Open Data Standards for Public Health Scotland & the Scottish Health and Social Care Open Data Platform

Document Control Sheet

Key information

Detail	Answer
Title	Open Data Standards for Public Health Scotland & the Scottish Health and Social Care Open Data Platform
Date issued	23/08/2021
Version	1.0
Author	Open Data Team
Comments to	**phs.opendata@phs.scot**

Revision history

Version	Date	Summary of changes
0.1	11/08/21	Version 0.1 for public-facing standards
1.0	23/08/21	Adjust style and branding

Table of Contents

1. Introduction to NHSScotland Open Data

The **Scottish Health and Social Care Open Data Portal** provides a single location for all health and social care organisations in Scotland to release open data. Health and social care organisations can use it to release and share any open dataset that meets the minimum requirements for open data as set out in the Scottish Government Open Data Strategy.

A single platform for the Scottish Health and Social Care data makes it much easier for users to explore the rich health and care data landscape that exists in Scotland. Users of open data are from a wide range of backgrounds - from individual citizens, health and care professionals, government and public sector, academic researchers and private businesses.

The Scottish Health and Social Care open data portal is managed by the Open Data team who are based within Public Health Scotland. Data is released on this platform under the UK Open Government Licence.

The Scottish Health and Social Care Open Data Portal is located at **https://www.opendata.nhs.scot/**.

What is open data?

Open data is publicly released information which is made freely available to everyone to use and reuse in any way they like. Open data offers all those who want to make use of its potential an opportunity to make better decisions, improve and innovate. The released information is structured, able to be linked with other data and in a machine readable format that enables programmatical use of the data. Open data should never be patient identifiable and it should not disclose any personal information about individuals.

Data sharing

Making data open is not the same as data sharing. Data sharing is where an individual or an organisation wishes to access data relating to individuals, derived from information relating to individuals or data that is deemed in anyway sensitive.

Policy context

In February 2015, the Scottish Government set out an **Open Data Strategy** that seeks to "create a Scotland where non-personal and non-commercially sensitive data from public services is recognised as a resource for wider societal use and as such is made open in an intelligent manner and available for re-use by others". The Scottish Health and Social Care Open Data Portal makes an important contribution to this Scottish Government strategy by facilitating the release of open data.

Legal Framework

The basis for making open data available is underpinned by the **Re-use of Public Sector Information Regulations 2015** which requires public sector bodies to make information available, where possible, in open format and machine-readable format together with metadata. The 2015 Regulations harmonise and relax the conditions of re-use for public sector information.

According to the **European Directive on the Public Sector Information** open data policies that encourage the wide availability and reuse of public sector information for private or commercial purposes, with minimal or no legal, technical or financial constraints, can play an important role in:

- kick-starting the development of new services, based on novel ways to combine and make use of such information,
- stimulating economic growth, and
- promoting social engagement.

2. What makes up an open dataset submission?

For every open dataset you wish to publish, you will need to prepare a number of different components.

Component	What is it?	Why is it important?
Data File	The data must be a machine-readable CSV file, and formatted according to uniform technical standards for import to the Scottish Health and Social Care Open Data Portal.	Data in an accessible format allows it to be readily used and re-used by citizens, businesses, researchers, journalists, developers, government and others to process, trend, innovate, and inform.
Metadata	Metadata describes characteristics and attributes of the data (eg who, what, where, why, how). It contextualises the data and explains what it represents.	Metadata provides important structural and contextual information about the data – without it people won't be able to make the best use of your data file. It gives users a comprehensive understanding of the dataset, who collects and maintains it, the methodology, statistical and analytical issues and any limitations regarding the use of the data. It makes finding content and data faster and easier. Metadata facilitates data discovery and linkage across relevant and different data sources.

