GRIFFITH UNIVERSITY

Local Protocol

Griffith Data Trust - Data Extraction and Sanitisation Protocol

1.0 Purpose

This Local Protocol establishes standardised procedures for the secure and compliant removal of data, safeguarding sensitive information during the extraction process. It aligns with organisational security policies, regulatory requirements, and industry best practices to maintain data integrity and confidentiality.

2.0 Scope

This Local Protocol applies to all staff and authorised users involved in data extraction and sanitisation activities within the Griffith Data Trust. Understanding and adhering to these guidelines is essential to safeguarding the facility's data assets.

This Local Protocol applies subject to any requirements of a Data Sharing Agreement, which will prevail over this document.

3.0 Local Protocols

3.1 Data Extraction

A data extraction policy from the data provider must be accepted by all parties before data is brought into the facility.

If the data is provided under the DATA scheme, extraction can only occur if expressly permitted in the Registered Data Sharing Agreement. Data extraction is subject to the data sharing agreement with the Commonwealth entity allowing it. This ensures personal information is protected against loss, unauthorised access, use, modification, disclosure, and other misuse.

Data providers may have stricter extraction requirements than the Data Sanitisation Policy but never more lenient. For example, if the rule of 10 applies, a data providing party might request a minimum of 20 contributing numbers but never fewer than 10.

3.1.1 Data Extraction procedure

- The sanitisation/data extraction procedure required to remove data from the facility or the research zone to a teaching zone within the facility is as follows:
  1. Users ensure data has been appropriately sanitised in accordance with the appropriate data extraction policy.
  2. Users submit a ticket, requesting data extraction.
  3. Administrative staff check the data to ensure it meets the appropriate data extraction policy.

4. Administrative staff copy approved data to a removable drive.
5. Administrative staff archive removed data to secure storage for historical evidence.
6. Administrative staff closes the ticket, logging the actions taken.

3.2 Data Sanitisation

Prior to data being exported out of the facility it needs to be sanitised, aggregated, and anonymised to prevent re-identification of individual contribution data points. The Griffith Data Trust follows ABS DataLab input and output clearance rules for secure data export. Refer to Input and output clearance | Australian Bureau of Statistics (abs.gov.au).

3.2.1 Output type rule

The most common types of analysis and their applicable rules are listed below. Other output types will be assessed based on similar principles.

Output type	Applicable Rules
Frequency tables (counts, percentages)	Rule of 10 Group disclosure
Magnitude statistics (means, sums, ratios)	Rule of 10 Group disclosure Dominance
Quantiles (percentiles, medians)	Minimum contributors for quantiles
Minimums, maximums, ranges	Minimum contributors for quantiles
Models incl. regressions	Degrees of freedom Model- specific rules
Charts (graphs, plots, and histograms)	Chart clearance
Microdata	Not appropriate for output
Synthetic microdata	Not appropriate for output

3.2.2 Rule of 10

- Applies to most outputs including counts, percentages (both numerator and denominator), means, sums, and other statistics. It is the minimum number of contributors required for each cell or statistic. The underlying (unweighted) count of observations must meet this threshold, and evidence must be provided.

- When producing multiple data tables, differences of less than ten should not be able to be calculated by combining the tables.
- To protect the confidentiality of data, use methods such as suppression of small counts, category aggregation, or data perturbation. If a suppressed cell can be derived from other outputs, suppress additional values to protect the primary suppressed cell's value from being worked out.

3.2.3 Dominance

- Prevents re-identification of units contributing a large percentage of a cell's total value, which could reveal information about individuals, households, or businesses.
- Follow the (1,50) and (2,67) rules, where the largest contributor cannot account for more than 67% of the total value.
- Where a variable can take both positive and negative values, the negative values should be replaced with absolute values to determine if the (1,50) rule is met, and the sum of the two largest absolute values to check the (2,67) rule.
- If the dominance rule fails and a cell is suppressed but can be derived from other outputs, suppress additional values to protect the primary suppressed cell from being worked out.
- Dominance must be checked if any mean, total, or similar statistic is calculated for continuous or magnitude variables. It does not apply to counts.

