SAIL – Person-level Export File Structure & Data Transfer Guidance Document

NOTE	UNCONTROLLED WHEN PRINTED

VERSION	24
CLASSIFICATION	Open
OWNER	Cynthia McNerney
DATE ISSUED	12 October 2022

1. Introduction

The SAIL Databank is a central repository containing anonymised person-level and address-level data drawn from operational and national systems. Using a novel anonymisation process, the SAIL Databank links datasets together to form a rich information base which is a national resource for e-health research and evaluation. All datasets are securely transferred into SAIL using the "Split-file" process with the support of the Digital Health and Care Wales our trusted third party. During this process person-level demographics are translated to an Anonymous Linking Field (ALF). This document describes the required file formats for person-level data and methods of data transfer.

2. Split-file Process

The original dataset is split into two types of files:

1. "File 1" dataset containing sensitive person-level demographics data which is sent to Digital Health and Care Wales **and not sent to the SAIL Team**. "File 1" data is processed by Digital Health and Care Wales who match and anonymise the data and then send it to us.
2. "File 2" containing clinical data or other non-identifiable data and is sent directly to us **and not sent to Digital Health and Care Wales**.

File 1: (Person Identifiable Data)
This contains unique person-level information.

The table below describes the required file structure for the "file 1" that will be sent to Digital Health and Care Wales for matching and anonymisation. (*To be delivered to Digital Health and Care Wales only)*

Field Name	Data Type	Description
SYSTEM_ID	varchar(50) (Unique)	Unique identifier
NHS_NUMBER	varchar(10)	NHS Number (with no spaces)
SURNAME	varchar(50)	Surname
FORENAME	varchar(50)	Forename
ADDRESS_1	varchar(255)	Address
ADDRESS_2	varchar(255)	Address
ADDRESS_3	varchar(255)	Address

Field Name	Data Type	Description
ADDRESS_4	varchar(255)	Address
ADDRESS_5	varchar(255)	Address
POSTCODE	varchar(8)	Post Code, where possible in formal space separated format i.e. **4 & 3 = "YYYY ZZZ"** **3 & 3 = "YYY ZZZ"** **2 & 3 = "YY ZZZ"**
DATE_OF_BIRTH	varchar(10) dd.mm.yyyy (EUR format)	Date of Birth: **dd.mm.yyyy** (EUR format)
GENDER_CD	char(1) Where 1=male, 2=female, 8= not specified, 9=unknown	Gender in codified format as **1=male, 2=female, 9=unknown**
CREATE_DATE	Varchar(20) Ideally dd.mm.yyyy (EUR format) (without time) But flexible e.g. will take text version of date such as '10th August 2007'	File creation date: **dd.mm.yyyy** (EUR format)
FIELD_1	varchar(50)	Not used. Consent Start Date if requested
FIELD_2	char(1)	Not used
FIELD_3	varchar(50)	Not used. Consent End Date if requested
FIELD_4	varchar(50)	Not used

SYSTEM_ID is the unique join key that will be used to link the final two files back together. This join key can be generated as part of the split or an existing unique field can be used. This field can be generated by simply creating a unique number for each row. It is essential that the SYSTEM_ID field is unique in the "File 1".

If there is a field in the "File 1" specification for which no data is available, the field must be included with the data cells left blank. For text fields where the source data contains commas, please remove the commas or replace the commas with a different character. Do not surround any text fields in quotes (""). The only exceptions to this are the Address_1 to Address_5 fields.

When submitting the "File 1" file to DHCW, please exclude the column headers. Please note that if the above structure is not followed then the file will not get loaded.

File 2: (Clinical or other non-identifiable Data)
This comprises of a delimited extract for all the tables containing clinical information or other non-identifiable data. Preferably presented as CSV or a delimited text format. Please avoid uploading Excel files.

