KnowFife Data Quality and Disclosure Control Overview

Background

This document sets out background information regarding data quality and data disclosure which was used to help create the written procedures for the KnowFife Dataset. It also outlines the basic awareness and understanding all data administrators should adhere to when adding data to the system and an associated guide outlining the key steps that must be completed during this process (this 'KnowFife Dataset Admin Procedures' document is available separate to this paper) . Further to this, a set of 'frequently asked questions' (FAQs) for KnowFife Dataset users has been created and will appear on the KnowFife Dataset website. This document was borne out of a project carried out during the Summer of 2014 by an Intern within NHS Fife's Public Health department, examining best practice and the application of data quality and disclosure methods for the data held within the KnowFife Dataset.

Many of the statistics used by the KnowFife Dataset are official statistics and have been subject to the UK Statistics Authority Code of Practice as well as the Scottish Government Statistician Group guidance. However, steps still need to be taken to ensure the data is high quality and proper disclosures are in place so as to maintain the privacy of individuals to whom the data relates and the confidence of KnowFife Dataset users. What follows in this paper is an overview of some of the key considerations and practices that are implemented to ensure the high quality and accuracy of data loaded to the KnowFife Dataset.

Data Quality

Examining the information from the procedures already done within KnowFife Dataset, the literature reviewed and practices in other organisations highlight that the components of data quality outlined below are key for maintaining high quality data.

Component 1: Timeliness

- It is important that the KnowFife Dataset provides the most current data available for the indicators that it holds. The exact time periods this will cover will vary depending on the nature of the indicator; for example some indicators update monthly, the majority yearly but others every two or four years.
- Information on the timeliness of each indicator should be included in the metadata. Using the metadata fields 'Date available', 'date next version due', 'time period' and 'frequency' will provide users of the KnowFife Dataset re-assurance that the data is current and also enable the KFD co-ordinator to have an overview of when data should update.

Component 2: Completeness

- In order for the KnowFife Dataset to function properly to provide data for different geographies in Fife it is important that all datasets loaded into the system are complete. This is particularly relevant for data that is loaded at datazone level which is then aggregated upwards to higher geographies.
- To ensure completeness datasets should be checked that they contain all expected figures and statistics, and no duplicates or blanks are present.

Component 3: Accuracy

- For the KnowFife Dataset to be a reliable source of information about Fife residents it is essential that the data it provides is accurate. Inaccurate figures published in the Dataset could lead to significant negative consequences for users of Dataset and the KnowFife Dataset itself. The data uploading procedures in theKnowFife Dataset and other data repositories such as SNS provide potential for error as a result of building data from smaller geographies, copying and pasting data from other sources and the specific labeling of indicators.
- To help maintain the accuracy of information in the KnowFife Dataset a series of simple checks can be undertaken on the data during and after uploading including performing logic checks for sub and overall totals, checking the spelling of labels and headings, checking figures against other publications of the same data source and comparing with previous period data to identify any large changes in figures.

Component 4: Interpretability

- It is important the KnowFife Dataset users are able to present and describe the data they have obtained from KFD in the correct way. Making full use of the metadata for each indicator will allow users of the Dataset to know exactly what the information is they are using, what statistics has been used to present the information, what time period it covers and if there are any caveats to be taken into account when using the information.

Data Disclosure

A number of different disclosure control methods are used in the KnowFife Dataset and other organisations to maintain the privacy of individuals. The method chosen will depend on the type of data that has been collected and how for that specific indicator. Below is a description and example of common disclosure control methods. If a method of disclosure control has been applied to data within the system (either at source or by KnowFife Dataset administrators) this should be stated in the indicator metadata.

Small cell suppression

This is the most common form of data disclosure used in data uploaded to the KnowFife Dataset. If a cell in a dataset is less than 5 (which is most common, but it could be less than 10, 50, or even 100 depending on the data), it will be rounded down to 0. This is designed to protect the privacy of individuals at small geographies. Typically this method is used for data at the datazone level.

Example indicator: Job-seekers Allowance data at datazone level. Values less than or equal to 2 in a given datazone are suppressed to 0 to protect the privacy of the individuals in the data.

Geographical suppression

Some indicators contain data that is sensitive in nature or the number of events are very small. If the geography was also small it would be easier to identify individuals. Therefore, the geographical suppression disclosure method can be used to limit the geographic boundaries at which the data can be presented.

Example indicator: Life Expectancy (interzone upwards) or cancer deaths (Area Committees, SIMD quintiles)

Rounding

Similar to small cell suppression, counts may be rounded to the nearest 5 or 10 for some indicators at low geographical levels.

Example indicator: The Department for Work and Pensions (DWP) takes this approach which can be seen in many benefit related datasets including 'Children living in poverty' and Working age key benefit claimants'.

Record swapping

Record swapping is a method used to swap records from one geography with a similar number of records from another geography of the same type. In this way, a similar cell count is displayed, but it will not be the exact number for that specific indicator at that geographic level.

Example indicator: Scotland Census 2011 data at datazone level.

Time period total

This method is used to display information about a topic that may have a small number of events due to the nature of the data or the geography at which it is presented or a combination of both or is sensitive information. The counts of the events are totaled across years, often 3 or 5 year periods and presented as a count for that period. The time period is fixed so the end user cannot see the counts within a specific subset of the time period (e.g. if it is a three year total the user cannot select a specific year within that range).

Example indicator: All cancer deaths by area committee 2009-11 (3 year total)

Averaging across time

As above but the figure shown is the total number of events averaged across a defined period. For example the total number of events in three years (2010, 2011 and 2012) would be divided by three and presented as the average number of events 2010-12.

Multiple disclosures methods

There will be some instances where more than one disclosure method may have been applied to the data. For example within the KnowFife Dataset it will be common to find data for a three year period (time period total) only available for larger geographies such as Area Committees or SIMD quintiles (geographical suppression).

Example indicator: CHD deaths by area committee 2009-11 (3 year total)

Further information

Other data quality and disclosure control policies:

Eurostat Quality Assurance Framework

UK's Code of Practice for Official Statistics

Scottish Government Statistician Group Quality Guide

Scottish Neighbourhood Statistics FAQ