Table of Contents

[Overview 2](#_Toc115405258)

[Intended Audience 2](#_Toc115405259)

[Skills and tools 2](#_Toc115405260)

[Examples of Standardizing VOTER file data fields 3](#_Toc115405261)

[Example of searching for odd entries in PREFIX field 3](#_Toc115405262)

[Example of searching for alphanumeric data (e.g. letters) within Street Number field 4](#_Toc115405263)

[Example of searching for variations in specifying “PO Box” 5](#_Toc115405264)

[Example of searching for variations in spelling of the (postal) town name within the residence address fields 6](#_Toc115405265)

[Appendix A - How to create the collection of files called a DISKFILE and extract the set of compressed files within 7](#_Toc115405266)

[Importing the VOTER “text” data file into EXCEL 10](#_Toc115405267)

[Appendix B - How to create column headers (names of fields) from the REPORT.HTML file within DISKFILE 13](#_Toc115405268)

[Appendix C - How to confirm the data is stored in proper columns within EXCEL (no extra commas or other delimiters have made more columns for some rows than others) 17](#_Toc115405269)

[Appendix D - How to use the Filter options and the filtered search from the DATA tab/menu 20](#_Toc115405270)

# Overview

Over many years the quality of data in any system can decline as protocols and conventions used by different personnel maintaining the data drift without being rigidly controlled by the data entry system(s). This can cause difficulties over time, and in particularly can create messes when exporting or converting data to other systems.

Intended Audience

Generally this document will be most useful to Connecticut Registrar’s of Voters, their staff, SOTS Election Division staff, and the Regional Election Monitors or other assistants in the data maintenance. While the document uses examples found in individual town’s “DISKFILE’s, similar review and analysis can be made by those with access to the state level CDFILE such as parties, candidates, and SOTS itself.

Skills and tools

This document assumes readers’ id & password authorized access to CT’s voter registration system (CVRS) in use still in 2022, some experience with Microsoft OFFICE spreadsheet tool “EXCEL”, and familiarity with terms used to describe data, such as “numeric”, “alphanumeric”.

Those who are not already familiar with how to create a DISKFILE for their town from CVRS will need to learn from Appendix section that explains how to create the collection of files called a DISKFILE, extract the set of compressed files with, and then usually most further work is mostly with the “VOTER” text file and spreadsheets made from it.

Readers who do not already have a list of the fields (column headings) can refer to Appendix section for how to “lift” the column headers (names of fields) from the REPORT.HTML file within DISKFILE. Thereafter they can likely re-use or modify a column header file received from others or that they made earlier.

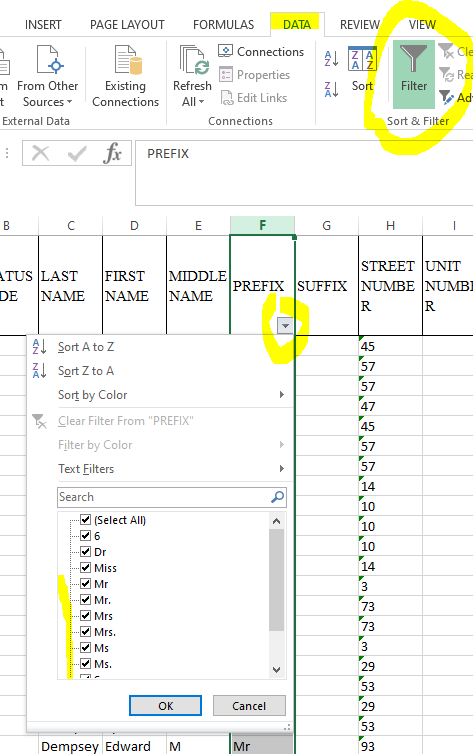
Readers already comfortable with extraction of a DISKFILE, labeling the columns, and use of the EXCEL DATA tab’s FILTER tool, ***or*** *who use the assistance of those that already do know this background material*, can proceed to the examples of **how to confirm the data is stored in proper columns within EXCEL** (no extra commas or other delimiters have made more columns for some rows than others), how to clean that up any extra delimiters existing within data fields.

Some readers will need practice with spreadsheet techniques in Appendix such as highlight/select, delete, insert, and **especially… from the DATA tab/menu use the Filter options and the filtered search.**

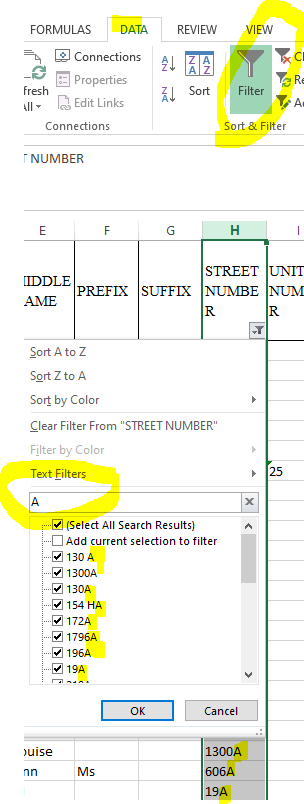
If readers’ data already has all rows with same number of columns… then such readers can work directly from examples of data cleanup provided (prefixes, street numbers and town names, etc.).

# Examples of Standardizing VOTER file data fields

## Example of searching for odd entries in PREFIX field

 In this example there are several variants of common prefixes – some with periods, some without. If this lack of standardization bothers you, or bothers a data conversion process you need to comply with, you can find out which voters need such fields edited using a filter like this one.

