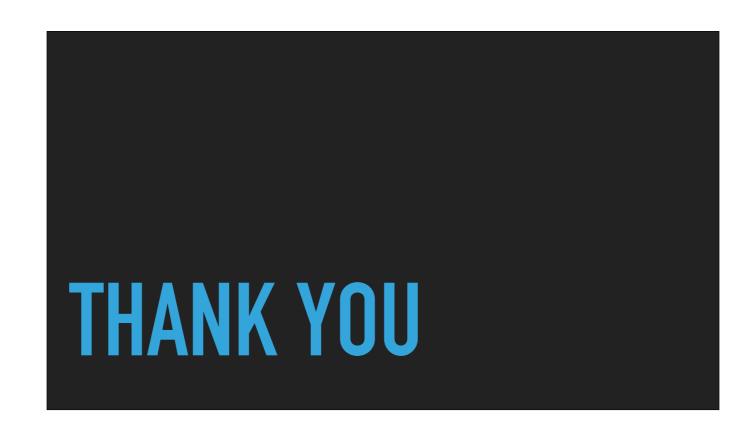
Facets are awesome

Why can't I do this in SAS?

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

```
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.



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! :)

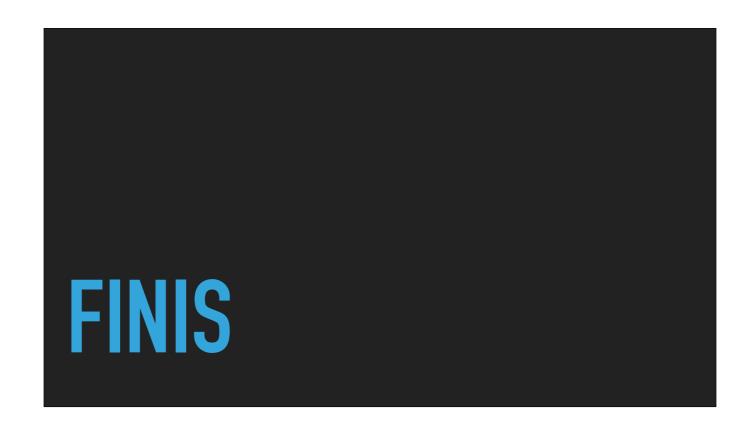
Questions?

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

email: <u>jwaller7@elon.edu</u>

know your data:)

Any questions?



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.