Anonymizer QuickStart Guide

The Anonymizer is a tool designed for use in anonymizing structured data-sets so that data values remain consistent after anonymization. It can be configured for use for almost any structured data-set, although it is not designed to work directly with more complicated formats, such as OpenOffice or Microsoft Excel. Fortunately, these tools allow data to be saved out in formats that are compatible with the Anonymizer.

This tool was originally designed to allow an organization conducting surveys to provide researchers with safely anonymized data. It is particularly useful when the data that would normally be removed is of interest to the researchers, such as names or locations when trying to examine network effects.

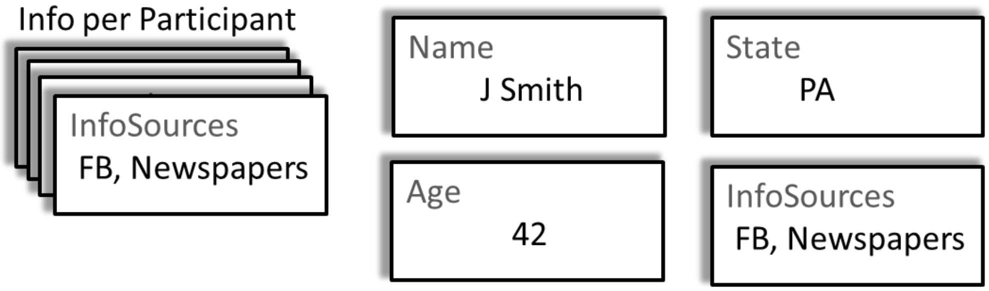
This document is intended to introduce functions of the Anonymizer and help other researchers configure it usefully for their own work.

Supported File Formats

The Anonymizer natively supports TSV (tab-separated values) data that has a header row, with each survey participant on their own row. Through configuration, any consistent data-separation value can be used, such as CSV (comma-separated values) if TSV is not available or practical. We prefer tab-separated values because most spreadsheet programs use tab to move across cells.

The header row is used to make sense of the data, and essentially the Anonymizer generates a set of note-cards for each participant.

Name	Age	State	InfoSources
J Smith	42	PA	FB,Newspapers



The data-column “InfoSources” also explains why we prefer tab-separated values. “FB” and “Newspapers” are two separate sources of information, but this context could be lost.

More information on how to configure the Anonymizer to read many types of files will be discussed in the section “Configuring the Anonymizer”.

MS Excel and Tab-Separated Value (TSV) Data

A Microsoft Excel spreadsheet can be converted to a tab-separated file which can be processed by the Anonymizer. The resulting file, which will also be a tab-separated file, can also be read by Microsoft Excel. This section will give the steps necessary to both output a tab-separated value, and to read a tab-separated value into Excel.

Converting Excel to TSV Data

1. Start Excel
2. Open your data-file
3. Click the "File" tab
4. Click "Save As", a dialog "Save As" will pop up
5. Near the bottom of the dialog, there will be a field called "Save As Type"
6. Select "Text (Tab Delimited) (*.txt)"
7. Click "Save"

Reading TSV Data in Excel

By default, Excel will automatically open CSV or Excel files properly, but text files will not be read natively by Excel (instead, the file will be opened by MS Word or another program), so follow these directions to open a text file or other file in Excel.

Method One: Use "Open With" from the File Window

1. Navigate to where your file is located on your file system
2. Right-Click on the File
3. Select "Open With >"
4. Select "Microsoft Excel"
5. The file should open correctly

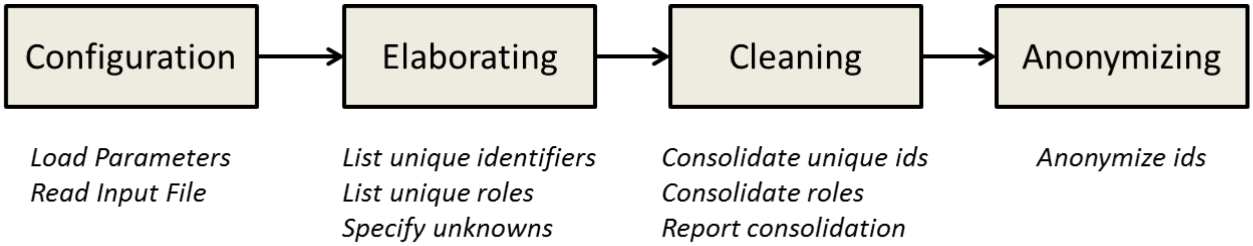
If the file fails to open correctly using Method One, switch to Method Two

Method Two: Opening the File in Excel

1. Start Excel
2. Navigate through directories to where your data-file should be
3. At the bottom of the "Open" dialog, there is a combo-box that indicates what files will be shown, change the filter from "All Excel Files" to "All Files", now your data-file should be visible!
4. Select your data-file
5. A new dialog, called "Text Import Wizard – Step 1 of 3" will come up. The choices are "Delimited" or "Fixed Width". Choose "Delimited", and click "Next"
6. Excel will scan your file and try to guess what the delimiters are – make sure only "Tab" is selected and click "Finish"

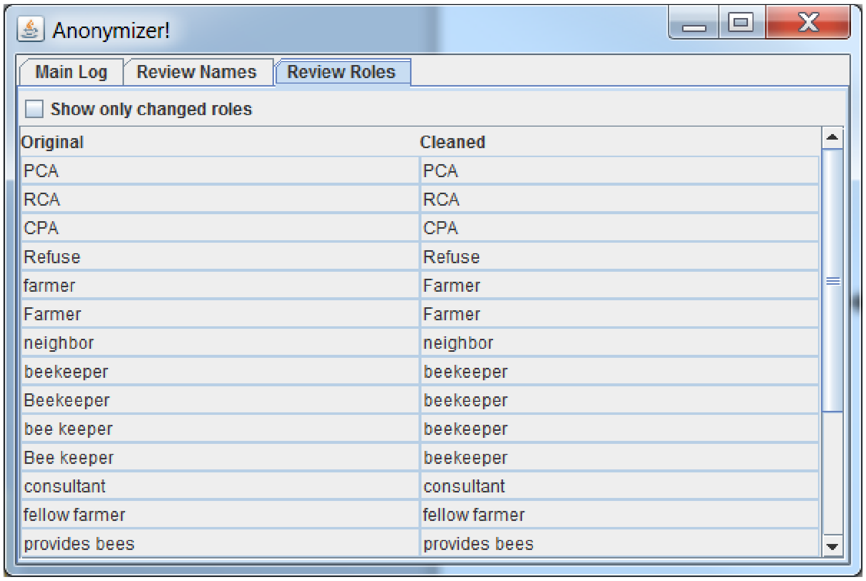
The Anonymizer GUI

The Anonymizer, after you specify an input file, will produce a simple GUI intended to provide you with information about the anonymization and cleaning work. The main window is a textual log, informing you, the User, of the state of the system. The Anonymizer works through four states.



These four states are configuration, elaborating, cleaning, and anonymizing. What happens at each state is listed in the graphic above in italics beneath its state.

All process results are reported to the main log, but we have added review panels for cleaning operations for names and roles, so that users can quickly evaluate the consolidation output and determine whether the configuration provided is appropriate to their needs. Are important unique organizations being lost in the cleaning process? Are things that are probably the same thing still listed as two separate things afterward? Both of these outcomes must be balanced, but we suggest in favor of preserving information for later manual cleaning than for more destructive consolidation. The appropriate settings will depend on your needs.



□ Show only changed roles

Original	Cleaned
PCA	PCA
RCA	RCA
CPA	CPA
Refuse	Refuse
farmer	Farmer
Farmer	Farmer
neighbor	neighbor
beekeeper	beekeeper
Beekeeper	beekeeper
bee keeper	beekeeper
Bee keeper	beekeeper
consultant	consultant
fellow farmer	fellow farmer
provides bees	provides bees

The review panel can be adjusted to show only values that have changed, or to show all entries. In the graphic above, all entries are shown. Note that beekeeper has been cleaned from 4 variations to a single entry “beekeeper”.

