# Overview

This documents outlines the data management and release policy for the collaborative research center (CRC) 1182 (SFB1182). The goal of this document is to provide instructions on how data is to be stored, described, accessed and disseminated.

Within this document, the term “data” is primarily used to refer to sequencing data generated as part of research projects within the CRC1182. Additional data types may be adequately covered by the general instructions below. If existing rules seem unsuitable for a given type of data, amendments to this document under Section 3 shall be made in dialogue with the INF project.

# Roles and responsibilities

The following roles and associated responsibilities are currently defined:

## University compute center

The compute center (CC) and Kiel University provides the IT infrastructure required for the CRC data management.

The CC ensures that services are maintained and expanded as required by the CRC.

## CRC INF project

The INF project is responsible for implementing data life cycles and workflows. Most of the “primary” data will be handled by the INF project and processed according to the guide lines outlined in the present data policy (Section 3 and 4).

## CRC Z2/Z3 project

The Z2 and Z3 project are responsible for providing relevant metadata to the INF project for data generated as part of their support activities.

## Data producers / PIs

Principle investigators within the CRC are responsible for communicating their data generation plans to the INF project, or the Z2/Z3 project, to ensure that all data can be captured in the data management system of the CRC.

The INF project, where possible, shall be designated a data recipient for all sequencing projects.

PIs must ensure that metadata is captured prior to e.g. sequencing and passed on to the INF project. The Z2/Z3 project can provide appropriate forms for standard applications.

## All users

All users of the data management system must comply with data access and release policies outlined in Section 4.

# Storage and annotation

## iRODS archive storage

All raw (sequencing) data generated within the CRC1182 (i.e. financed through the CRC) must be deposited in the CRC data management solution (iRODS@RZ).

If sequencing is not coordinated with the INF project for automatic data delivery/archiving, it is the users’ responsibility to deposit the data with all relevant metadata (Section 3.2).

Analytical results based on these raw data may also be submitted, together with a description on how this information was generated. For further details regarding data types, see Section 3.

Data not generated as part of the CRCs’ activities must not be submitted to the CRC iRODS.

Individual data sets consisting of multiple files should be combined into one “tar” archive and uploaded as a single object. A thusly created tar archive shall represent a given state within the research project (e.g. raw data or analysis results). Different data stages may not be mixed into one tar object (e.g. data and analysis results) but should reference each other via metadata keys.

Data shall be deposited in the iRODS folder of the corresponding research project the data was generated for.

Data access shall be set in accordance with guide lines described under Section 4.

Data will be stored within the CRC iRODS system for the life time of the CRC, but at minimum 3 years from initial submission into the system.

Storage beyond this time should be sought through publication to international data repositories (Section 4.2).

## Metadata annotation

All data submitted to the data management system must be annotated with relevant metadata to enable data discovery across projects within the CRC and to help with submission to downstream repositories. The current metadata standard for the CRC1182 is described under:

<https://github.com/marchoeppner/metaorganism-research/tree/master/metadata>

These specifications are based on the EBI MIxS standard (<http://www.ebi.ac.uk/ena/submit/mixs-checklists>). Metadata keys not covered by this specification should be used only after consulting with the INF project so they can be added to the metadata vocabulary.

# Data types

## Short read files (fastq)

Short reads generated on e.g. the Illumina platform of machines are the main data type within the CRC1182.

Short reads shall be deposited within iRODS in their raw form (no trimming, no filtering), with their original file name as given by the sequencing provider intact.

For efficient usage of the storage system, read files shall be compressed using the GZIP algorithm, following the naming convention “.fastq.gz”, and be accompanied by an md5sum.

Where multiple read files belong to one sample (i.e. paired-end reads), the compressed files shall be group into one tar archive. The tar archive shall be named after the sequencing library (usually the common root of the filenames to be grouped).

The tar archive shall contain the corresponding fastqc statistics (one statistics file per sequence file).

## Analytical results (generic)

Finalized analyses (tabular counts or similar) may only be stored in iRODS if the analysis process is included with the data so that any user may be able to accurately reproduce the workflow. Analysis and the “analysis trace” shall be submitted together as tar archive and must be annotated with appropriate meta data. Specifically, it must be clear which raw data was used in the analysis by reference to the sample or library name(s) within iRODS.

The format of the analytical result must be discussed with the INF project prior to submission to help identify common standards and develop best practices.

# Access and release

## Data sharing within the CRC1182

Data and metadata from the individual projects shall be readable by all CRC1182 members to aid in data discovery.

Data shall only be writable by members of the respective projects.

Data belongs to the user it was generated for, as indicated by the mandatory metadata key “MAIN\_CONTACT\_NAME”.

Any use of the data prior to its publication needs permission by the owner.

## Release of data into the public domain

Data generated within the CRC1182 shall be made available to the public upon publication, or after a period of 3 years (measured from the date of submission into iRODS@RZ) – whichever event occurs first.

Data shall be published through upload into suitable, widely accepted repositories (e.g. NCBI or EBI archives).

Data published to international repositories shall be annotated within the CRC iRODS using the resulting digital object identifier (DOI) in the remote database, if possible/applicable.

# Revision history

Draft 1 generated in Summer 2016.

Release 0.1 generated in November 2016.