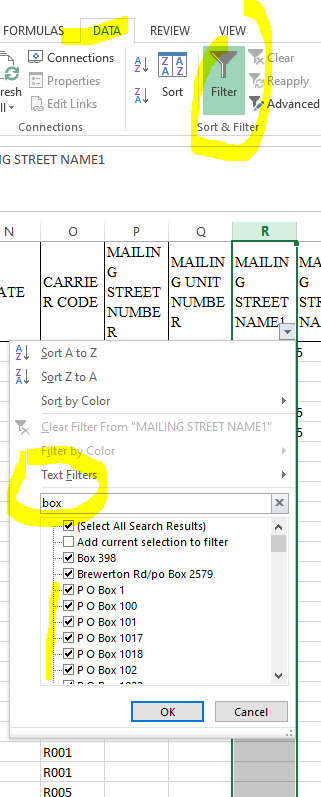
# Example of searching for alphanumeric data (e.g. letters) within Street Number field

Here notice the use of the Text Filter search box with example of “A” here though other non-numeric data values could be found.

An alternate approach to test the street number field would be to create a formula that consisted of “=VALUE(columnrowofcelltotest)” and copying that down for all rows. Numeric values will show as a number if the value of the formula, but alpha numeric values will produce an error value.

There is good value in cleaning up the street number fields for both residence and mailing addresses as this helps their sort in the voter check in lists, adds in correct zip code determination, and is essential for any street that is split in redistricting by odd or even values, or a range. For example, 16 is even, not odd. 16 is less than 20 and more than ten. But 16 and an apartment of unit number such as “16B” is neither odd, nor even, nor clearly higher of lower than a number. This leads to ZIP CODE and redistricting problems, and likely will be a data conversion concern ahead.

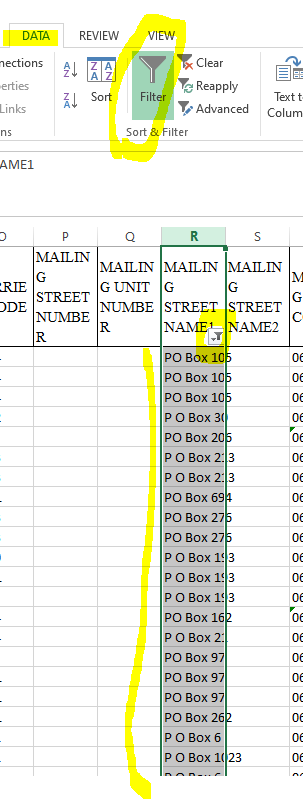
# Example of searching for variations in specifying “PO Box”



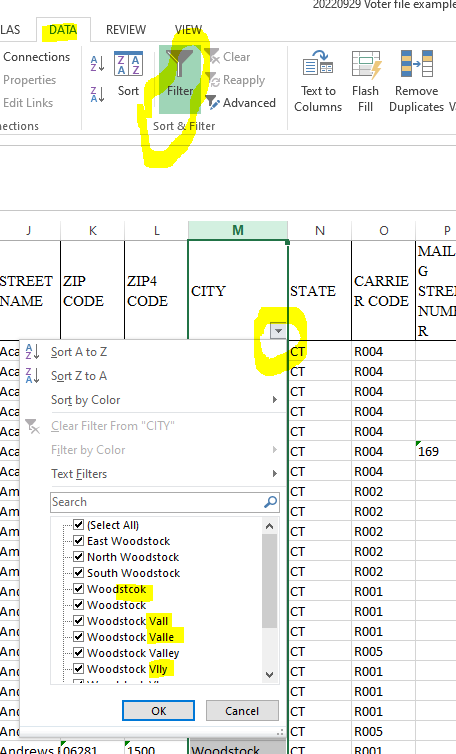
Note here the wide range of ways PO Box was spelled (with spaces, with periods, without, etc).

ZIP code software responds better to conformance with USPS preference for specifying the “PO BOX” note all capital letters, one space.

Here’s an example of varying spellings:

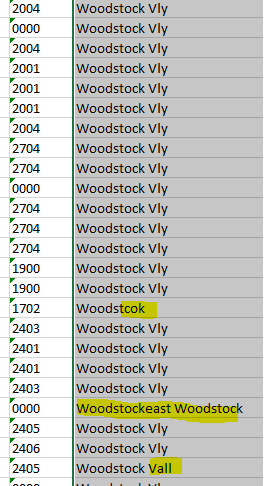


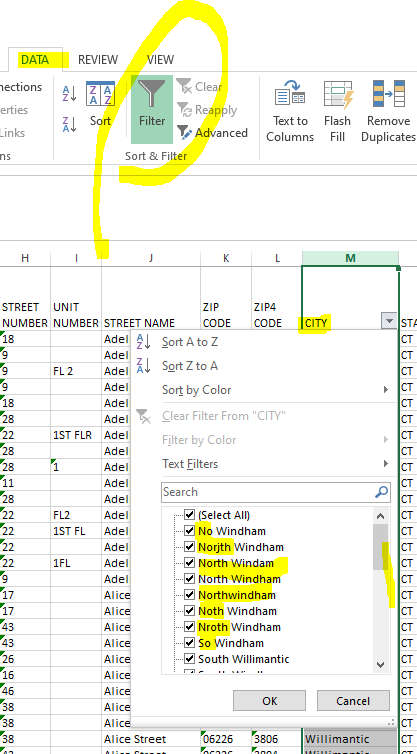
# Example of searching for variations in spelling of the (postal) town name within the residence address fields



These variations can cause ZIP CODE software to become confused, and are a common cause of troubles for candidates using our data.

