

## **Anonymizer**

### **The ICPSR Qualitative Data Anonymization Program**

#### **Purpose**

Anonymizer is a UNIX based program to anonymize qualitative data – such as interview transcriptions, typed field notes, or other records. Anonymizer is a result of a collaborative partnership among ICPSR, DSDR, and The University of Pennsylvania Population Studies Center.

The program is designed as a tool to streamline the process of removing and/or changing identifiers to maintain respondent confidentiality. Instead of a data producer reviewing each document for identifiers, Anonymizer uses a name key created by the user with a list of names and corresponding pseudonyms to change identified names to specified pseudonyms. Such a tool decreases the amount of time connected to reviewing each document.

The description provided here outlines the steps in using the Anonymizer program.

#### **Program Requirements:**

- Anonymizer runs on a Unix platform.

#### **Input:**

- Documents to be anonymized need to be in Rich Text Format or ASCII text.
- User generated Name Key needs to be in tab-delimited ASCII text.

#### **Output:**

- Anonymized document: [document name]\_anon.rtf or [document name]\_anon.txt (output will be in the same format as the input document).
- Anonymizer summary report: anonymizer\_rep.txt.

### **Running Anonymizer**

#### **Step 1: Create a name key.**

- The name key should include all the names that the researcher would like to anonymize.
- We suggest that the name key be converted from Excel into tab-delimited ASCII text.
- The name key file name should not contain any spaces (e.g., NameKey.txt or Name\_Key.txt, not Name Key.txt).
- The first column includes original names; the second column specifies the replacement names.
- First and last names and nicknames need separate entries. In the example below, Farai Zelda is the first and last name, each are entered separately and assigned pseudonyms
- Users can include multiple spelling variations of the same name to be changed to anticipate typographical errors. For example, Phillip is assigned the pseudonym John,

however, the user can anticipate that Phillip might be misspelled as Philip, and can enter Philip as another variation to be changed to John.

Example:

Agatha	Abigelo
Ajusa	ALFESO
Amadi	ABATA
Fei	FAUSTA
Sevi	SAINET
Deborah	DAILES
DeClerks	D.
Edie	ELENI
Emma	Rose
Ernest	ELIAM
Fanizo	Fred
Farai	FOUNTAIN
Zelda	ZUNDA
Hanock	HELLINGS
Jambo	JAFIFI
Lovemore	LONJEZANI
Mwenyemanzi	MAJONGOTI
Malikita	MABVUTO
Ngaurye	NAMBEWE
Oga	OLINI
Patricia	Eveness
Phillip	RAPHAEL
Philip	RAPHAEL
Runes	REDESI
Muhone	NDEGE
Seleman	SAIDI
Sharai	SAWSYEJE
Simba	SEREVASI
Wilford	Geoffrey
Yuda	YOUNGSON

Convert name key to a tab-delimited ASCII text file.

If converting from Excel: File → Save As → Text (Tab delimited) (\*.txt)

**Step 2: Invoke Anonymizer program.**

Run Anonymizer in the folder where both your name key and the documents to be anonymized are located.

*At the UNIX prompt type: anonymizer*

**Step 3: Choose output form.**

Anonymizer provides a choice of two outputs:

1. An annotated document with original names included.

2. A final anonymized document.

Anonymizer will ask:

*Annotate output with original names? (N/y):*

If you want an annotated transcript identifying changes, type y.  
If you do not want an annotated transcript, type N or just hit return.

In the annotated transcript, the replacement name will be followed by the original name in brackets: Agatha [Abigelo]. In RTF, [Abigelo] will be highlighted in red. In Text, **[\*\*Abigelo\*\*]**

**Step 4: Specify input files.**

Next, you will need to identify the name key file to be used:

*Name Key file (e.g., namekey.txt):*

The next step is to identify the document you would like anonymized

*Transcript file (e.g., GeorgeInterview.rtf):*

As Anonymizer runs, it will identify whether a name specified in the name key was found and the number of times each identified name was changed. For example:

```
Agatha changed to Abigelo 1 time(s)...  
Ajusa changed to ALFESO 1 time(s)...  
Akley not found...  
Amadi changed to ABATA 13 time(s)...  
Annie not found...  
Beatrice not found...  
Chande not found...  
Daffen not found...  
Deborah changed to DAILES 2 time(s)...  
DeClerks changed to D. 24 time(s)...  
Eddie changed to ELENI 1 time(s)...  
Emma changed to Rose 2 time(s)...  
Ernest changed to ELIAM 5 time(s)...  
Fanizo changed to Fred 3 time(s)...  
Farai changed to Fountain 19 time(s)...  
Fei changed to FAUSTA 2 time(s)...  
Hanock changed to HELLINGS 12 time(s)...  
Hawa not found...  
Jambo changed to JAFIFI 12 time(s)...
```

At the completion, Anonymizer will provide the name of the anonymized transcript (e.g., GeorgeInterview\_anon.rtf) and identify the title of the summary report:

*Done! Anonymized transcript is GeorgeInterview\_anon.rtf*

*See anonymizer\_rep.txt for summary*

### **Step 5: Review summary report and anonymized transcript**

The report and anonymized documents are located in the file in which you ran Anonymizer.

