May 2023

Technical documentation ANCHDA datasets/indicators

prepared by the data science team

Contents

[Chapter 1. Introduction 3](#_Toc135041413)

[Background Information 3](#_Toc135041414)

[The Extract, Transform, and Load (ETL) Process 3](#_Toc135041415)

[Purpose of the Technical Report 4](#_Toc135041416)

[Chapter 2. Data Collection – the “Extract” phase 5](#_Toc135041417)

[Data Sources 5](#_Toc135041418)

[Data Collection/ Data Acquisition Methods 5](#_Toc135041419)

[Data Retrieval from APIs 5](#_Toc135041420)

[Data Purchase and Licensing 5](#_Toc135041421)

[Data Quality Assessment 5](#_Toc135041422)

[Data Transfer and Storage 5](#_Toc135041423)

[Data Privacy and Legal Considerations 5](#_Toc135041424)

[Compliance with Data Protection Laws 5](#_Toc135041425)

[Anonymization and Pseudonymization Techniques 5](#_Toc135041426)

[Data Security and Confidentiality 5](#_Toc135041427)

[Chapter 3. Data cleaning – the “Transformation” phase 6](#_Toc135041428)

[Objectives 6](#_Toc135041429)

[Data Transformation 6](#_Toc135041430)

[Standardisation and Consistency 6](#_Toc135041431)

[Data Quality Assurance 6](#_Toc135041432)

[Documentation and Reproducibility 6](#_Toc135041433)

[Specifications for cleaning the data 6](#_Toc135041434)

[Data cleaning workflow and tools 9](#_Toc135041435)

[Data cleaning workflow 9](#_Toc135041436)

[Data Cleaning Tools and Software 9](#_Toc135041437)

[Validation and Quality Control 10](#_Toc135041438)

[Cell Suppression 10](#_Toc135041439)

[Temporal Correspondence 11](#_Toc135041440)

[Chapter 4. Data Documentation and Metadata 12](#_Toc135041441)

[Metadata Standards and Formats 12](#_Toc135041442)

[Documentation of Data Sources and Collection Methods 12](#_Toc135041443)

[Variable Descriptions and Data Dictionary 12](#_Toc135041444)

[Chapter 5. Data Storage and Organization 13](#_Toc135041445)

[Database Design and Schema 13](#_Toc135041446)

[Data Indexing and Retrieval Methods 13](#_Toc135041447)

[Data Backup and Recovery Strategies 13](#_Toc135041448)

[Data Versioning and Change Tracking 13](#_Toc135041449)

[Chapter 6. Data Validation and Verification 14](#_Toc135041450)

[Data Validation Techniques 14](#_Toc135041451)

[Conclusion 15](#_Toc135041452)

[Summary of Data Collection and Preparation Process 15](#_Toc135041453)

[Challenges and Lessons Learned 15](#_Toc135041454)

[Recommendations for Future Data Collection Efforts 15](#_Toc135041455)

[References 16](#_Toc135041456)

[Citations for External Resources 16](#_Toc135041457)

[Glossary of Terms 16](#_Toc135041458)

# Chapter 1. Introduction

## Background Information

The Australian National Child Health and Development Atlas is an ambitious project aimed at providing comprehensive insights into the health and development of children across Australia. Spearheaded by the Australian Research Data Commons (ARDC), this project brings together a wide range of data sources and expertise to create a robust and multidimensional atlas. The overarching goal is to facilitate evidence-based decision-making and policy development to improve child health outcomes.

The project revolves around the collection, integration, and analysis of diverse data sets, including demographic information, health indicators, socio-economic factors, and educational outcomes. By combining these data sources, the atlas offers a holistic view of the factors that influence child health and development, allowing researchers and policymakers to identify patterns, trends, and areas of concern.

