Feb 09, 2022

Cleaning, Aggregating, and Filtering CMU Libraries Open Science and Data Collaborations Program Data

Patrick Campbell[1], Huajin Wang[1], Melanie Gainey[1], Sarah Young[1], Katie Behrman[1]

[1]Carnegie Mellon University Libraries

dx.doi.org/10.17504/protocols.io.b29gqh3w

renuka.s

This document describes the process and tools used to clean, aggregate, and filter the data resources collected and maintained by University Libraries' Open Science and Data Collaborations program at Carnegie Mellon University (CMU). This data cleaning protocol is used to support program evaluation and strategy development.

DOI

dx.doi.org/10.17504/protocols.io.b29gqh3w

Patrick Campbell, Huajin Wang, Melanie Gainey, Sarah Young, Katie Behrman 2022. Cleaning, Aggregating, and Filtering CMU Libraries Open Science and Data Collaborations Program Data. **protocols.io** https://dx.doi.org/10.17504/protocols.io.b29gqh3w

Data Harmonization, User Data, Open Science, Program Evaluation, Data Cleaning, Academic Libraries

protocol ,

Dec 24, 2021

Feb 09, 2022

56328

Designing Final Data Model

1 Each row in the final master dataset represents a unique user and columns represent

attributes related to their identity and activity. The data is organized using the data model pictured below (Figure 1).

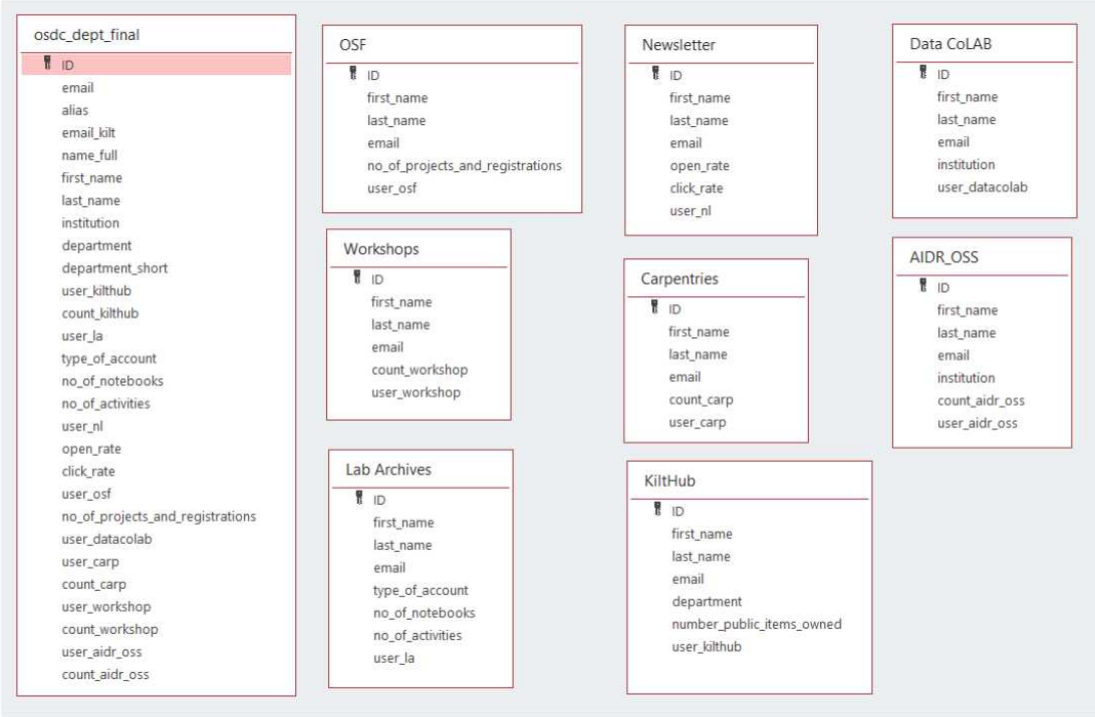

Figure 1. Data model for the Open Science and Data Collaborations program master dataset. The leftmost table (osdc_dept_final) presents the set of fields included in the final dataset. This dataset combines and merges records and fields from each of the additional data tables present to the right, including tables for OSF, Workshops, LabArchives, KiltHub, Newsletter, Carpentries, dataCoLAB, and AIDR-OSS Symposia.

Cleaning

2

The steps included in this section were taken to correct minor spelling and organizational errors that were present in the component datasets in order to prepare them for matching, merging, and other more advanced operations.

First, import all the component datasets (OSF, Workshops, LabArchives, KiltHub, Newsletter, Carpentries, dataCoLAB, and AIDR-OSS) into separate worksheets in a single Excel workbook. Refer to Data Collection Methods in the manuscript for more information about these datasets.

Then, for each worksheet, where necessary:

3 Separate first and last names into separate columns called *first_name* and *last_name*, respectively.

Citation: Patrick Campbell, Huajin Wang, Melanie Gainey, Sarah Young, Katie Behrman Cleaning, Aggregating, and Filtering CMU Libraries Open Science and Data Collaborations Program Data https://dx.doi.org/10.17504/protocols.io.b29gqh3w

4 Use the Proper function to edit all name fields to use proper capitalization.

This step requires inserting a new column, placing the formula in that new column to return the desired values, then replacing the old column with this new one.