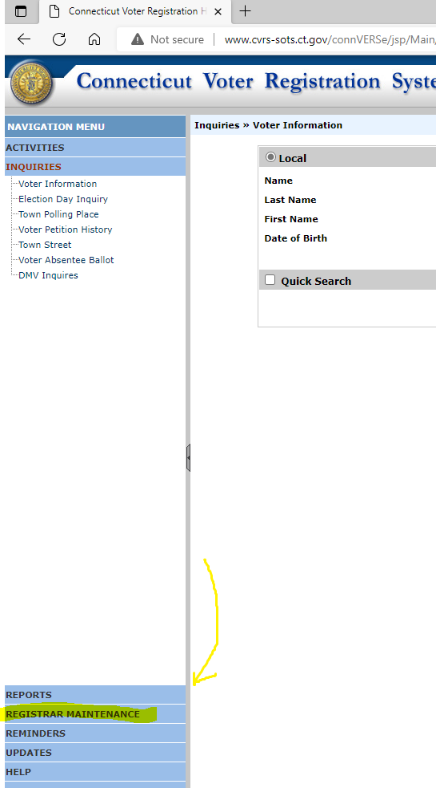
Here’s a subset:

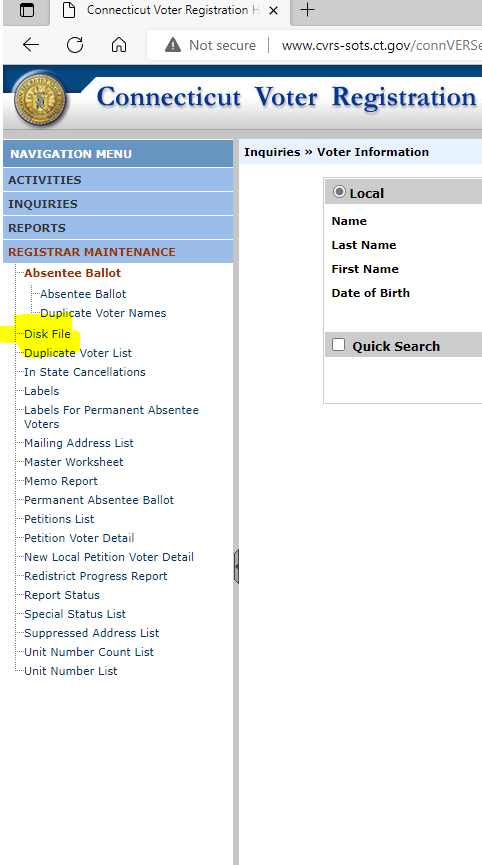
 and another town:



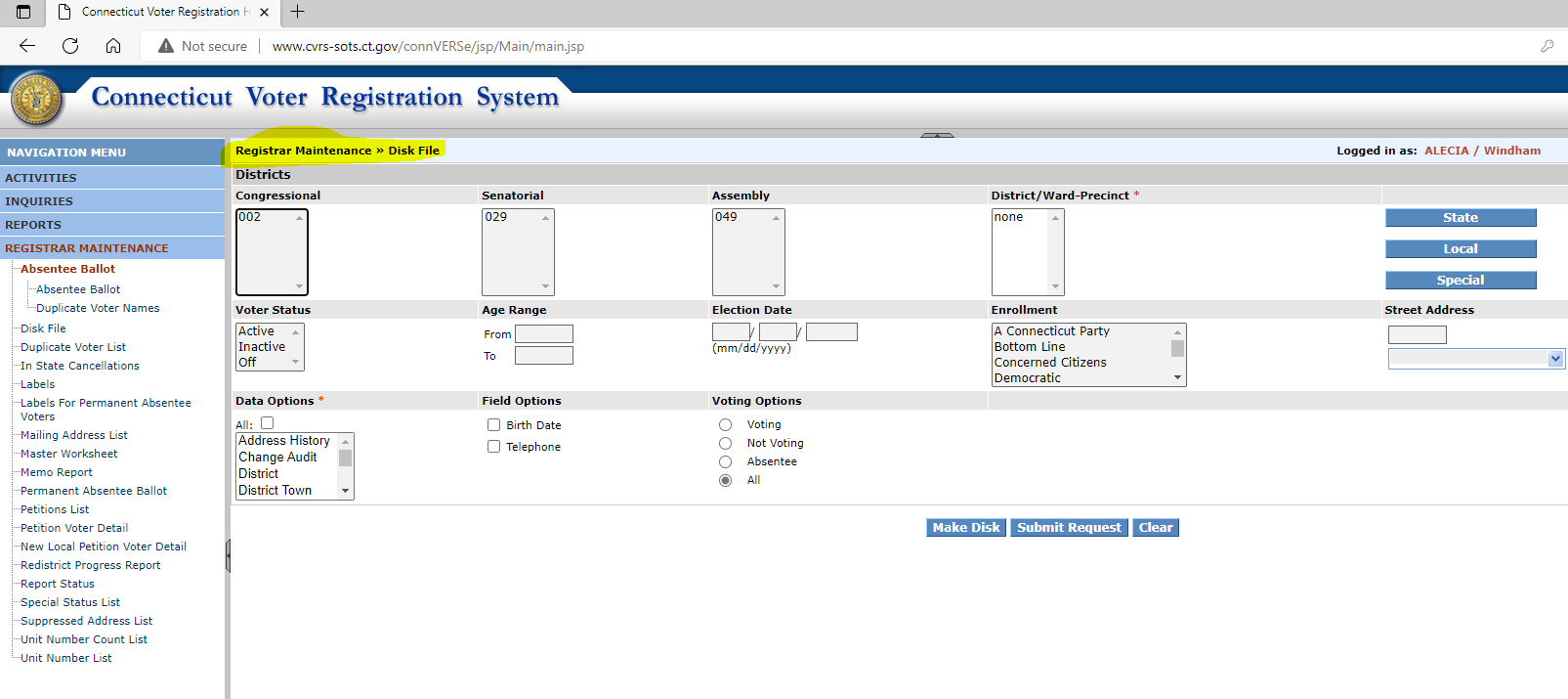
# Appendix A - How to create the collection of files called a DISKFILE and extract the set of compressed files within

Use the menu in the **lower left corner** of the CVRS system

click on **"maintenance"**

·    then click on**DISK FILE**

Should bring up the “criteria page” for making a DISKFILE:



Be as all inclusive as possible. Include all districts, all parties, all statuses, include election history, phones, birth dates. Often takes more than one try because the selection approach with high lighting and "x's" mixed is a bit confusing, and no one gets enough practice. Often highlighting options excludes others so select only as much as needed to get the “full town, all voters all statuses”.