[placeholder here – an overview of indicators and domains in anchda]

The Australian National Child Health and Development Atlas aims to provide a user-friendly platform that enables stakeholders, policymakers, and researchers to explore the data and gain insights into child health and development outcomes at various geographic levels. The interactive nature of the atlas allows for in-depth analysis, comparison, and visualization of data, empowering users to identify areas for intervention and policy formulation.

With its comprehensive approach, the atlas serves as a valuable resource for researchers, policymakers, and organizations working in the field of child health and development. By shedding light on the diverse factors that influence child well-being, the project contributes to evidence-based decision-making, enabling targeted interventions and policies to promote healthier and more prosperous futures for children across Australia.

## The Extract, Transform, and Load (ETL) Process

“ETL” stands for Extract, Transform, and Load, which is a process commonly used in data integration and data warehousing. ETL refers to the steps involved in extracting data from various sources, transforming it into a consistent and usable format, and loading it into a target system or database.

The first step, "Extract," involves retrieving data from multiple sources such as databases, files, APIs, or web scraping. The data is extracted in its raw or original form, without any modifications.

The second step, "Transform," focuses on cleaning, restructuring, and standardizing the extracted data. This includes tasks such as data cleaning, data validation, removing duplicates, handling missing values, and converting data types. Data transformation also involves applying business rules, calculations, and aggregations to prepare the data for analysis.

The final step, "Load," involves loading the transformed and processed data into a target system or database. This could be a data warehouse, a data mart, or any other storage platform designed for efficient data querying and analysis. The data is structured and organized in a way that facilitates easy access, retrieval, and analysis by end users or applications.

**The ETL process is crucial for ensuring data quality, consistency, and reliability. It helps to integrate data from various sources, harmonize different data formats, and provide a unified view of the data. ETL pipelines are often automated to streamline the process and ensure regular updates of the data.**

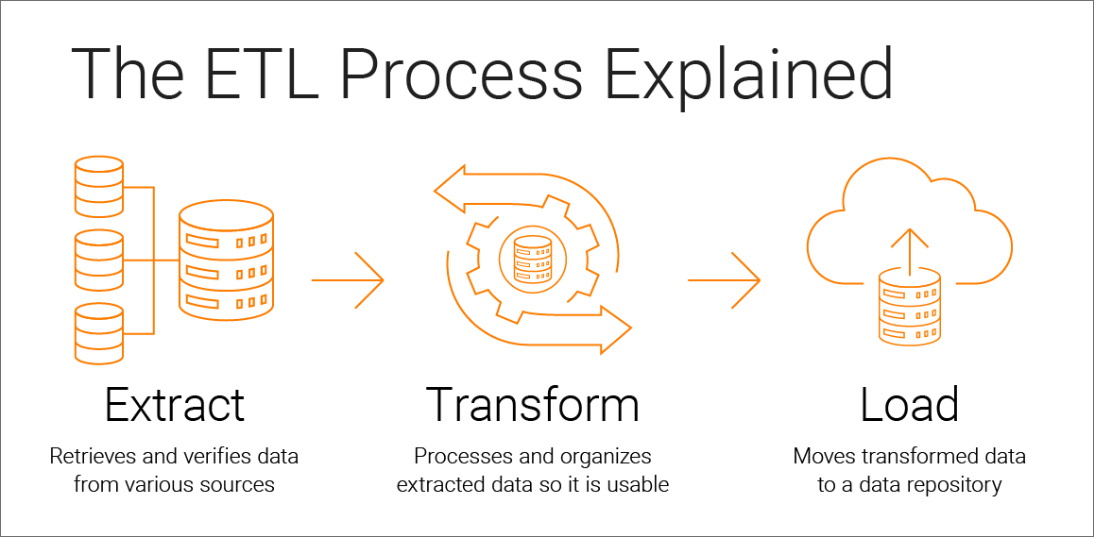


Figure 1 – Visual representation of the Extract, Transform, Load process

## Purpose of the Technical Report

The purpose of this technical report is to provide a comprehensive overview of the data collection, data cleaning, and data processing activities conducted during the project, with a particular focus on the “Extract” (Chapter 2) and "Transform" (Chapter 3) phase of the Extract, Transform, Load (ETL) process.

