DATA MANIPULATION FOR ILS MIGRATION

CIL 2018 - JERRY WALLER
SYSTEMS LIBRARIAN, ELON UNIVERSITY
Greetings, Earthlings!
Thank you...
A Little About Me
ILS Migration
Goals
What they are
Why I used them
Strategies
WHEN I WAS A PROGRAMMER...
Phase I Clinical Trials
Adverse events
Side effects
FIRST

Collected

Transcribed
THIRD

Cleaned
WHAT DOES BIOSTATISTICS HAVE TO WITH ILS MIGRATIONS?

Both depend on clean data.
ETL:

Extract
Transform
Load
Extract
Why do we clean data?
“DATA COLLECTED BY STATISTICAL AGENCIES MAY CONTAIN MISTAKES MADE DURING THE ACQUISITION, TRANSCRIPTION AND CODING PROCESS.”

For Big-Data Scientists, ‘Janitor Work’ Is Key Hurdle to Insights

By STEVE LOHR • AUG. 17, 2014
CLEANING DATA MEANS:

Ensuring the data conform to a standard.
Know your data!
Know what “good” data look(s) like
your use-case

your requirements
DISCLAIMER

I was a SAS intern
Why SAS?
I know how to use it.
SAS is primarily a procedural language.
Very granular.
Specify the order of code execution.
SQL Syntax.
SAS IS GREAT FOR:

Patron records
SAS IS GREAT FOR:

Circulation data
SAS IS GREAT FOR:

Item records
Acquisitions Data
SAS IS GREAT FOR:
Massive batch conversions of conditional data.
What does that mean?
Millennium → WMS
For example:
<table>
<thead>
<tr>
<th>Name</th>
<th>Fawkes, Guy J.</th>
</tr>
</thead>
<tbody>
<tr>
<td>DOB</td>
<td>04-13-1570</td>
</tr>
</tbody>
</table>
### WMS INPUT FORMAT

<table>
<thead>
<tr>
<th>First</th>
<th>Guy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Middle</td>
<td>J.</td>
</tr>
<tr>
<td>Last</td>
<td>Fawkes</td>
</tr>
<tr>
<td>DOB</td>
<td>1570-04-13</td>
</tr>
</tbody>
</table>
http://www.jerrywaller.org

SAS Programs for Academic and Public Libraries
Fawkes, Guy J.
Fawkes, Guy J.

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

result: p=7
n = length(name);
result: n = 14
Fawkes, Guy J.

\texttt{lname=substr(name,1,(p-1))};

result: \texttt{lname=Fawkes}
Fawkes, Guy J.

\[ \text{fm\_name} = \text{substr(name, (p+1), (n-p))} \]

result: \( \text{fm\_name} = \text{Guy J.} \)
fname=scan(fm_name, 1);
result: fname=Guy
mname=scan(fm_name, 2);
result: mname=J
<table>
<thead>
<tr>
<th>fname</th>
<th>Guy</th>
</tr>
</thead>
<tbody>
<tr>
<td>mname</td>
<td>J</td>
</tr>
<tr>
<td>lname</td>
<td>Fawkes</td>
</tr>
</tbody>
</table>
260 Outliers
To Re-emphasize...
CONS:

Learning Curve
CONS:

MARC Records
CONS:
Expensive
I am not a cataloger.
Why even mess with the marcs?
aren’t they universal?
No.
CREDIT WHERE CREDIT IS DUE:

Terry Reese is smarter than me.
MarcEdit has OpenRefine import/export functionality.
Windows
Mono
Export as tab-delimited text
OpenRefine is free.
OpenRefine is open source software.
OpenRefine is a Java application.
Relies on JRE.

CURRENT VERSIONS (2018-04-18)

OpenRefine: 2.8

Oracle JRE: 8
Because OpenRefine is Java-based...
OpenRefine’s interface is consistent across platforms.
Highly recommended data management practices.
Always make a copy of your data.
Always work on the copy.
GOOD NEWS EVERYONE!

OPENREFINE MAKES A COPY OF YOUR DATA!
These copies are kept in a project.
A power tool for working with messy data.

Create a project by importing data. What kinds of data files can I import?

