ID-DarkMatter-NCD deliverable report

D7.1 DMP (Data management plan)

Lead beneficiary	1 - MUW	Due Date	30 June 2024
WP no	7	New due date (if delay)	n.a.
Task no	T7.3	Actual Delivery Date	28 June 2024
Dissemination level	PU – Public	Status	Submitted

Authors

Authors	Partner no	Partner organisation	Name of author
Main author	1	MUW	Thomas Vogl
Contributing author(s)	3	EUT-RS	Alberto Fresolone

Review

Authors	Partner no	Partner organisation	Name of author
Technical review	3	EUT-RS	Alberto Fresolone
Language review – *if applicable*	n.a.		

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101136582 as well as the Swiss State Secretariat for Education, Research and Innovation (SERI).

Document history

Date	Version	Chapters affected	Description of change	Author	Document status
24 May 2024	0.1	All	Setup of first draft	Thomas Vogl, MUW	DRAFT
19 June 2024	0.2	All	Minor revisions	Alberto Fresolone, EUT-RS	FINAL DRAFT
27 June 2024	1	n.a.	Final version & PDF conversion	Alberto Fresolone, EUT-RS	FINAL

List of abbreviations

BCR/TCR	B & T Cell Receptor
DNA	Deoxyribonucleic Acid
DoA	Description of Action
DOI	Digital Object Identifier
GA	General Assembly Meeting
HLA	Human Leukocyte Antigen
IR-NCDs	Immune Related Non-Communicable Diseases
ML	Machine Learning
NGS	Next Generation Sequencing
PhIP-Seq	Phage Immunoprecipitation Sequencing
WP	Work Package

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101136582 as well as the Swiss State Secretariat for Education, Research and Innovation (SERI).

Table of Content

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101136582 as well as the Swiss State Secretariat for Education, Research and Innovation (SERI).

Publishable Executive Summary

The Horizon Europe Model Grant Agreement requires that a data management plan ('DMP') is established and regularly updated. The official template recommended for Horizon Europe projects has been used. In completing the sections of the template, the requirements for research data management as outlined in Section 1.2.6 of the Description of Action (DoA), were addressed.

Description of deliverable

This deliverable of the DMP is related to task T7.3 Data management in the DoA (lead: MUW, participants: all partners for coordination). MUW, supported by the relevant WP leaders, summarizes the procedures and documentation required for the proper handling of research data of the project, as set out in management of data, according to the FAIR principles. This Data Management Plan (DMP, deliverable 7.1) specifies comprehensively all relevant data. This first version is prepared until M6, representing a living document that may be extended and revised to accommodate future developments.

1. Data Summary

Will you re-use any existing data and what will you re-use it for? State the reasons if re-use of any existing data has been considered but discarded.

We have already published datasets on antibody responses using the phage immunoprecipitation sequencing (PhIP-Seq) methodology also applied in this project (e.g. (Andreu-Sánchez et al., 2023; Bourgonje et al., 2023; Klompus et al., 2021; Leviatan et al., 2022; Vogl et al., 2021, 2022)). We may use this existing data for benchmarking comparisons. This data is publicly available via the journals' websites and public repositories. Similar considerations apply to genetic, metabolomics, and metagenomics data outlined in the DoA.

What types and formats of data will the project generate or re-use?

Primarily standard formats for protein/DNA sequences such as FASTA, standard Illumina next generation sequencing (NGS) data. Data from proprietary assays will be simply saved as .csv files. This is mostly relevant for the PhIP-Seq methodology, where we have already made data publicly available and have hence experience in using shareable formats (Andreu-Sánchez et al., 2023; Bourgonje et al., 2023; Klompus et al., 2021; Leviatan et al., 2022; Vogl et al., 2021, 2022). Other formats such as HLA analyses do not require specific formats (e.g. simple numerical values for promiscuity).

What is the purpose of the data generation or re-use and its relation to the objectives of the project?

The main use is for benchmarking, to compare immune responses of healthy individuals and other disease cohorts to the IR-NCDs, that encompass the focus of this project (for example (Vogl et al., 2021)).