5 Remove any periods after initials in name fields.

6 Remove nonsensical characters.

7 Reconcile equivalent emails (capitalization, etc.).

8 Add institutional addresses for CMU students (*institution* = "Carnegie Mellon University") with gmail addresses using CMU directory.

9 Use conditional formatting in the *first_name*, *last_name*, and *email* fields to identify duplicate entries. Merge and delete duplicate records.

This step may require additional investigation to determine whether records with matching first and last names but with non-matching email addresses refer to the same person. If you're not able to resolve this through further investigation, assume the records refer to distinct individuals and do not merge them.

10 Separate institution and department values into separate columns named *institution* and *department*, respectively (dataCoLAB table only).

11 Use Find/Replace to replace all missing values (NA, Null, etc.) with an empty cell.

Citation: Patrick Campbell, Huajin Wang, Melanie Gainey, Sarah Young, Katie Behrman Cleaning, Aggregating, and Filtering CMU Libraries Open Science and Data Collaborations Program Data https://dx.doi.org/10.17504/protocols.io.b29gqh3w

12 Add a column named *user_<ToolName>* that takes 1 or 0 values. Assign a 1 to all users to indicate that they are users of this tool or event. (0's will be added after the Aggregation and merging section for non-users of each tool or event.)

13 Add another column and name it *count_<ToolName>*. (Note that there could be more than one count columns for some datasets, eg. LabArchives has both "count_notebooks" and "count_activities.")

Aggregation and merging

14 Create a new worksheet in the Excel workbook and name it "Master_Merged."

15 Add the following fields in this order: *ID, first_name, last_name, email, institution, department, user_kilthub, count_kilthub, user_la, type_of_account, no_of_notebooks, no_of_activities, user_nl, open_rate, click_rate, user_osf, no_of_projects_and_registrations, user_datacolab, user_carp, count_carp, user_workshop, count_workshop, user_aidr_oss, count_aidr_oss*.

> Additional fields shown in the model above, including *alias*, *email_kilt*, *name_full*, and *department_short*, will be added in the steps below.

16 Append the data from every component dataset into the new master worksheet "Master_Merged") shifting the values over as necessary to match the field name at the top.

> **Note**: This will result in a lot of empty/unused fields for records from certain datasets. That's ok. By the end, you should have a table containing all the chosen fields from all of the component datasets.

17 Use conditional formatting to identify duplicate users using the *first_name*, *last_name*, and *email fields* across the different data sets. One by one, manually merge the data from all the rows for each user, deleting the duplicate records as you go.

> **Note**: Be very careful during this step not to misplace or change values as you're copying them. I recommend using the first instance of every duplicate user to collect all the values from each subsequent instance.
>
> **Caution:** It's a good idea to use version control, or work with a copy of the original file.

Citation: Patrick Campbell, Huajin Wang, Melanie Gainey, Sarah Young, Katie Behrman Cleaning, Aggregating, and Filtering CMU Libraries Open Science and Data Collaborations Program Data https://dx.doi.org/10.17504/protocols.io.b29gqh3w

18 Insert an empty column to the right of the *institution* column and name it *institution2*. In the top cell of the new column, use the IF function to test whether the user's email contains either a CMU (@andrew.cmu.edu, @cmu.edu) or University of Pittsburgh (@pitt.edu) email extension. If the test returns TRUE, insert the name of the corresponding institution ("Carnegie Mellon University" for @andrew.cmu.edu and @cmu.edu extensions, "University of Pittsburgh" for @pitt.edu extensions). If the test returns FALSE, return the current value of the original *institution* column for that record.

19 Highlight the contents of the *institution2* column and use the Copy/Paste Values function to replace the contents of the column with the text values. Delete the original *institution* column and rename the *institution2* column *institution*.

Filtering

20

In this stage of the data cleaning process, we removed all the records where there was no email address to identify the user and also those records where the user was not affiliated with either CMU or the University of Pittsburgh.

Caution: It's a good idea to use version control or work with a copy of the original file, especially before deleting records.

For the steps in this section, first make a copy of the "Master_Merged" worksheet and rename it "Master_CMU_Pitt_only." In the new worksheet, do the following:

Use the Filter tool to sort all records by their email address. Select all records with missing values for email and delete them. Turn the filter off.

21 Use the Filter tool to sort all records by their institution. Select all records where *institution* does not equal either "Carnegie Mellon University" of "University of Pittsburgh" and delete them.

