



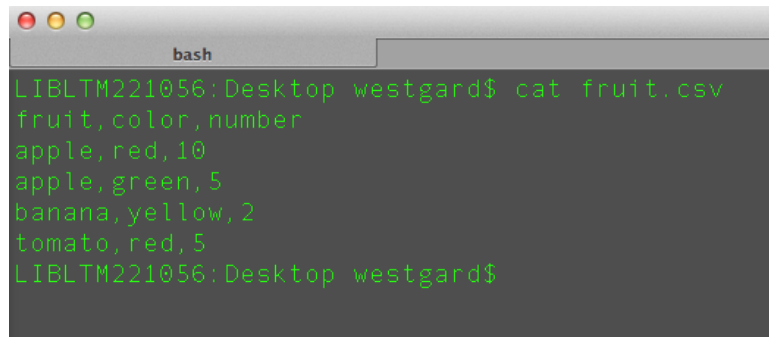
CSV Validation for Metadata Wrangling

***A Libraries Research and Innovative Practice Forum
Lightning Talk***

Joshua Westgard
2015-06-04

The Problem

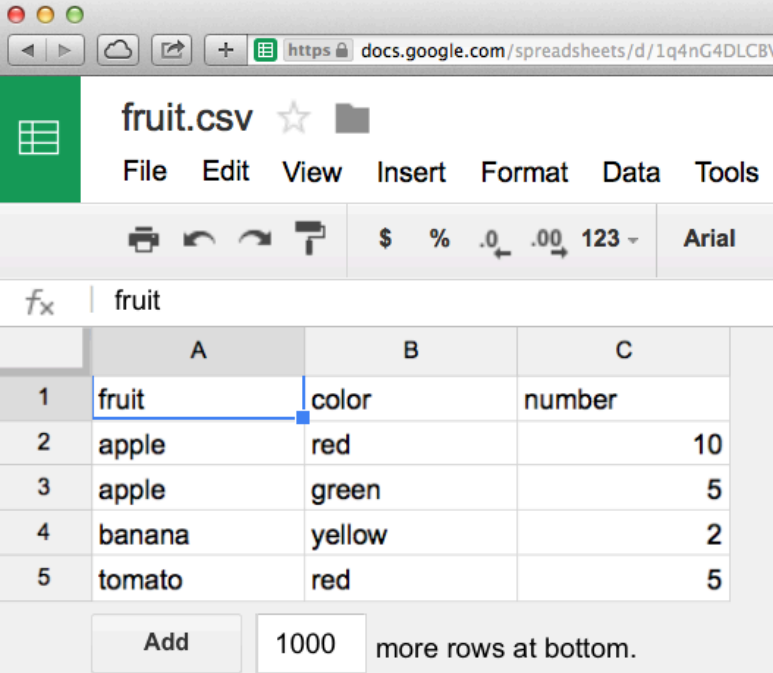
- Libraries have data stored in legacy systems or spreadsheets;
- For moving data between systems or into/out of a spreadsheet, the CSV file is a *de facto* standard;
- CSV = Comma Separated Values; AKA "the king of tabular file formats";
- The flexibility and ease of generating a CSV also means it is not very error tolerant and can be easily mangled;
- Import failures can precipitate a cycle of file ping-pong that is frustrating for both data creators and systems maintainers.



```
LIBLTM221056:Desktop westgard$ cat fruit.csv
fruit,color,number
apple,red,10
apple,green,5
banana,yellow,2
tomato,red,5
LIBLTM221056:Desktop westgard$
```

Issues with CSV

- The comma is a special character (field separator);
- The line break is a special character (record separator);
- Character encoding problems and differences in line breaks (Windows vs. Mac OS vs. Unix/Linux);
- Raw CSV is difficult to read and therefore difficult to correct when problems crop up;
- Tools exist, such as AWK or CSVkit, but may require Unix command line skills;
- Google Sheets/Open Refine are more user-friendly, but not as beloved as Excel.



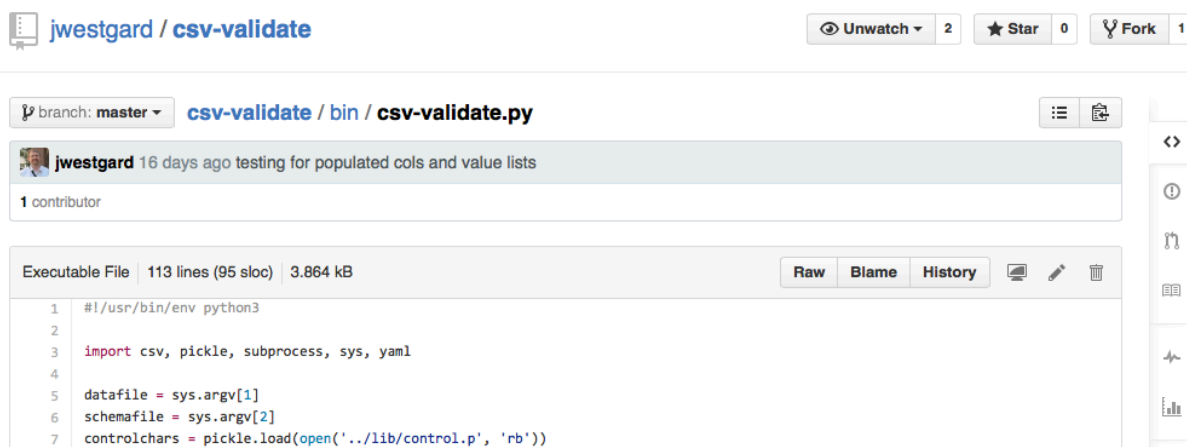
The screenshot shows a Google Sheets interface for a spreadsheet named 'fruit.csv'. The spreadsheet contains the following data:

	A	B	C
1	fruit	color	number
2	apple	red	10
3	apple	green	5
4	banana	yellow	2
5	tomato	red	5

At the bottom of the spreadsheet, there is an 'Add' button, a text input field containing '1000', and the text 'more rows at bottom.'.

