NLM-Scrubber User Manual

Windows Version 19.0411W

Natasha Raja
Pamela Sagan, RN
Justin Jones, MS
Ming Way, MS
Mehmet Kayaalp, MD, PhD

May 3, 2019

Lister Hill National Center for Biomedical Communications
U.S. National Library of Medicine
8600 Rockville Pike, Bethesda, Maryland 20894

Table of Contents

A. How to Run NLM-Scrubber

1. Unzip the downloaded file
2. Double click on the application file: **scrubber.19.0411W.exe**
3. The application should open the graphical user interface window called **NLM-Scrubber Execution Panel**. Please note at the top of the window the version of NLM-Scrubber, which is v.19.0411W for the example in Figure 1.

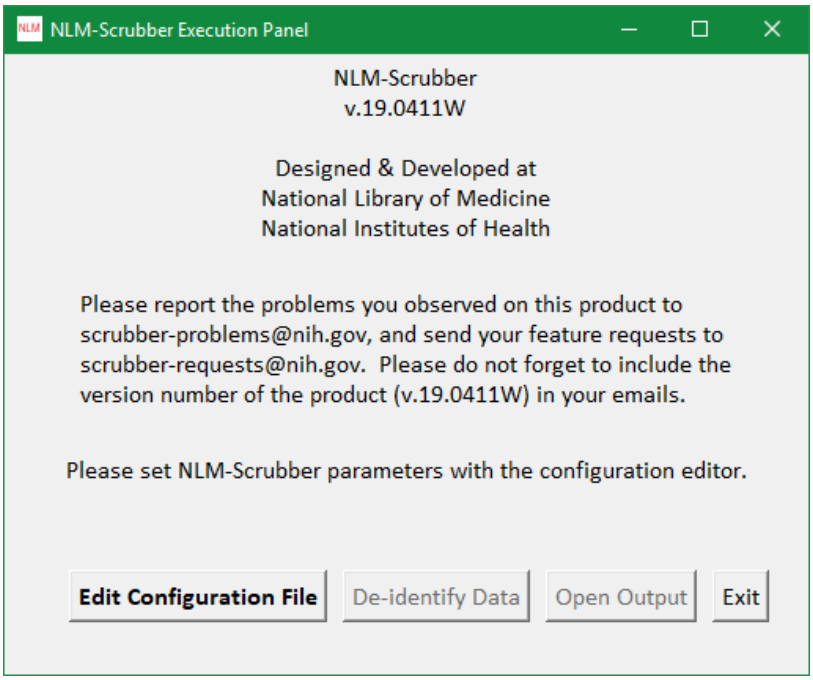

Figure 1 NLM-Scrubber Execution Panel

4. To run NLM-Scrubber, you need to have a configuration file. Pressing on the **Edit Configuration File** button would bring you the NLM-Scrubber **Configuration Editor** window.

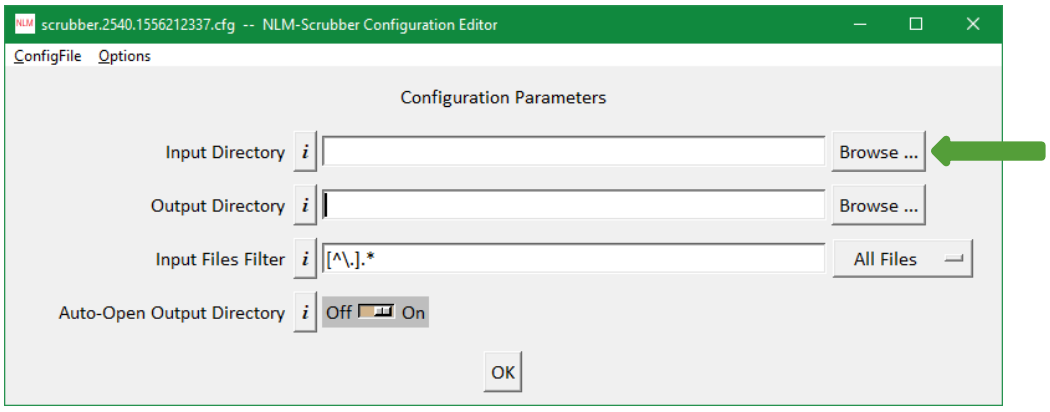

Figure 2 NLM-Scrubber Configuration Editor

To select the files you would like NLM-Scrubber to de-identify, click on the **Browse** button for the **Input Directory**. This will open another window labelled **Browse for Folder**. Once you have located your desired input file, click on it and press the **OK** button at the bottom of the window.

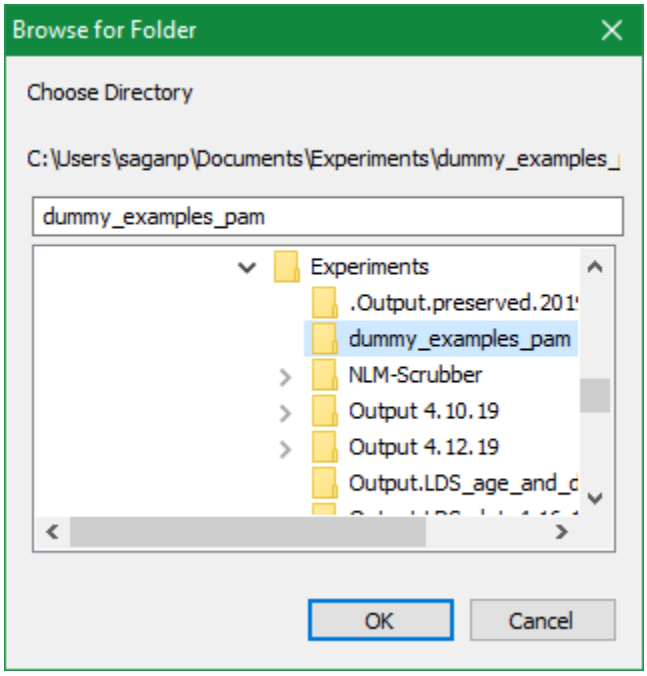

Figure 3 Browse for Folder Window

This will take you back to the **NLM-Scrubber Configuration Editor**. You will now see the input files location as the **Input Directory** (see Figure 4).

5. To select the location in which Scrubber will place the de-identified files, click on the **Browse** button for the **Output Directory**.

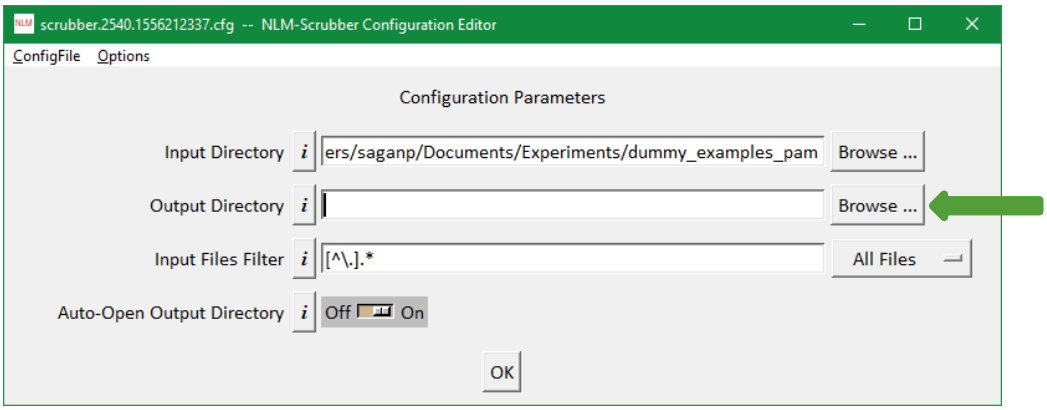

Figure 4 Selecting the Output Directory

This will open another window labelled **Browse for Folder** for you to select the directory visually.

Note, you can also change the desired directory's location both Input and Output by manually typing new text in the corresponding text box. Once both Input and Output files have been selected, press the **OK** button at the bottom of the window or move to the optional steps described below.

6. Users can customize input files type (1) by selecting the **Input Files Filter** by pressing the **All Files** button and choosing a preset filters, or (2) by editing the regular expression in the text box (e.g., ^ds.*\.txt filters in files with prefix ds and suffix .txt).

   Users can choose how the Output Directory is opened by moving the **Auto-Open Output Directory** switch to either "On" or "Off". The default for Windows operating system is "On" which automatically opens the Output file directory after it has finished de-identifying the data; whereas, the default is "Off" for Linux systems (per users' requests).

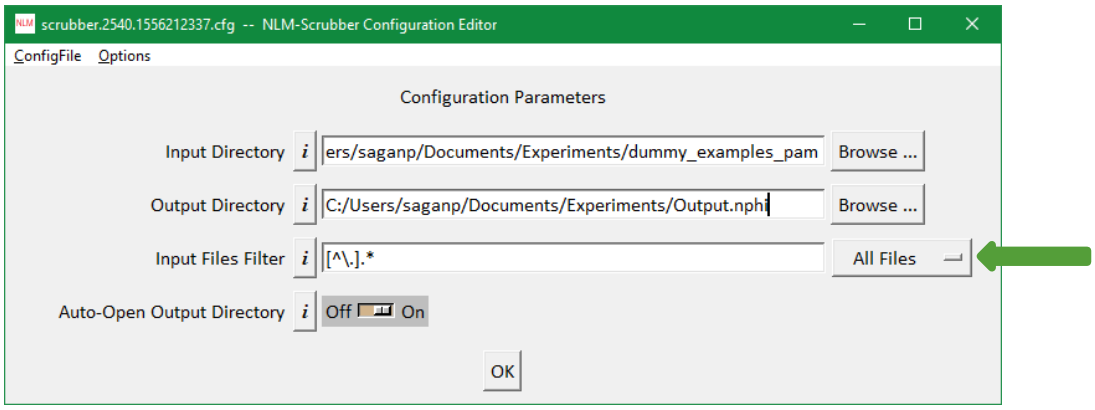

Figure 5 Filtering Input Files

7. Pressing the **OK** button on the configuration editor would close the editor and take you back to the **NLM-Scrubber Execution Panel** window (see Figure 1). To begin the de-identification process, press the **De-identify Data** button. As soon as NLM-Scrubber begins running, the **NLM-Scrubber Execution Panel** will inform you that the libraries are loaded (see Figure 6.a). This may take a couple of minutes depending on your machine's CPU, type of the secondary storage and the bus speed between them.

