Ref. Ares(2019)2891262 - 30/04/2019

Project Acronym: Nutrishield
Grant Agreement number: 818110 (H2020-SFS-2018-IA)
Project Full Title: Fact-based personalised nutrition for the young

DELIVERABLE

D1.2 – Data Management Plan

Dissemination level	PU - Public
Type of Document	Report
Contractual date of delivery	30/04/2019
Deliverable Leader	ALPES LASERS
Status & version	Final, V1.0 – 30/04/2019
WP responsible	ALPES LASERS
Keywords:	DMP

Deliverable Leader:	ALPES LASERS
Contributors:	Cranfield University, All
Reviewers:	Alpes Lasers
Approved by:	Full name(s) (organisation)

Document History			
Version	**Date**	**Contributor(s)**	**Description**
v0.1	05/03/2019	ALPES	Draft
v0.9	29/04/2019	CU	First complete version
v1.0	30/04/2019	INTRA, CU, ALPES	Final complete version

This document is part of a project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 818110. It is the property of the NUTRISHIELD consortium and shall not be distributed or reproduced without the formal approval of the NUTRISHIELD Management Committee. The content of this report reflects only the authors' view. EC is not responsible for any use that may be made of the information it contains.

Executive Summary

This deliverable summarizes the Data Management Plan (DMP) for Nutrishield. This is the first version of a living document, its content will be updated as the project progresses.

Table of Contents

Definitions, Acronyms and Abbreviations

Acronym	Title
DMP	Data Management Plan

1. Data Summary

What is the purpose of the data collection/generation and its relation to the objectives of the project?
What types and formats of data will the project generate/collect?
Will you re-use any existing data and how?
What is the origin of the data?
What is the expected size of the data?
To whom might it be useful ('data utility')?

This project focuses on developing a novel platform for personalised nutrition for the young, based on mobile and Web platform as well as machine learning. Our approach will be based on integrating genomic profiling of patients in three disease groups (diabetes, lactating mothers for milk alregy babies, and obesity in children) done as part of a clinical study at the hospital setting within some of our project partners workplace. The project methodology in itself well in line with the European Commission (Horizon 2020) guidelines and all the necessary ethical and clinical approvals will be obtained in advance of the clinical study by the involved partner(s) prior to the study. However, in order to enable fast and efficient data maintenance and sharing of the project results. Anonymised genotypic and clinical profile data will be provided by all the involved partners to Cranfield University, who will be in charge of providing the hardware and software infrastructure for this purpose.
We will perform genotyping by sequencing followed by variant calling for our patient subject using Illumina HiSeq paired-end reads. Raw sequence reads will be generated in FASTQ format, while assembled genome sequence will be developed in FASTA format. Genome functional annotation will be generated in TSV and GFF3 formats. Genotyping by Sequencing (GBS) will be performed to identify variants among patients and correlate them with the relevant disease group. Variant calling files will be generated in VCF format (Expected to be ~1Gb file per patient, our clinical study will include ~120 individuals in total

2. FAIR data

2.1.Making data findable, including provisions for metadata

Are the data produced and/or used in the project discoverable with metadata, identifiable and locatable by means of a standard identification mechanism (e.g. persistent and unique identifiers such as Digital Object Identifiers)?
What naming conventions do you follow?
Will search keywords be provided that optimize possibilities for re-use?
Do you provide clear version numbers?

What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

Data Standards

Genomic sequence data be generated in FASTA format while Variant Calling files will be in VCF format. All patient details and metadata will be anonymised before being sent to Cranfield as per the partner clinical regulation. Additional clinical data and metabolic profile will be mostly in a tabular format (Anonymised patient IDs as rows and biomarkers are columns) in MS Excel or CSV format (relatively small files of <500kb each).

Nutrishield experimental results (genotypes and clinical/metabolic profiles) will be hosted on a RAID expansion server which provides 120 TB of raw storage capacity. The expansion server will be added to our existing HPC storage facility hosted at Cranfield University. The storage facility is equipped with enterprise-grade storage hard drives configured on a RAID10 system which provides data protection during multiple drives failure. Data stored on the RAID expansion server will be supported for a period of 10 years beyond the course of this project.