By documenting these essential steps, the report aims to ensure transparency, reproducibility, and accountability in the data analysis process. It serves as a reference document for the project team, stakeholders, and future researchers, outlining the methodologies, tools, and techniques employed to transform the raw data into a format suitable for analysis.

In Chapter 2, the report covers the "Extract" phase, which involves gathering data from various sources and consolidating it into a unified dataset. It highlights the data sources utilized, the methods employed to extract the relevant information, and any data acquisition challenges encountered during this phase. The report also addresses the considerations taken to ensure data integrity and reliability during the extraction process.

In Chapter 3, the report focuses on the "Transform" phase, which involves cleaning, standardizing, and enhancing the raw data to make it suitable for analysis. It discusses the data cleaning techniques applied, such as handling missing values, outlier detection, and data validation. The report also outlines the data transformation processes, including data normalization, encoding, and feature engineering, which aim to improve the quality and usability of the data. It highlights the tools, software, and algorithms used for these transformations, providing insights into the decision-making processes involved.

Furthermore, the report showcases the utilization of specific tools and software, such as R and its associated packages, for data transformation. It discusses the custom scripts and functions developed to automate and streamline the transformation process, ensuring efficiency and consistency. The report also highlights the documentation practices employed, such as using R Markdown to create detailed and reproducible reports that capture the transformations applied to the data.

By providing a detailed account of the "Extract" and "Transform" phases in the ETL process, the report enables stakeholders to understand the methodologies and techniques employed to extract and prepare the data for analysis. It facilitates the assessment of the data quality, integrity, and reliability, allowing for confident decision-making and meaningful analysis.

# Chapter 2. Data Collection – the “Extract” phase

## 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/ Data Acquisition Methods

### Data Retrieval from APIs

### Data Purchase and Licensing

* 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 Quality Assessment

* Source Reliability and Trustworthiness
* Data Completeness and Accuracy
* Data Consistency and Timeliness

## 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 Privacy and Legal Considerations

### Compliance with Data Protection Laws

### Anonymization and Pseudonymization Techniques

## 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.

# Chapter 3. Data cleaning – the “Transformation” phase

## Objectives

The overarching goal of the data transformation process in this project is to generate clean and standardized CSV files for each indicator. These CSV files are specifically designed to be easily imported into a database system, enabling seamless data integration, and facilitating subsequent analysis. By achieving this objective, we ensure that the CSV files are formatted consistently, free from errors or inconsistencies, and optimized for efficient querying, exploration, and analysis. This standardized approach to data cleaning not only enhances the usability of the data but also contributes to the overall reliability and integrity of the project's outcomes. To achieve this goal, we have defined the following objectives:

### **Data Transformation**

The first objective is to transform the raw data into a structured format that adheres to predefined guidelines and specifications. This includes standardising variable formats, resolving inconsistencies, and handling missing values. By transforming the data, we ensure its compatibility with the database system and facilitate efficient data integration.

### **Standardisation and Consistency**

The second objective is to standardise the data and ensure consistency across variables. This involves harmonizing variable names, categorizations, and coding schemes. By establishing consistent conventions, we enhance the usability and reliability of the dataset, enabling accurate and meaningful analysis.

### Data Quality Assurance

Another objective is to ensure the overall quality and integrity of the data. This includes identifying and resolving data anomalies, such as outliers and data entry errors. By implementing validation and quality control measures, we verify the accuracy and reliability of the data, minimising potential biases and inconsistencies.

Documentation and ReproducibilityThe final objective is to document the data cleaning process thoroughly and ensure reproducibility. By providing clear documentation of the steps taken, transformations applied, and decisions made during data cleaning, we promote transparency and enable future replication of the process. This enhances the reliability and trustworthiness of the data for both internal and external stakeholders.

