



DATA MANAGEMENT PLAN EATRIS-PLUS PROJECT

Authors : Peter-Bram 't Hoen (Radboud university medical centre, Nijmegen, Netherlands, RUMC), Anna Niehues (RUMC), Casper de Visser (RUMC), Eliis Keidong (EATRIS central coordination office, Amsterdam, Netherlands, EATRIS), Toni Andreu (EATRIS), Anne-Charlotte Fauvel (EATRIS), Lukas Najdekr (Institute of Molecular and Translational Medicine, University Hospital Olomouc, Palacky University, Olomouc, Czech republic, UP), Marian Hajduch (UP)



CONTENTS

Executive summary	3
INTRODUCTION Project objectives	3
DATA MANAGEMENT PLAN	4
1. Data Summary.....	4
2. FAIR data	7
3. Data security	10
Abbreviations.....	10
Adjustments made.....	12



EXECUTIVE SUMMARY

This data management plan is for the demonstrator study in the EU-funded EATRIS-Plus Project, involving the reuse of genomic data and the generation, analysis and integration of multi-omics data from a cohort of 127 healthy blood donors. Data has been generated by five project partners in five different EU countries. The purpose of the data generation is to establish multi-omics reference values based on high-quality reference protocols of a well-characterised population cohort of healthy individuals.

A solid Data Management Plan (DMP) is an important basis for good data management in a multi-beneficiary research project reusing and generating new data. This Data Management Plan for EATRIS-Plus project describes the data management life cycle for the data to be collected, processed and generated during the project life cycle. As part of making research data findable, accessible, interoperable and re-usable (FAIR), this DMP considers important elements such as the handling of research data during and after the end of the project, data collection and generation. It further outlines the methodologies and describes which standards have been applied, as well as how the data has been shared, and made openly accessible (including curation and preservation during and beyond the life cycle of the project).

All data generating project partners were consulted in preparation of this deliverable. These partners are University of Uppsala (UU), Institute of Molecular and Translational Medicine, University Hospital Olomouc, Palacky University (UP), Institute for Molecular Medicine Finland, University of Helsinki (FIMM), the Madrillienian Health Service (SERMAS) and Radboud university medical center (RUMC).

The first iterations of the DMP was reviewed by the project's Ethics Committee as a part of the Ethics Committee's Annual Report 2020 and revised based on the Ethics Committee's suggestions in the following three areas: 1) description of the security measures that will be implemented to prevent unauthorised access to personal data or the equipment used for processing; 2) description of the anonymisation/pseudonymisation techniques that will be implemented; and 3) labelling of data using definitions of GDPR. Later versions included improvements and further detailing of the data FAIRification.

INTRODUCTION PROJECT OBJECTIVES

The flagship project EATRIS-Plus aims to build further capabilities and deliver innovative scientific tools to support the long-term sustainability strategy of EATRIS as one of Europe's key European research infrastructures for Personalised Medicine.

The main goals of the EATRIS-Plus will be to:

- Consolidate EATRIS capacities in the field of Personalised Medicine (particularly omics technologies) to better serve academia and industry and augment the number of EATRIS Innovation Hubs with large pharma;
- Drive patient empowerment through active involvement in the infrastructure's operations;



- Expand strategic partnerships with research infrastructures and other relevant stakeholders, and
- Further strengthen the long-term sustainability of the EATRIS financial model.
- Develop a Multi-omic Toolbox for researchers.

The purpose of the multi-omics profiling study of a cohort of 127 healthy blood donors is to establish high-quality reference protocols for data generation, analysis and integration, to establish reference values for a large set of omics features in the general population, and to produce reproducible workflows for data analysis and integration that can be incorporated in The Multi-omics Toolbox (MOTBX): <https://motbx.eatris.eu>

DATA MANAGEMENT PLAN

1. DATA SUMMARY

Purpose of the data collection and generation and its relation to the objectives of the project

The purpose of the data generation is to establish multi-omics reference values of a well-characterised population cohort. This will serve as a demonstrator for high-quality reference protocols for multi-omics technologies and will be part of a multi-omics toolbox serving as reference for multi-omics studies.

Types and formats of data

Different types and formats of data have been generated in the project. Standardised data formats have been used as much as possible. Formats of raw and processed files for the different data types are listed in Table 1.