Guiding Principles

- It is neither feasible nor desirable to eliminate Excel and CSV from our workflows.
- Solution needs to make things easier, so an onerous learning curve is not an option.
- Data creators are properly the ones to do data validation and cleanup.
- Solution has to be general-purpose, not narrowly conceived.



The screenshot shows a GitHub repository page for 'jwestgard / csv-validate'. At the top, there are buttons for 'Unwatch' (2), 'Star' (0), and 'Fork' (1). Below this, the repository path is shown as 'branch: master / csv-validate / bin / csv-validate.py'. A commit by 'jwestgard' is displayed, dated '16 days ago', with the message 'testing for populated cols and value lists'. Below the commit, the file 'csv-validate.py' is shown as an 'Executable File' with '113 lines (95 sloc)' and a size of '3.864 kB'. The code is displayed in a monospaced font with line numbers 1 through 7. The code includes a shebang line, imports for 'csv', 'pickle', 'subprocess', 'sys', and 'yaml', and assignments for 'datafile', 'schemafile', and 'controlchars'.

```
1 #!/usr/bin/env python3
2
3 import csv, pickle, subprocess, sys, yaml
4
5 datafile = sys.argv[1]
6 schemafile = sys.argv[2]
7 controlchars = pickle.load(open('../lib/control.p', 'rb'))
```

The Solution

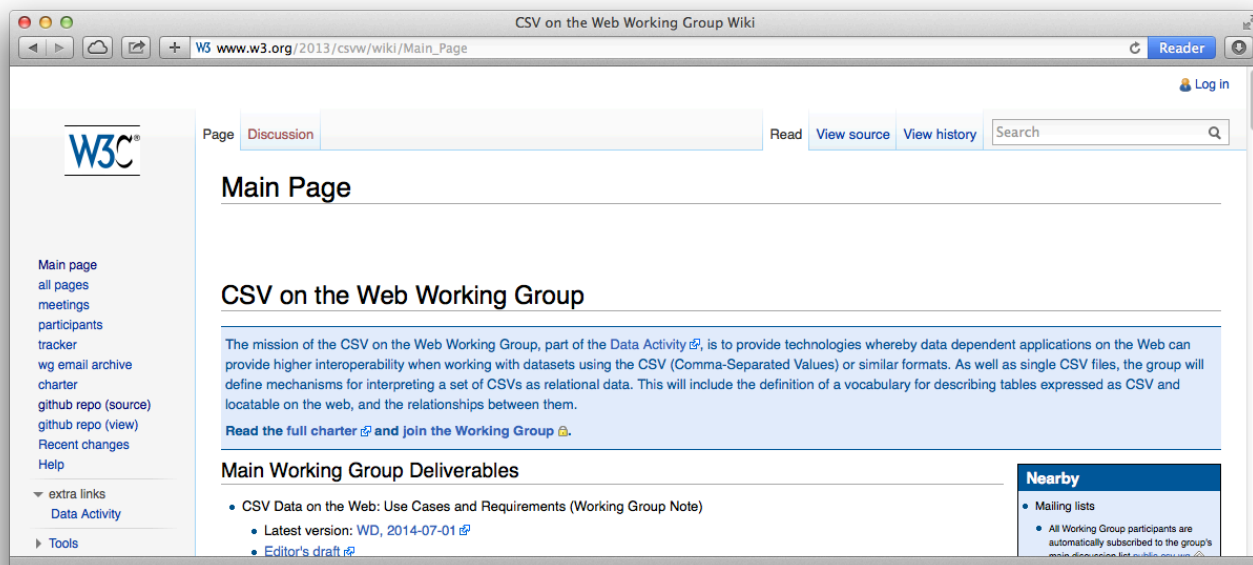
- CSV-Validate is a Python project that seeks to put control of data quality back in the hands of data creators.
- Allows for the creation of a human-readable CSV schema in the form of a YAML file.
- This file can be applied to a CSV with a single-line command:

```
python3 csv-validate.py  
schema.yaml data.csv
```
- CSV-Validate will read the YAML file and the CSV, and evaluate the latter against the former, producing a report that identifies violations of the schema by line and column.
- It can also work directly from an Excel file.

```
-  
  name:      Accession  
  required:  True  
-  
  name:      AlternateTitle  
  required:  True  
-  
  name:      ArchivalCollection  
  required:  True  
-  
  name:      AspectRatio  
  required:  True  
-  
  name:      Box  
  required:  True  
-  
  name:      Century  
  required:  True  
-  
  name:      Color  
  required:  True  
-  
  name:      Continent  
  required:  True
```

Future Plans

- Add data cleanup features to facilitate common data cleaning tasks (similar to Microsoft Office's spell check);
- Convert the python script to a GUI application;
- Explore opportunities to bring the application into alignment with new standards such as CSV on the Web (http://www.w3.org/2013/csvw/wiki/Main_Page).



Thank you!

Joshua Westgard (westgard@umd.edu)



1 / 7

Go to slide:

Go