TSV, CSV, *SV, Excel (.xls and .xlsx), JSON, XML, RDF as XML, and Google Data documents are all supported. Support for other formats can be added with OpenRefine extensions.

Get data from

- This Computer
- Web Addresses (URLs)
- Clipboard
- Google Data

Locate one or more files on your computer to upload:

Choose Files | No file chosen

Next »
Create a project by importing data. What kinds of data files can I import?

TSV, CSV, *SV, Excel (.xls and .xlsx), JSON, XML, RDF as XML, and Google Data documents are all supported. Support for other formats can be added with OpenRefine extensions.

<table>
<thead>
<tr>
<th>Get data from</th>
<th>Locate one or more files on your computer to upload:</th>
</tr>
</thead>
<tbody>
<tr>
<td>This Computer</td>
<td>Choose Files</td>
</tr>
<tr>
<td>Web Addresses (URLs)</td>
<td>No file chosen</td>
</tr>
<tr>
<td>Clipboard</td>
<td></td>
</tr>
<tr>
<td>Google Data</td>
<td></td>
</tr>
</tbody>
</table>

Version 2.7-rc.2 [TRUNK]
<table>
<thead>
<tr>
<th>IP Address</th>
<th>UserID</th>
<th>DateTime</th>
<th>GMT Offset</th>
<th>Method</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>134.173.78.2</td>
<td>-</td>
<td>[01/Mar/2016:00:00:00 -0500]</td>
<td>GET</td>
<td><a href="http://web.a.observabo.com:80/content1?1=10.1007/s00220-016-2898-y">http://web.a.observabo.com:80/content1?1=10.1007/s00220-016-2898-y</a></td>
<td></td>
</tr>
<tr>
<td>134.173.78.2</td>
<td>-</td>
<td>[01/Mar/2016:00:00:00 -0500]</td>
<td>GET</td>
<td><a href="http://web.a.observabo.com:80/content1?1=10.1007/s00220-016-2898-y">http://web.a.observabo.com:80/content1?1=10.1007/s00220-016-2898-y</a></td>
<td></td>
</tr>
<tr>
<td>134.173.78.2</td>
<td>-</td>
<td>[01/Mar/2016:00:00:00 -0500]</td>
<td>GET</td>
<td><a href="http://web.a.observabo.com:80/content1?1=10.1007/s00220-016-2898-y">http://web.a.observabo.com:80/content1?1=10.1007/s00220-016-2898-y</a></td>
<td></td>
</tr>
<tr>
<td>134.173.78.2</td>
<td>-</td>
<td>[01/Mar/2016:00:00:00 -0500]</td>
<td>GET</td>
<td><a href="http://web.a.observabo.com:80/content1?1=10.1007/s00220-016-2898-y">http://web.a.observabo.com:80/content1?1=10.1007/s00220-016-2898-y</a></td>
<td></td>
</tr>
<tr>
<td>134.173.78.2</td>
<td>-</td>
<td>[01/Mar/2016:00:00:00 -0500]</td>
<td>GET</td>
<td><a href="http://web.a.observabo.com:80/content1?1=10.1007/s00220-016-2898-y">http://web.a.observabo.com:80/content1?1=10.1007/s00220-016-2898-y</a></td>
<td></td>
</tr>
<tr>
<td>134.173.78.2</td>
<td>-</td>
<td>[01/Mar/2016:00:00:00 -0500]</td>
<td>GET</td>
<td><a href="http://web.a.observabo.com:80/content1?1=10.1007/s00220-016-2898-y">http://web.a.observabo.com:80/content1?1=10.1007/s00220-016-2898-y</a></td>
<td></td>
</tr>
<tr>
<td>134.173.78.2</td>
<td>-</td>
<td>[01/Mar/2016:00:00:00 -0500]</td>
<td>GET</td>
<td><a href="http://web.a.observabo.com:80/content1?1=10.1007/s00220-016-2898-y">http://web.a.observabo.com:80/content1?1=10.1007/s00220-016-2898-y</a></td>
<td></td>
</tr>
<tr>
<td>134.173.78.2</td>
<td>-</td>
<td>[01/Mar/2016:00:00:00 -0500]</td>