Table 1: Data types and formats

Type of data	Partner where data is generated	Raw data format	Processed data format(s)
Genomics	UP/IMTM	.fastq.gz	.bam, .vcf, .tsv
Proteomics	UP/IMTM	.raw	mzTab, .csv (peptide/protein lists from ProteomeDiscoverer)
miRNA-seq	UH/FIMM	.fastq.gz	.bam additional file formats in the pre-processed data are: .html, .txt, .zip, .csv, .sf, .tsv, .pdf, .r and .gtf
RNA-seq	UH/FIMM	.fastq.gz	.bam

			additional file formats in the pre-processed data are: .html, .txt, .zip, .csv, .sf, .tsv, .pdf, .r and .gtf
miRNA qRT-PCR	SERMAS	.txt	.txt and .xls following NCBI GEO database standard
EM-Seq	UU	.fastq	.bam .bedGraph
Metabolomics	RUMC	.raw	After raw data conversion to open format: .mzML After pre-processing: .tsv files (metabolite annotation file, MAF)

Re-use of existing data

Genome sequencing data already generated by UP has been re-used (for data origin, see below).

Origin of the data

All data is derived from human subjects from a cohort of healthy individuals recruited in UP. Participants have provided the required informed consent covering the future use for research purposes.

The WP1 genome sequencing data is coming from an existing research project from partner UP and derives from the same cohort.

The pseudonymisation procedure is only performed by UP, all other project partners/analysing sites receive the information already in pseudonymised format. Any patient data transferred to the analysis sites in EATRIS-Plus project for the sample analysis is done in a pseudonymised format associated with a unique ID only.

Patient/study participant data/ case report forms, are accessible only through ClinData software. Primary personal data, such as first name, last name, DOB, address, national personal ID, phone number, email, coming from patient/study participant sample analysis are associated with a unique ID only (pseudonymised).

In ClinData software, personal information is visible only to the authorised users. Users can be authorised in accordance with GDPR practices. Every attempt to get personal data must be justified and approved. Access rights of the users can be controlled in very detailed manner, so the ClinData can control the approach to every single record.

To other users, patients/study participants data are visible as a system assigned Patient ID only, which does not relate to any personal information provided. System assigned Patient ID is unique across all studies and registries in the ClinData system. Different users can be assigned read-only access, read only for chosen part of the data, write/data entry access, admin access.

There are no anonymisation techniques used, as in case of discovery of clinically relevant results, those are reported to the patient/study participant.

New data

For WP1, various omics data on samples from the UP cohort has been produced by different partners, as indicated in the table below.

Methods	Partner
Genomics: Genome sequencing (existing whole genome sequencing data) WGS) and arrayCGH	UP
Epigenomic DNA methylation Enzymatic Methylation sequencing (EM-seq)	UU
Transcriptomic RNA Sequencing (RNA-seq)	UH
MicroRNA sequencing (miRNA-seq)	UH
MicroRNA qRT-PCR panel	SERMAS
Proteomic analysis	UP
Metabolomic analysis	RUMC (targeted) UP (untargeted)

Expected data size

The sizes of the data of different -omics types is listed in Table 2. The total expected size of raw data based on single measurements of 100 samples is estimated to be ~12 TB. The size of pre-processed data will be approximately the same. There will be additional data for integrated omics analyses. This data will be small in comparison to the raw data.

Table 1: Size of the data generated within the project

Type of data	Partner where data is generated	Expected size of raw data (per sample if not stated otherwise)	Expected size of processed data (per sample if not stated otherwise)
Genomics	UP/IMTM	6.4 TB (all samples)	6.3TB (including BAM files of all samples)
Proteomics	UP/IMTM	10 GB	167MB
miRNA-seq	UH	>2 GB	1 GB
RNA-seq	UH	>10 GB	5 GB
miRNA-seq	UH	2 GB	2 GB

MicroRNA qRT-PCR panel	SERMAS	~50 kB	~50 kB
EM-seq	UU	~55 GB	75 GB (aligned BAM) or 45 GB (sorted aligned BAM with duplicate alignments discarded), 250 MB (bedGraph)
Metabolomics	RUMC	~1 GB per sample ~130 GB per analytical batch (~135 samples)	~800 MB total size for .tsv files

2. FAIR DATA

2. 1. MAKING DATA FINDABLE, INCLUDING PROVISIONS FOR METADATA

Extensive human and machine-readable metadata will allow the data sets to be findable through data set searches. Data are findable by including metadata in the cBioPortal catalogue: https://cbioportal.imtm.cz/study/summary?id=mixed_imtm_2023. We are planning to make the metadata available and the data accessible through a FAIR Data Points (FDP)¹ or a Beacon².

Dedicated research data repositories hosted by European Bioinformatics Institute (EBI) provide stable and unique identifiers for each submitted data set. If possible, data sets will be assigned a Digital Object Identifier (DOI), which will make them easily citable. One possibility for automatically assigning a DOI is submission to Zenodo (<https://about.zenodo.org/principles/>).