Data Dictionary	The Data Dictionary defines and explains the columns that make up each dataset. A data dictionary template is automatically generated on the platform after uploading a dataset.	The Data Dictionary is critical in that it clearly and consistently defines the columns and the characteristics of the data elements contained within the columns. A consistent, standardised vocabulary helps the end-user to fully understand the information contained within the dataset.
Supplementary documents	This could include a report specific to the data, explanatory documentation, a description of research benefits, survey findings, or information on data collection tools.	Additional documentation is useful to end users. For example it can explain how the data might be used, aid with interpretation and provide additional insight to complex data.
Approval Form	Before making a dataset available, the Open Data team at Public Health Scotland must receive a completed approval form, signed by all required parties.	It is important for the data provider to demonstrate that their data has been approved for public release and to confirm compliance with data protection and confidentiality.

Each of these should be provided for each new dataset prior to first publication on the Scottish Health and Social Care Open Data Portal.

3. Open Data standards - Producing your open data file

3.1 Format

- Files should be produced in CSV format.

- Individual files should be kept compact by avoiding duplication which will reduce the time needed to download the data. One way of doing this is to exclude information that is already provided on the platform in the reference files such as geographies, hospitals and specialities.

- Datasets containing multiple independent categories should be split by those categories, for example 'by age', 'by deprivation', 'by council area'. This helps to avoid accidental double-counting and allows better understanding of the data.

3.2 Column header names

Column headers should fulfil the following requirements with examples given below:
- Column field names must be meaningful, distinct, clear and in plain English.

- Column field names must be 48 or less characters in length, shortening should not result in misinterpretation.

- Avoid use of abbreviations or cryptic names unless these are widely recognised abbreviations, such as NHS (National Health Service). All abbreviations have to be fully explained in the data dictionary.

- Use **Camel case** – the practice of writing compound words or phrases such that each word or abbreviation in the middle of the phrase begins with a **capital letter**, with no intervening spaces or punctuation.

- Do not use spaces, underscores or hyphens in column names.

- Do not use special characters such as '%' or '+'. You may write the full word 'Percentage' or use common abbreviations such as 'Pc'. If the special character is part of a construct, such as '50+' or '5%' it should be considered as one phrase and therefore translate to '50plus' and '5pc'.

Examples of acceptable and non acceptable column headers:

3.2.1 Clarity

Good column header should be clear to all

☑	ProjectName	ContactTeam	ContactAddress
	Open Data Platform	Open Data Team	phs.opendata@phs.scot

These column names would be seen as cryptic and would not be accepted

☒	Prj_Nme	Ct_Tm	Ct_Ttl
	Open Data Platform	Open Data Team	phs.opendata@phs.scot

3.2.2 Camel case

All words should be capitalised with no exceptions, spaces or underscores

☑	CountryOfResidence
	Afghanistan

These examples would not be acceptable

☒	COUNTRY_OF_RESIDENCE
	Afghanistan

☒	country_of_residence
	Afghanistan

☒	countryofresidence
	Afghanistan

3.2.3 Use of acronymns to shorten column headings

It is acceptable to use acronyms to shorten long names, particularly where the acronynm is a well understood; noting that all acronynms must be fully documented in your data dictionary.

This acronym is accepted when documented in the dictionary

☑	**EASRNumberOfNewPatients**
	2052

Not acceptable as column name is now too long

☒	**EuropeanAgeSexStandardisedRatesNumberOfNewPatients**
	2052

3.3 Column header row

The top row of data should be the header row containing column headers (variable names) for each column. Data should only contain one header row. Multi-row headers are not acceptable. The data should be populated from the second row downwards.

3.3.1 Merged Cells

You must not include merged cells in your file. They cannot be reproduced in a CSV. Each column needs to have a unique header, contained in the top row, defining the data.

Acceptable – single-row header, column names in title case, underscores used instead of spaces

☑	**ProjectName**	**ContactTeam**	**ContactAddress**
	Open Data Platform	Open Data Team	phs.opendata@phs.scot

Not Acceptable – multi-row header, merged cells and column names in uppercase

☒	**PROJECT**		**CONTACT**	
	NAME	**TEAM**	**ADDRESS**	
	Open Data Platform	Open Data Team	phs.opendata@phs.scot	

3.4 Variable types

A variable/column can only contain one data type. This can be either numeric or a string. A column should never consist of a mixture of numeric cells with text cells. This means special characters cannot be used in numeric data columns. If a dataset contains missing and/or suppressed values these should be displayed as empty cell and coded in a separate 'Qualifier column'. Please see the section 3.11 for additional information.

Numeric values should always be plain numbers such as 12 or 1400.35. Do not use commas in values to increase readability for humans. As commas are used to separate columns in a CSV file values with a comma are exported as text field.

Acceptable – Numeric data without commas

☑	HBT	NumberOfPatients	AverageCasesPerMonth
	S08000015	20500	1750.38
	S08000016	14750	980.79

Not Acceptable – Numeric data containing commas

☒	NHS_Health_Board	Number_of_patients	Average_cases_per_month
	S08000015	20,500	1,750.38
	S08000016	14,750	980.79

3.5 Ensure consistent coding convention

Include standard codes and descriptions. The data dictionary needs to include the link to existing resource/look-up file that explains the labels of the code. This is good practice for open data as it ensures consistent labelling across all data on a platform or from an organisation. A data file should never contain codes and labels.

- Standard nine digit geography codes can be found at
  https://www.opendata.nhs.scot/dataset/geography-codes-and-labels

and

http://www.gov.scot/Topics/Statistics/sns/SNSRef/StanGeoCodeRegister

- Specialty codes, location codes etc. lookups can be found at:

  https://www.opendata.nhs.scot/group/reference-files

Acceptable – The header for health boards and other geography levels needs to encompass:

- Type of health board indicated, eg health board of residence (HBR), health board of treatment (HBT) etc.

- Configuration is specified in the data dictionary and mentions the configuration used: health board 2014, health board 2006, health board time of event (which can be a mixture of both 2006 and 2014 configurations).

- The corresponding health board names can be found in the reference files.

☑	HBR	NumberOfCases	NumberOfReferrals
	S08000015	20	14
	S08000016	42	20
	S08000017	18	8
	S08000018	62	30

Not Acceptable – Type of health board not indicated. Labels are used instead of codes for health boards.

☒	NHSHealthBoard	NumberOfCases	NumberOfReferrals
	Ayrshire and Arran	20	14
	Borders	42	20
	Dumfries and Galloway	18	8
	Fife	62	30

3.6 Layout and structure

Files should not be specially formatted before creating a CSV as this can lead to errors in the file structure. Style elements cannot be stored in CSV files. Avoid drop down lists and filters.

3.7 Vertical rather than horizontal orientation

Vertical orientation should be used wherever possible. Vertical datasets are more easily understood, sortable, as well as more useful for creating visualisations, and improve consistency between the datasets. Datasets with a horizontal data orientation should be restructured to vertical.

Acceptable – vertical data orientation

☑	Year	HBR	NumberOfCases
	2016	S08000015	0
	2016	S08000016	1
	2016	S08000017	0
	2016	S08000018	2

Not Acceptable – horizontal data orientation

☒	Year	AyrshireAndArran	Borders	DumfriesAndGalloway	Fife
	2016	0	1	0	2

This specifically applies to datasets containing data by year, especially numerous years. Years should have their own rows, rather than columns, in the data. Presentation, consumption, ease of use, and refresh of the data become much more difficult as data files get wider with numerous columns.

Horizontal orientation restricts the user`s ability to create visualisations and makes it difficult to perform analytics and observe time-based trends in a single view. Vertical data orientation not only makes the data machine-readable, but also human readable.

Acceptable – vertical data orientation, additional years can be appended to the data file as needed

☑	HBR	HiringYear	NumberOfNewHires
	S08000015	2014	700
	S08000015	2015	1000
	S08000015	2016	1200

Not Acceptable – horizontal data orientation, each subsequent year of data requires the addition of another column

☒	HealthBoard	2014NewHires	2015NewHires	2016NewHires
	Ayrshire and Arran	700	1000	1200

Vertical orientation provides maximum flexibility for reuse and application. Data can be sorted by year or to create visualisations with the data rolled up by year. The benefit of this more vertical orientation is that any grouping of variables is possibly facilitating complex analyses of the data. A user can use any number of third party business intelligence and data analysis visualisation tools to create pivot tables or graphs to identify patterns and trends over time.