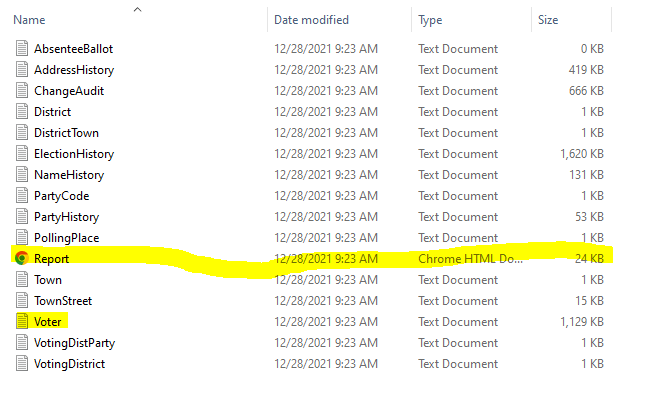
**Submit the request**

In your report queue you get back a few minutes later **a single "ZIP file" that contains several text files.** The **“VOTER”** should be large – about as many lines as all voters when opened in Notepad or EXCEL for your town.

Store the DISKFILE is a directory/folder for the current effort by town and date. Then, from within Windows Explorer, right click on the “zipped” files and “Extract all”. Often the default suggested location is adequate.

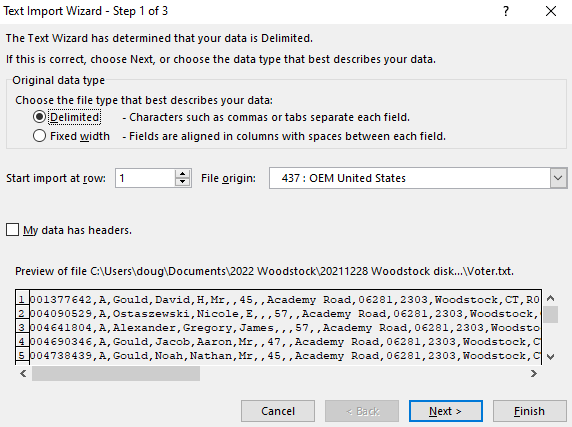
Within the extracted uncompressed files you will see many “relational database” files. All the criteria you selected, and the files extracted are described in REPORT.HTML which opens in your browser. And lays out all the files, their fields, and the criteria used for this DISKFILE.

You can grab the names of the fields for any file within DISKFILE from this REPORT.HTML. Until the data changes shape, you can reuse such a list of field names – or make a new one as described in another Appendix.

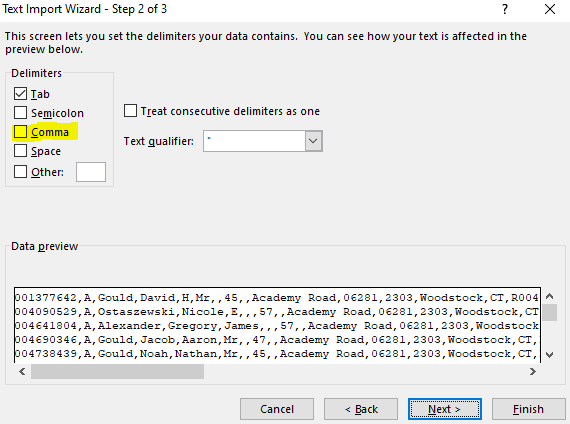


## Importing the VOTER “text” data file into EXCEL

From within EXCEL, go to the FILE tab, then “OPEN”, navigate to your files within the DISKFILE and select the VOTER file:

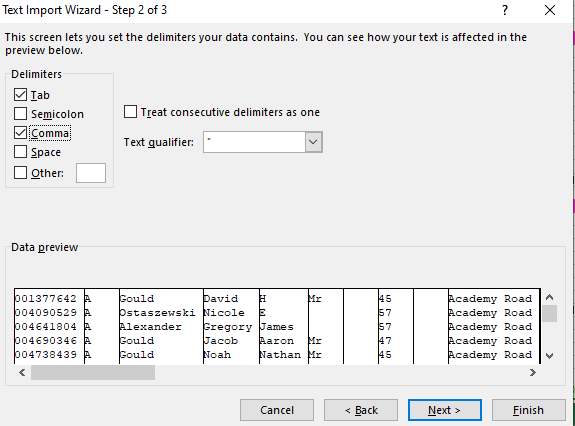


The DISKFILE is arranged as a comma separated values (CSV) file, delimited by commas, so select radio button “Delimited” then “Next”.

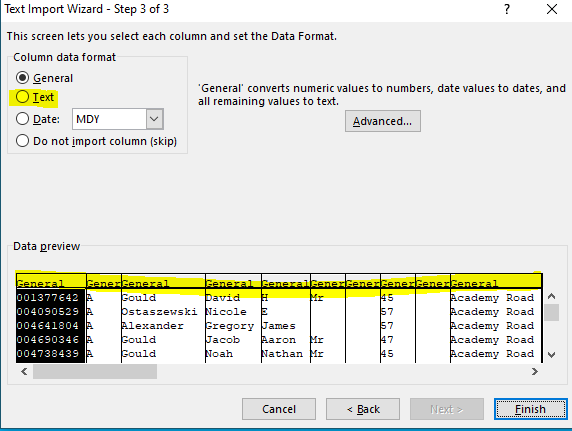


Then click in the “Comma” option box,

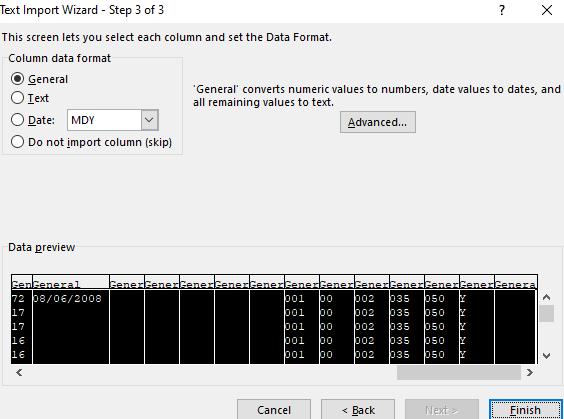
Should look like this:



