Data Quality Guideline

Section 1 - Purpose

(1) This guideline provides guidance to staff on assessing and improving data quality at RMIT and implementing RMIT's data quality standard in practice.

Section 2 - Authority

(2) Authority for this document is established by the Information Governance Policy.

Section 3 - Scope

(3) This guideline applies to all RMIT staff, including temporary employees, contractors, visitors and third parties globally who create, manage or use RMIT data, with the exception of research data as defined by the Research Policy.

Section 4 - Guideline

Overview

(4) RMIT must continuously and proactively define, assess, improve and monitor the quality of its data.

(5) High quality data supports University strategy and enables quality insights and decision-making, driven by data that is trusted.

(6) The data quality resources i.e. the Data Quality Guideline, the Data Quality Standard and tools and templates are part of RMIT's Data Quality Framework, and support the establishment of Data Quality Management as a key capability at RMIT.

Assessing and Improving Data Quality

(7) Improving enterprise data quality is a collaborative effort involving Information Trustees and Information Stewards as well as data producers and consumers across the University, working together to raise the standard of enterprise data quality at RMIT.

(8) Data that is of high-value or high risk to RMIT (e.g. master data) must be identified and prioritised for data quality improvement. The Master Data Management Standard provides additional guidance on the management of master data. Risks associated with poor quality data must be escalated via risk and governance processes.

(9) Data quality assessment and improvement activities must be undertaken for all high priority data. These activities will be guided by the Information Stewards, Information Trustees, the Information Governance Board and the Data and Analytics team.

(10) Improving enterprise data quality is not a one-time process, but a continuous and iterative process that is

described below.

Step 1: Define Data Quality

(11) Define and agree upon the data quality goals, expectations, rules and requirements based on business criticality, in collaboration with all stakeholders.

a. Identify data assets, their supporting attributes and their priority based on business impact.
b. Clearly define and document what acceptable or good enough data quality looks like using RMIT's Data Quality Dimensions - completeness, validity, accuracy, consistency, timeliness and uniqueness, as per the Data Quality Standard.
c. Engage with all stakeholders impacted by the quality of the data across the University.

Step 2: Assess and Measure Data Quality

(12) Data quality must be systematically assessed against data quality rules and expectations from Step 1 so that it can be improved.

(13) Enterprise data quality metrics must be developed and implemented for all high priority data at RMIT via the RMIT Data Quality Dashboard. This dashboard is a resource available to all staff that enables RMIT to assess, measure and monitor enterprise data quality.

Step 3: Identify and Plan Improvements

(14) A Data Quality Improvement Plan for RMIT's high priority data must be developed in collaboration with the Information Stewards responsible for managing their data quality and endorsed by the Information Trustee accountable for the domain.

(15) The plan will provide details on the assessment of current data quality, the desired state of data quality, take into consideration data quality issues and identify improvement actions (e.g. changes to roles and responsibilities, improved processes, new tools or technology, training, introduction of standards or naming conventions, data cleansing activities, etc).

(16) The Data Quality Improvement plan must be developed in consultation with the Information Custodians and impacted stakeholders and endorsed by the Information Trustee accountable for that domain.

(17) If the improvement actions require an enterprise-funded project, a project should be initiated via Enterprise Projects and Business Performance processes, and a Project brief created.

Step 4: Execute Improvement Plans

(18) The endorsed Data Quality Improvement Plan must be executed and the status and progress of the plan monitored and reported.

Step 5: Control Data Quality

(19) The following controls will ensure that the desired data quality improvement is achieved and maintained over time.

a. Data quality will be monitored by the Information Stewards, Information Trustees and the Data and Analytics team.
b. Data quality will be tracked using the RMIT Data Quality Dashboard.
c. The status and progress of Data Quality Improvement Plans will be communicated regularly to stakeholders via Governance forums ie., the Data Quality Stewards Group and the Information Governance Board.

d. The Information Stewards will periodically review and reassess data to ensure it is consistent with the data quality expectations that were identified in Step 1, and will periodically review Data Quality Improvement Plans.

Step 6: Addressing Data Quality Issues

(20) Data quality issues impacting high priority data must be registered in the RMIT Data Quality Issues Register. This ensures that these issues can be effectively triaged, analysed and addressed by the Information Stewards.

(21) The Data and Analytics team will work with the Information Stewards to provide guidance on the Data Quality Issues Management Process including undertaking root cause and impact analyses.