What is the expected size of the data that you intend to generate or re-use?

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101136582 as well as the Swiss State Secretariat for Education, Research and Innovation (SERI).

Reuse of antibody data of up to 2,000 individuals (Andreu-Sánchez et al., 2023; Bourgonje et al., 2023; Klompus et al., 2021; Leviatan et al., 2022; Vogl et al., 2021, 2022), generation of new antibody datasets for approximately 7,000 individuals. Similar extends for BCR/TCR germline and HLA analyses, and hundreds of samples for metabolomics and metagenomics.

What is the origin/provenance of the data, either generated or re-used?

For re-used data see publications cited above, new data will be generated by the PhIP-Seq assay on the samples outlined in the grant agreement (blood), as well as germline BCR/TCR, metabolomics, metagenomics.

To whom might your data be useful ('data utility'), outside your project?

Immunologists/clinicians interested in genetic/environmental factors in IR-NCDs.

2. FAIR data

2.1. Making data findable, including provisions for metadata

Will data be identified by a persistent identifier?

Yes, DOIs will be assigned and we will make sure that the databases and repositories, where data will be deposited, are compatible with the FAIR principles of findability, accessibility, interoperability, and reusability including data search possibilities as well as quality and legacy controls.

Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

Metadata standards do not really exist for PhIP-Seq data, but this can be viewed rather as other “omics” data sets. In the past, we have published an open access paper explaining the cohorts and different meta data available making it directly available (e.g. conventional blood tests, that were correlated with antibody repertoires (Vogl et al., 2022)). We will follow this approach in this project, with metadata including standard blood tests, cohort info (age/sex, sample collection date) and metagenomic microbiome sequencing. This will also be applied for other data such as BCR/TCR germline sequencing (where no clear default format is avaiablle).

Will search keywords be provided in the metadata to optimize the possibility for discovery and then potential re-use?

Yes, we will use similar key words as for the publications in scientific journals, although other researchers will rather indirectly access our data via journal websites and the articles we will publish (as it had been done in the past with various datasets: (Andreu-Sánchez et al., 2023; Bourgonje et al., 2023; Klompus et al., 2021; Leviatan et al., 2022; Vogl et al., 2021, 2022)).

Will metadata be offered in such a way that it can be harvested and indexed?

Yes, as outlined above and exemplified with (Vogl et al., 2022).

This project has received funding from the European Union’s Horizon Europe research and innovation programme under grant agreement no. 101136582 as well as the Swiss State Secretariat for Education, Research and Innovation (SERI).

2.2. Making data accessible

Repository:

Will the data be deposited in a trusted repository?

Yes, all raw data will be stored in public repositories (such as the EMBL Nucleotide Sequence Database, or Zenodo/Figshare, HAL for general purposes).

Have you explored appropriate arrangements with the identified repository where your data will be deposited?

We have used these repositories in the past and do not expect any new arrangements to be necessary (Andreu-Sánchez et al., 2023; Bourgonje et al., 2023; Klompus et al., 2021; Leviatan et al., 2022; Vogl et al., 2021, 2022).

Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object?

Yes, this is a standard feature of these repositories.

Data:

Will all data be made openly available? If certain datasets cannot be shared (or need to be shared under restricted access conditions), explain why, clearly separating legal and contractual reasons from intentional restrictions. Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if opening their data goes against their legitimate interests or other constraints as per the Grant Agreement.

Yes, all data will be made available.

If an embargo is applied to give time to publish or seek protection of the intellectual property (e.g. patents), specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

As outlined in the consortium agreement (further details therein):

Prior notice of any planned publication shall be given to the other Parties at least 45 calendar days before the publication.

Any objection to the planned publication shall be made in accordance with the Grant Agreement by written notice to the Coordinator and to the Party or Parties proposing the dissemination within 30 calendar days after receipt of the notice.