<td>GET</td>
<td><a href="http://web.a.observabo.com:80/content1?1=10.1007/s00220-016-2898-y">http://web.a.observabo.com:80/content1?1=10.1007/s00220-016-2898-y</a></td>
<td></td>
</tr>
<tr>
<td>134.173.78.2</td>
<td>-</td>
<td>[01/Mar/2016:00:00:00 -0500]</td>
<td>GET</td>
<td><a href="http://web.a.observabo.com:80/content1?1=10.1007/s00220-016-2898-y">http://web.a.observabo.com:80/content1?1=10.1007/s00220-016-2898-y</a></td>
<td></td>
</tr>
<tr>
<td>134.173.78.2</td>
<td>-</td>
<td>[01/Mar/2016:00:00:00 -0500]</td>
<td>GET</td>
<td><a href="http://web.a.observabo.com:80/content1?1=10.1007/s00220-016-2898-y">http://web.a.observabo.com:80/content1?1=10.1007/s00220-016-2898-y</a></td>
<td></td>
</tr>
<tr>
<td>134.173.78.2</td>
<td>-</td>
<td>[01/Mar/2016:00:00:00 -0500]</td>
<td>GET</td>
<td><a href="http://web.a.observabo.com:80/content1?1=10.1007/s00220-016-2898-y">http://web.a.observabo.com:80/content1?1=10.1007/s00220-016-2898-y</a></td>
<td></td>
</tr>
<tr>
<td>134.173.78.2</td>
<td>-</td>
<td>[01/Mar/2016:00:00:00 -0500]</td>
<td>GET</td>
<td><a href="http://web.a.observabo.com:80/content1?1=10.1007/s00220-016-2898-y">http://web.a.observabo.com:80/content1?1=10.1007/s00220-016-2898-y</a></td>
<td></td>
</tr>
<tr>
<td>134.173.78.2</td>
<td>-</td>
<td>[01/Mar/2016:00:00:00 -0500]</td>
<td>GET</td>
<td><a href="http://web.a.observabo.com:80/content1?1=10.1007/s00220-016-2898-y">http://web.a.observabo.com:80/content1?1=10.1007/s00220-016-2898-y</a></td>
<td></td>
</tr>
<tr>
<td>134.173.78.2</td>
<td>-</td>
<td>[01/Mar/2016:00:00:00 -0500]</td>
<td>GET</td>
<td><a href="http://web.a.observabo.com:80/content1?1=10.1007/s00220-016-2898-y">http://web.a.observabo.com:80/content1?1=10.1007/s00220-016-2898-y</a></td>
<td></td>
</tr>
<tr>
<td>134.173.78.2</td>
<td>-</td>
<td>[01/Mar/2016:00:00:00 -0500]</td>
<td>GET</td>
<td><a href="http://web.a.observabo.com:80/content1?1=10.1007/s00220-016-2898-y">http://web.a.observabo.com:80/content1?1=10.1007/s00220-016-2898-y</a></td>
<td></td>
</tr>
<tr>
<td>134.173.78.2</td>
<td>-</td>
<td>[01/Mar/2016:00:00:00 -0500]</td>
<td>GET</td>
<td><a href="http://web.a.observabo.com:80/content1?1=10.1007/s00220-016-2898-y">http://web.a.observabo.com:80/content1?1=10.1007/s00220-016-2898-y</a></td>
<td></td>
</tr>
</tbody>
</table>

**Parse data as**
- CSV / TSV
- separator-based files
- Line-based text files
- Fixed-width field text files
- PC-ASCII text files
- JSON files
- MARC files
- RDF/NS files
- Wikitext
- XML files
- Open Document Format

**Character encoding**

- CSV (CSV)
- TSV (TSV)
- Fix (fix)
- Custom (custom)
- Escape special characters with \\n
**Update Preview**

- Ignore first 0 line(s) at beginning of file
- Parse next 1 line(s) as column headers
- Discard initial 0 row(s) of data
- Load at most 0 row(s) of data
- Parse cell text into numbers, dates, ...
- Quotation marks are used to enclose cells containing column separators
- Store blank rows
- Store blank cells as nulls
- Store file source (file names, URIs) in each row