## Specifications for cleaning the data

The following section outlines the specific requirements for cleaning the data in this project, with the ultimate goal of generating a clean and structured CSV file for each indicator derived from the raw data. These specifications ensure that each CSV file contains accurate, consistent, and well-formatted information that is ready for further analysis.

To achieve this goal, several key considerations are taken into account during the data cleaning process. Firstly, field names are standardized and descriptive, allowing for easy identification and interpretation of the variables within each CSV file. Data types are appropriately assigned to reflect the nature of the data, ensuring compatibility with subsequent analysis tasks.

Data validity and integrity are crucial aspects of the cleaning process. Each field's values are subjected to predefined rules and constraints, ensuring that the data is accurate and reliable. Missing values are addressed, either through imputation or appropriate treatment, to maintain the completeness of the dataset. Inconsistencies and outliers are identified and resolved, ensuring the overall integrity and quality of the data.

Furthermore, the order and structure of the data are carefully arranged to facilitate efficient analysis. Chronological or logical ordering of time-related information, for example, helps provide a coherent framework for interpreting the data.

By adhering to these specifications, the data cleaning process guarantees that each indicator derived from the raw data is transformed into a well-structured and standardized CSV file. This ensures that the subsequent analysis can be performed seamlessly, allowing for reliable insights and informed decision-making.

In the next sub-sections, we present a detailed overview of each field name and its corresponding format as per the specifications for cleaning the data. This comprehensive exploration sheds light on the specific variables present in the cleaned datasets, their data types, and the formatting guidelines employed to ensure consistency and accuracy.

#### 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:

* Australia
* STE\_CODE16
* SA2\_CODE16
* SA3\_CODE16
* SA4\_CODE16
* LGA\_CODE16

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 | 5-5 |
| Year 1 | 6-6 |
| Year 2 | 7-7 |
| Year 3 | 8-8 |
| Year 4 | 9-9 |
| Year 5 | 10-10 |
| Year 6 | 11-11 |
| Year 7 | 12-12 |
| Year 8 | 13-13 |
| Year 9 | 14-14 |
| Year 10 | 15-15 |
| Year 11 | 16-16 |
| Year 12 | 17-17 |

#### 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 "NULL" to indicate that the data is not available.

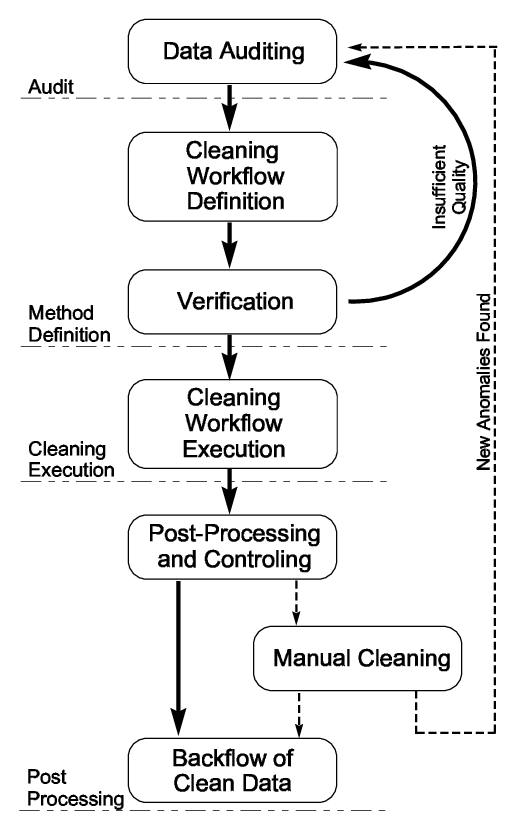
**Cell suppression** occurs when a cell contains sensitive or confidential information that cannot be released due to privacy concerns. When cell suppression is applied, the value in the cell is replaced with a 9999999 which indicate that the value is suppressed. Cell suppression is also applied when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

## Data cleaning workflow and tools

The following sections details the comprehensive data cleaning workflow and the tools utilised to ensure the integrity and quality of the datasets.