An objection is justified if

a) the protection of the objecting Party's Results or Background would be adversely affected, or

b) the objecting Party's legitimate interests in relation to its Results or Background would be significantly harmed, or

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101136582 as well as the Swiss State Secretariat for Education, Research and Innovation (SERI).

c) the proposed publication includes Confidential Information of the objecting Party.

The objecting Party can request a publication delay of not more than 90 calendar days from the time it raises such an objection. After 90 calendar days the publication is permitted, provided that the objections of the objecting Party have been addressed.

Will the data be accessible through a free and standardized access protocol?

n.a.

If there are restrictions on use, how will access be provided to the data, both during and after the end of the project?

n.a.

How will the identity of the person accessing the data be ascertained?

n.a.

Is there a need for a data access committee (e.g. to evaluate/approve access requests to personal/sensitive data)?

n.a.

Metadata:

Will metadata be made openly available and licenced under a public domain dedication CC0, as per the Grant Agreement? If not, please clarify why. Will metadata contain information to enable the user to access the data?

See above 2.1, metadata will also be made openly available.

How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available?

As we are using the same storage options (which do not have any annotated expiry dates) for both data/metadata this is not an issue.

Will documentation or reference about any software be needed to access or read the data be included? Will it be possible to include the relevant software (e.g. in open source code)?

If custom code should be used this will be made available (as for example in (Vogl et al., 2022) in a GitHub repository, together with a copy of the data and metadata).

2.3. Making data interoperable

What data and metadata vocabularies, standards, formats or methodologies will you follow to make your data interoperable to allow data exchange and re-use within and across disciplines? Will you follow community-endorsed interoperability best practices? Which ones?

We will mostly use standard life sciences formats as outlines in section 1.

In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies,

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101136582 as well as the Swiss State Secretariat for Education, Research and Innovation (SERI).

will you provide mappings to more commonly used ontologies? Will you openly publish the generated ontologies or vocabularies to allow reusing, refining or extending them?

As outlines above, if custom code should be used this will be made available (as for example in (Vogl et al., 2022) in a GitHub repository, together with a copy of the data and metadata)

Will your data include qualified references[1] to other data (e.g. other data from your project, or datasets from previous research)?

Potentially yes, as oultined in section 1 on data re-use.

2.4. Increase data re-use

How will you provide documentation needed to validate data analysis and facilitate data re-use (e.g. readme files with information on methodology, codebooks, data cleaning, analyses, variable definitions, units of measurement, etc.)?

We have so far published peer-reviewed articles on our data, that requires stringent documentation on data reuse (Andreu-Sánchez et al., 2023; Bourgonje et al., 2023; Klompus et al., 2021; Leviatan et al., 2022; Vogl et al., 2021, 2022) and we will follow the same approach in this project (e.g. see the GitHub and journals readmes of (Vogl et al., 2022)).

Will your data be made freely available in the public domain to permit the widest re-use possible? Will your data be licensed using standard reuse licenses, in line with the obligations set out in the Grant Agreement?

Data will be made freely available using standard licences alongside open access publication of the corresponding research articles.

Will the data produced in the project be useable by third parties, in particular after the end of the project?

Yes, see the previous section on expiry of deposited data for details.

Will the provenance of the data be thoroughly documented using the appropriate standards?

n.a.

Describe all relevant data quality assurance processes. Further to the FAIR principles, DMPs should also address research outputs other than data, and should carefully consider aspects related to the allocation of resources, data security and ethical aspects.

As outlines above, DOIs will be assigned and we will make sure that the databases and repositories, where data will be deposited, are compatible with the FAIR principles of findability, accessibility, interoperability, and reusability including data search possibilities as well as quality and legacy controls. All raw data will be stored in public repositories (such as the EMBL Nucleotide Sequence Database, or

[1] *A qualified reference is a cross-reference that explains its intent. For example, X is regulator of Y is a much more qualified reference than X is associated with Y, or X see also Y. The goal therefore is to create as many meaningful links as possible between (meta)data resources to enrich the contextual knowledge about the data. (Source: https://www.go-fair.org/fair-principles/i3-metadata-include-qualified-references-metadata/)*

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101136582 as well as the Swiss State Secretariat for Education, Research and Innovation (SERI).

