

6.5. Data quality

A lot of mechanisms are used to guarantee data quality. You will find below a brief explanation of all the mechanisms used and a link to the documentation where you can see them in action.

Data dictionary

Definition

The data dictionary contains all the terms, data objects and fields that are used in the context of the RNCV. The goal of the data dictionary is keep communication clear, consistent and meaningful for all involved parties.

Implementation

[8. Data architecture](#)

[9. API Proposal](#)

Inconsistent formats

Definition

Variability in formats, like dates or numbers, can lead to misinterpretation or processing errors. For instance, a date might be entered as "dd/mm/yyyy" in one place and "mm-dd-yyyy" elsewhere.

Implementation

[8. Data architecture](#)

[9. API Proposal](#)

Duplicate data

Definition

Duplicates often occur when data comes from multiple sources or overlapping imports. These duplicates can inflate datasets and introduce biases if not properly managed.

Implementation

[6.2. Address ingestion](#)

[6.3. Versioning, approvals and audit logs](#)

[6.9. Manual reviews and audits](#)

Missing or incomplete data

Definition

Empty fields or missing essential information reduce the dataset's completeness. This may result from data entry errors, system limitations, or gaps in data collection processes.

Implementation

[8. Data architecture](#)

Inaccurate or incorrect entries

Definition

Errors from manual entry or measurement inaccuracies can lead to faulty data. This includes misspelled names, transposed digits, or incorrect values.

Implementation

Inconsistent data standards

Definition

Lack of adherence to common standards (e.g., different units of measurement, terminology variations) makes it challenging to compare or aggregate data across sources.

Implementation

[8. Data architecture](#)

[9. API Proposal](#)

Poorly defined data

Definition

Ambiguous labels, unclear fields, or undefined variables can limit the data's usability by complicating its interpretation.

Implementation

[8. Data architecture](#)

[9. API Proposal](#)

Lack of data integrity

Definition

Missing links between related records or absence of primary keys in relational data can fragment datasets, making cohesive analysis difficult.

Implementation

[8. Data architecture](#)

[9. API Proposal](#)

Incorrectly classified data

Definition

Mislabelling categories or misclassifying items within datasets can skew analysis. For example, categorising a purchase as "corporate" instead of "personal" could mislead marketing analysis.

Implementation

Limited number of free text fields

Definition

It was decided to limit the number of free text fields to the minimum. Each free text field is defined in the data model alongside its justification and approval.

Implementation

[8. Data architecture](#)

Field validation rules

Definition

Every field has a well defined type and where possible an associated validation rule that limits the valid values that can be inputted. All mandatory fields are marked as such in the data model and the validation process will enforce these rules and return an error if any required fields are missing.

Implementation

[8. Data architecture](#)

Automatic validation processes

Definition

On top of the validation rules at the field level, where possible automated validation processes are put in place to detect and prevent invalid inputs.

E.g. If two equipments are too far away from each other to be connected by a cable, that physical link is not allowed by the system.

E.g.2: If an address is being created that already exists the system will return an error

E.g.3: If an address is created but similar addresses already exist (typo) the list of similar addresses is return for validation before proceeding with the creation of the new entry.

Implementation

[6.2. Address ingestion](#)

[6.3. Versioning, approvals and audit logs](#)

[6.8. Automatic data approvals and deletion](#)

Manual Reviews

Definition

Even though automated processes are in place to ensure high quality standards, manual reviews of the data by experts can help identify and correct errors that automated systems might miss.

Implementation

[6.2. Address ingestion](#)

[6.3. Versioning, approvals and audit logs](#)

[6.9. Manual reviews and audits](#)

Cross-referencing

Definition

Comparing data from multiple sources can help identify discrepancies and validate the accuracy of the data. Cross-referencing can be particularly useful for ensuring the consistency and reliability of data.

Implementation

[6.2. Address ingestion](#)

Approval processes

Definition

Since the data ingestion process is 100% manual (usually performed by technicians on site) we need to consider the human error factor. From the discussions with the operators, the data produced by field technicians is considered as being highly qualitative and is fully trusted.

Nonetheless, human error can occur, therefore we created an approval process that redirects data ingested by field technicians to an Approver from his organisation that can perform sanity checks on the produced data records.

Implementation

[6.3. Versioning, approvals and audit logs](#)

Address database quality monitoring

Definition

Address ingestion into the RNCV database follows a process that is designed to keep the address database accurate and up to date. This process consolidates data from various datasources and approves addresses submitted by **Editors**.

Since the addresses submitted by the **Editors** are not considered as valid / approved but need to be validated by the ingestion process at a later stage, it could happen that some addresses are invalid and never get validated.

To cover such cases, a monitoring will be set up. This monitoring will be configured to detect address entries that have not been validated within a reasonable amount of time. When such entries are detected, the Application Administrators and / or Approvers will receive an alert indicating that the entry needs to be manually validated.

Implementation

[6.2. Address ingestion](#)

Versioning

Definition

Vertical cabling physical links between two equipments are crucial information whose quality needs to be guaranteed. Once a physical link dataset is produced it cannot be deleted anymore, the produced data can then be validated or rejected.

If a problem is detected with a dataset, after it has been validated/rejected, this decision can be undone at a later stage by Administrators, effectively reverting the approved version to the previously approved version.

Implementation

[6.3. Versioning, approvals and audit logs](#)

Audit logs

Definition

Audit logs of every action performed on the system are kept. The audit logs are mainly kept for accountability, but can also be used to analyse drops in data quality. Furthermore since all actions performed are stored, the audit logs could be used as last resort to manually correct unintended or malicious actions.

Implementation

[6.3. Versioning, approvals and audit logs](#)

Systematic reviews

Definition

Audits involving systematic reviews ensure that datasets align with quality benchmarks and organizational standards. For example, checking for duplicates or data not conforming to predefined rules.

Implementation

Mark old data as deleted

Definition

Once data is inserted in the VC database it won't be deleted anymore (sites, blocks, units, equipments, physical links) instead, it will be marked for deletion, and will go through a validation process. Once the data deletion is validated by the **Approver** or **Organisation Approver**, the data is marked as Deleted and the system will not allow new links to that data entry.

Note that user objects will also not be deleted from the system, but all personal data will be deleted (first name, last name, email, ...)

Implementation

Revision #4

Created 16 October 2025 12:35:53 by Sergio Sousa

Updated 25 November 2025 10:22:03 by Sergio Sousa