### Data cleaning workflow



Template to adapt..

### Data Cleaning Tools and Software

We chose to use R as the primary tool for data cleaning in this project due to its wide range of capabilities and its suitability for statistical computing and data manipulation tasks. R provides a comprehensive set of packages and libraries specifically designed for data cleaning, such as dplyr and tidyr, which offer efficient and intuitive functions for data wrangling and transformation. With its extensive ecosystem, R allows us to address various data cleaning challenges effectively, including handling missing values, outlier detection and treatment, data integration, and aggregation. Furthermore, R's scripting capabilities enable us to automate and document the entire data cleaning process, ensuring consistency and reproducibility. The ability to work with large datasets and integrate seamlessly with other statistical and visualization packages makes R a valuable tool for comprehensive data cleaning workflows. By utilizing R, we streamline our data cleaning tasks, improve the quality and accuracy of the prepared data, and enhance the overall reliability of our project results.

To facilitate the data cleaning process and promote collaboration, we established the ANCHDA Data Cleaning [repository on GitHub](https://github.com/of2/ANCHDA_Data_Cleaning/tree/main). This repository serves as a centralized hub for all the R scripts used in the data cleaning phase. Each indicator being cleaned has its own dedicated script, organised, and named according to the corresponding indicator code. This structured approach allows for easy navigation and access to the specific scripts required for cleaning individual indicators. By utilizing the ANCHDA Data Cleaning repository, team members can efficiently share, review, and update the R scripts, promoting consistency and reproducibility across the project.

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.

## Validation and Quality Control

To ensure the accuracy and quality of our data, we employ validation and quality control measures throughout the project. One of the key tools we utilize for this purpose is R Markdown. R Markdown provides us with a streamlined and automated approach to validate the formatting of our outputs and identify potential outliers in the data. By leveraging R Markdown's capabilities, we can define specific rules and standards to check the quality of our outputs. We develop custom functions and scripts within R Markdown to verify the formatting of the data, ensuring adherence to predefined guidelines and specifications.

Additionally, R Markdown allows us to create comprehensive RMD (R Markdown Document) files that serve as a systematic and reliable way to validate the analysis outputs. These RMD files contain a set of predefined rules and standards that we apply during the validation process. For example, we define a list of acceptable values for the first column and use functions to check if column names are in the appropriate snake case format. We also employ regex patterns to validate the formats of age\_group, year\_range, calendar\_year, and other relevant fields.

In addition to rule-based validation, the RMD files facilitate outlier detection in the data. We integrate outlier detection algorithms within the R Markdown framework to automatically identify any unusual or anomalous data points that deviate significantly from the expected patterns. This allows us to flag potential outliers for further investigation and ensure data integrity.

Moreover, the RMD files provide convenient features to facilitate the validation and quality control process. We utilize the RMD file's ability to establish a connection to the output file, set the working directory, and retrieve information about the CSV files in the designated folder. By opening a connection to the output file, we can easily access and validate the generated results. Additionally, we leverage the power of R Markdown to create loops that iterate through each data frame in the output files. Within these loops, we implement a series of checks to verify the presence and correct formatting of crucial columns such as age\_group, sex, calendar year, and year\_range. Furthermore, we utilize the RMD file's capability to print the unique values of age\_group, calendar\_year, year\_range, and sex, enabling us to gain insights into the distribution and characteristics of the data.

By employing R Markdown for validation and quality control, we establish a robust and systematic approach to ensure the accuracy, reliability, and consistency of our data analysis outputs. The combination of custom rules, outlier detection algorithms, and the flexibility of R Markdown's features provides us with a powerful framework to validate and verify our data. This comprehensive approach not only saves time and effort but also instils confidence in the integrity of the project outcomes, delivering reliable and trustworthy data to the VISER team.