Project funded by

Zenodo/Figshare, HAL for general purposes). See the Ethics section of the proposal for details on these aspects.

3. Other research outputs

In addition to the management of data, beneficiaries should also consider and plan for the management of other research outputs that may be generated or re-used throughout their projects. Such outputs can be either digital (e.g. software, workflows, protocols, models, etc.) or physical (e.g. new materials, antibodies, reagents, samples, etc.). Beneficiaries should consider which of the questions pertaining to FAIR data above, can apply to the management of other research outputs, and should strive to provide sufficient detail on how their research outputs will be managed and shared, or made available for re-use, in line with the FAIR principles.

We will make all measurements and results as well as code, scripts, and machine learning algorithms publicly available (in GitHub repositories). All raw data will be stored in public repositories (such as the EMBL Nucleotide Sequence Database, or Zenodo/Figshare, HAL for general purposes). Machine learning algorithms will be annotated and published in a GitHub repository (along scientific, peer-reviewed publications explaining their generation and use). Summary statistics and aggregated results (as well an insights on the biomarkers) will also be published in peer reviewed open access papers.

4. Allocation of resources

What will the costs be for making data or other research outputs FAIR in your project (e.g. direct and indirect costs related to storage, archiving, re-use, security, etc.) ?

In general, the databases we are using for nucleotide data etc. are free to use, also GitHub for code is free to use. Furthermore, we have reserved a designated budget for open access journals.

How will these be covered? Note that costs related to research data/output management are eligible as part of the Horizon Europe grant (if compliant with the Grant Agreement conditions)

See above.

Who will be responsible for data management in your project?

Overall, the Principal Investigator / Project Manager. Each partner will be responsible for ensuring that the formal agreements needed to access existing data from any external data controllers are in place. Each partner is also responsible for the uploading of these data to the common data storage platforms, and for all data management pertaining to their own data, uploaded or not.

However, the Principal Investigator will oversee the data management of the project.

How will long term preservation be ensured? Discuss the necessary resources to accomplish this (costs and potential value, who decides and how, what data will be kept and for how long)?

See sections above on expiry dates.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101136582 as well as the Swiss State Secretariat for Education, Research and Innovation (SERI).

5. Data security

What provisions are or will be in place for data security (including data recovery as well as secure storage/archiving and transfer of sensitive data)?

The overall strategy for data collection, storage, and access adhere to legislations and directives within the EU (e.g. password protected, project specific GitHub repositories etc.).

Will the data be safely stored in trusted repositories for long term preservation and curation?

Yes, as outlined above, we will make all measurements and results as well as code, scripts, and machine learning algorithms publicly available (in GitHub repositories). All raw data will be stored in public repositories (such as the EMBL Nucleotide Sequence Database, or Zenodo/Figshare, HAL for general purposes). Machine learning algorithms will be annotated and published in a GitHub repository (along scientific, peer-reviewed publications explaining their generation and use). Summary statistics and aggregated results (as well an insights on the biomarkers) will also be published peer reviewed open access papers

6. Ethics

Are there, or could there be, any ethics or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).

Only minor issues are expected in this regard. See the Ethics section of the DoA and D7.2 (Ethics Report – from the inaugural meeting of the EGC) for details on these aspects.

Will informed consent for data sharing and long term preservation be included in questionnaires dealing with personal data?

Since the cohorts have already been assembled and all samples collected, this issue has been taken care of, see the uploaded documents for the Ethics section in the DoA.

7. Other issues

Do you, or will you, make use of other national/funder/sectorial/departmental procedures for data management? If yes, which

n.a.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101136582 as well as the Swiss State Secretariat for Education, Research and Innovation (SERI).

Disclaimer

Acknowledgement of funding

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101136582 as well as the Swiss State Secretariat for Education, Research and Innovation (SERI).

Co-funded by the European Union

Project funded by
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Swiss Confederation
Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI