

Documenting the Undocumented Trail: The Hidden Process of Changing CMC into Data

Amber Warren & Jay Pfaffman

Indiana University, Bloomington | University of South Alabama

Abstract

It is widely acknowledged that one of the hardest parts of doing statistical analysis, especially from computer-generated data, is transforming raw data into a format suitable for statistical analysis (Wickham, 2014). Less widely recognized, however, is that readying for analysis raw data from a computer mediated communication (CMC) system is just as difficult. Manipulating text from computer mediated communications (CMC) environments into a form suitable for analysis with qualitative data analysis software (QDAS) is surprisingly difficult. Qualitative researchers typically lack programming skills or resources to get a programmer to convert raw online data to a form suitable for analysis. As a result, they are likely to fall back on tedious and error-prone work with word processing software. We argue that considering the acquisition and manipulation of CMC data is important part of the research process.

This poster describes the powerful potential of using Python and shell scripts for qualitative researchers studying CMC. A typical process for moving discussions from a learning management system like Sakai or Canvas to a QDAS like Atlas.ti⁷™ involves opening each conversational thread individually and printing each of them to a PDF file. For 13 weeks of discussion involving 86 threaded discussions, manipulating these files for anonymization while maintaining threading would take easily dozens or hundreds of hours. Even with Adobe's powerful PDF editing tools, replacing names with pseudonyms and other preparations for analysis is daunting. This poster provides sample code and a model for how a script can be developed to transform discussion forum data into text suitable for import into QDAS. Implications include the importance of developing examples for creating such code and possibilities for including simple coding alongside introduction of QDAS in qualitative methods courses.

Objectives

- Preparing CMC data for use in QDAS
 - Learn to access data from web in a format for easy manipulation
 - Learn to modify a shell script to anonymize data by replacing names with pseudonyms

