,

OpenPseudonymiser
Desktop Software

User Guide

Version No: 2.0.2b

Revision History

|  |  |  |
| --- | --- | --- |
| Revision date | Version | Summary of Changes |
| 17/09/2014 | 2.0.2 | Original |
| 18/09/2015 | 2.0.2a | Added information about the 128 random character generation of salt to Appendix A |
| 20/02/2016 | 2.0.2.b | Amends to Licence section 3 |

OpenPseudonymiser documentation by the Julia Hippisley-Cox, University of Nottingham is licensed under a Creative Commons Attribution-NoDerivs 2.0 UK: England & Wales License. [www.openpseudonymiser.org](http://www.openpseudonymiser.org)

The OpenPseudonymiser software is issued under the GNU General Public License. University has made reasonable enquiries regarding granted and pending patent applications in the general area of this technology and is not aware of any granted or pending patent in Europe which restricts the use of this software. In the event that University receives a notice of perceived patent infringement, then University will inform users that their use of the software may need to or, if appropriate, must cease in the appropriate territory. University does not make any warranties in this respect and each user shall be solely responsible for ensuring that they do not infringe any third party patent.

# Overview

The University Of Nottingham has created an Open Source standalone windows desktop application called OpenPseudonymiser which is available for download at www.openpseudonymiser.org

The program is designed for processing Comma Separated Values ‘CSV’ files. It is used to remove patient identifiable fields such as NHSNumbers, ID’s etc and replace them with a pseudonymised field called the ‘digest’.

The digest is a pseudonymised identifier which can be consistently applied to different datasets so that record linkage can be undertaken without disclosure of patient identifiable data.

The hashing algorithm used to create the digest a one way process. This means the digest cannot be reverse engineered to reveal the identifiable data used to create it.

# Pre-requisites

The program runs on windows computers (XP, Vista, Win7, Win8, Server 2003/8) and uses the .NET framework. It works on both 32 and 64 bit versions of windows.

If the program file OpenPseudonymiser does not open correctly, verify you have the .NET framework is installed by going to <http://www.microsoft.com/net/>

Version 3.5 (or later) of the .NET framework is required.

# Licence

OpenPseudonymiser is free software: you can redistribute it and/or modify it under the terms of the GNUGeneral Public License as published by the Free Software Foundation, either version 3 of the License, or (at your option) any later version.

OpenPseudonymiser, including the website, software, documentation and key server technology, is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the GNU General Public License for more details.

Organisations who wish to make use of the OpenPseudonymiser technology have full responsibility for regarding information governance and security considerations relevant to their purposes. The Key Server is intended for demonstration purposes only. Organisations wishing to use OpenPseudonymiser for production purposes should deploy an instance of the software/key server etc suitable for their own purposes in order to satisfy their own information governance and security requirements.

You should have received a copy of the GNU General Public License along with OpenPseudonymiser. If not, see <http://www.gnu.org/licenses/>

OpenPseudonymiser makes use of the following Open Source libraries:

RSAEncryption Class Version 1.00 which is Copyright (c) 2009 DudiBedner

BigInteger Class Version 1.03 which is Copyright (c) 2002 Chew Keong TAN

NHSNumber-Validation which can be found at <https://github.com/pfwd/NHSNumber-Validation>

# Acknowledgements

We thank QResearch ([www.qresearch.org](http://www.qresearch.org)) for contributing to the costs of developing this software.

# Running OpenPseudonymiser

1. **To process a CSV datafile you will need two things ready:**
	1. The CSV datafile itself.
	2. An encrypted salt file.
	Encrypted salt files can be created at the companion website: www.openpseudonymiser.org
	See Appendix A for more information about encrypted salt files.
2. **Install OpenPseudonymiser and run it**The latest version is available on the site: http://www.openpseudonymisation.org/OpenPseudonymiser\_Software.aspxThe first time the program is run you will be asked to read and agree to the licence
3. **Screen 1: “Choose Data File”**
	1. **Select an input file**



This will open a file browser where you will be able to pick the CSV file that you want to work with.
Example CSV files are provided on the site: www.openpseudonymiser.org. The files provided are dummy data, provided only to illustrate how the software can be used.
The rest of this document shows screenshots when working with the example file “Example Input File 1.csv”

1. **Screen1: OpenPseudonymiser performs some basic checks**



OpenPseudonymiser will perform some basic checks on the file that you select
The ‘Next’ button will light up if the checks are all OK

1. **Screen 2: Select Salt**



The digest created by OpenPseudonymiser uses a one way hashing algorithm. Adding extra data to the input column(s) ensures that another party with the same set of input data (list of NHS Numbers for example) could not create the same digest.