Naming conventions

Standardized terms (ontologies) will be used as much as possible. Existing metadata schema used include the study-level Data Catalog Vocabulary (DCAT) and more detailed assay and sample-level metadata from Investigation-Study-Assay (ISA). Standardized naming conventions are taken from the FAIR genomes project (ZonMw 846003201)³, the [Netherlands X-omics Initiative](#), the [Global Alliance](#)

¹ FAIR Data Points can be used to describe data sets in a FAIR way, using standard metadata and make them available through simple WWW protocols, <https://www.fairdatapoint.org/>

² Fiume, M., Cupak, M., Keenan, S. et al. Federated discovery and sharing of genomic data using Beacons. Nat Biotechnol 37, 220–224 (2019). <https://doi.org/10.1038/s41587-019-0046-x>

³ <https://www.zonmw.nl/nl/onderzoek-resultaten/geneesmiddelen/programmas/project-detail/personalised-medicine/fair-genomes-a-national-guideline-to-promote-optimal-reuse-of-ngs-data-in-research-and-healthcare/verslagen/>

for Genomics & Health(GA4GH), the [Beyond 1 Million Genomes \(B1MG\) project](#), the [HUPO Proteomics Standards Initiative \(PSI\)](#). A codebook with definition of a metadata schema⁴ has been published. Data sets are equipped with a version number and/or time stamps.

Metadata

The FAIR Data Points exploits the DCAT metadata schema. We have used a standardized file format suitable for collection of complex metadata from omics-based experiments, i.e. [the Investigation/Study/Assay \(ISA\) tab-delimited \(TAB\) format](#). Rich metadata will not only optimize findability and re-usability, but will also facilitate integration of different omics-based data within this project. Our reusable template for multi-omics metadata can be found here: <https://github.com/EATRIS-Plus>.

2.2. MAKING DATA ACCESSIBLE

Sharing of identifiable, privacy-sensitive data is restricted by the consent given by the participants. All (raw) –omics data is considered privacy-sensitive under GDPR and will be made available to external users through a controlled access mechanism overseen by the EATRIS-Plus Data Access Committee (DAC, EATRIS-Plus-DAC@eatris.eu). The DAC of EATRIS-Plus includes all members of its Multi-Omics Task Force. All credible researchers with a solid research proposal will be given access to the data under conditions described in the Data Access Agreement (DAA). Provisions included in the DAA are that the requesting researcher should properly protect the data from leakage into the public domain, refrain from attempts to reidentify the participants in the study, refrain from distribution to third parties, and that EATRIS-Plus is properly acknowledged in any artefacts coming out the studies with the data. The identity of the person accessing the data will be ascertained through the signatures of institute's representatives on the DAA.

Phenotype information including Body Mass Index (BMI), gender, age, haematology, blood group, and smoking behaviour is stored in the ClinData database developed and hosted by UP and available through cBioPortal. The primary research data is saved in the IMTM PrivateCloud, which can be used for secure and convenient data accession. Object storage allows convenient access to the research data - with a direct link to the data accessible within ClinData. Scientific results will be communicated in peer-reviewed journals providing DOIs and following open access publishing.

Raw data formats for some data types require proprietary software to access. Where possible, we have used open formats that do not require proprietary software. Documentation about relevant software for preprocessing is included.

Fully processed data (counts for RNA, miRNA-seq, beta-values for EM-seq, intensities for miRNA-qRT-PCR, proteomics and metabolomics) are considered anonymous and will be made openly available from Zenodo and equipped with a DOI persistent identifier.

Data analysis tools and pipelines will be made publicly available via the EATRIS-Plus GitHub repository: <https://github.com/EATRIS-Plus> and equipped with an MIT open access license. Additionally, software will be registered with a software registry (<https://bio.tools/>). We further use containerization,

⁴ https://github.com/fairgenomes/information/tree/master/fairgenomes_codebook_nictiz

workflow management systems and workflow registries (<https://workflowhub.eu/>) to enhance FAIRness of analysis workflows.

2.3. DATA INTEROPERABILITY

The generated data will be made available using standardized formats commonly used in the respective domain (see Table 1). Omics-specific metadata standards include standards of the Metabolomics Standards Initiative (MSI), Minimum Information About a Next-generation Sequencing Experiment (MINSEQE) guidelines, Minimum Information About a Proteomics Experiment (MIAPE), Minimum Information for Publication of Quantitative Real-Time PCR Experiments (MIQE), as well as guidelines from commonly used domain-specific data archives (EBI MetaboLights, EBI European Nucleotide Archive – ENA).

Standard vocabularies will be used for all data types where possible. See Table 3 for a list of used vocabularies and identifiers.

Table 3: Controlled vocabularies and ontologies used in EATRIS-Plus. For all gene-related identifiers, the human genome reference build GRCh38/hg38 is used.