Manipulating the Data

Replace names with pseudonyms

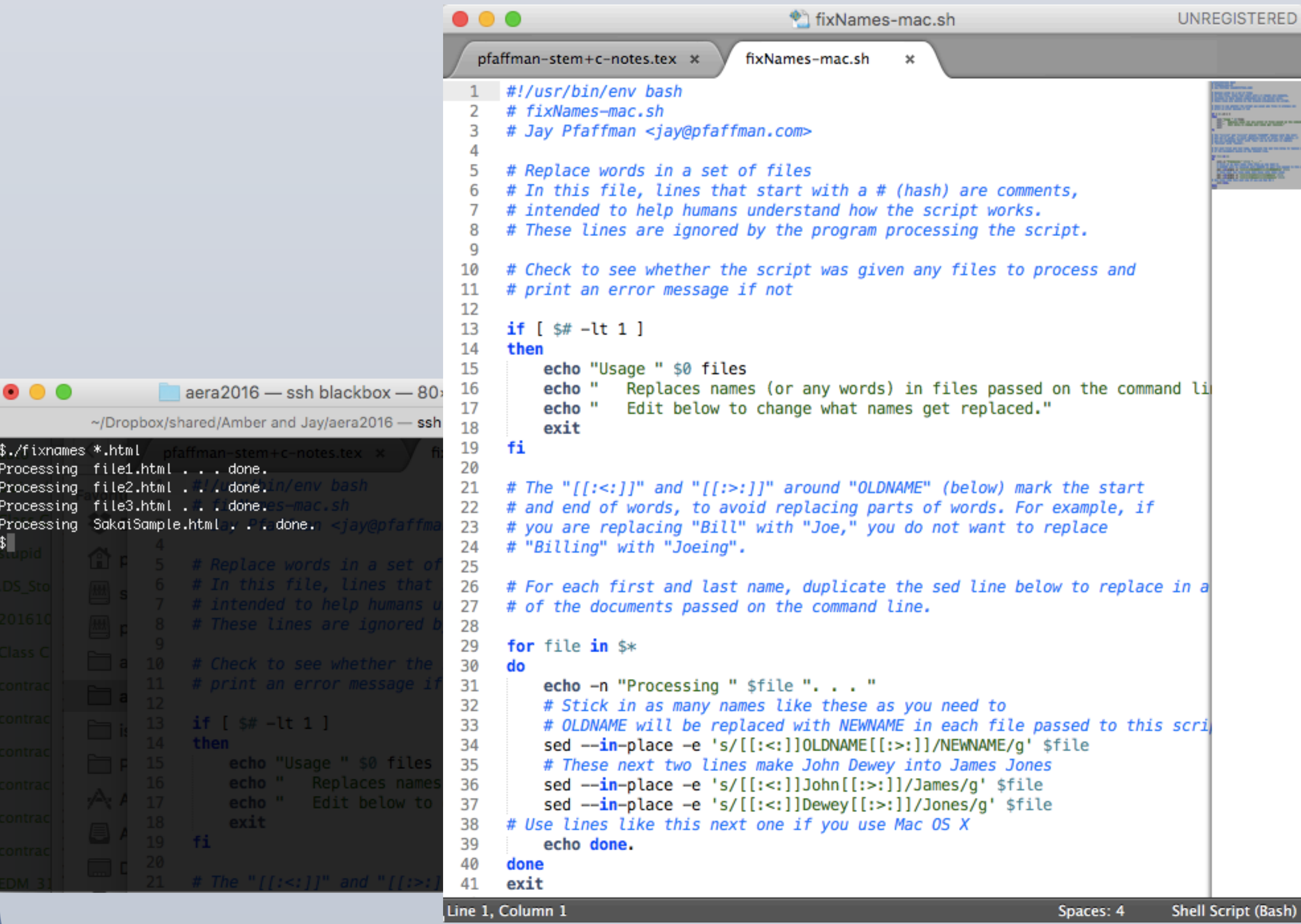
The following script will replace names with pseudonyms.

```
#!/usr/bin/env bash

# The "<" and ">" characters around OLDNAME below tells the sed
# command that replacement work only for whole words. For example, if
# you are replacing "Bill" with "Joe," you do not want to replace
# "Billing" with "Joeing".

# To use this script, open a terminal window on a Mac or a git bash
# terminal in Windows and enter the following to fix all .txt files:
#   ./fixNames *.txt
for file in $*
do
    echo -n "Processing " $file ". . . "
    # Stick in as many names like these as you need to
    # OLDNAME will be replaced with NEWNAME in each file passed to this
    # script
    sed --in-place -e 's/<OLDNAME>/NEWNAME/g' $file
    sed --in-place -e 's/<John>/James/g' $file
    sed --in-place -e 's/<Dewey>/Jones/g' $file
    # Use lines like this next one if you use Mac OS X
    # sed --in-place -e 's/[[:<:]]OLDNAME[[:>:]]/NEWNAME/g' $file
    echo done.
done
exit
```

See the Script in Action



Translating the Script

After you download the files from this presentation (see DOI below) you will be able to open it in a text editor (like Notepad) to manipulate the script. For this activity, we will anonymize the data by replacing old names with new names.

The "<" and ">" characters around the name <OLDNAME> below tells the sed command that replacement work only for whole words. For example, if you are replacing "Bill" with "Joe," you do not want to replace "Billing" with "Joeing".

Do this as many times as you need to replace each OLDNAME with a NEWNAME

THIS IS WHAT YOU WILL SEE:

sed --in-place -e 's/<OLDNAME>/NEWNAME/g' \$file
Example: sed --in-place -e 's/<John>/James/g' \$file
Example: sed --in-place -e 's/<Dewey>/Jones/g' \$file
Use lines like this next one if you use Mac OS X:
sed --in-place -e 's/[[:<:]]OLDNAME[[:>:]]/NEWNAME/g' \$file

Then, to run this script, open a terminal window on a Mac or a git bash terminal in Windows (complete instructions on our website) and enter the following to fix all .txt files: ./fixNames *.txt

Results

The script we provide here will allow a researcher with no programming skills to anonymize any CMC data saved as .html or .txt files (e.g., replace site names, participant names, nicknames, etc.). Calculations suggest that preparing the files for the project the script was initially developed for would take approximately 15 hours (86 files, 30 global replacements x 20 seconds/replacement = 51600 seconds or 14.333 hours). We also found that documenting this process encouraged us to reconsider traditional means of both acquiring and manipulating CMC data and suggest qualitative researchers may look to programming to automate such tasks in the future.

For more information:

To contact the developer: jay@pfaffman.com
To learn about his other work: <https://literatecomputing.com>
To contact the presenters: ambwarre@indiana.edu; jay@pfaffman.com
This poster and source code to all files available at the DOI below.