2.2.Making data openly accessible

Which data produced and/or used in the project will be made openly available as the default? If certain datasets cannot be shared (or need to be shared under restrictions), explain why, clearly separating legal and contractual reasons from voluntary restrictions.
Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if relevant provisions are made in the consortium agreement and are in line with the reasons for opting out.
How will the data be made accessible (e.g. by deposition in a repository)?
What methods or software tools are needed to access the data?
Is documentation about the software needed to access the data included?
Is it possible to include the relevant software (e.g. in open source code)?
Where will the data and associated metadata, documentation and code be deposited? Preference should be given to certified repositories which support open access where possible.
Have you explored appropriate arrangements with the identified repository?
If there are restrictions on use, how will access be provided?
Is there a need for a data access committee?
Are there well described conditions for access (i.e. a machine readable license)?
How will the identity of the person accessing the data be ascertained?

Genomic data will be submitted and made publically available on public sequence repositories such as the European Nucleotide Archive (EBI-ENA) and the Sequence Read Archive (SRA). Clinical finding and metabolic profiles will be submitted as appendices and tables within peer-reviewed journals and submitted to public repositories when possible.

2.3.Making data interoperable

Are the data produced in the project interoperable, that is allowing data exchange and re-use between researchers, institutions, organisations, countries, etc. (i.e. adhering to standards for formats, as much as possible compliant with available (open) software applications, and in particular facilitating re-combinations with different datasets from different origins)?
What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?
Will you be using standard vocabularies for all data types present in your data set, to allow inter-disciplinary interoperability?
In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies?

API/REST services will be integrated within the Nutrishield platform, to allow import/export of genomic sequence data to/from public repositories such as the European Nucleotide Archive (EBI-ENA) and the Sequence Read Archive (SRA).

2.4.Increase data re-use (through clarifying licences)

How will the data be licensed to permit the widest re-use possible?
When will the data be made available for re-use? If an embargo is sought to give time to publish or seek patents, specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.
Are the data produced and/or used in the project useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why.
How long is it intended that the data remains re-usable?
Are data quality assurance processes described?
Further to the FAIR principles, DMPs should also address:

N/A

3. Allocation of resources

What are the costs for making data FAIR in your project?
How will these be covered? Note that costs related to open access to research data are eligible as part of the Horizon 2020 grant (if compliant with the Grant Agreement conditions).
Who will be responsible for data management in your project?
Are the resources for long term preservation discussed (costs and potential value, who decides and how what data will be kept and for how long)?

Current computing facilities include three high-performance computing (HPC) server nodes (total 220 core processors and 1TB RAM on each node) used mainly for *de-novo*genome assembly tasks and RNA-Seq analysis and two application servers used for hosting multiple Web applications and online research frameworks. Storage capability includes a QNAP Enterprise grade hybrid application server (TDS-16489U), providing a total of 160 TB of raw storage expandable up to 1 PB. The processing power and memory for this server is well-suited for this project, however, additional storage space was required as part of Cranfield University budget to manage and store the sequencing data generated. We request sum for a 3U QNAP RAID expansion server (REXP-1620U-RP); this will be connected to our existing platform and will provide additional 80 TB of raw storage. The expansion server will be configured at RAID10, which can resist multiple hard drive failures and offers real-time disaster recovery functionality.

4. Data security

What provisions are in place for data security (including data recovery as well as secure storage and transfer of sensitive data)?
Is the data safely stored in certified repositories for long term preservation and curation?

Our data storage facility will be behind the Cranfield University firewall and all the security patches are kept up-to-date as per the institute ICT standards. The data will be anonymised before being sent by the clinical partner to be stored on our facility to ensure that any patient-sensitve information is physically located within the clinical setting. Regular invasion tests are carried out on all of Cranfield University servers to ensure their compliance.

5. Ethical aspects

Are there any ethical 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).
Is informed consent for data sharing and long term preservation included in questionnaires dealing with personal data?

Ethics forms related to the clinical studies will be submitted separately by the relevant partners as part of the project delivery

6. Other issues

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

N/A