Data type	Identifiers, ontologies, controlled vocabularies
Metabolites	HMDB ID, ChEBI ID (Chemical Entities of Biological Interest)
Measurements and units identification	Units of Measurement Ontology (UO)
Phenotypes	Human Phenotype Ontology (HPO), National Cancer Institute Thesaurus (NCIT)
Genes, transcripts	HUGO Gene Nomenclature (HGNC), Ensembl Gene ID, Ensembl Transcript ID
Gene annotation	Gene Ontology (GO)
Genomic coordinates	Unique identifiers based on chromosome (referred to via GenBank ID and version) and genomic coordinate (GRCh38) corresponding to, e.g., CpG site
Peptides and proteins	UniProtKB Sequence and UniProtKB Accession Number
microRNAs (sequencing and qRT-PCR)	miRbase ID (release 20)
Protocols, methods, sample metadata	Following ISA schema, including Ontologies: Ontology for Biomedical Investigations (OBI), Chemical Methods Ontology (CHMO), Experimental Factor Ontology (EFO), NCI Thesaurus OBO Edition, Metabolomics Standards Initiative Ontology (MSIO), PRIDE Controlled Vocabulary
Roles and contributions	CRO - Contributor Role Ontology

In case uncommon or project specific ontologies or vocabularies are used we will provide mappings to more commonly used ontologies. For example, the X-omics ontology already includes other ontologies.

2.4. INCREASE DATA RE-USE (THROUGH CLARIFYING LICENSES)

To permit the widest re-use possible all researchers with a solid research plan (evaluated by the Data Access Committee) that can process the data in a secure and protected institutional environment will be provided access.

The data will be made available at the time of the first publication describing the data and the data will remain usable for 10 years.

Reuse is restricted to investigators (academic or industry) with a solid research plan and after signing a terms and conditions document that states that distribution to third parties is prohibited and that the privacy of participants can be guaranteed through storage in a secure, institutional server or data cloud that can only be accessed by the investigators.

Quality assurance

Quality Control measures will be published together with the data. The integrity of data files will be secured by storing an md5 checksum (or related measure) with the data.

3. DATA SECURITY

Personally identifiable information will be stored in the ClinData system developed by UP. It is specifically designed to provide a secure environment for storing and accessing this kind of data. The data are protected by two-phase authentication and authorization. There is no possibility of getting unauthorized access to the server even for some demonstration purposes, with all user accounts needing to be created by the administrator. The data are backed up daily to another facility, also located on the premises of Palacky University Olomouc under surveillance and with controlled access. The transferred data are fully encrypted. All user operations are logged.

Currently, there is no dedicated Data Protection Impact Assessment (DPIA) done on ClinData. However, the system is compliant with all legislation required to manage clinical data, which includes extensive regulation of personal data management. Also, the system architecture has been stress tested by the Czech army – as the architecture of ClinData is the foundation of a nationwide used system CovIT (database system for electronic management of laboratory samples tested for the presence of SARS-CoV-2 virus, serving as a full-fledged replacement for laboratory information management system LIMS).

ABBREVIATIONS

In the order of appearance:

WP – Work Package

DMP – Data Management Plan

FAIR - findable, accessible, interoperable and re-usable

DOB – Date of Birth

GDPR - General Data Protection Regulation

WGBS - Whole Genome Bisulfite Sequencing

DNA - Deoxyribonucleic Acid

RNA - Ribonucleic Acid

qRT-PCR - Real-Time Quantitative Polymerase Chain Reaction

TB – tera byte

TBD – to be determined

GB – giga byte

KB – kilo byte

MB – mega byte

BAM - Binary Alignment Map

DRE – Digital Research Environment

DOI - Digital Object Identifier

FDP - FAIR Data Points

ISA - Investigation/Study/Assay

DAC – Data Access Committee

BMI – Body Mass Index

EGA - European Genome-phenome Archive

DAA - Data Access Agreement

MSI - Metabolomics Standards Initiative

MINSEQE - Minimum Information About a Next-generation Sequencing Experiment

MIAPE - Minimum Information About a Proteomics Experiment

MIQE - Minimum Information for Publication of Quantitative Real-Time PCR Experiments ()

ChEBI ID - Chemical Entities of Biological Interest

UO - Units of Measurement Ontology

HPO - Human Phenotype Ontology

NCIT - National Cancer Institute Thesaurus

HGNC - HUGO Gene Nomenclature

GO - Gene Ontology



OBI - Ontology for Biomedical Investigations

CHMO - Chemical Methods Ontology

EFO - Experimental Factor Ontology

MSIO - Metabolomics Standards Initiative Ontology,

CRO - Contributor Role Ontology

DPIA – Data Protection Impact Assessment

ADJUSTMENTS MADE

This section will reflect the versioning of this deliverable.

HISTORY OF CHANGES		
Version	Publication date	Change
1.0	30.09.2020	▪ Initial version
1.1	21.12.2021	▪ Version 1.1
2.0	29.02.2024	▪ Publishable version