3.2.4 Group disclosure

- Applies to frequency tables when all or nearly all units with one feature have another feature. When individual units may appear protected based on other rules, a previously unknown attribute of a unit may be disclosed based on the attributes of the group.
- Assess risk when any cell contains more than 90% of the total number of units in the row or column.

3.2.5 Minimum contributors for quantities

- Quantiles and other ranks must be based on a minimum number of contributors depending on the precision. Provide underlying unweighted counts when reporting quantiles in the outputs. Required contributors for quantiles are:

Quantile	Minimum contributors
Medians (0.50)	10
Quartiles (0.25, 0.5, 0.75)	20
Quintiles (0.2, 0.4, 0.6, 0.8)	25
Deciles (0.1, 0.2, 0.3 ... 0.9)	50
Vigintiles (0.05, 0.1, 0.15 ... 0.95)	100
Percentiles (0.01, 0.02 ... 0.99)	500

- Minimums and maximums are generally unsafe to output. The following percentiles are safe options if the minimum contributors' rule is satisfied:
    - 1st and 99th percentiles
    - 5th and 95th percentiles
    - 10th and 90th percentiles

3.2.6 Degrees of freedom

- Calculate degrees of freedom by subtracting the number of parameters and model restrictions from the total observations that contribute to the model. Models and regressions must have a minimum of 10 degrees of freedom, with evidence provided.

3.2.7 Model-specific rules

- Specific models have additional rules.
- For ordinary least squares regressions, the R-squared should be lower than 0.9. If higher, the constant may need to be suppressed to prevent predictions.
- For ordinary least squares regressions with a continuous dependent variable with only categorical independent variables, the regression will approximate the tabular means. To reduce disclosure risk, add a continuous independent variable or suppress the intercept. Otherwise, apply the rule of 10 and dominance rules.
    - For survival curves, each step change should represent at least 10 data subjects.
    - Correlation and Gini coefficients should be calculated based on a minimum of 10 contributors.
    - For classification and regression trees, underlying unweighted counts must meet the rule of 10.

- For other models, provide evidence that no estimates or parameters are derived from fewer than 10 underlying contributors and explain why the output is non-disclosive.

3.2.8 Chart clearance

- All graphs, plots, and other charts are subject to the output rules of the underlying output type. The data used in the chart must be provided, along with any relevant supporting evidence that it meets output rules.
- Charts plotting characteristics of fewer than 10 units will not be cleared.

4.0 Definitions

For the purposes of this policy and related policy documents, the following definitions apply:

Data refers to raw, unorganised facts such as numerical figures, words, or characters. This term may occasionally be used interchangeably with the term ‘information’.

Data provider is an organisation that provides internal data assets to external parties.

Data Sanitisation is the disciplined process of deliberately, permanently, and irreversibly removing or destroying the data stored on a memory device to make it unrecoverable.

Microdata is unit-level data obtained from sample surveys, censuses, and administrative systems.

Synthetic data is annotated information that computer simulations or algorithms generate as an alternative to real-world data.

5.0 Information

Title	Data Extraction and Sanitisation Protocol
Document number	Provided by relevant Policy team
Purpose	This Local Protocol establishes standardised procedures for the secure and compliant removal of data, safeguarding sensitive information during the extraction process. It aligns with organisational security policies, regulatory requirements, and industry best practices to maintain data integrity and confidentiality.
Audience	GDT Staff
Category	Academic
Subcategory	Research
UN Sustainable Development Goals (SDGs)	This document aligns with Sustainable Development Goal/s: 16: Peace, Justice and Strong Institutions
Approval date	October 2024
Effective date	October 2024
Review date	2025
Policy advisor	GDT Policy and Compliance Coordinator
Approving authority	Director, Griffith Data Trust

6.0 Related Policy Documents and Supporting Documents

Legislation	*Data Availability and Transparency Act 2022* (Cth) (DAT Act)

Policy	Griffith Data Trust Policy
Procedures	Griffith Data Trust Procedure
Local Protocol	Griffith Data Trust Committee Terms of Reference Griffith Data Trust Data Governance and Management Protocol Griffith Data Trust Security Management Protocol Griffith Data Trust Human Resource Skills and Capability Management Protocol
Forms	General Staff Agreement Form Academic Researcher Agreement Form