Configuring the Anonymizer

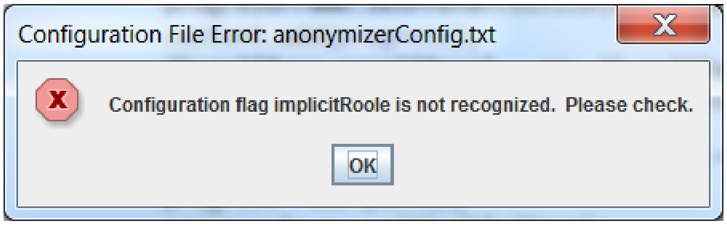
The Anonymizer can be configured to support nearly any structured data-file by specifying appropriate values in a configuration file. The configuration file uses comma-delimited files. The configuration flags and their default values are:

Configuration Flag	Default Value
columnsToAnonymize	Quest14a,Quest14b,Quest14c, Quest14d, Quest14e
columnsIndicatingRespondent	respondent,opername
dataDelimiter	“\t”
columnHasRole	True
entryDelimiter	:
implicitRole	Grower
cleaningThresholds	3,8,12

Specifying new values for each of these configuration flags requires specification in a configuration file. By default, this file is assumed to be next to the Anonymizer JAR file, and called “anonymizerConfig.txt”. The file-name and location can be changed by using a specific flag when called by Java. The file assumes that both the flag and its value are on the same row, separated by a colon “:”. The default values, specified in the config file, would look like so:

```
columnsToAnonymize: Quest14a,Quest14b,Quest14c,Quest14d,Quest14e
columnsIndicatingRespondent: respondent,opername
dataDelimiter: “\t”
columnHasRole: true
entryDelimiter: :
implicitRole: Grower
cleaningThresholds: 3,8,12
```

If you specify a flag that is not used by the Anonymizer, it warns you that it does not understand a given configuration flag, like so:



Configuration File Error: anonymizerConfig.txt

Configuration flag implicitRoole is not recognized. Please check.

Columns To Anonymize – Getting the Behavior You Want

The most significant element of the Anonymizer that may not be expected behavior is whether your data has separate elements within an entry to anonymize. The tool was originally used in a study where the survey giver entered a person’s name and role in the same field of the data, like so “Tanisha Banks:Grower”, where the first element, the name (“Tanisha Banks”), would need to be anonymized while the second element, the role (“Grower”) would not be. If you do not anticipate structuring your data in this way, then please make sure to use the `columnHasRole` configuration flag and set it to false.

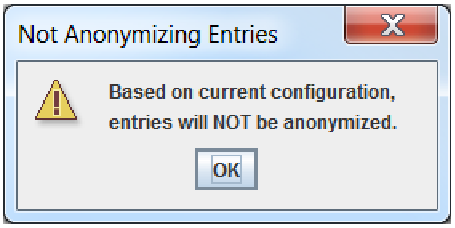
By default, if `columnHasRole` is set to true, then an entry without the `entryDelimiter` will be assumed to be a role, which would **not be anonymized**.

This implementation decision was made because a respondent may feel uncomfortable providing a name but would feel comfortable providing a role.

Again, if you want to make sure the entire field is anonymized, simply set `columnHasRole` to false. There will be empty role columns in the output data, but those can be quickly discarded.

Configuration Parameter: anonymize

This setting is a master-switch for turning on or off anonymization behavior. By default, this is assumed to be set to true, so a user has to intentionally add the “anonymize” parameter to their configuration file. If the anonymize parameter is included, and is set to “false”, then the data-file will be cleaned but no entries will be anonymized. This provokes a warning from the tool itself:



Based on current configuration, entries will NOT be anonymized.

We include this warning at run-time to make sure that the tool is being used as intended. Other parameters, such as “columnsToAnonymize” and “columnsIndicatingRespondent” are still interpreted, processed, and cleaned, but the final anonymization step will not be taken.

The output data file will read “cleaned_<original_file_name>.txt” instead of “anonymized_<original_file_name>.txt”.