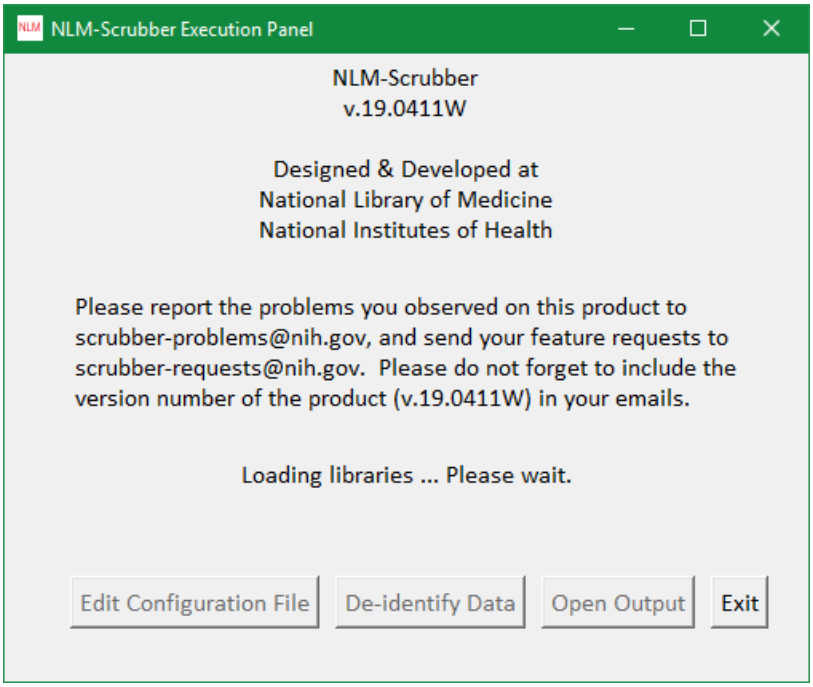

Figure 6 (a) NLM Scrubber Initialization

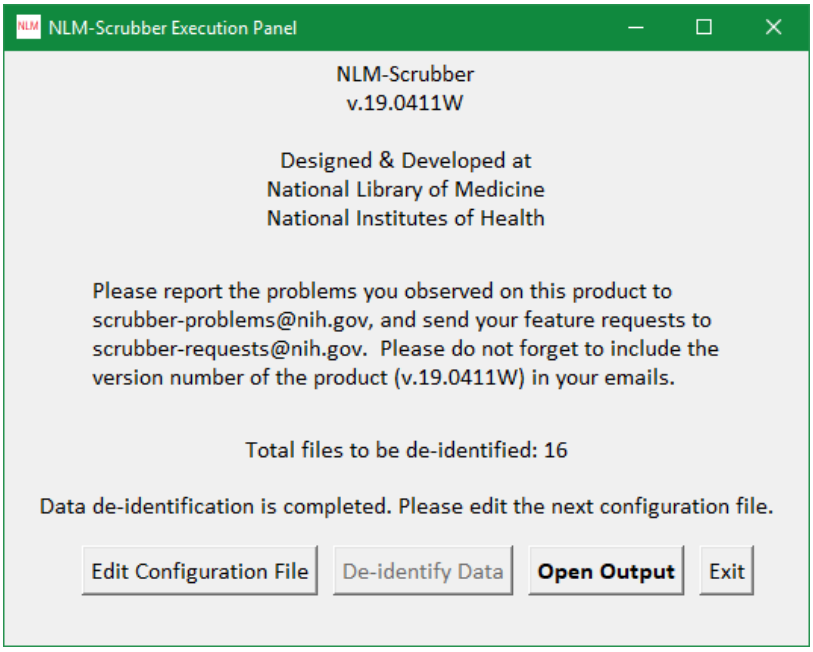

Figure 6 (b) Completion of De-identification

8. If the **Auto-Open Directory** was "On", once NLM-Scrubber has completed de-identifying the files, a new window would pop up, showing the de-identified files in the output directory. If you selected the **Auto-Open Directory** "Off", you can manually open this window by pressing the **Open Output** button on **NLM-Scrubber Execution Panel** (see Figure 6.b).

9. Once you completed de-identifying first batch of files, you can either start with your second set by pressing **Edit Configuration File** or exit the application by selecting the **Exit** button (see Figure 6.b).

B. De-Identification Options

Users can select **Preserve Terms File** (preserving scientific information from a created "whitelist"), select **PII Terms File** (redacting certain words ensuring the desired de-identification from a created "blacklist") and also create **Limited Data Sets**. Test files referenced in this section can be found under Examples.1904/Dummy_input/ and Examples.1904/Dummy_list/ in Examples.1904.zip on the NLM-Scrubber website.

1. Preserved Terms File Option

Using the example in Figure 7 below (test file "nci_test_003"), the user would like to preserve the text "**Betty Ourisman Foundation"** (see Figure 7.a) that Scrubber de-identifies as [PERSONALNAME] (see Figure 7.b).

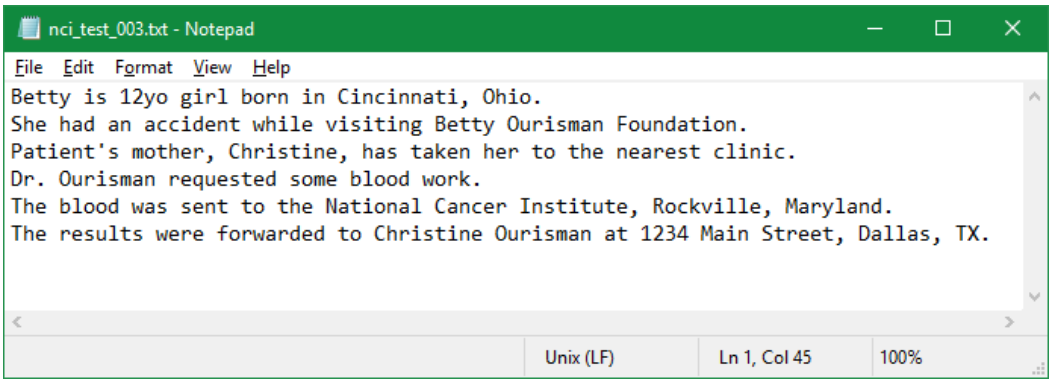

Figure 7(a) Example of test file for Preserved and PII options

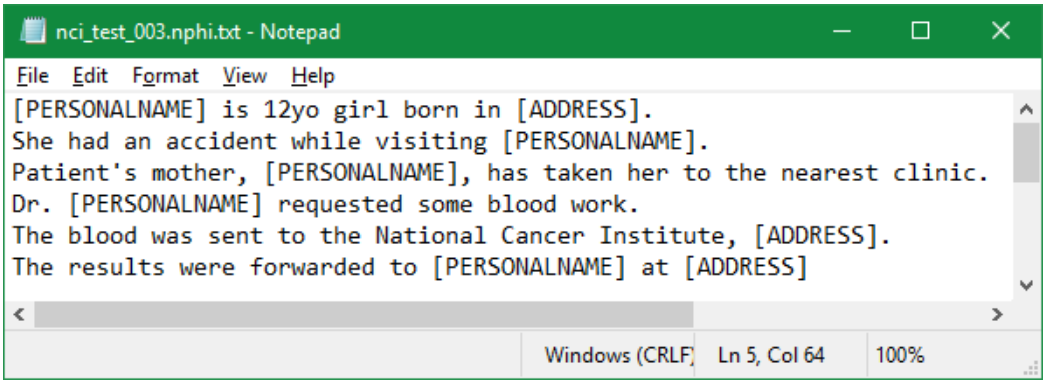

Figure 7(b) Scrubber's de-identification of test file above

a. After selecting the desired input and output directories as described in Part A, select the **Options** button.