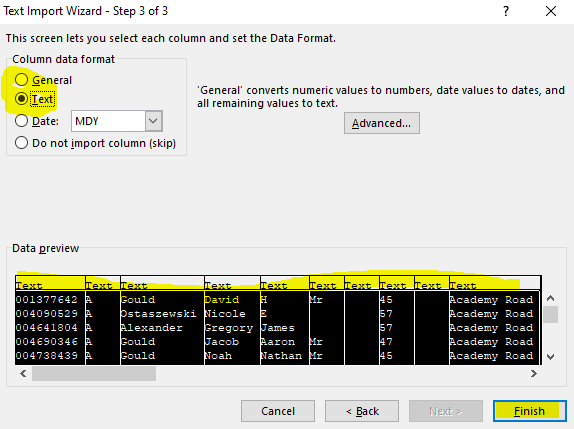
then click “Next”, which then brings up this box:



We’ll want to “select” all the columns, left to right, by dragging the slide bar to far right, pressing SHIFT while clicking ofn the “end of range” right most column to select all columns. Should look like:



Once all columns are highlighted, click in the radio button “text” to change the field type from the default “General” to type “Text”. Should then look like this:



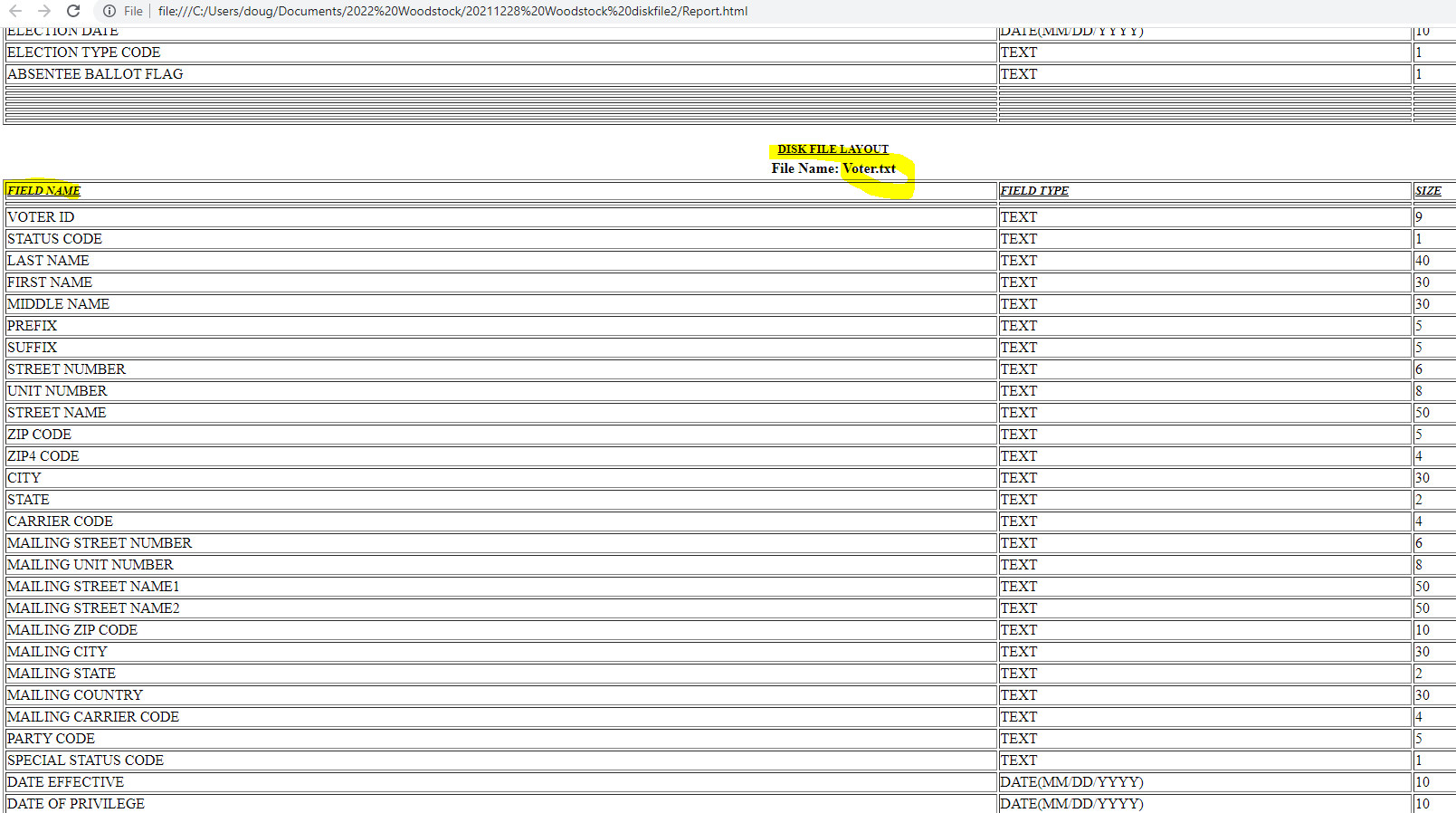
Once all the columns are type “text”, click on “Finish” to process the import into EXCEL.