## Cell Suppression

In this project, we employ the practice of cell suppression to protect sensitive information and ensure data privacy. Cell suppression involves obscuring or removing specific cells in a dataset that may contain personally identifiable information. This technique is crucial when working with datasets that require careful handling of privacy concerns.

To implement cell suppression, we adopt a standardized approach in this project. Count and proportion values below a certain threshold are replaced with the value **9999999**, indicating that the data has been suppressed. The specific rules for cell suppression depend on the level of data privacy required. In Table 2, we provide a comprehensive overview of the cell suppression rules applied to the datasets and indicators prepared for ANCHDA.

As a general guideline, we suppress count or proportion values when the number of observations (n) falls below 5. This threshold ensures that individual identification is prevented. Additionally, if high uncertainty is detected in the data, we also apply cell suppression to the corresponding count or proportion values.

Cell suppression is performed during the final stage of the data preparation process. After the indicators have undergone thorough review and the data format has been validated, the reviewer applies the cell suppression function to the dataset. This step ensures that sensitive information is adequately protected while maintaining the integrity and quality of the data.

By implementing cell suppression techniques, we prioritize data privacy and adhere to ethical considerations, safeguarding the confidentiality of individuals within the dataset. This practice enhances the trustworthiness and reliability of our analyses and findings, allowing us to confidently present meaningful insights while respecting privacy regulations and obligations.

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 xxxx | ASGS 2011 | ASGS 2016 |
|  |  |  |
|  |  |  |

# Chapter 4. Data Documentation and Metadata

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 Standards and Formats

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

## Documentation of Data Sources and Collection Methods

In this project, meticulous documentation of data sources and collection methods plays a critical role in ensuring transparency, traceability, and accountability throughout the data cleaning process. To achieve this, we have employed an Excel spreadsheet as a comprehensive tool to record and track the status of each dataset. The Excel spreadsheet is securely saved on a OneDrive platform, ensuring that it is accessible only to the project team members. To maintain data confidentiality and protect sensitive information, the OneDrive is safeguarded with a login email and password authentication system. This security measure ensures that only authorized individuals within the project team can access and contribute to the documentation process, enhancing data privacy and maintaining strict control over the documentation contents.

The Excel spreadsheet serves as a centralized repository to capture essential information related to the datasets. It includes details such as the dataset name, custodian contact information, the date when the data was received, and any specific instructions or requirements provided by the custodian. This comprehensive documentation enables us to maintain a clear record of the dataset's origin and the custodian's involvement.

Furthermore, the Excel spreadsheet allows us to document any problems or issues encountered during the data cleaning process. This includes the detection of data anomalies, missing values, or inconsistencies. We record the nature of the problem, the steps taken to address it, and the individual responsible for cleaning the data. By documenting these details, we ensure that data cleaning tasks are properly assigned, and accountability is maintained throughout the process.

Moreover, the Excel spreadsheet serves as a log to track the progress of each dataset. It records the dates when the data was cleaned, reviewed, and approved. This information helps in establishing a timeline of activities, facilitating coordination among team members, and ensuring timely completion of the data cleaning process.

By using the Excel spreadsheet for documentation, we establish a structured and organized approach to track and monitor the status of each dataset. This promotes transparency and accountability, allowing us to maintain a clear record of the data sources, collection methods, problems encountered, and the individuals responsible for cleaning and reviewing the data. Ultimately, this documentation process contributes to the overall reliability and integrity of the project's data.

Add link to project tracker

## Variable Descriptions and 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

# Chapter 5. Data Storage and Organization

## Database Design and Schema

## Data Indexing and Retrieval Methods

## Data Backup and Recovery Strategies

## Data Versioning and Change Tracking

# Chapter 6. Data Validation and Verification

## Data Validation Techniques

* Expert review / peer-review – pilot and testing phase

# Conclusion

## Summary of Data Collection and Preparation Process

## Challenges and Lessons Learned

## Recommendations for Future Data Collection Efforts

# References

## Citations for External Resources

## Glossary of Terms

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)