Harmonization Using the KiltHub Data Model

22

In this stage of the data cleaning process, we used the data model from the KiltHub dataset to identify and harmonize any remaining duplicate records. These duplicates were not addressed in step 7 of the first cleaning and aggregation stages because they were identified by distinct email addresses (the most common of these were alias accounts for CMU staff and faculty members). Because KiltHub accounts are created for all CMU researchers automatically using their primary institutional email

Citation: Patrick Campbell, Huajin Wang, Melanie Gainey, Sarah Young, Katie Behrman Cleaning, Aggregating, and Filtering CMU Libraries Open Science and Data Collaborations Program Data https://dx.doi.org/10.17504/protocols.io.b29gqh3w

(@andrew.cmu.edu), we were able to use these to standardize the email addresses used to identify each user and harmonize them with the KiltHub dataset. Secondary email addresses were preserved in a new field that we named “alias.”

Filter the Institution column to only show CMU records.

23 Copy the most recent complete version of the KiltHub data (see Data Collection Methods in the manuscript) into a separate sheet and delete all fields except the *email* and *group* (department) fields. Rename the sheet “KiltHub data model” to distinguish it from the KiltHub dataset that is already contained in your workbook.

24 Back in the main worksheet (“Master_CMU_Pitt_only”), create three empty columns to the right of the *department* field. Name them *department_kilt*, *department_merged*, and *department_cleaned*, respectively.

25 In the new *department_kilt* column, use the VLOOKUP function to search the email fields in the “KiltHub data model” sheet and return the *department* field (labeled group) for each matching email address from the master dataset (“Master_CMU_Pitt_only”). Use the Copy/Paste Values function to replace the formulas in the *department_kilt* column with just values.

26 In the *department_merged* column, use the IF function to test if the value for *department_kilt* is null. If the value is not null, return that value from the *department_kilt* column. If the value is null, return the value of the original *department* field instead. Use the Copy/Paste Values function to replace the formulas in the *department_merged* column with just values.

27 Copy the values from the *department_merged* column into the *department_clean* column. Sort the records in reverse alphabetical order (Z-to-A) by the values in the *department_clean* column.

28 Compare the values from the *department_cleaned* column with the values from the *department_kilt* column. Identify corresponding department names between the two columns (e.g., “Biological Sciences” and “Department of Biological Sciences”) and use the Find/Replace function to replace the value in the *department_clean* column (e.g., “Biological Sciences”) to match the value in the *department_kilt* column (e.g., “Department of Biological Sciences”).

This step is to harmonize the departmental names to match the KiltHub terminology.

29 For the remaining non-matching records, refer to the CMU Academics webpage to determine the closest matching department from the *department_kilt* column and replace the value with that department name.

Citation: Patrick Campbell, Huajin Wang, Melanie Gainey, Sarah Young, Katie Behrman Cleaning, Aggregating, and Filtering CMU Libraries Open Science and Data Collaborations Program Data https://dx.doi.org/10.17504/protocols.io.b29gqh3w

In some cases, this will result in replacing a more specific value, e.g., Information Security, with a more general value, e.g., Heinz College.

If departments cannot be harmonized using this method, leave the values as they are.

30 In the “Master_CMU_Pitt_only” worksheet, create a new column to the left of the *first_name* field and name it *name_full*. Use the concatenate function (=concatenate([first_name],” “, [last_name])) to combine the values from the *first_name* and *last_name* columns.

31 In the “KiltHub data model” sheet, use the Find/Replace function to eliminate all periods (.) from the *Author name* field.

Author name in the KiltHub data model worksheet corresponds to the *name_full* field in the “Master_CMU_Pitt_only” worksheet.

32 Repeat steps 21 through 26 using the *name_full* field.

This series of steps is designed to return any missing departments from the KiltHub data set where the records match for the name field but not for the email field.

33 Remove every department column except for the *department_clean* column. Rename the *department_clean* column as *department*.

34 In the main sheet, use conditional formatting on the *name_full* field to identify any remaining duplicate entries. Merge the values from the duplicate records using the primary (institutional) email address as the primary record, copy/paste the secondary email address into the *email_alias* column, and delete the duplicate record.

35 Use the Filter tool to sort records by department. For any record where the value in the department field is the name of an administrative office rather than an academic department or research center, replace that value with “Other”. Do the same thing for any record where the value in the *department* field is “Carnegie Mellon University” or “Faculty Other.”

36 Insert a new column to the right of *department* and name it *department_short*. For any department name in the *department* column containing extraneous elements (e.g., “Department of,” “School of”, etc.), copy just the non-extraneous elements from into the new *department_short* column. For any department name in the *department* column still

Citation: Patrick Campbell, Huajin Wang, Melanie Gainey, Sarah Young, Katie Behrman Cleaning, Aggregating, and Filtering CMU Libraries Open Science and Data Collaborations Program Data https://dx.doi.org/10.17504/protocols.io.b29gqh3w

containing three or more words after removing all extraneous elements, copy just the acronym into the *department_short* column.

37 Turn off all filters and correct any remaining obvious errors.

38 Select “Master_CMU_Pitt_only” worksheet and save the file as "osdc_dept_final.csv".

Citation: Patrick Campbell, Huajin Wang, Melanie Gainey, Sarah Young, Katie Behrman Cleaning, Aggregating, and Filtering CMU Libraries Open Science and Data Collaborations Program Data https://dx.doi.org/10.17504/protocols.io.b29gqh3w