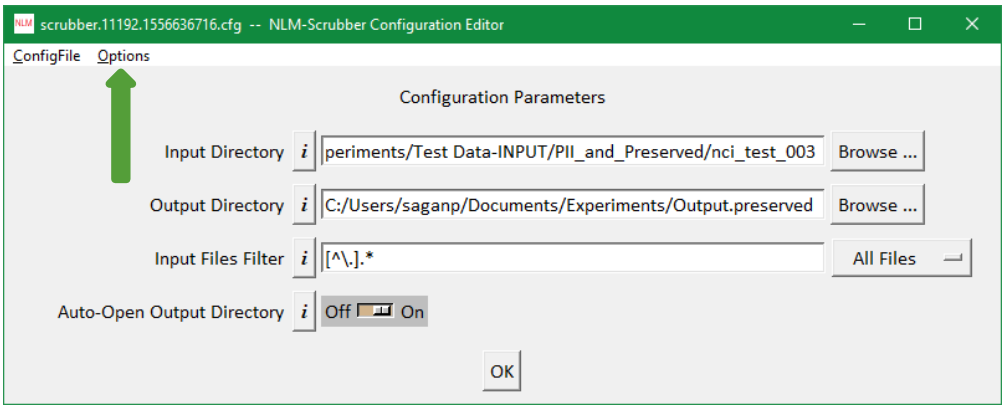

Figure 8 NLM-Scrubber Configuration Editor Options location

b. To select the option of **Preserved Terms File** from the drop down menu click on it. See the checkmark appear once the desired option is selected and the **Preserved Terms File** option now appears under the Configuration Parameters on the **Configuration Editor** window.

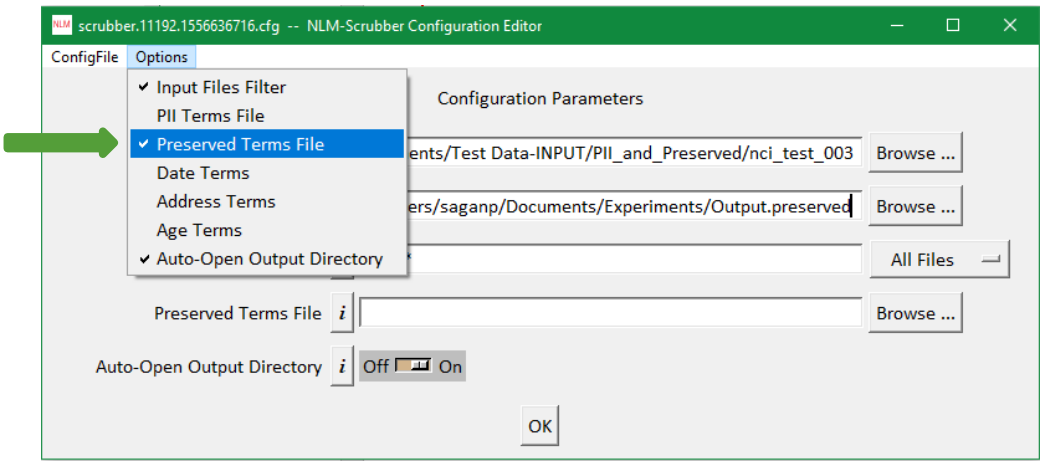

Figure 9 Preserved Terms File option selected from menu

c. To select the file whose terms you would like Scrubber to preserve, click on the **Browse** button for the **Preserved Terms File**.

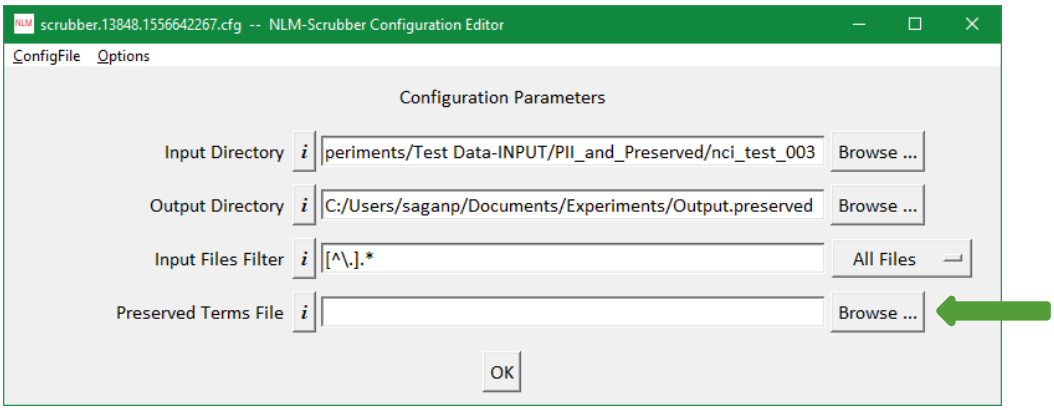

Figure 10 Browse for Preserved Terms File

d. This will open another window similar to Figure 3. Once you have located your desired file, double click to select it. Note that the contents of the **Preserved Terms File** are not case-sensitive.

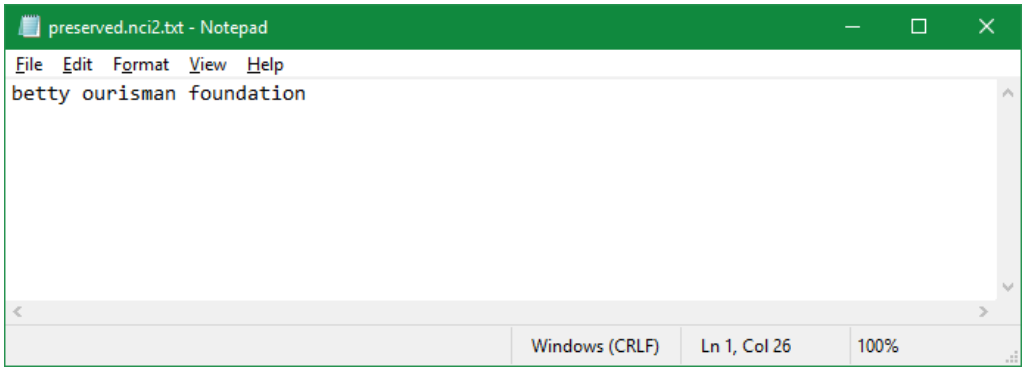

Figure 11 Contents of Preserved Terms File

e. Once you have selected your desired **Preserved Terms File** (test file list "preserved.nci2") you are taken back to the **Configuration Editor** screen. Press the **OK** button at the bottom of the window.

