# Authors Data Cleanup Documentation

## CVS File authors\_processed.csv

The authors\_processed.csv is a pre-processed version of author.csv (the original file from Neo4j) via a Python script so that author names that are potentially a duplicate are identified. There are a total of **10,916 author entries** in this format in which ***n.name*** is the author’s name and ***indicator*** are flagged ***similar*** if the author’s name is deemed to be a possible duplicate and ***dissimilar*** if not:

A table of names

Description automatically generated with medium confidence

## Data Comparison

Using Excel’s functionalities of filtering and conditional formatting, I further simplified the data set to manually review for duplicate author names.

1. I identified that there were **848 author entries** marked with the indicator ***similar***
2. I manually reviewed and noticed entries marked ***similar*** compares to the entry above to determine if they are duplicates (authors\_processed\_manual\_test excel tab)
3. I created a conditional formatting that would highlight the entry with the ***similar*** indicator as well as the entry above it (authors\_processed\_data\_cleanup excel tab)
4. I manually checked between the two ways of viewing the data to ensure my highlight conditional formatting of the potential duplicate “pair” (*two or more author name entries sharing similarities and thus potential duplicates)* is accurately captured
   * I manually looked through all similar indicated entries and highlight marked only entries that are indeed duplicates in the authors\_processed\_manual\_test excel tab
   * The conditional formatting highlights all potential duplicate entries and I highlight marked to identify actual duplicates is the authors\_processed\_data\_cleanup excel tab
5. Now, there is **1,535 author entries** to look at for manual review

|  |  |
| --- | --- |
|  |  |

## Understanding Duplicate Scenarios

As I went through manually reviewing potential author name duplicates, I tried to find common situations and trends to further decrease manual review work and a method to ensure all actual duplicate author names were captured. While there are commonalities, I believe manual review and human judgment from here on out is required to compare the similarities between author names in each potential duplicates, in which some strings represent duplicate author names and others do not, and also to compare and ***cross-reference author details as needed***.

## Cleaning Up Potential Duplicates to Determining Actual Duplicates

1. After data set is adjusted to capture **1535 potential duplicate author entries**, I identified those that could actually be duplicates
2. In the first review, I filtered for those I identified as actual duplicates and investigated for correction of author name and eliminated duplications
3. In the second review, I checked through for any I may have missed the first time
   * Duplicate situations I came across include:
     1. Using middle names as an abbreviation or written out in full
     2. Existing more complete versions of a name
     3. Formatting names differently (e.g. spacing, capitalization)
     4. Spelling differences including differences due to use of special characters (e.g. è)
4. Methods I used to determine actual author name duplications include personal judgement, google and ChatGPT.
   * Personal judgement was used for duplications due to different uses of middle names, a more complete version of a name, and differences in formatting
   * I used Google to compare names that were very similar and compare author identities by affiliations and research interests to determine if they were indeed the same person
   * I used ChatGPT to decode encoding issues and identify names with special characters (took out special character in duplicates)
   * Below is a screenshot showing use of ChatGPT to help identify author name duplications as well as other identified duplicate author names put in authors\_processed\_data\_cleanup excel tab:

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

## Identified Actual Duplicate Author Names and Finalized Authors List

Out of **1535 potential duplicate author entries** there are **392 duplicate author entries**. I made a new excel tab called authors\_processed\_copying\_over to copy over the nonduplicate author names into one column (Cleaned Ver of All Author Names). Because it is a large excel file with filters on certain cells, using the copy and paste function causes issues in properly migrating the data to the correct cells. Thus, I used an excel formula to ensure accurate migration of nonduplicate author names without disrupting the changes I made for duplicate author names. Furthermore, I cross-referenced to check the **10,916 author entries** total count and **392 duplicate author entries** count to ensure data integrity.

A screenshot of a computer

Description automatically generated

Finally, I created an excel tab called authors\_processed\_final\_ver to hold all the author names with the cleaned up duplicate author names. All formatting of cells is cleared to prevent any issues with converting the excel file back to a csv file.

A screenshot of a computer

Description automatically generated