This extra data is called the ‘salt’. The website [www.openpseudonymiser.org](http://www.openpseudonymiser.org) allows you to create encrypted salt files. Encryption provides another level of security by removing knowledge of the salt data from the users of software. See Appendix A for more information.

Salt files can be selected from the file system, or by connecting to a KeyServer.

If you have created an account on the OpenPseduonymiser web site ([www.openpseudonymiser.org](http://www.openpseudonymiser.org)) then the salt files you create will be available via the OpenP public KeySever (<https://keyserverapi.openpseudonymiser.org/api/>)

1. **Screen3: Select columns**



In the above screenshot note that we have:

1. Elected to use a single column ‘NHSNumber’ from the Example Input File 1.csv to create the digest.
2. Unticked NHSNumber and age from ‘Use in Output ‘. This will ensure the resulting CSV file will not contain these two columns.
3. Specified that the column called NHSNumber is our NHSNumber field using the drop down at the top of the screen.

The third column ‘Process as Date’ can be used to turn dates into less identifiable dates. The day and month are turned into 1st Jan, the year is kept intact.

**(Step E. Column selection screen continues.. )**

**A note about columns selected for Digest:**

All spaces are stripped from any data if a column is selected for use in a digest. This is a very important point and means OpenPseudonymiser should not be used to create digests for things like full names, postcodes etc.

**A note about NHS Numbers:**

If a column is specified as the NHSNumber field using the drop down:



Certain actions are performed on the data before it is used in the digest creation, namely:

* All non-numeric characters are removed before the data is processed. i.e only characters 1,2,3,4,5,6,7,8,9 & 0 are left in the data, everything else is removed.
* Checksum validation is performed on the data. A final column is added to the processed data set “validNHS” which will contain a “1” if the NHS Number passes the checksum test.

The NHS Checksum validation test is described here:

<http://www.datadictionary.nhs.uk/data_dictionary/attributes/n/nhs_number_de.asp>

A summary is also added to the “RunLog” file (see output files) The summary has the count of Passed (1’s), Failed (0’s) and Missing(-1’s)

**A note about “Process as Date”:**

If the “Process as Date” box is ticked then OpenPseudonymiser will attempt to interpret the data for this column as a date. If it can read the date format it will strip the Day and Month from the date leaving only the Year.

E.g. 29/11/1971 will become 1/1/1973.

OpenPseudonymiser supports the following date formats:
"yyyyMMdd", "dd/MM/yy", "dd/MM/yyyy", "dd.MM.yy", "dd.MM.yyyy"

**A note about data in quotes:**

Version 0.9.8 introduced a change where data that contains commas can be wrapped in quotes.

So the following row of data

123,456,”789,10”,11

Would be split into 4 columns:

123

456

789,10

11

This gives the app a bit more flexibility when dealing with textual data.

A rows of data that does not conform to the number of columns declared on the first line of the CSV file is treated as a “Jagged Line” and is not processed or present in the output file.

The counts of lines can be checked in the “RunLog” file (see H below)

1. **Screen4: Review the summary and start the process**

The final screen will let you change the output location, which is set to the same folder as the input file location by default.

It will also allow you to save a settings file which contains the column selections you made in the previous section.



1. **Press run and wait for the file to be processed**

OpenPseudonymiser will process the file.
Large files will show a progress bar and give you the option to cancel at any time.



1. **Inspect the output files**


Two files will be created in the location you specify:

* 1. The ‘RunLog’. A text file that shows the person with whom you are sharing the data the settings you used for the processing. This will have the same name as your input file, but will have a file extension of “OpenPseudonymiserRunLog”
	Note: this file is a record of processing, not a settings file.
	A settings file can be created by pressing the Save Setting File button.
	2. The processedCSV file which will have the same name as your input file but with the word “OpenPseudonymised\_” before it.

# Appendix A – Encrypted Salt Files

To create an encrypted salt file for use with the OpenPseudonymiser desktop software you will need to login to the website: [www.openpseudonymiser.org](http://www.openpseudonymiser.org). The site is free to use.

Navigate to the section entitled ‘Salt Files’.

Two fields are required, the file name, and the salt phrase or word:



The system will store your encrypted salt file. You can now share it with other users of the OpenP website, or you can use it in your own projects either by downloading it or by logging into the KeyServer via the Batch Processor (screen 2)

**A technical note on salt creation:**

The system will always create a 128 character salt. The phrase you enter will make up the first part of this 128 character salt, the rest of the 128 characters will be random. Therefore if you want to create “deterministic” salt (i.e. salt files that are the same each time you create them) then you will need to use a salt phrase that is exactly 128 characters long. This is potentially useful if you are using the batch processor in conjunction with the DLL implementation and need to recreate the same results in both systems.