This is especially important for datasets which have not yet got the exact same number of columns (fields) across for every row.

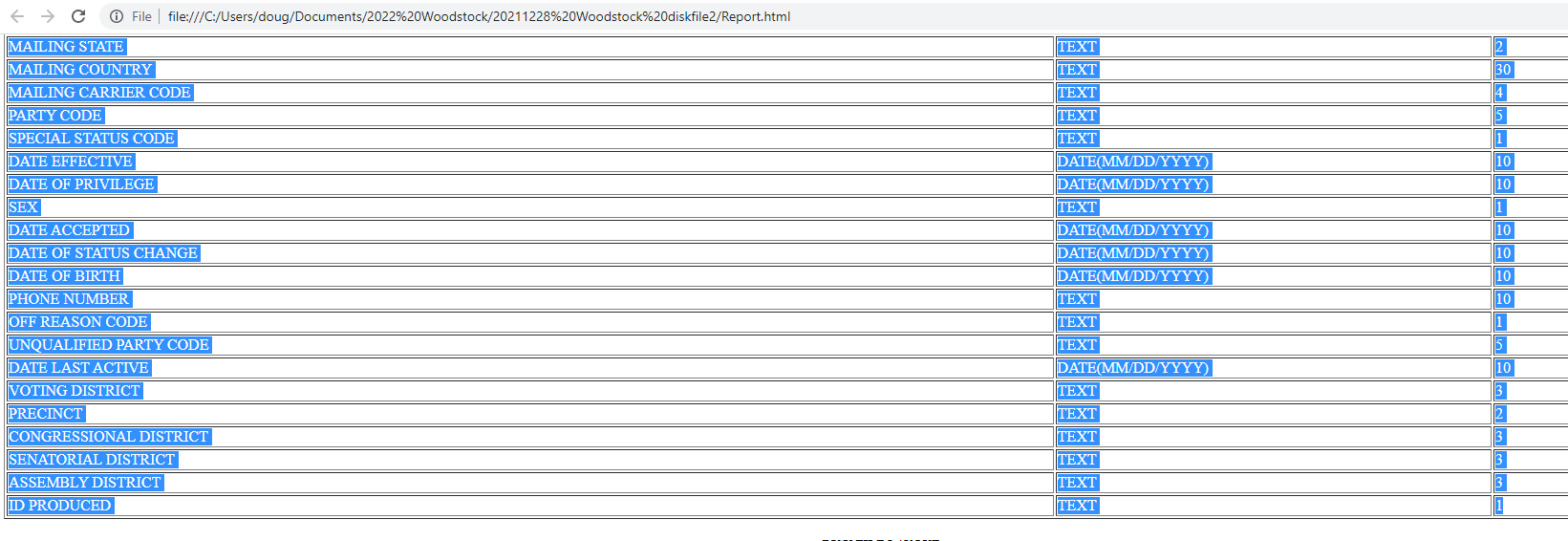
# Appendix B - How to create column headers (names of fields) from the REPORT.HTML file within DISKFILE

The REPORT file within DISKFILE can be opened (by double clicking or right click, “open”) in your browser – just like a web site. All the criteria you selected, and the files extracted are described in REPORT.HTML which opens in your browser. And lays out all the files, their fields, and the criteria used for this DISKFILE.

You can grab the names of the fields for any file within DISKFILE from this REPORT.HTML. Here we’ll look at how to do this for the most commonly used file, “VOTER”.

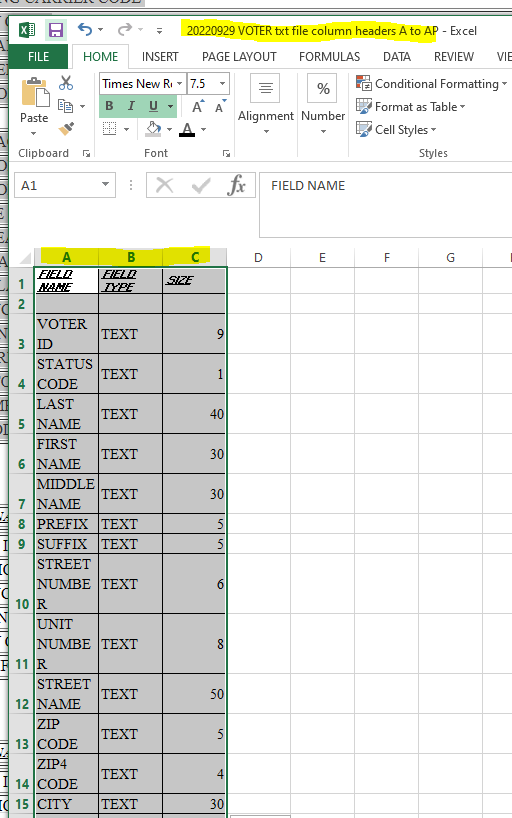


To “copy” the information about the VOTER file, click first on the “Field Name” and scroll to the end of the file description (likely another full page down) and press SHIFT *and then also* click on the size of the last (“ID produced”) field. This should select all three columns for every field/row of the file description.



Then use CTRL+C or right click and “copy” to get this information in your buffer.