#### **Additional Options:**

You can also run anonymizer directly by using specified UNIX commands:

Usage: anonymizer [-k name key] [-a] [-h] [-q] [-t transcript]

```
-t: Specify single transcript filename to be anonymized
-a: Annotate output with original names
-h: Help; display this message
-k: Specify name key filename
-q: Quiet; only report name replacements
```

### **Example Output**

#### **Annotated Document**

Here is an excerpt from an annotated transcript:

YOUNGSON [Yuda] was saying this while shaking his head disappointedly and that was showing that he had already started recalling some of the things he was doing with girls at that time. Then MABVUTO [Malikita] said, "I am sure that none of the guys from Muriel Mine who were at our age at that time will be spared with that AIDS because that was the time when we shared that disease at that place."

The original names are enclosed in [ ] preceded by the pseudonym as assigned in the name key.

#### **Fully Anonymized Document**

Here is an excerpt from a document without name annotations:

YOUNGSON was saying this while shaking his head disappointedly and that was showing that he had already started recalling some of the things he was doing with girls at that time. Then MABVUTO said, "I am sure that none of the guys from Muriel Mine who were at our age at that time will be spared with that AIDS because that was the time when we shared that disease at that place."

## Summary Report

The summary report identifies the number of times a name has been changed in the transcript:

---



---

### ANONYMIZATION RESULTS

Date : Thu Aug 10 11:05:07 EDT 2006  
 Original transcript : George4.rtf  
 Name Key : NameKey.txt  
 Anonymized transcript : George4\_anon.rtf  
 Annotated : Yes

---

NUMBER	ORIGINAL NAME	PSEUDONYM
-----	-----	-----
1	Agatha	Abigelo
1	Ajusa	ALFESO
0	Akley	Pepulani
13	Amadi	ABATA
0	Annie	Becky
0	Beatrice	Luwiza
0	Chande	Jafeti
0	Daffen	Maulida
2	Deborah	DAILES
24	DeClerks	D.
1	Edie	ELENI
2	Emma	Rose
5	Ernest	ELIAM
3	Fanizo	Fred
19	Farai	Fountain
2	Fei	FAUSTA
12	Hanock	HELLINGS
0	Hawa	Ngawechinyalo
12	Jambo	JAFAFI
6	Lovemore	Lonjezani
22	Malikita	MABVUTO
1	Muhone	NDEGE
1	Mwenyemanzi	MAJONGOTI
1	Ngaurye	NAMBEWE
1	Oga	OLINI
1	Patricia	Eveness

### **Batch Anonymizing**

A user can also anonymize several documents at once using a specified name key.

The batch option is invoked using the -d option, followed by the name of the directory containing the batch of transcripts to be anonymized.

Usage: anonymizer [-k name key] [-a] [-d batchdir] [-h] [-q] [-t transcript]

Use only one of the following two options:

- d: Specify directory with batch of transcripts to be anonymized
- t: Specify single transcript filename to be anonymized

Additional options:

- a: Annotate output with original names
- h: Help; display this message
- k: Specify name key filename
- q: Quiet; only report name replacements

In addition, if you just enter `-d` and a directory name in batch mode, Anonymizer will prompt you for the rest of the options.

### Examples:

Anonymize and annotate a batch of transcripts in a directory called BATCH using the name key file called NameKey1.txt:

```
% anonymizer -d BATCH -k NameKey1.txt -a
```

Anonymize and annotate a batch of transcripts in the current directory (.) using the name key file called NameKey1.txt:

```
% anonymizer -a -k NameKey1.txt -d .
```

Anonymize a single transcript called GeorgeInterview.rtf using the name key file called NameKey1.txt:

```
% anonymizer -t GeorgeInterview.rtf -k NameKey1.txt
```

None of the original files will be replaced and all anonymized output will be written to a separate subdirectory labeled ANONYMIZED. In batch mode, it will try to anonymize all files in the specified directory except the specified name key file and the summary report file (anonymizer\_rep.txt). Obviously this could produce unexpected results if there are extraneous files in the directory.