3.8 Dates

The way the dates are formatted may depend on the structure of your data. Please keep in mind that open data should be as granular as possible. The following date formats can be used noting that a date should not be spread over multiple columns.

Format	Structure
YYYYMMDD	This is the preferred format for full dates when the exact date is available.
YYYYMM	This should be used if your data is available monthly - eg 201712.
YYYY	This should be used if the data is presented annually - eg 2017.
YYYYQN	Quarters should be represented in the format $1 \leq N \leq 4$ – eg 2017Q1.
YYYY/YY	Financial years - eg 2016/17

3.9 No multiple data items in a cell

A cell may contain only one item of information; multiple lines within a cell will cause the import process to fail. Cells that contain collections of data are impossible to evaluate and could cause problems for end-users of your data.

Depending upon each organisation's preference, this may also be 'highly vertical', especially if the inclusion of additional data concerning status (see example below) is being planned. This will serve to minimise manual intervention for refresh.

Best practice - highly vertical alternative

☑	ProjectName	Status	StatusDate
	ABC	Started	20130103
	ABC	Design Complete	20130112
	ABC	Development Started	20130115
	ABC	Development Complete	20130120

Acceptable – horizontal, one item of information per field

☑	ProjectName	Started	DesignComplete	DevelopmentComplete
	ABC	20130103	20130112	20130120

Not Acceptable – collection of data in one field

☒	Project Name	Status
	ABC	01/03/2013 – Started
		01/12/2013 – Design Complete
		01/20/2013 – Development Complete

3.10 Ensure there are no formulas within the cells – component values should always accompany calculated fields

These data fields should be expanded to include each data component, especially when the creation of visualisations will rely upon this data. For example, if two figures were added together to create a summary value, you should include three columns: one for the first added value, a second column for the second added value, and a third column for the sum of the two.

Acceptable – components included with calculated field

☑	FacilityName	Year	TurnoverItemA	TurnoverItemB	TotalTurnover
	XYZ Medical Care Center	2013	18048	8709	26757

Not Acceptable – calculated field but missing components

☒	Facility Name	Year	Total Turnover
	XYZ Medical Care Center	2013	26757

3.11 Qualifiers - Blank, 'N/A', or other unknown cells

Blank fields, when left unexplained, risk confusion – particularly when the column is numeric.
- If the blank field represents zero, then the field should be zero.
- If the blank field represents 'not collected' or 'unknown', then this should be a blank field and explained in a separate qualifier column, adjacent to the data column this is referring to.

Equally important is consistency.
- If data is numeric, it should not also be presented as text within the same column. Blank fields should therefore not contain space " ".
- Fields containing "N/A", "-" or "unknown" should not be mixed within a numeric column as this will make sorting and analysis difficult.

Our advice is to follow the **guidance of the Office for National Statistics** on symbols in tables. This suggests the use of an extra column next to the data column the qualifier refers to. The column should have the identical name of the column it is referring to with the suffix QF (ie NumberOfPatients – NumberOfPatientsQF or NewHires - NewHiresQF). A full list of qualifiers can be found on **https://www.opendata.nhs.scot/dataset/statistical-qualifiers**, some of the most important qualifiers are shown here.

Symbol	Guidance
c – confidential	The small letter 'c' should be used for suppressed data.
: – not available	A colon is used for data that is not available.
z – not applicable	The small 'z' indicates data that is not applicable.
r – revised	The figure stated has been revised since it was first published.
d – derived	A figure is derived for custom groupings and for totals; this includes totals for Geographies (ie Scotland when data is otherwise presented at council area level) and specialities (ie all cancers, all alcohol conditions) to prevent double counting.