**The Preview Screen.**
Parse data as

**CSV / TSV / separator-based files**
- Line-based text files
- Fixed-width field text files
- PC-Axis text files
- JSON files
- MARC files
- RDF/N3 files
- XML files
- Open Document Format spreadsheets (.ods)
- RDF/XML files
- Excel files

**Character encoding** UTF-8

- Columns are separated by
  - commas (CSV)
  - tabs (TSV)
  - custom \t
- Escape special characters with \n
**Options**
- Ignore first 0 line(s) at beginning of file
- Parse next 1 line(s) as column headers
- Discard initial 0 row(s) of data
- Load at most 0 row(s) of data
- Parse cell text into numbers, dates, ...
- Quotation marks are used to enclose cells containing column separators
- Store blank rows
- Store blank cells as nulls
- Store file source (file names, URLs) in each row
## Parse data as

**CSV / TSV / separator-based files**

- Line-based text files
- Fixed-width field text files
- PC-Axis text files
- JSON files
- MARC files
- RDF/N3 files
- XML files
- Open Document Format spreadsheets (.ods)
- RDF/XML files
- Excel files

### Character encoding

- **UTF-8**

### Columns are separated by

- **Commas (CSV)**
- **Tabs (TSV)**
- Custom \t

### Escape special characters with \n
### Options

- Ignore first 0 line(s) at beginning of file
- **Parse next 1 line(s) as column headers**
- Discard initial 0 row(s) of data
- Load at most 0 row(s) of data
- **Parse cell text into numbers, dates, ...**
- Quotation marks are used to enclose cells containing column separators
- **Store blank rows**
- **Store blank cells as nulls**
- **Store file source (file names, URLs) in each row**
UTF-8
bêçãüsè diåčcritićš
Coffee
Split .mrc files into manageable file sizes.
Group .mrc files into like categories.
AN ASIDE:

1.237 GB .mrc file
The trick...
is defining the individual marc records
Steps
Blank rows
marc records don’t have headers
No quotation marks
The “rows” view
Take me to your LDR.

(I’m so, so sorry)
Facet the first column for “LDR”
Star or Flag those rows
Add a column based on the LDR column
Unique Identifier for Each Record.
rowIndex + 1
Close the facets
Move the new column to the beginning
Click into "records" view
You’ve got MARC
<table>
<thead>
<tr>
<th>All</th>
<th>key</th>
<th>Column 1</th>
<th>Column 2</th>
<th>Column 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>LDR</td>
<td>01908nam2200421i4500</td>
<td></td>
</tr>
<tr>
<td></td>
<td>001</td>
<td></td>
<td>ocm00999999</td>
<td></td>
</tr>
<tr>
<td></td>
<td>003</td>
<td></td>
<td>OCoLC</td>
<td></td>
</tr>
<tr>
<td></td>
<td>005</td>
<td></td>
<td>20151122075022.0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>006</td>
<td></td>
<td>m\o\d\\\\</td>
<td></td>
</tr>
<tr>
<td></td>
<td>007</td>
<td></td>
<td>c\c\n||||||</td>
<td></td>
</tr>
<tr>
<td></td>
<td>008</td>
<td></td>
<td>141031t20142014luwil\o\\\\\000\0\eng\d</td>
<td></td>
</tr>
<tr>
<td></td>
<td>020</td>
<td>\</td>
<td>$z9781629500041</td>
<td></td>
</tr>
<tr>
<td></td>
<td>020</td>
<td>\</td>
<td>$a9781629500058 (e-book)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>020</td>
<td>\</td>
<td>$z9781629500409 (EPUB)</td>
<td></td>
</tr>
</tbody>
</table>
Facets are awesome
Why can’t I do this in SAS?
proc freq data=alpha;
tables col1;
run;
Review
Tools
What they are
Why I used them
Strategies
THANK YOU
Thank you.
Questions?

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

email: jwaller7@elon.edu
References

- https://github.com/OpenRefine/OpenRefine/wiki/GREL-String-Functions
- https://www.oclc.org/support/services/ezproxy/documentation/cfg/logformat.en.html
- https://github.com/OpenRefine/OpenRefine/wiki/Recipes
References


