Interreg IT-AT 2021-2027

Data Management Plan SENECA

Contributors:

- Christian Weichenberger (contact person for data management)
  christian.weichenberger@eurac.edu, 0000-0002-2176-0274
  Roles: Contact Person, Data Manager, Project Member, Researcher
  Affiliation: Institute for Biomedicine
- Claudio Brancolini
  claudio.brancolini@uniud.it, 0000-0002-6597-5373
  Roles: Data Collector, Project Leader, Supervisor
  Affiliation: University of Udine
- Fritz Aberger
  fritz.aberger@plus.ac.at, 0000-0003-2009-6305
  Roles: Data Collector, Project Member, Supervisor
  Affiliation: University of Salzburg
- Fulvia Felluga
  ffelluga@units.it, 0000-0001-8271-408X
  Roles: Data Collector, Project Member, Supervisor
  Affiliation: Università degli Studi di Trieste

HISTORY OF CHANGES		
Version	Publication date	Changes
0.1	04.04.2023	▪ Initial version for internal discussion
1.0	18.04.2023	▪ Ready for attachment to grant submission
1.1	25.06.2024	▪ Changes after project approval and for publication on blog

Project
SENECA (Interreg Italia-Österreich 2021-2027 ITAT-11-018-SENECA)

Based on
Common DSW Knowledge Model (v2.4.4) and the EU Horizon DMP template (v1.0, 5.5.2021)

Project phase
Before Submitting the Proposal

Created by
Christian Weichenberger (christian.weichenberger@eurac.edu) and
Andrea Minio (andrea.minio@eurac.edu)
Institute for Biomedicine, Eurac research, Italy

Generated on
Mar 28th, 2023 (start of writing, for more information, see versioning history)

This data management plan (DMP) was initially created with Data Stewardship Wizard «ds-wizard.org» and has been subsequently refined, taking the EU Horizon data management plan structure as a template.

1. Abstract

Explore the effect of an aged immune system on the (epigenetic) landscape of the micro-environment of certain cancer types, as for example colorectal cancer.

Note

This DMP is designed for the grant application of project EPIC-S and has been submitted as an attachment to the coheMON platform. In this version, the plan describes our intention for data generation and use and will be updated accordingly over time, in case the project is accepted by the program.

2. Data Summary

In this project, we will generate new data from laboratory experiments that are tailored towards our research questions. Already existing datasets will be integrated in the analysis of the newly generated sets, maximizing use of available knowledge. Our primary databases for obtaining reusable datasets are:

- NCBI human genome hg38 and transcripts (doi: 10.25504/FAIRsharing.7ad252)
- Encyclopedia of DNA Elements (doi: 10.25504/FAIRsharing.v0hbjs)
- The Cancer Genome Atlas (doi: 10.25504/FAIRsharing.m8wewa)
- Gene Ontology (doi: 10.25504/FAIRsharing.6xq0ee)
- Reactome pathways (doi: 10.25504/FAIRsharing.tf6kj8)
- STRING protein interactions (doi: 10.25504/FAIRsharing.9b7wvk)

A pivotal part of the project will be collection of gemomic data within well-defined experimental setups in the laboratory. These involve *in-vitro* cellular experiments in which various conditions of cancer cells are compared to each other by analyzing their gene expression profiles. Reused data will be either retrieved directly from the dataset provider or will be included in analysis software packages. In the latter case we are documenting the package versions to allow reproducibility of results. Existing datasets are used to annotate our internally generated dataset and facilitate analysis and interpretation of results. Only in rare cases, we will carry out analyses on web services, as they are frequently updated and are usually not a proper choice for reproducible workflows.

Genomic sequencing data are usually delivered as files in FASTQ format with several ten GB of file size. These are kept in this format, as the processing pipelines rely on them and output standard bioinformatics data file formats (SAM/BAM, BED, and VCF). Therefore, no “object store” or relational database systems will be used to store these data. Instead, the samples will be organized in a file system with files and folders. There will be a tabular separated file describing the experimental conditions for each FASTQ file. Data will be backed up to a dedicated resource and published on public archives to prevent accidental loss.

The main reason to generate sequencing datasets is to answer research questions related to cellular senescence and its relationship to cancer in the context of the immune system. Each generated dataset will be associated with at least one publication. Disseminating the dataset alongside the publication is common practice, as it allows to reproduce the published results and to test other hypotheses from either the ageing or cancer research community.

3. FAIR data

3.1. Making data findable, including provisions for metadata

We plan to deposit our data with the EMBL-EBI genomic archives; in particular we are going to use the European Nucleotide Archive (ENA, doi: 10.25504/FAIRsharing.dj8nt8), which guarantees high visibility and easy access. ENA is part of the International Nucleotide Sequence Database Collaboration, which further consists of the DNA DataBank of Japan (DDBJ) and GenBank at NCBI. All three organizations exchange data on a daily basis. Each submission to ENA will receive a unique identifier. Metadata are created as part of the submission process at ENA, making use of their checklist, and therefore following a well-defined standard for sequencing reads deposition. ENA metadata checklists follow the minimum information about any (x) nucleotide sequence (MIxS) standard for describing genomes, supplying datasets with rich, ontology-based metadata. Furthermore, the ENA allows programmatic access to their metadata repository, providing a rich set of query options for automated processing.

