December 22, 2018

Sampling Data, a Primer

Jennifer Eustis, *University of Massachusetts Amherst*

SAMPLING DATA
A PRIMER

UNDERSTAND YOUR DATA

Creating Context

To identify any trends and possible solutions for your data set, it is first necessary to create your data set's narrative. Can you answer the how, what, why, when, or what?

What format is your data set (xlsx, csv, etc.)?
Who created the set and how?
Why was the set created?
Is this part of a project?

Asking these questions and more will allow you to provide context to your data set.

PROFILE YOUR DATA I

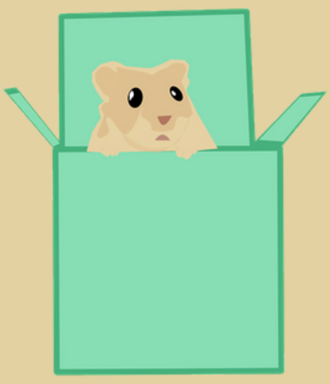
Imagining Structure

Once you understand your data set, then it's necessary to dig into the data itself. This means finding how the data is structured and the content. If you have a spreadsheet, are there column headers and what type of data are found in the columns and/rows? For xml, is the document associated with a schema? Knowing this information will allow you to know what to expect in terms of content.

PROFILE YOUR DATA II

Imagining Content

Knowing the structure of your data set and whether or not it is associated to a schema/data model/etc. will help you anticipate the type of content that should be in the data set. For instance, should dates be in ISO 8601, w3dtf, or a local standard? When in doubt, refer to the person who created the data set or a readme file.



IDENTIFY TRENDS & EXCEPTIONS

Thinking Out of the Box

Once you have painted the picture of the structure and content of the data set, you can more easily see commonalities and differences. Are some formats mismatched, i.e. bib record says book but the item record says audiocd? Are the number of inconsistencies or inaccuracies such that the data needs to be corrected manually or can it be done by a batch process?

CREATE SOLUTIONS

Knowing What Tools To Use

Determining how to correct inaccuracies (manual or batch) is made easier with certain tools. Applications such as OpenRefine, Excel, Excel's Virtual Basic, Python, or MarcEdit all have features to help identify trends, find exceptions, and determine the best way to fix issues. This is also true for creating data sets. Do you query the Oracle database or run a report? Once you've profiled (aka sampled your data), do you use a batch process in your system or change each record one at a time?

Jennifer Eustis

SAMPLING DATA TOOLS

EMPLOYEE OF THE MONTH

CERTIFICATE

OpenRefine

FREE OPEN SOURCE DATA PROFILING & CLEANING TOOL

OpenRefine can be downloaded for free from its website. It is a tool that can be used to either profile data (a data sampling) or clean messy data. OpenRefine works particularly well with spreadsheets but can be used with other formats such as json or xml.

On its website, there are a number of excellent video tutorials to how to use OpenRefine. Other tutorials can be found on YouTube.

HTTP://OPENREFINE.ORG/

OpenRefine Tips

TIPS AND TRICKS

OpenRefine is "like" Excel on steroids. Like Excel, there are some tips to getting the most out of it.

- If your data is just strings or textual data and your original data set is a spreadsheet, highlight the entire file and ensure the cell format is Text. For library data, this is essential so as not to change barcodes into scientific notation for example.

- You can allocate more memory. Just remember not to allocate more memory that what you have on your computer.

-Use "GREL" or String functions supported by OpenRefine Expression Language (https://github.com/OpenRefine/OpenRefine/wiki/GREL-String-Functions)

-List of Tips and Tricks: https://gist.github.com/netsensei/5016312c06fd3a08cb69

HTTPS://GITHUB.COM/OPENREFINE/OPENREFINE/WIKI/DOCUMENTATION-FOR-USERS

Excel

MICROSOFT'S SPREADSHEET PROGRAM

Excel is common to have installed on many computers. It is meant to deal with rows and columns of data.

Tutorials can be found on YouTube.

Tips & Tricks:
- Get a sample of your data by looking at a small amount at the beginning, middle, and end of your spreadsheet. You can also select every 10th row for the same effect.

- Text to Columns (Data Menu)

- Highlight Duplicates (Conditional Formatting in the Home Menu)

-Filter out (and remove) blank cells (Data Menu)

- Macros (VBA) (Developer Menu).

HTTPS://WWW.EXCEL-EASY.COM/VBA.HTML

MarcEdit

FREE AND OPEN SOURCE SUPER TOOL

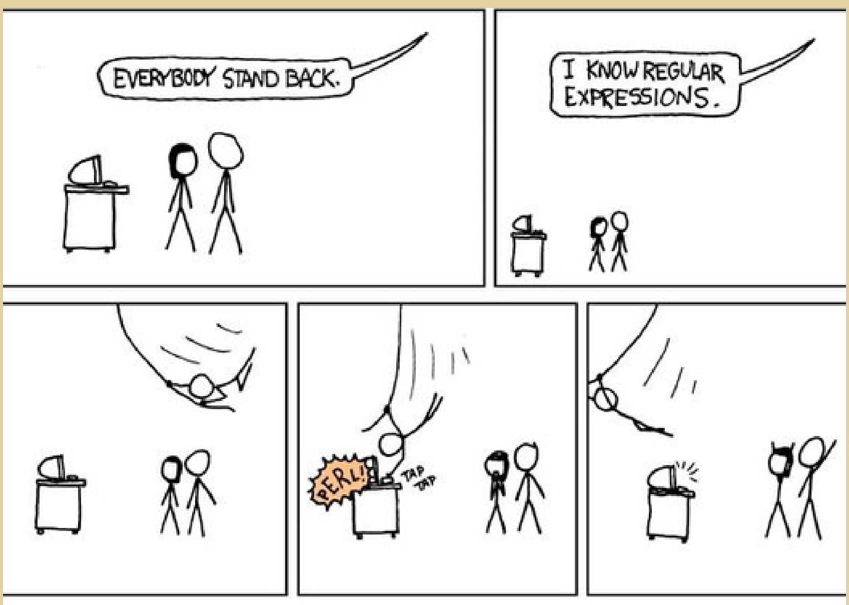
MarcEdit can be downloaded for free from its website. It is a tool that can be used to either profile data (a data sampling) or clean messy data. MarcEdit was developed with Marc in mind. However, this tool can be used to harvest data (OAI-PMH), convert data from MARC to MODS or a variety of other standards, download records from WorldCat, and so much more..

-Field and Material Counts

-Create "Tasks"

-Validate Marc records

HTTPS://MARCEDIT.REESET.NET/



Regex or Regular Expressions

A Regular Expression is a sequence of characters that define a search pattern. Regular Expressions are included in many applications and tools such as OpenRefine, Python, or MarcEdit. Original regexs were (and still are) used in Unix, The programming language Perl added more complex ones and a regex library.

Tutorials exist online. Lynda.com and Code Academy provide good introductions.

To test your regexs in multiple languages, use regex101 at https://regex101.com/.

Jennifer Eustis