The required file structure for the "file 2"s that will be sent to the SAIL team. (*To be delivered to SAIL only)*

Field Name	Data Type
SYSTEM_ID	varchar(50) This should correspond to a value in File1 SYSTEM_ID
……	Data provider specific information

For the file 2, the most common problem's which prevent our upload software from processing the file are:

- Text fields – if the file is in csv format, it is best to enclose address fields in quotes in case there are commas in the text. For all other fields, please avoid using quotes. For larger text fields, the extract code should ensure that no return characters are included in the text since return characters are normally used to separate individual records and having return characters in text fields would cause a problem for any processing software.
- Dates and times – Time should be in the format hh:mm:ss. Date fields should include all parts numerically (e.g. dd/mm/yyyy or yyyy-mm-dd or dd.mm.yyyy), and years should include the century component. Examples of acceptable dates are 18/05/2006, 2012-05-23 and 16.02.1984. Examples of dates that aren't acceptable are 18/05/06, 02-19-1947 (i.e. American day & month order) and 10-mar-05
- Dates and times combined – we can take fields which include both date and time together so long as the date and time components follow the advice given in the above point. Examples of acceptable datetime values are: 2016-05-02 18:15:21, 02/05/2016 18:15:21
- Date/DateTime formats – Different formats should not be used within a file.
- Empty data cells – best left as an empty string. Often, we get a code to denote an empty cell like NULL, -, *, and a dot(.). This causes problems with fields that are a numeric or date data type in that it prevents the database from assigning a numeric or date data type to the field. This makes it much harder for analysts to analyse the data.

Unlike the File 1 files, please include the column headings in all File 2 files. Where possible, clear and descriptive column names help to reduce delay if the team checking data quality can easily understand the nature of the data.

Filename formatting preferences:

1. File Names should follow the following naming convention. Please contact your project lead for SAIL Project Number. *<tablename>_<sail project number>_<todays dateYYYYMMDD>.csv e*.g. GPREADCD_0230_20140112.csv
2. Data presented in csv (comma delimited file) file format. For massive data quantities, this format is most suitable.
3. Character fields enclosed in double quotes
4. NHS Number should not contain any spaces. If your exported csv file has spaces please open it in excel and use this code below, to remove spaces on all rows.
    a. Open excel, Select NHS number column -> then go to Home -->
    b. Find & Select --> Replace
    c. In "Find What" box enter single space --> Keep "Replace with" box empty --> Replace All
    d. Save file.

4. Data Transfer

Never send data files via e-mail. E-mail is not sufficiently secure for transferring potentially sensitive personal information.

Method 1:

For **File 1**, secure electronic data transfer facility at Digital Health and Care Wales is available.

If your organisation is on the NHS DAWN / NHS network in England, use website: https://nwdss.wales.nhs.uk/NwdssSFU/
If your organisation is not within NHS network use website: https://www.nwdss.wales.nhs.uk/NwdssSFU/sfuLogin.aspx

The Data Acquisition Team at Digital Health and Care Wales can set up an account for new users to upload demographic data – please email PDIT@wales.nhs.uk

For **File 2**, using secure file upload you can send "File 2" data files directly to SAIL. An account will be created for you, and you can login and upload files. See **Section 6 – Key Contact Details** for the information required.

Method 2:

If your organisation has a secure file download service, we could login to your website and download relevant data from there.

Method 3:

Using SFTP we can provide a login for you to deposit files on our server.

If none of the above methods are possible please contact the SAIL team to discuss alternative secure methods of file transfer.

5. Data Dictionary

If a data dictionary is available, and has not already been supplied to the SAIL team, please attach a copy to the email you send to helpdesk@chi.swan.ac.uk when requesting account details for upload.

6. Key Contact Details

FOR FILE 1: (Identifiable Data)

Digital Health and Care Wales

Email : PDIT@wales.nhs.uk

FOR FILE 2: (Clinical Data or other Non-Identifiable Data)

SAIL Databank

Email: helpdesk@chi.swan.ac.uk with the following details:

1. Your SAIL project number or dataset name.
2. The PDM reference number provide by the Data Acquisitions team.
3. A name, email address and phone number for the official contact within your organisation, regarding delivery of SAIL data.