3.2. Making data accessible

The ENA repository is part of the ELIXIR Core Data Resources, which are a set of European data resources of fundamental importance to the wider life-science community and the long-term preservation of biological data. Data deposition with ENA is free and without restrictions on data size and can be done at any time without further arrangements of the service provider. Each study will be assigned a unique identifier and will be grouped into sequence datasets furnished with metadata that are used for characterizing each individual sequence dataset when downloaded by a user.

We will publish all raw sequencing datasets on ENA and the data processing pipeline along with the documentation of any third party software will be made accessible via Github. Any interested individual can therefore download data from the ENA archive utilizing a variety of methods, including web access, FTP access, or programmatic access via an application interface. The public archive has detailed documentation on how to access and retrieve these data. Currently we do not foresee withholding data; if during the course of the project we are going to obtain potentially identifiable samples directly from human donors, we will update this document appropriately - if needed, the datasets will be uploaded to the European Genome-Phenome Archive (EGA) to guarantee access control.

In parallel with the metadata published on ENA, we are going to announce the availability of datasets on our home page on a project-specific, dedicated place. This page will describe the dataset and provide links to the ENA-deposited data. In addition, we are going to furnish the website with metadata such that web crawlers will be able to collect them and incorporate them into their indexes.

3.3. Making data interoperable

Data from sequencing experiments will be stored and processed according to the latest standards from the bioinformatics community. Sequencing data are typically delivered in FASTQ format from the sequencing center/core facility, which are then further processed using bioinformatics pipelines. The project pipelines will build on publicly available open source software to guarantee reproducibility of results and facilitate later data re-use. Files in FASTQ format represent an atomic data type without any need for ontologies. Metadata will comply with the ENA metadata model[†] and will be updated and/or extended when required, as for example when a manuscript has been published.

In case external datasets are integrated into our research, we will list them on our dedicated project website in addition to specifying their unique identifiers in the accompanying manuscript.

[†] https://ena-docs.readthedocs.io/en/latest/submit/general-guide/metadata.html (page visited: April 3rd, 2023)

On the SENECA web portal, a dedicated section will collect all the links for an easy access to the raw data, as well as providing an extended and coherent metadata collection of the samples analyzed for the project.

3.4. Increase data re-use

Data will be uploaded to the ENA public archive and will be made available to the public at the time of publication at the latest. Submission will be fully open to the public without embargo. This release of data to the public is intended for as long as the public archive exists, which will be long after the project has finished. Our goal here is to maximize access of data, reproducibility of results and re-use of the datasets we have generated to support future studies in this field.

Data quality assurance will be addressed in a dedicated work package, “Data management and analysis”. Experiments will be designed according to standards in the field of next generation sequencing (experimental protocol, number of replicates, read depth, etc.) to yield reliable, accurate, and interpretable results. The data processing pipelines to be implemented will investigate the quality of sequence reads and perform consistency checks prior to any analysis steps.

Analysis pipelines, data accessibility information, and experimental protocol details will be made available as repositories on the SENECA webpage. Publicly available through Github, the web portal will allow transparent access to the latest updates on the project, to the data produced, to download and to apply the developed pipelines.

The public will be able to reproduce the results obtained from this project as well as apply the same procedures to other projects and to other datasets. This aims to promote the standardization of analysis protocols, which is essential for comparing research results.

Detailed instructions on how to perform the procedures will be delivered alongside the generated pipelines and code. Taking advantage of the version control system at the base of the portal, pipeline updates will be constantly accessible to the public guaranteeing the execution at the latest standards set in the bioinformatics community, making use of the most recent software available at the time of analysis, and keeping a record of the changes as they are implemented for the improvement of methodology.

4. Other research outputs

The research workflow will be made fully transparent by submitting a publication on the respective topic, which includes clear descriptions of the experimental protocols used to obtain and cultivate cells. Data will be made available in support of FAIR principles on the ENA public archive and the analysis pipelines will be made available on a suitable software repository such as Github. Biological samples will be available to all members of the research network and to the public on request.

5. Allocation of resources

Costs for making data FAIR are included in the grant proposal under work package “Data management and analysis”, which will implement the procedures described above. The project will make use of public sequencing data archives to guarantee long-term findability, accessibility, interoperability, and reusability of generated data. These archives have long-term funding for data preservation, such that data will remain accessible far beyond the end of the project.

The responsible person for data management in this project will be Christian X. Weichenberger (Eurac

Research, Italy).

6. Data security

During research phase, data will be stored on internal servers with a dedicated backup storage solution financed by the grant. In the final phase of research, prior to submission of a manuscript, data will be uploaded to public archives (e.g. ENA) to make them available upon manuscript submission. The archival policy of the data repository will guarantee long-term accessibility of the datasets, including metadata that describe experimental setup and analysis steps.

7. Ethics

In this phase, we intend to generate datasets from experiments on cell models and don’t foresee any ethical or legal actions. Our plans for data sharing and long-term preservation are not subject to informed consent. In case we obtain samples from human donors, this DMP will be updated appropriately.