Data Management Plan

yvonne_kallberg_test_DMP

Contributors	**Sam Smith** (sam.smith@example.com) Uppsala University (PI) **Yvonne Kallberg** (yvonne.kallberg@nbis.se) Stockholm University (**contact person**, researcher)
Based on	SciLifeLab Science Europe / VR DMP, 3.2.3 (SciLifeLab:SLL-SE-VR-DMP:3.2.3)
Generated	2024-03-18 using Data Stewardship Wizard
Created by	**Yvonne Kallberg** (yvonne.kallberg@nbis.se)

Project

This data management plan describes the data and work in the following project.

SARS-CoV-2 genomic surveillance in Italy

RNA sequencing of nasopharyngeal swab samples taken at a hospital in Italy as part of genomic surveillance of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2).

Start date: 2022-11-13
End date: 2023-06-13

Funding	Vetenskapsrådet (applied)

Section A: Description of data – reuse of existing data and/or production of new data

1. How will data be collected, created or reused?

Instrument datasets

The following instrument dataset(s) will be acquired in the project:

- **RNA sequences**
  This dataset will be produced by National Genomics Infrastructure (NGI).

  For this dataset, the following instrument(s) are used:

  - **Illumina NextSeq 500** – Next generation sequencing instrument

Re-used datasets

We will use the following reference dataset(s):

- **SARS-CoV-2 genome from isolate Wuhan-Hu-1** (https://www.ncbi.nlm.nih.gov/nuccore/NC_045512.2)
  We will use version "Version 2" of this dataset. Even if a new version becomes available during the project, we will stay with the old version.

2. What types of data will be created and/or collected, in terms of data format and amount/volume of data?

We will be using the following data formats and types:

- **FASTQ Sequence and Sequence Quality Format** - We will have only a small amount of data stored in this format.

Section B: Documentation and data quality

3. How will the material be documented and described?

We will use an electronic lab notebook (ELN) to make sure that there is good provenance of the collected data.

In order to keep track of what has been run during analysis, we will:

- Use software tools that are professionally maintained, with version control
- Use standard workflow engines and automatic workflows
- Use interactive notebooks

We will use the following metadata standards to describe our datasets:

- ENA virus pathogen reporting standard checklist ERC000033

In addition, the following metadata will be provided as documentation: File organisation and naming conventions will be documented in a README text file that will be put in the root folder of the project. Also, a dictionary of the sample metadata will be created.

Storage and file conventions

This project has a documented strategy for file and folder organization. This project has a documented file and folder naming convention.

There will be a README file in the top folder, describing the folder structure and naming conventions.

4. How will data quality be safeguarded?

Instrument datasets

- **RNA sequences**
  The equipment used to produce this dataset is very well described and known.

  We will be using the following quality processe(s):

  Repeated measurements
  Instrument calibration

Section C: Storage and Backup

5. How is storage and backup of data and metadata safeguarded during the research process?

All project data will be in backed-up storage systems during the project.

All essential data is also stored elsewhere to prevent a total loss of data: The raw reads will be submitted to ENA upon arrival.

The project data will be shared using the following work space service:

- NAISS center Uppmax

6. How is data security and controlled access to data safeguarded?

Only project members have read/write access to the data.

All data will geographically be stored at:

- A national service hosted on national servers in Sweden

Project members will not store data or software on computers in the lab or external hard drives connected to those computers. They can carry data with them on encrypted data carriers and password-protected laptops. Project members have been instructed about both generic and specific risks to the project.

The possible impact to the project or organization if information is lost is small. The possible impact to the project or organization if information is leaked is small. The risk of corrupted information in the project or organization is acceptably low.

Section D: Legal and ethical aspects

7. How is data handling according to legal requirements safeguarded?

Data we reuse

For the reference and non-reference datasets that we reuse, conditions are as follows:

- **SARS-CoV-2 genome from isolate Wuhan-Hu-1** – freely available for any use (public domain or CC0).

All data will be owned by the university.

8. How is correct data handling according to ethical aspects safeguarded

Data we collect

None of the collected datasets are personal.

Section E: Accessibility and long-term storage

9. How, when and where will research data or information about data (metadata) be made accessible?

We will be working with the philosophy *as open as possible* for our data. Some of our data cannot be openly shared directly after production:

- Papers needs to be submitted first

Data that is not legally restrained will be released after a fixed time period: 1 year.

The datasets will be accessible as follows:

- **RNAseq** (RNA sequences of swab samples)
  - Using a domain-specific repository: European Nucleotide Archive
  - The dataset will be available under the following license: Freely available for any use (public domain or CC0).

10. In what way is long-term storage safeguarded, and by whom?

The long-term plan for the produced datasets is as follows:

- **RNAseq** (published)
  - The dataset will be stored using storage provided by the institute.
  - This dataset will be kept available as long as technically possible.
  - The metadata will be available even when the data no longer exists.

11. Will specific systems, software, source code or other types of services be necessary in order to understand, partake of or use/analyse data in the long term?

Long-term suitability of the data formats used:

- **FASTQ Sequence and Sequence Quality Format**
  It is a standardized format. This is a suitable format for long-term storage. No special software or tool is needed to use this material.

12. How will the use of unique and persistent identifiers be safeguarded?

- Dataset **RNAseq** will have a persistent identifier

Section F: Responsibilities and Resources

13. Who will be responsible for data management?

Yvonne Kallberg is responsible for overall data management including implementing the DMP, and ensuring it is reviewed and revised.

14. What resources will be required for data management?

To execute the DMP, additional specialist expertise is required which will be accomplished by training existing staff.

None of the repositories or long-term storage providers used charge for their services.