Acceptable – text excluded in numeric field

☑	Company	Year	WorkRelatedInjury	NumberOfCases	NumberOfCasesQF
	ABC	2009	Skin Disorders	27	
	ABC	2009	Respiratory Conditions		c
	ABC	2009	Poisoning	0	
	ABC	2009	Eye irritations		:

Not Acceptable – text included in numeric field

☒	Company	Year	Work Related Injury	Number of Cases
	ABC	2009	Skin Disorders	27
	ABC	2009	Respiratory Conditions	suppressed

	ABC	2009	Poisoning	0
	ABC	2009	Eye irritations	Missing value

3.12 Suppression of small numbers

The Open Data team expect you to follow all necessary disclosure control protocols that are available within your organisation, this includes the suppression of small numbers. If you require help with disclosure control please get in touch with the Statistical Governance Team on **phs.statsgov@phs.scot**.

It is necessary that disclosure controls are applied to your data before this leaves your department and is sent to us.

3.13 Categories

To support best use of your data, every effort should be taken to avoid the risk of double counting by users. Ensuring that columns don't contain multiple categories will help to reduce that risk. In most cases the total of each category would be identical. These categories should be split up into separate files.

Acceptable – two separate files, one per category

☑	Gender	NumberOfPatients
	M	50
	F	40

AgeGroup	NumberOfPatients
20-24 years	60
20-29 years	30

Not Acceptable – multiple categories in one column

☒	Category	SubCategory	NumberOfPatients
	Gender	M	50
	Gender	F	40
	Age Group	20-24 years	60
	Age Group	20-29 years	30

This rule also applies to Geographies. Various types of Geography should not be listed in a shared column. Data should be presented in an as granular way as possible with the smallest possible geography being used, considering confidentiality and suppression of small numbers.

Where it makes sense to present the data on multiple levels, this should be done in separate files. The exception to this rule is the Scotland total. However, if this is included, it has to be made clear with the use of the qualifier 'd'.

Acceptable – data split in 2 files, one per geography and the Scotland total is indicated as derived

☑	CA	CAQF	NumberOfPatients
	S12000005		50
	S12000006		40
	S92000003	d	90

	HBT	HBTQF	NumberOfPatients
	S08000015		60
	S08000016		30
	S92000003	d	90

Not Acceptable – Multiple Geographies in one Column

☒	Geography	NumberOfPatients
	S12000005	50
	S12000006	40
	S08000015	60
	S08000016	30
	S92000003	90

An option to include more than one level of geography is to nest them. Nesting means to provide information on which higher geography a lower level belongs to as shown in the example below.

Acceptable – Nested geographies

☑	HBT	CA	NumberOfPatients
	S08000020	S12000020	80
	S08000020	S12000033	27
	S08000020	S12000034	89

	S08000022	S12000017	102
	S08000022	S12000035	68

3.14 No blank rows and columns

Care must be taken to ensure that any blank rows/columns have been removed prior to creating your CSV files. Empty rows and empty columns among the data will not be accepted. Empty rows adversely impact sorting and analysis, and empty columns adversely impact initial import and refresh.

If you click outside the active data and save it Excel will allow for blank rows and columns at the end of dataset. If you click outside the border of the active data and save it as CSV, Excel will read these blank rows and columns as part of the data and save the blank cells. This may not appear to be part of the spreadsheet but will result in blank cells being included when saving as CSV.

An easy way to see if blank rows or columns are present in Excel:
- Press [Ctrl]+[End] and see if this takes you beyond your data in the spreadsheet.
- If it does, the data file contains blank rows and columns that need to be deleted. Delete the blank rows and columns, save the Excel file, and then press [Ctrl]+[End] again – this should now take you to the last row and column in your data.

Care must be taken when importing data from, for example, reports which may have blank rows. If blank rows are present the data should be cleansed to ensure that any blank rows and columns have been removed prior to creating your CSV files.