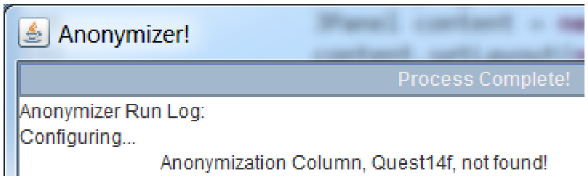
Configuration Parameter: columnsToAnonymize

Use this setting to indicate what columns in your data you want processed. This should be the full name of the column with correct capitalization, although white-space around the entry is ignored. If you want to anonymize the column “Quest10_Name”.

Entry	Outcome
“Quest10_Name”	Correct!
“Quest10_name”	Wrong! Capitalization error.
“Quest10Name”	Wrong! Missing the “_”
“Question10_Name”	Wrong! Question does not equal Quest
“Quest10_ Name”	Wrong! Spaces inside the text is not removed
“ Quest10_Name “	Correct! White Space around the text is removed

Any number of columns can be handled, separate each column entry with a comma. As stated in the prior section, columns included here will be assumed, by default, to have two sub-sections, a name and a role section. If this is not the case, make sure to also set the `columnHasRole` parameter to “false” – you can also leave this setting blank and use the `columnsIndicatingRespondent`, but, while functionally correct, that is semantically awkward.

If you specify a column (Quest14f, in this case) that is not found in the resulting data, the Anonymizer will warn you with a message like so:



```
Anonymizer Run Log:
Configuring...
                 Anonymization Column, Quest14f, not found!
```

Configuration Parameter: columnsIndicatingRespondent

Use this setting to indicate the respondent ID that needs to be anonymized. This was separated, functionally, from the prior field because there was no assumption that a role value would be included in the entry. As with the prior field, it's very important the field would be capitalized and spelled correctly to match the data.

As before, the Anonymizer will also warn you if a field specified here is not included in the data file.

Configuration Parameter: dataDelimiter

Use this setting if you are not using a tab-separated data sheet. Any arbitrary but consistent separation string may be used, including:

Sample Delimiters
,
;

__,__
-----!-----

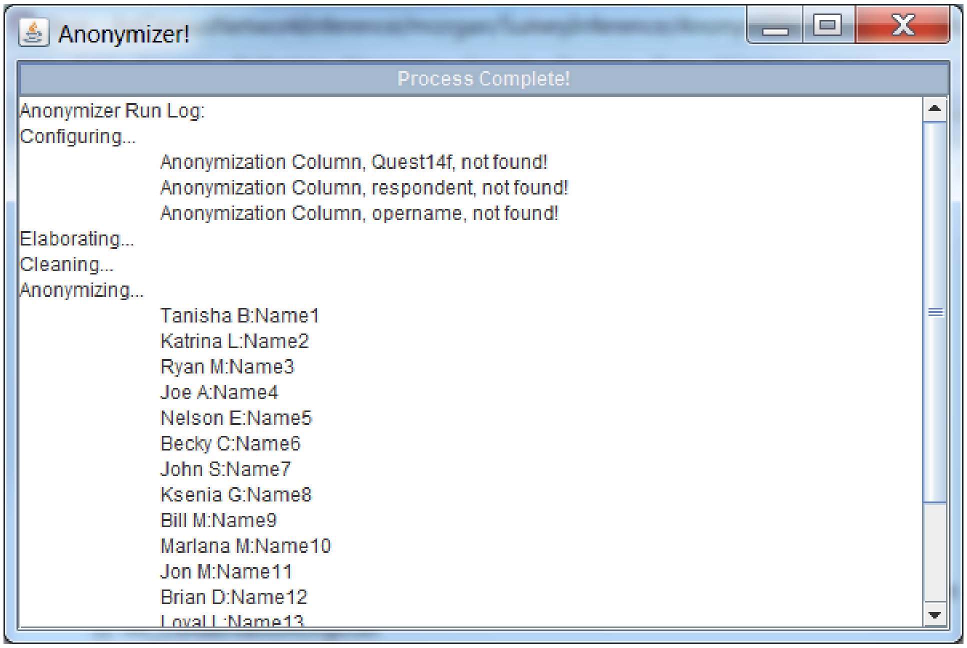
Avoid delimiters that involve quotation marks, except for the default "\t" option. Due to peculiarities with file output, quotations MUST be included when using the default delimiter, but are stripped when reading the file.