(22) Data quality issues must be remediated at their source i.e., the point of entry, creation, collection or corruption, and data must be corrected prior to use for purposes such as reporting and analytics in accordance with the Data Quality Standard.

Data Quality Improvement Practices

Collecting and Creating Data

(23) Errors may occur at the source of data collection, particularly during manual data entry. Human error and data quality issues must be expected and mitigated by putting in place pre-emptive checks and validation throughout systems and processes to reduce the occurrence of data quality issues.

(24) Validation and data quality checks must be implemented at the point where data is manually created or collected and throughout the data lifecycle, particularly when data is moved, transferred or changes state.

(25) Systems should be designed and implemented to incorporate validation and automation to reduce errors.

(26) Data quality that does not meet expectations must be identified and corrected at the source so that incorrect data is not propagated further. Data must be corrected prior to its use for purposes such as analytics and reporting.

(27) Staff should be trained to follow the correct processes and made aware of the importance of the quality of the data.

Promoting a shared understanding of data quality

(28) Clearly defined and agreed definitions of data, its use and its quality, enables data to be meaningful, clear and not left open for misinterpretation by different groups across RMIT.

(29) Staff must collaborate across groups to achieve a common understanding of data. The Information Stewards Group is a forum that can be leveraged to collaborate, define and agree on the meaning and use of data which can then be documented in the Information Domain Register to enable it to be discoverable and understood.

(30) Key terms, metrics, data quality rules, reports and high priority data assets should be catalogued in RMIT's Information Domain Register, which is available to all RMIT staff to promote their discoverability, consistency and reuse.

Using Data Quality Rules

(31) Data quality rules must be clearly defined and implemented addressing all applicable data quality dimensions. These rules may cover:

a. Mandatory or required fields to ensure completeness of data.
b. Key attributes where conformance is required (for data types, format, precision or specific business rules or domain) to ensure the validity of data.

c. Acceptable levels of latency (how often data should be refreshed) to ensure data is timely and up-to-date.
d. Uniqueness so that duplication is minimised.
e. Identifying where verification with the authentic reference or the actual entity is required to ensure the accuracy of the data.

(32) Data must be validated against these rules and validation automated and embedded into systems and processes.

(33) Corrective actions, such as data cleansing should be undertaken where the data does not align with defined rules prior to using the data, where appropriate. Rules may also be used to identify, ignore or correct records where appropriate.

(34) Rules should be centrally documented and stored, accessible by staff and where possible managed in a system. These should be reviewed regularly and updated.

Enabling People to Improve Data Quality

(35) It is important for all staff to understand the importance of data quality and the significant impacts that poor data quality can have, as well as how to improve it.

(36) All staff should have access to clearly defined, visible and discoverable data quality expectations and should ensure that they understand the expected level of data quality.

(37) All staff should be enabled to ensure data quality through effective training and operational documentation, with data quality practices embedded in those resources.

Embedding Data Quality into Processes

(38) Processes should be embedded with data quality practices at all stages of the data lifecycle, from creation to destruction.

(39) Systems should be enabled to automate processes and data quality checks against the data quality rules where practicable.

(40) Processes should be in place to:

a. Validate and correct manually created data against data quality rules.
b. Ensure that data created or manipulated is complete, valid, accurate and timely to an agreed level.
c. Correct identified data quality issues and improve data quality.
d. Provide feedback between staff producing and staff consuming data as well as impacted stakeholders to enable reporting of issues and corrective actions.

(41) Data quality practices should be embedded into operational processes, documentation and training material. Processes should be documented using enterprise tools so that they are discoverable.

(42) Continuous feedback and review processes should be implemented to ensure that any gaps or coverage issues are discovered, resolved and monitored.