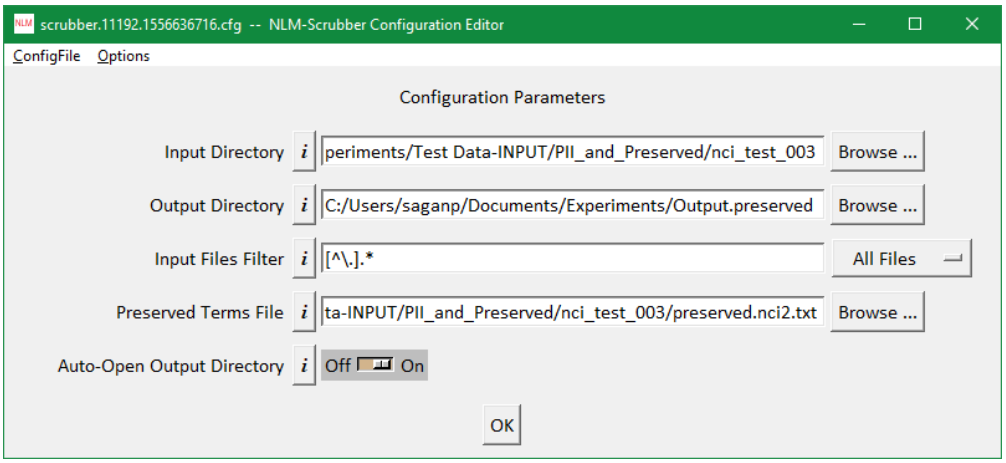

Figure 12 NLM-Scrubber Configuration Editor with Preserved Terms File Option

f. Pressing the **OK** button on the configuration editor closes the editor and takes you back to the **NLM-Scrubber Execution Panel** window. To begin the de-identification process with the **Preserved Terms File** option, press the **De-identify Data** button. The process is the same as Figures 6(a) and 6 (b) with the NLM-Scrubber initialization and completion of de-identification.

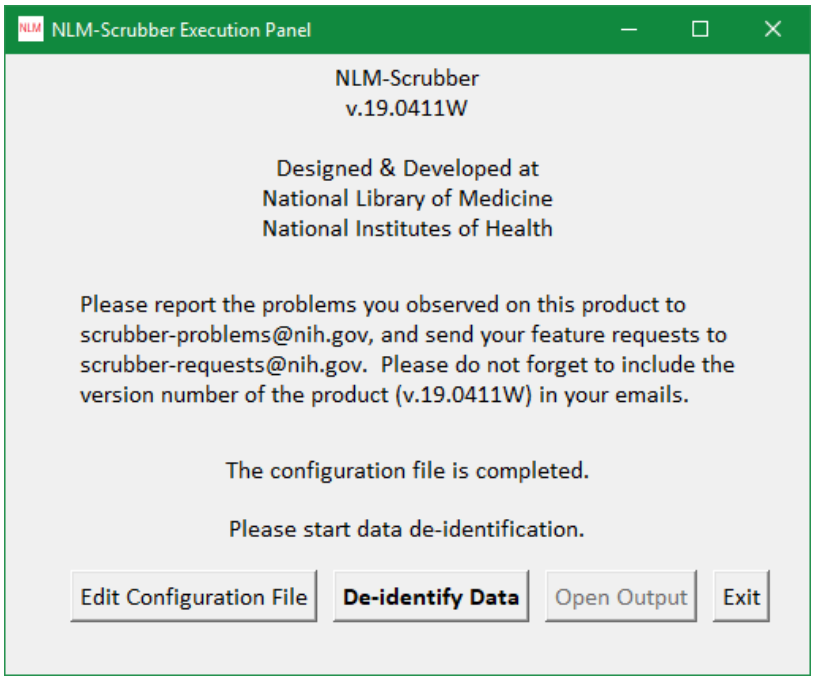

Figure 13 NLM-Scrubber Execution Panel Configuration file completed

g. If the **Auto-Open Output Directory** was "On", once NLM-Scrubber has completed de-identifying the files, a new window would pop up, showing the content of the output directory window showing the de-identified **Preserved Terms File**. Click on the output file to view the de-identified file. The text "**Betty Ourisman Foundation**" has now been preserved in the file and not de-identified by Scrubber.

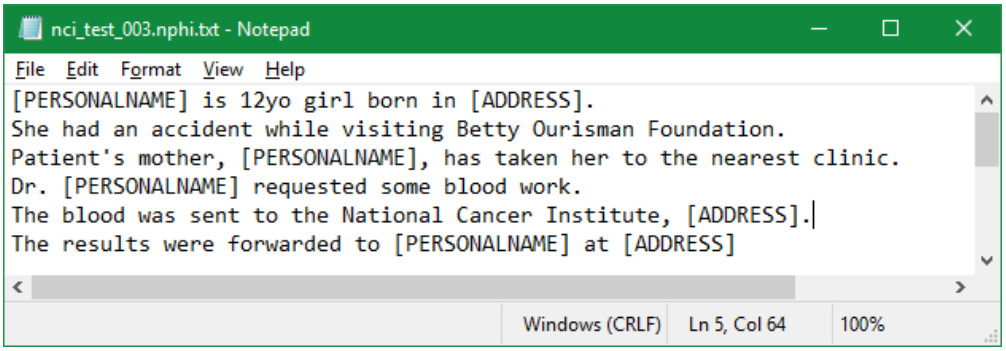

Figure 14 De-identified file with Preserved Terms preserved

h. The configuration file for the **Preserved Terms File** option.

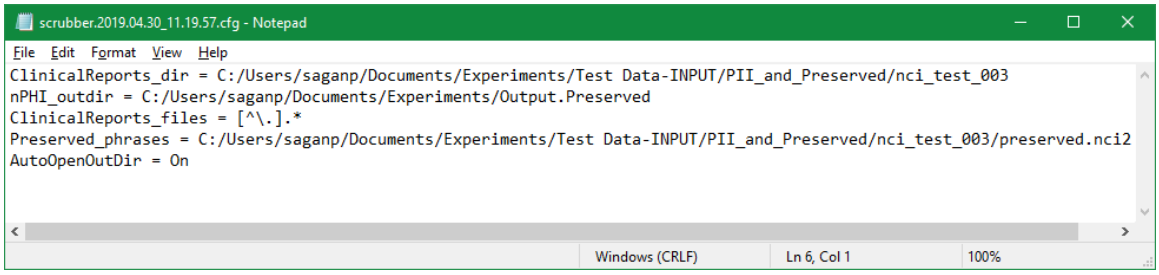

Figure 15 Configuration File for Preserved Terms Option

2. PII Terms File Option

Using the examples in Figure 7 (test file "nci_test_003"), the user would now like to redact "**National Cancer Institute"** which would not by de-identified by Scrubber.

a. After selecting the desired input and output directories as described in Part A, select the **Options** button.

