Technical documentation: ANCHDA indicators

prepared by the data science team

2023

Contents

[Data acquisition (drafting) 2](#_Toc134536312)

[Data Sources 2](#_Toc134536313)

[Data Collection 2](#_Toc134536314)

[Data Transfer and Storage 2](#_Toc134536315)

[Data Security and Confidentiality 2](#_Toc134536316)

[Data cleaning 3](#_Toc134536317)

[R Scripts 5](#_Toc134536318)

[Quality Control 5](#_Toc134536319)

[Cell Suppression 6](#_Toc134536320)

[Temporal Correspondence 7](#_Toc134536321)

# Data acquisition (draft section)

## Data Sources

* Description of the data sources, including data provider, data type, and data format
* Explanation of how the data sources were selected
* Overview of the data quality and completeness

## Data Collection

* Details of the data collection process, including the timeframe, frequency, and methods used
* Description of any tools or equipment used for data collection
* Explanation of how the data were verified and validated

## Data Transfer and Storage

* Explanation of how the data were transferred to the data management system
* Description of any data storage platforms used, including one drive, folders, subfolders, or other data management tools
* Overview of any data management protocols, or standards used, including data dictionary, unique IDs, tracking sheets, and other documentation

## Data Security and Confidentiality

* Overview of data security measures used to ensure data confidentiality and privacy
* Description of any encryption or safety protocols used during data transfer and storage
* Explanation of any ethical or legal considerations related to data security and confidentiality.

# Data cleaning

## Objective of formatting rules [drafting this section]

The objective of formatting rules is to establish a uniform and standardized structure for the data that enables its easy integration and analysis in the VISER database architecture. Our conversations with the VISER team revealed that the database structure relies heavily on consistent and well-formatted data to ensure accurate analysis and reporting.

Having clear formatting rules for each field ensures that data is organized in a consistent way, which reduces the risk of errors and discrepancies when importing it into the database.

[…] continue with overview of the csv importer

## Specifications for cleaning the data

The upcoming section outlines the requirements for each field that needs to be present in the cleaned datasets.

#### Geography

The "geography" field is used to indicate the geographic area to which the data pertains. This field should be named using one of the following options: "STE\_CODE16", "SA2\_CODE16", "SA3\_CODE16", "SA4\_CODE16", "LGA\_CODE16", or "Australia". The values in this field should be numeric.

For "SA2\_CODE16", "SA3\_CODE16", "SA4\_CODE16", and "LGA\_CODE16", the codes are always composed of 5 digits.

For "STE\_CODE16" and "Australia", the codes are always composed of 1 digit. If the raw data includes character values for States, the following conversion rules should be applied: 1 = NSW, 2 = VIC, 3 = QLD, 4 = SA, 5 = WA, 6 = TAS, 7 = NT, 8 = ACT.[[1]](#footnote-1)

Finally, if the "geography" field is set to "Australia", the value should always be coded as "0".

[…] add something about the use of the ASGS 2016?? And temporal correspondence??

#### Sex

The "sex" column is used to indicate the sex of the individuals in the dataset. This column should include categories for male and female. If the raw dataset does not include disaggregation by gender, a new column should be created and filled with "all" for all values.

#### Age\_group

The "age\_group" field is used to indicate the age range of the individuals in the dataset. This column should include categories in the format of a hyphen between two numbers (e.g. "18-24" or “0-4”).

If the raw dataset does not include disaggregation by age, the data analyst should contact the data custodian or the project manager to determine what proxy can be used instead.

If the data is presented with school year, the following conversion table can be used to populate the "age\_group" column:

Table 1 - conversion between school year and age groups

|  |  |
| --- | --- |
| School year | Corresponding age\_group |
| Prep | 5-6 |
| Year 1 | 6-7 |
| Year 2 | 7-8 |
| Year 3 | 8-9 |
| Year 4 | 9-10 |
| Year 5 | 10-11 |
| Year 6 | 11-12 |
| Year 7 | 12-13 |
| Year 8 | 13-14 |
| Year 9 | 14-15 |
| Year 10 | 15-16 |
| Year 11 | 16-17 |
| Year 12 | 17-18 |

#### Time frame

There are two options for indicating the time frame of the collected data. The first is to use a **"calendar\_year"** field which should be in numeric format with the values following the pattern of four digits (e.g. "2022"). The second option is to use a **"year\_range"** field which should also be in numeric format with the values following the pattern of four digits separated by a dash (e.g. "2010-2011").

#### Uncertainty

Uncertainty field: this field serves to indicate the level of uncertainty present in the raw data. A value of 0 should be assigned if there is no uncertainty, while a value of 1 or 2 should be assigned if there is some level of uncertainty indicated by one or two asterisks respectively (\* or \*\*).

#### IRSD

The field "irsd\_quintile" is a variable that shows the socio-economic status of a specific population or geographic region in terms of their Index of Relative Socio-economic Disadvantage (IRSD) quintile. This field is not available in all datasets. In case it is present in the raw data, the records having the lowest IRSD quintile (quintile 1) should be assigned a value of 1, while those in quintile 2 should be assigned a value of 2, and so on up to a value of 5 for the highest quintile.

#### Indicator(s) and indicator levels

In case the indicator is divided into different categories, the cleaned dataset must have:

* A column for the levels of the indicator, with a name based on the indicator's title in the indicator tracking sheet. The column name should be easy to understand and display the categories or levels for the indicator in question. The levels should be listed as separate rows in the indicator level column.
* Column(s) for n and p values (if the indicator doesn't need "p" values, include "n" values only.). These columns should indicate the number and percentage of individuals in each category or level.
* Both n and p values must be rounded to two decimal places.