Configuration Parameter: columnHasRole

This parameter controls how the Anonymizer interacts with each field specified in `columnsToAnonymize`. If `columnHasRole` is set to true, then the Anonymizer assumes that there will be two elements to each entry, a name, which must be anonymize, and a role, which should not be anonymized. Each entry is separated using the `entryDelimiter`. If the `entryDelimiter` is not found, then the Anonymizer assumes that only **a role is given, and thus the entry should not be anonymized**.

Setting `columnHasRole` changes that behavior, which assumes that the entire field represents a name. This is useful when your data entry implementation is not asking for multiple values in a single answer.

Note that the person operating the anonymizer will be able to see what is being anonymized, and should be able to cross-check from the input file whether the right things are being anonymized.



```
Anonymizer Run Log:
Configuring...
                Anonymization Column, Quest14f, not found!
                Anonymization Column, respondent, not found!
                Anonymization Column, opername, not found!
Elaborating...
Cleaning...
Anonymizing...
                Tanisha B:Name1
                Katrina L:Name2
                Ryan M:Name3
                Joe A:Name4
                Nelson E:Name5
                Becky C:Name6
                John S:Name7
                Ksenia G:Name8
                Bill M:Name9
                Marlana M:Name10
                Jon M:Name11
                Brian D:Name12
                Loyal L:Name13
```

Configuration Parameter: entryDelimiter

The entryDelimiter value is used to separate entries into a name and a role. The table below gives examples using that `columnHasRole` is set to true. If `columnHasRole` is set to false, then this parameter is ignored.

Value	Name	Role	Outcome
Geoff M:Developer	Geoff M	Developer	Good!
Developer	Unknown	Developer	Good!
Geoff M:Developer:Specialist	Geoff M	Developer	Information Lost!
Geoff M	Unknown	Geoff M	Data not anonymized!

Configuration Parameter: implicitRole

This interacts with the fields listed in `columnsIndicatingRespondent`. Fields listed here will have a field added that matches the value given in `implicitRole`.

Field Name	Role Field
opername	opername_role
respondent	respondent_role

Configuration Parameter: cleaningThresholds

This is one of the most important features of the Anonymizer. Survey responses are often imperfect, with spelling mistakes. Thus, the Anonymizer allows for fields to be evaluated and closely-similar names evaluated the same name. How “similar’ two strings are is determined by its Levenstein Distance. Every insert, deletion, or modification of a character adds 1 to the distance. The shortest possible distance is found, and that is used to compare the two strings.

String1	String2	Distances	Total Distance
Jon Aiken	John Aiken	Insert 1	1
Jon Aiken	Jon Aken	Delete 1	1

Jon Aiken	Jon Siken	Change 1	1
Jon Aiken	John Sifken	Insert 2, Change 1	3

The Anonymizer, by default, uses different thresholds for similarity based on the length of the string. Note that the Anonymizer now enforces a symmetry condition – the threshold must be met for both strings. Jan and Jane, for example, would evaluate their closeness differently with our default configuration. Both would report a change distance of 1, but while that is acceptable to turn Jane into Jan, that is not acceptable for Jan to turn into Jane, therefore, both unique strings would be retained.

Length of String	**Threshold**
3	0 (exact match)
8	1
12	2
> 12	3

But you can change this behavior! Here are some examples and how the Anonymizer will behave.

Value	**Anonymizer Behavior**
3,8,12	Default behavior
3	Fields of length 3 or less will be evaluated exactly, longer fields will use a ‘1’ threshold for evaluation
8	Fields of length 8 or less will be evaluated exactly, longer fields will use a ‘1’ threshold for evaluation
8,8	Fields of length 8 or less will be evaluated exactly, longer fields will use a ‘2’ threshold for evaluation
<Blank Value>	All fields will be evaluated exactly