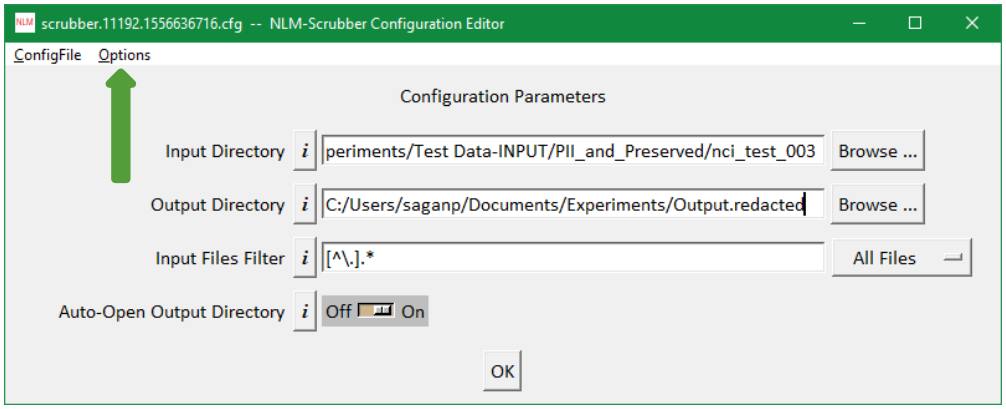

Figure 16 NLM-Scrubber Configuration Editor Options location

b. To select the option of **PII Terms File**, from the drop down menu click on it. See the checkmark appear once the desired option is selected and the **PII Terms File** option now appears under the Configuration Parameters on the **Configuration Editor** screen.

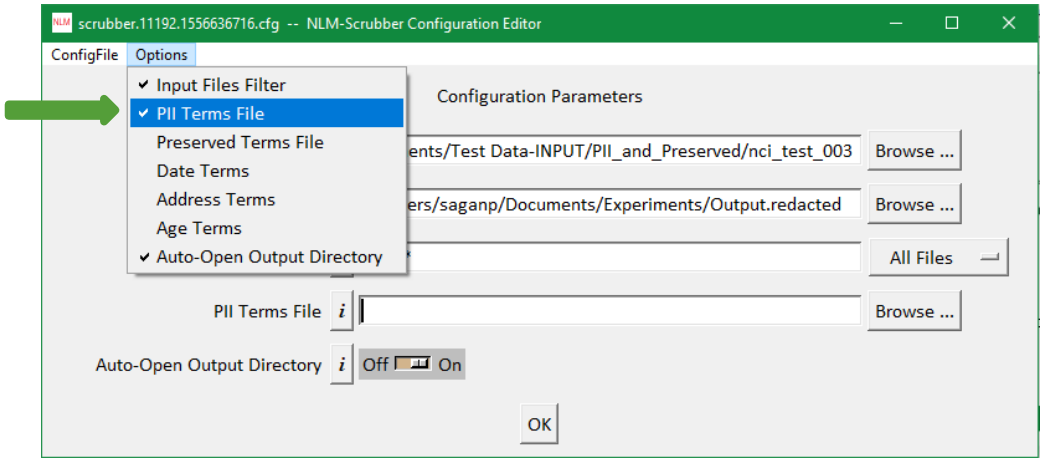

Figure 17 PII Terms File option selected from menu

c. To select the file whose terms you would like NLM-Scrubber to redact, click on the **Browse** button for the **PII Terms File**.

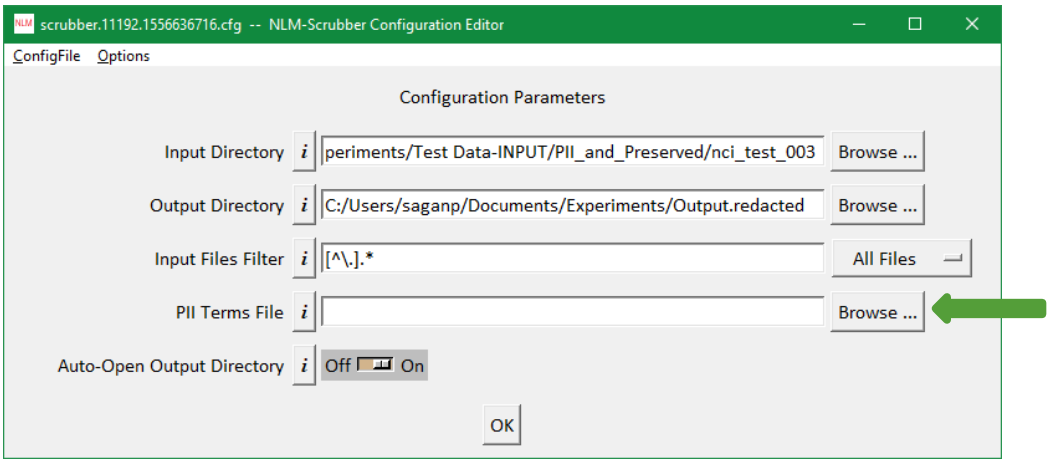

Figure 18 Browse for PII Terms File

d. This will open another window similar to Figure 3. Once you have located your desired file, double click to select it. Note that the contents of the **PII Terms File** are not case-sensitive.

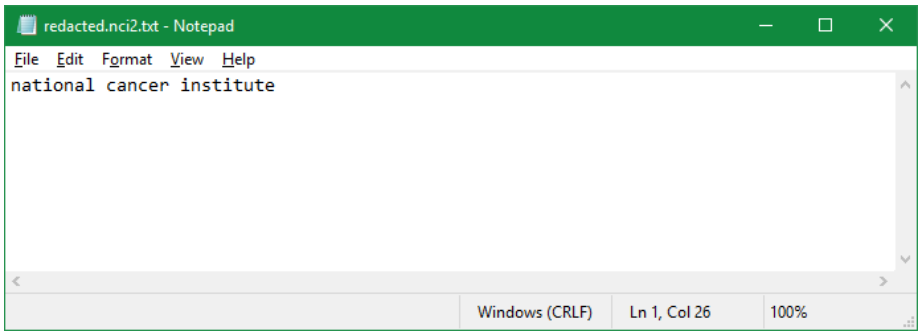

Figure 19 Contents of PII Terms File

e. Once you have selected your desired **PII Terms file** (test file list "redacted.nci2") you are taken back to the **Configuration Editor** window. Press **OK** at the bottom of the window.