Examples of Data Quality Issues and Ways to Mitigate them

(43) This section provides common examples of data quality issues, recommended practices that can be implemented to address them and ensure the quality of data for each dimension

Dimension	Example of Data Quality Issue and its Impact	Suggested Mitigation
Completeness Data must contain values for all expected attributes and instances identified as mandatory.	When a student's phone number is missing in the phone number field in the student admissions systems, then the admissions team will not be able to contact a student for administration purposes.	• Assign data quality rules for mandatory fields to ensure that data created by users is complete. • Put checks in place at the point of data creation for required or mandatory fields that are critical for operational processes. • Implement proactive data quality monitoring with ongoing data quality checks against specific business rules for completeness e.g. using quality dashboards and reports.
Validity Data must contain values that conform to the defined data type, format and precision as required by domain specific business rules.	Valid email addresses consists of a prefix (before the `@` symbol) and an email domain (after the `@`symbol). When attempting to submit `person.com` into an e-mail address field, a warning or error should be raised due to the format being invalid (i.e. there is no `@` symbol) otherwise, the data will enter the system of record and be assumed as true and accurate. If the University wants to generate a contact list to send an email to students, this email address would result in errors and may prevent those with invalid emails from receiving the communications.	• Understand what qualifies as valid data and implement measures to ensure validity. • Put in place validation checks at the point of data creation to ensure values entered conform with data type, format, and precision as required by domain or business rules.
Accuracy Data must correctly represent the true value of the real-world concept or event being described.	If a staff member capturing student details misspells a student's name, then this doesn't represent the real world. This will have a negative impact on further communications and engagement with that student and the inaccurate name will also be propagated across the other areas.	• Understand what qualifies as accurate data and implement measures to ensure accuracy. • Put in place checks, particularly for data created manually which is prone to errors such as typos, data entered in the wrong fields, spaces and blanks.
Consistency Data must be absent of difference, when comparing two or more representations of a thing against a definition.	When a change is being made to the University's application system to capture `Country Code` and `Phone Number` as two separate fields. Any records created prior to the change will show as having a blank `Country Code`, with the country code contained in `Phone Number`. Whereas any new record created would appear as expected - split between the two fields.	• Put in place checks to ensure that data values are consistently represented within a data set and between data sets which may require comparison against multiple sources. • Put in place checks to ensure the consistent size and composition of data sets between systems or across time.
Timeliness Data must represent reality from the required point in time.	A prospective student attends RMIT Open Day in August of 2022 when they are in Year 12. They provide their information on a registration form, indicating they are currently studying Year 12, and consent to receipt of marketing and promotional materials. This data is accurate as at the time of submission, but in 2023 it may no longer be correct to include this person in marketing campaigns targeted at current Year 12 students.	• Ensure data is regularly refreshed so that the most up-to-date version of data is available. • Be aware of the time between when the data was created and when it is made available for use.

Dimension	Example of Data Quality Issue and its Impact	Suggested Mitigation
Uniqueness No record exists more than once within a data set or between multiple data sets.	When multiple deferral records are created for an individual student in the application that captures students' application data. As a result, the analytics team cannot identify the correct record and cannot determine when the student is likely to return to the University. This negatively impacts on reporting (e.g double-counting students) and also leads to communication issues such as sending duplicated or conflicting communications to the same student.	• Perform checks when creating a record to ensure that it does not already exist. These checks could be embedded into manual processes or automated in systems. • Put in place processes and checks to identify duplicates within a single data set and across data sets. Data cleansing and de-duplication activities can be undertaken to remediate the duplicated records.

Section 5 - Data Quality Roles and Responsibilities

(44) All Staff are responsible for:

a. ensuring the data they create, manage and use is of high quality, in accordance with the Data Quality Standard.
b. taking reasonable steps to resolve or request the resolution of data quality issues; and ensuring that data quality issues are registered in the RMIT Data Quality Issues Register.

(45) Information Trustees are:

a. accountable for decisions impacting data quality and ensuring data quality is improved in their Information Domain.
b. accountable for ensuring data quality improvement initiatives are adequately prioritised and resourced, and targets achieved for their Information Domain.
c. responsible for the endorsement of Data Quality Improvement Plans for data in their Information Domain.

(46) Information Stewards are responsible for:

a. leading data quality assessment and improvement activities including the development of Data Quality Improvement Plans.
b. making recommendations to the Information Trustees for data quality improvement.
c. championing data quality and supporting staff in best practices for data quality in their teams.

(47) Information Custodians are responsible for providing access to systems where data quality improvements need to be made.

Status and Details

Status	Current
Effective Date	23rd November 2023
Review Date	14th March 2028
Approval Authority	Senior Policy Advisor
Approval Date	18th November 2023
Expiry Date	Not Applicable
Policy Owner	Fiona Notley Chief Operating Officer
Policy Author	Nonna Milmeister Chief Data and Analytics Officer
Enquiries Contact	Data Management and Governance