If the indicator isn't divided into categories, the cleaned dataset should include:

* An indicator column with a human-readable name (the tracking sheet includes indicator names) and a prefix n\_ or p\_ to indicate whether the value represents a count or a proportion.
* If the rate or proportion values are expressed as whole numbers in the raw data, they must be converted to decimals (by dividing by 100) in the cleaned data.

#### Missing values and suppressed cells

When **missing data** is encountered, missing values should be recoded as "NA" to indicate that the data is not available.

**Cell suppression** occurs when a cell contains sensitive or confidential information that cannot be released due to privacy concerns. When cell suppression is applied, the value in the cell is replaced with a 9999999 which indicate that the value is suppressed. Cell suppression is also applied when there is too much uncertainty associated with a particular value.

Table 2, in section Temporal Correspondence provides a summary of the dataset names and their corresponding indicator codes. The table also indicates the cell suppression rules that were applied to the datasets.

## Preparation of indicators with R and R scripts

R scripts are an essential part of the data cleaning process, as they are used to transform the raw data into a format that is suitable for import into the VISER database. These scripts are written in the R programming language and contain a series of commands and functions that manipulate the data.

The ANCHDA Data Cleaning [repository on GitHub](https://github.com/of2/ANCHDA_Data_Cleaning/tree/main) contains R scripts for each indicator that is being cleaned. These scripts are organized by indicator and are named according to the corresponding indicator code.

The R scripts use various functions to reshape the data into tables that are suitable for importing into the database, and to aggregate and summarize the data as necessary.

The use of R scripts helps to ensure consistency and reproducibility in the data cleaning process. By documenting each step of the data cleaning process in the scripts, it is easier to review and audit the process, and to identify and correct errors. Additionally, the scripts can be easily modified or updated as needed, which helps to streamline the data cleaning process and improve its efficiency.

## Quality Control

Quality Control is a critical step in the data analysis process that ensures the accuracy and integrity of the data. The process involves reviewing and validating the outputs of the analysis to identify formatting errors and inconsistencies in the data. A recommended way to review the outputs is through a report markdown file (RMD).

The ANCHDA Data Cleaning [repository on GitHub](https://github.com/of2/ANCHDA_Data_Cleaning/tree/main) contains RMD scripts for each indicator that is being reviewed. These scripts are organized by indicator and are named according to the corresponding indicator code.

The RMD file provides a systematic way of validating the analysis outputs by providing a set of rules and standards to check the data quality. For instance, the reviewer can define a list of acceptable values for the first column, a function to check if column names are in snake case format, regex patterns to check the age\_group format, year\_range, and calendar\_year, and so on.

Additionally, the RMD file enables the reviewer to open a connection to the output file and set the working directory and print the number and names of the CSV files in the folder. After initializing an empty list to store the data frame, the reviewer can read each CSV file into a data frame and store it in the df\_list, appending the data frame to the list using the file name as the key.

Finally, the reviewer can loop through each data frame in the list and perform a set of checks. They can check if age\_group, sex, calendar year, or year\_range columns exist, and if their values are in the correct format. They can also check if the geography column is one of the acceptable values and print the unique values of age\_group, calendar\_year, year\_range, and sex. By performing these checks, the reviewer can ensure the accuracy and reliability of the data analysis outputs.

## Cell Suppression

Cell suppression refers to the practice of obscuring or removing certain cells in a dataset to protect sensitive information or ensure data privacy. This technique is commonly used when analysing datasets that contain personally identifiable information.

In this project, cell suppression is performed by replacing count and proportion values with 9999999 to indicate that the data is suppressed. The exact approach to cell suppression depends on the level of data privacy required. Table 1 outlines the cell suppression rules applied to the datasets and indicators prepared for ANCHDA.

The general rule is to suppress where n < 5, meaning that if a count or proportion value is based on less than 5 observations, it should be suppressed to prevent identification of individuals.

In addition, if high uncertainty is detected in the data, the corresponding count or proportion values are also suppressed.

Cell suppression occurs at the final stage of the data preparation process. Once the indicators have been reviewed, and the format of the data has been validated, the reviewer applies the cell suppression function to

Table 2 - overview of cell suppression rules applied

|  |  |  |
| --- | --- | --- |
| Dataset name | Indicator names and code | Cell suppression rules |
| Dataset 1 | Indicator 1 | n < 5 |
|  |  |  |
|  |  |  |

## Temporal Correspondence

Temporal correspondence refers to the process of ensuring that datasets are compatible with a common temporal reference framework. In this project, we have converted datasets from different temporal frameworks to the Australian Statistical Geography Standard (ASGS) 2016 to ensure that they are comparable and can be used for spatial analysis.

Explain what the R script used for conversion do – add link here

Table 3 - overview of datasets converted to ASGS 2016

|  |  |  |
| --- | --- | --- |
| Dataset Name | Original ASGS | Converted to ASGS |
| Dataset 1 | ASGS 2011 | ASGS 2016 |
|  |  |  |
|  |  |  |

# Data documentation

Data documentation is a critical component of any data analysis project. In this project, several steps were taken to ensure that the data is well-documented and can be easily understood and reproduced by others.

## Metadata

A metadata table was created for each dataset used to generate indicators. The metadata tables include information such as the source of the data, date of collection, data format, location of the data {other??}. This metadata is important for tracking the provenance of the data and understanding any potential limitations or biases in the data.

Add link to metadata folder

## Data dictionary

The data dictionary serves as a reference table that lists all the variable names from the datasets created throughout the data cleaning process, along with a brief description of what each variable represents.

Add link to data dictionary file

1. If the dataset includes rows with "Australia" or "0" in the "STE\_CODE16" field, these rows should not be included in the state dataset. Instead, they must be exported to a new table and saved as an *Australia cleaned dataset*. [↑](#footnote-ref-1)