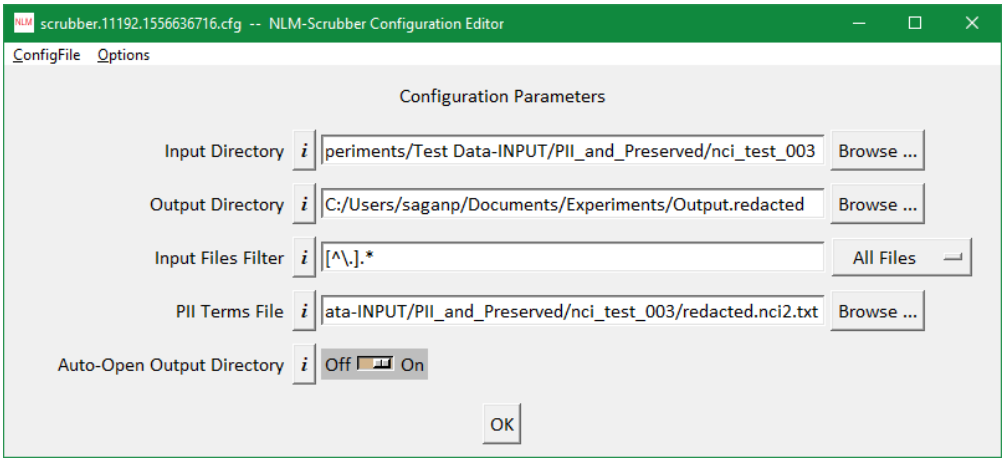

Figure 20 NLM-Scrubber Configuration Editor with PII Terms File Option

f. Pressing the **OK** button on the **Configuration Editor** closes the editor and takes you back to the **NLM-Scrubber Execution Panel** window. Follow the same steps for **Preserved File Terms** for NLM-Scrubber initialization and completion of de-identification (Figure 6.a and 6.b).

g. If the **Auto-Open Output Directory** was "On", once NLM-Scrubber has completed de-identifying the files, a new window would pop up, showing the content of the output directory window showing the de-identified files. Click on the output file to view the de-identified file. The text "**National Cancer Institute**" has now been redacted from the file and replaced with [PII].

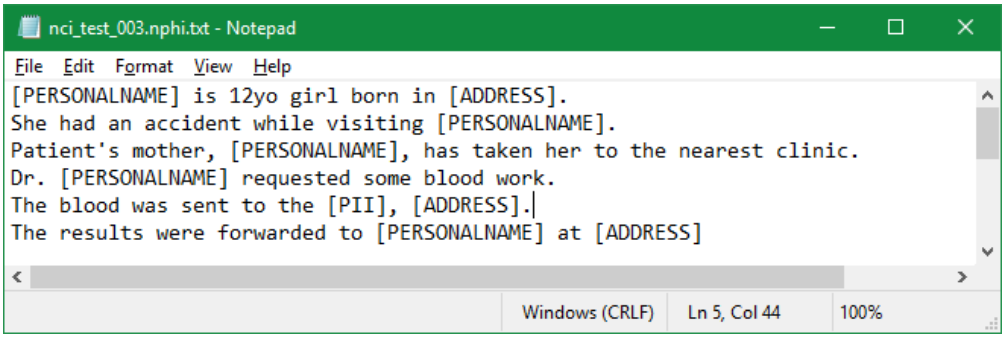

Figure 21 De-identified file with PII Terms redacted

h. The configuration file for the **PII Terms File** option.

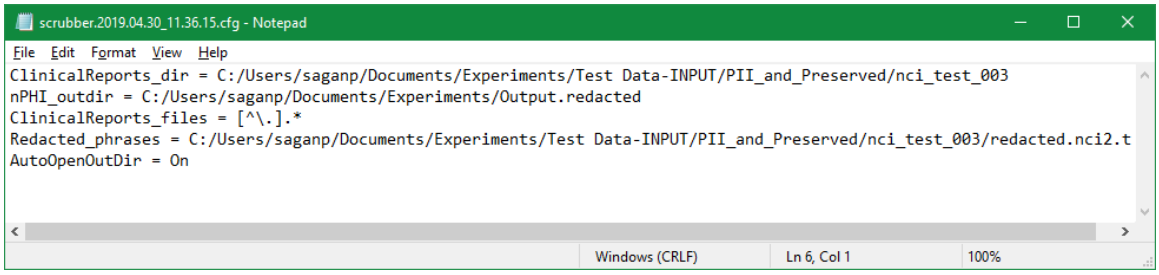

Figure 22 Configuration File for PII Terms Option

3. Limited Data Sets

Limited Data Sets allow users to keep certain identifiers such as **Full Date**, **Address** and **Age (for ages above 89)** in the files for their research.

a. After selecting the desired input (test file “dummy_date_2”) and output directories as described in Part A, select the **Options** button and click on the desired options(s). See the checkmark appear once the desired option(s) is selected and those options now appear under the Configuration Parameters on the **Configuration Editor** screen