Open a **new blank** spreadsheet or a new tab in an existing spreadsheet. Go to cell A1 and then “paste” (CTRL+V):



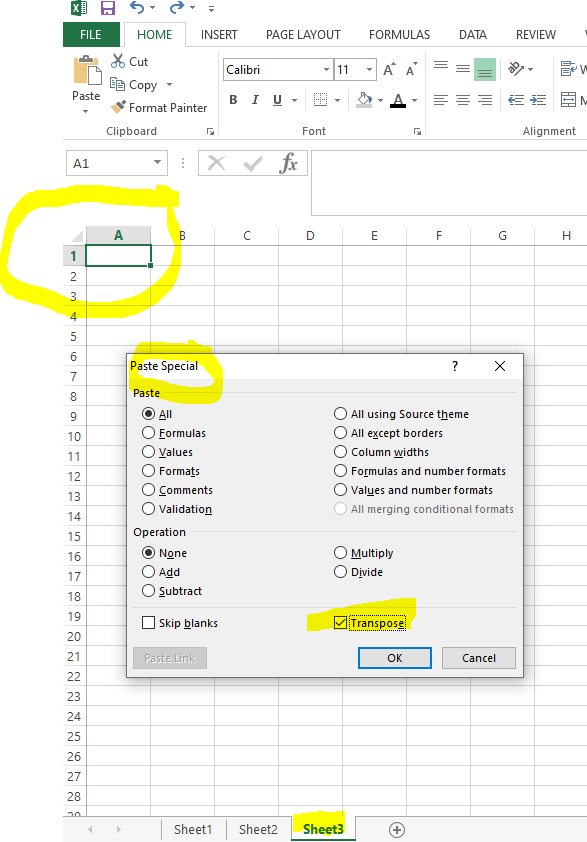
Notice that while we have all the field names now, we also have the field type (text, date, etc) and field size in the database.

While that is nice to know… we can clean this up a bit to get “just” the column headings, and make them list left to right like the VOTER file shows the fields.

Start by highlighting all three columns. Then CTRL+C (or any other method to indicate “copy”).

Open another blank empty tab in this or another a spreadsheet.

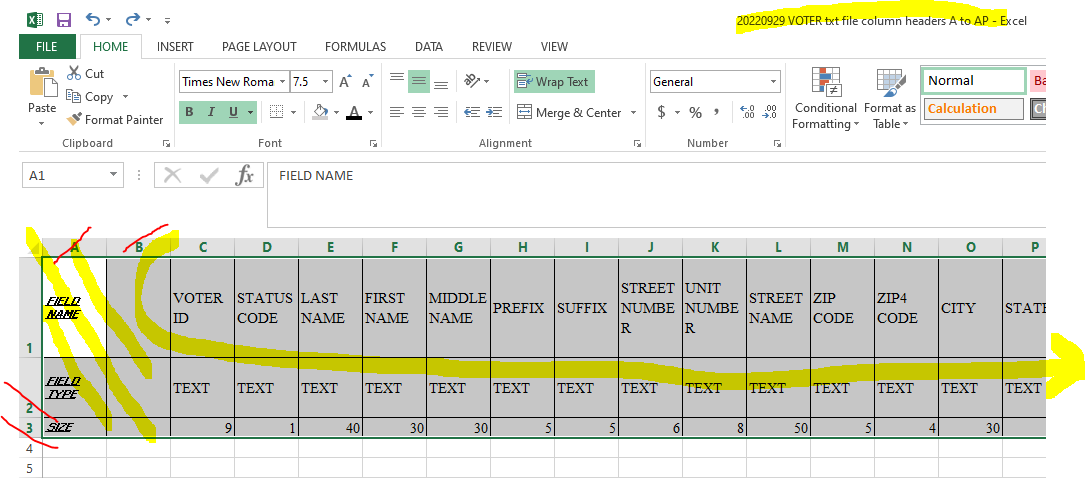
Then we’ll paste the column headings “transposed” (left to right instead of top to bottom).



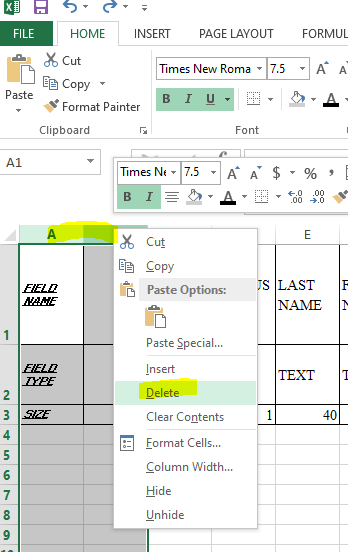
Select cell A1. Right click, select “Paste Special” from the menu.

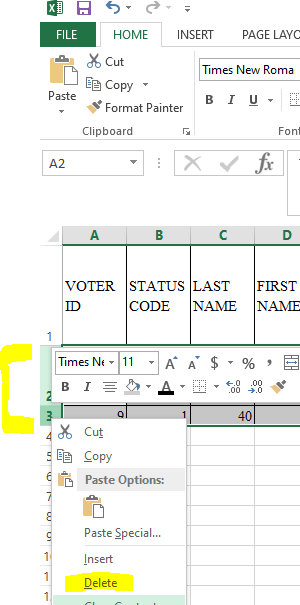
Click “Transpose”,

Then enter/OK.



Notice that we have some extra columns (A&B) that we don’t need, and two rows (“2 & 3”) we don’t need. Delete these.

