**SCReeD Dataset Preparation Guidelines**

# **Overview**

This document presents the guidelines for SCReeD users regarding dataset preparation, in particular, how to sanitise datasets before uploading.

# **The Sanitisation Process**

Before initiating the dataset upload process, it is imperative to sanitise the dataset(s) in accordance with the following guidelines.

**a) Dataset Information:**

* Dataset Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
* Description: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
* Date of preparation: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
* Owner/Uploader: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
* Purpose of Dataset: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**b) Data Source:**

* Data Source(s): \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
* Data Collection Method(s): \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**c) Data Cleaning:**

* [ ] Remove duplicate records.
* [ ] Check missing values and decide on an appropriate strategy (e.g., impute or remove).
* [ ] Correct any inconsistent or erroneous data.
* [ ] Normalize or standardise data where necessary.
* [ ] Address data outliers as per project requirements.
* [ ] Convert data types to appropriate formats (e.g., date, numeric, categorical).

**d) Data Privacy, security and Legal Compliance**

* [ ] Ensure compliance with all applicable data privacy regulations [e.g. GDPR, HIPAA].
* [ ] Verify that you have the legal rights or permissions to use/upload the data.
* [ ] Implement encryption or access controls if necessary.
* [ ] Ensure that any sensitive or personally identifiable information (PII) is anonymised or removed.
* [ ] Obtain necessary approvals from relevant stakeholders for dataset usage/uploading.

**e) Data Quality Assurance/Assessment:**

* [ ] Conduct data quality checks and validation (e.g., data consistency, accuracy & completeness).
* [ ] Validate data against domain knowledge or business rules.
* [ ] Document any data quality issues and their resolutions.
* [ ] Validate the dataset with relevant use cases and scenarios.

**f) Data Documentation:**

* [ ] Create a data dictionary with variable descriptions.
* [ ] Document any transformations or modifications made to the dataset.
* [ ] Include metadata such as creation date, update frequency, and data owner.
* [ ] Document how the dataset is intended to be used.
* [ ] Specify any restrictions or limitations on dataset usage.
* [ ] Establish procedures for ongoing data monitoring and maintenance.
* [ ] Regularly review and update the dataset as needed.

**g) Data Retention and Disposal:**

* [ ] Define data retention policies and adhere to them.
* [ ] Establish backup procedures for dataset preservation.
* [ ] Document data recovery protocols in case of data loss.
* [ ] Safely dispose of data when it is no longer needed.

**h) Data Review and Approval:**

* Data Owner/Provider: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
* Data Steward: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
* Data Scientist/Analyst: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
* Review Date: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

I acknowledge that the dataset has been appropriately prepared, thoroughly reviewed in compliance with all guidelines and that all relevant stakeholders have been duly informed and are ready for upload.

Signature: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Date: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_