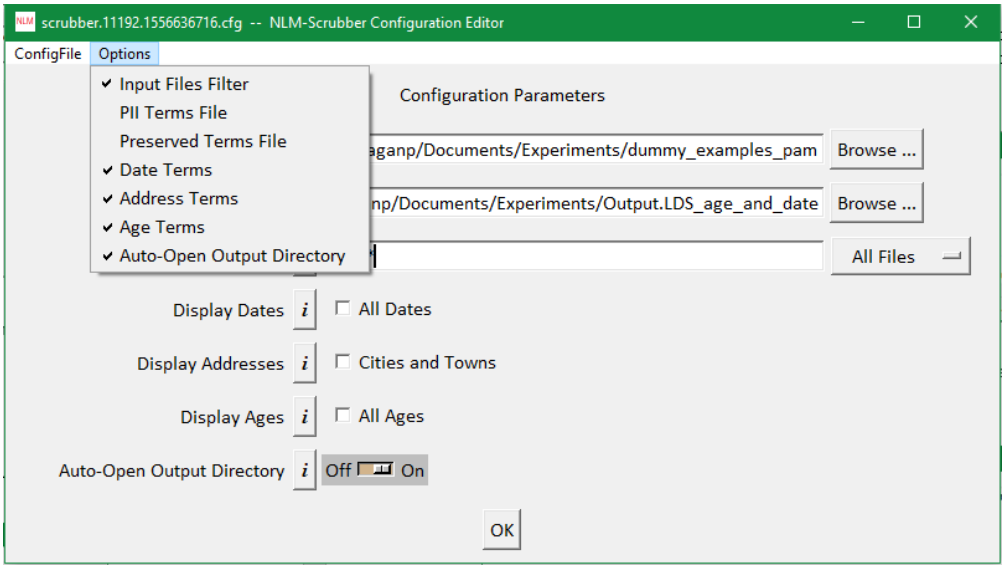

Figure 23 Limited Data Set options selected from menu

b. Once the desired options are selected, they will appear on the **NLM-Scrubber Configuration Editor**. Check off the desired LDS terms and select the **OK** button.

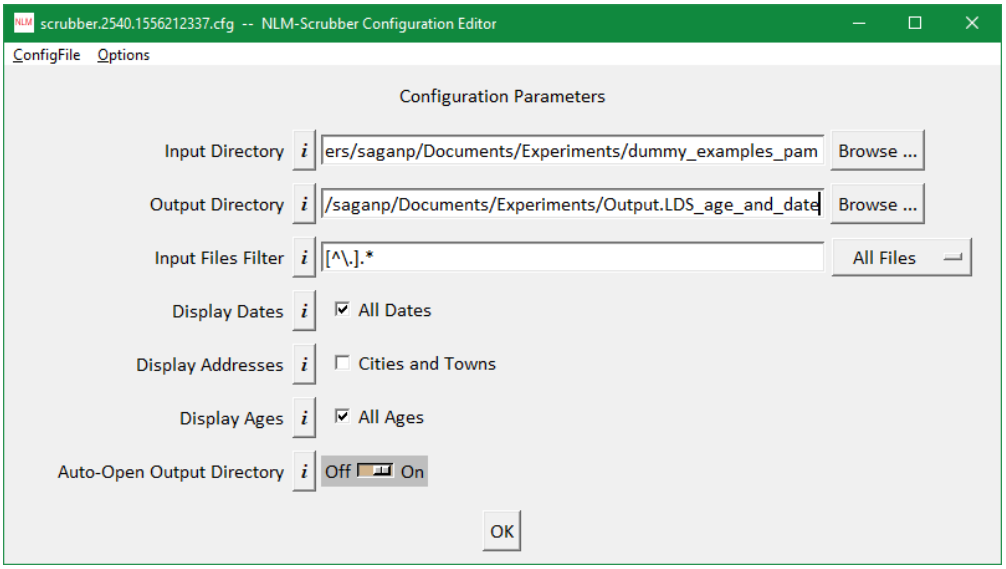

Figure 24 Limited Data Set options checked off under Configuration Parameters

c. Pressing the **OK** button on the **Configuration Editor** would close the editor and take you back to the **NLM-Scrubber Execution Panel** window. Follow the steps for NLM-Scrubber initialization and completion of de-identification (Figure 6).

d. If the **Auto-Open Output Directory** was ON, once NLM-Scrubber has completed de-identifying the files, a new window would pop up, showing the content of the output directory window showing the de-identified LDS files. Click on the output file to view the de-identified file. Note that the desired terms (age and date) have been preserved in Figure 25 below.

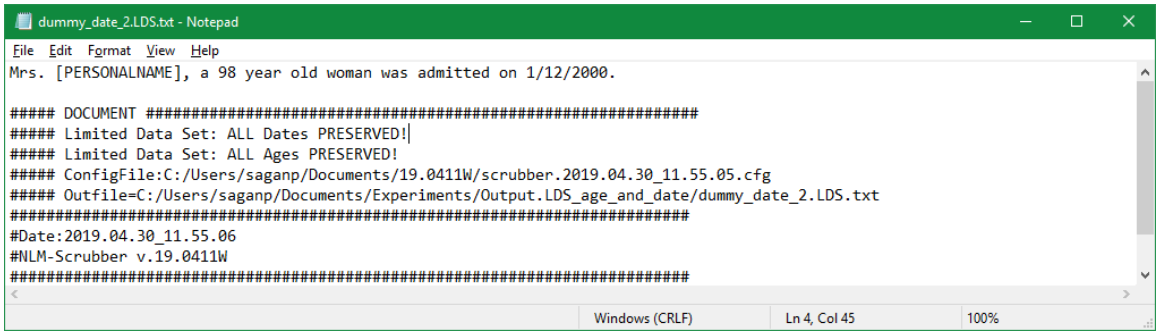

Figure 25 Contents of LDS age and date file

e. The configuration file for the **Limited Data Set.**

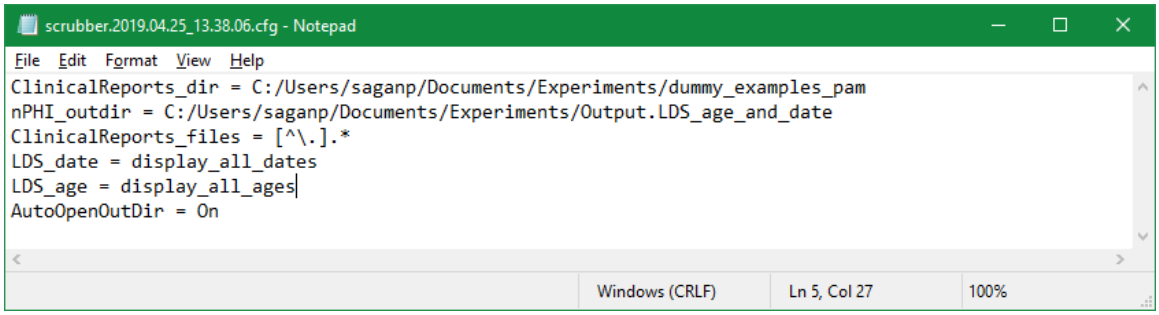

Figure 26 Configuration File for Limited Data Set above