Not Acceptable – empty rows and empty columns

☒	Company	Year		WorkRelatedInjury
	ABC	2009		Skin Disorders
	ABC	2009		Respiratory Conditions
	ABC	2009		Poisoning

	XYZ	2009		Skin Disorders
	XYZ	2009		Poisoning

4. Data dictionary

The data dictionary is an essential part of any open data dataset. It helps the users of your data to understand what it represents by giving a clear explanation for each of the column headers (variables) used. An empty data dictionary is created automatically after a resource is uploaded to the Scottish Health and Social Care Open Data Portal.

The following fields are created in the data dictionary:

- Column - The column headers (variables) you have used in your dataset are automatically populated in the data dictionary.
- Type - The variable type of the column (numeric or text).
- Description - A definition for each of the variables used in the data set. Where applicable, the description needs to contain all details on potentially used abbreviations and scales/rankings.
- Lookup- If a column contains codes, this field needs to be populated with the URL of the reference/look-up file. If the column does not contain codes this field can remain empty.

The data dictionary needs to be completed for each data resource as the fields may vary between resources. When you submit your data, each file needs to be accompanied by a data dictionary with the fields 'Column' a list of all column headers, 'Description' the description of the column headers and 'Lookup' URL to the reference file for coded columns.

5. Metadata

Metadata is an important feature of each dataset. Through metadata, users are given background information on the dataset which helps them to better understand and interpret the data increasing the use and re-use of the data. The following table gives an overview on the fields that are embedded on the platform. If you feel that a field is not applicable, you will need to add an explanation why this is the case.

Field	Description
Title	Short, descriptive and unique title for the dataset.
Contact address	Postal address of the dataset author.
Subject	Narrowing down from the theme eg Alcohol Misuse or Pharmacy services.
Frequency	Frequency of the release ie annually, quarterly, monthly, weekly or ad hoc.
Time frame of data and timeliness	Provide information on timing between the data extraction and publication eg four months in arrears.
Coverage	Spatial and/or temporal coverage of the data.
Completeness	Explain any limitations or changes to data or data sources that may have had an effect on completeness.
Accuracy	Explain any errors, revisions or other changes that may have had an effect on accuracy and what that effect was.
Continuity of data	Discuss factors that may have had an impact on continuity of data, for example methodological changes, geographical changes, framework changes, system/software changes, collection changes. Explain the potential impact that this may have had on the dataset.

Concepts and definitions	Terms and underlying concepts that help to better understand and interpret the data.
Disclosure	Disclosure control methods that have been applied to the data.
Revision statement	For example - These data are not subject to planned major revisions.
Official statistics designation	Official Statistics designation of the publication (National, Official, Experimental).
Relevance and key uses of the statistics	Explain what the relevance of the provided data is and known uses of the data.
Format	Format in which the datafile is provided, typically CSV.
Language	English
Links	Links to similar resources
Description	A brief description of the dataset. The description is displayed beneath the title and should provide a better understanding on what can be expected of the dataset.
Tags	Keywords that will help users to find the dataset in searches.
Licence	UK Open Government Licence
Theme	On CKAN, datasets belong to a 'Theme': Health and Care; Blood, tissues and cells; Health protection; Information technology; Practitioner services; Logistics; Procurement.
Source	A source where the data are extracted or derived from.
Version	Current version number of the document.

Author	Name of the team or organisation that created the resource.
Author email	Provide the email address for the author, to which queries about the data should be sent. This could be a generic team email address.
Maintainer	Team name of the maintainer if different from author - otherwise not necessary
Maintainer email	Team email of the maintainer for the dataset.

6. File naming convention

When you create open data files and save them as CSV, you should choose a name that describes your dataset. A good rule of thumb in naming is to imagine that if you have downloaded a file and looked at the name a week later, would you still remember what it stands for? Avoid using cryptic names with not widely known acronyms. A suggestion for naming the file would be the following:

- 'Dataset_Name'_'Dataset_Period'.csv – for example Prescribing_by_GP_Practice_March_2016.csv or GPPracticePrescriptionsMarch2016.csv
- 'Dataset_Name'_Metadata.csv
- 'Dataset_Name'_DataDictionary.csv