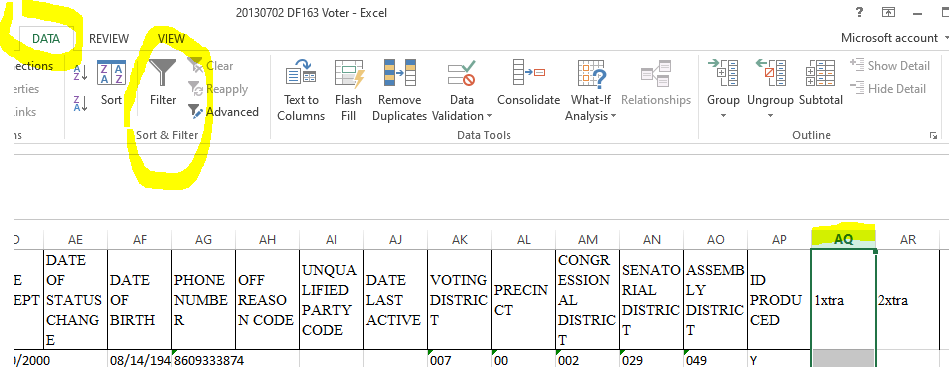


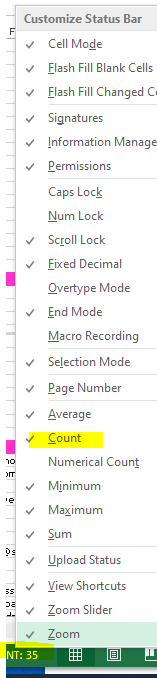


# Appendix C - How to confirm the data is stored in proper columns within EXCEL (no extra commas or other delimiters have made more columns for some rows than others)

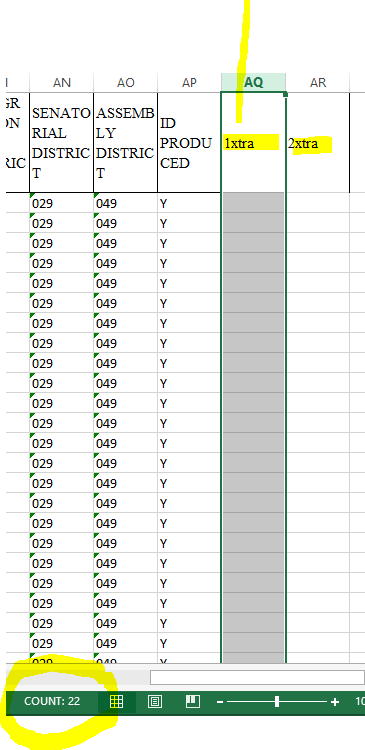
Usually the last column with data is “AP” – the “ID produced” column/field. But is some earlier fields have commas imbedded in a name, address, title or even a typos, this can cause the deliminted import techniques to create extra columns fo9r some rows which makes the data “not line up” in the same column for every voter.

We can check for this frustrating condition by adding some “xtra” columns to the right of column “AP”. Give them names.



To check if any data was pushed into the columns past “AP” (field “ID produced” ), make sure you have statistics turned on by hovering in the lower right of the border of the spreadsheet, right click, to control items shown in the bottom edge “status bar”.

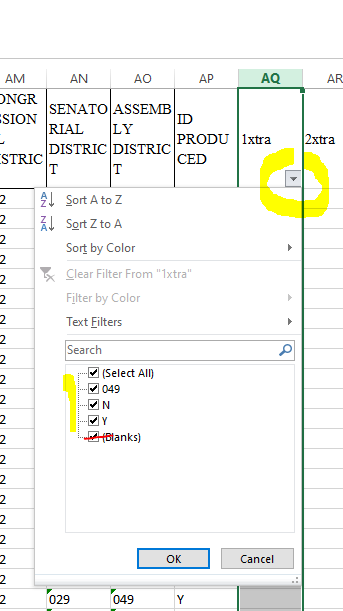
Then if you select a column (here “AQ”) and there are two or more data in the colum, a “count” of how many not empty cell there are will be shown. We don’r want any… but is a messy DISKFILE has so this will reveal it:



If you find a DISKFILE with such “extra columns” the cause is nearly certainly that some rows contain fields with commas in the data. When imported to EXCEL these commas injected another colum instead of being stored as data with in a field.

To identify which rows have such extra fields, use the DATA tab on the ribbon menu, then the FILTER tool while the column (AQ) is selected.

Here’s an example:



Unselect all the “blanks” rows, then click OK to reveal which rows have fields with extra commas imbedded in them.

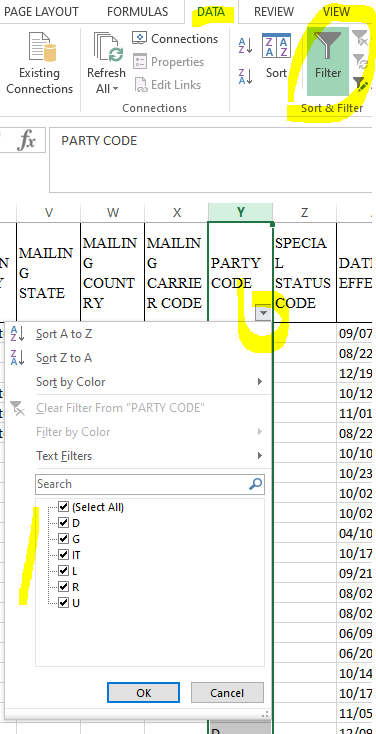
Generally the first field that is shifted to right is the second half of a field value that contained a comma. Go back to CVRS and check that field (usually a name or address field, but sometimes typos show up in prefix and suffix fields.

Once all rows have the same number of columns after a DISKFILE export, you can proceed to analyize the data for undesirable values in each field. Until then the “extra commas” causing undesired fields during import will distract you from the standardization work. **Solve extra commas first**.

# Appendix D - How to use the Filter options and the filtered search from the DATA tab/menu

Often we want to look at a subset of the data. The EXCEL ribbon menu’s DATA tab has a filter function iuseful for this purpose.

Select a column. Navigate to the DATA tab of the ribbon menu, then click on “Filter” Here’s I’ve shown the party code field for a town with only the most common parties:



Its fairly easy tool to use for one column at a time, but it is possible to filter on several columns at once.

The filter interface fist shows you several of the values found in the column, and lets you select or deselect them.

For more aggressive filtering, type a value in the search box.

Once you have items you want to see more of shown in this filter interface, click OK to return to the spreadsheet with only these rows showing.

To return to all rows showing, de-select the highlighted “FILTER” on the Data tab of the ribbon menu.