NEXTGEN PROTEINS

Bioconversion of Underutilized Resources into Next Generation Proteins for Food and Feed

Project start: 01 October 2019 **Project duration: 48 months**

Deliverable No 7.6.

Deliverable Title: Data Management Plan

Version 2.1

Lead author/editor: TTZ

Due Date of Submission: 31 March 2020
Submission Date: 04 May 2021

This Project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement no. 862704.
www.nextgenproteins.eu

0 Document Information

Document Data

Work package related	Work package 7: Stakeholder involvement, dissemination and exploitation of results
Task related	Task 7.5: Creating a Data Management Plan under the H2020 Open Research Data Pilot
Type	ORDP: Open Research Data Pilot
Dissemination level	Public
Keywords	Research data, Data Management Plan

Contributors

Authors	**Organisations name**	**E-Mail**
Marie Shrestha (MSh)	TTZ	mshrestha@ttz-bremerhaven.de
Birgir Smárason (BS)	MATIS	birgir@matis.is
Guðmundur Stefánsson (GSt)	MATIS	gst@matis.is
Margrét Geirsdóttir (MG)	MATIS	mg@matis.is
Rósa Jónsdóttir (RJ)	MATIS	rosa@matis.is
Björg Jónsdóttir (BJ)	MATIS	bjorg@matis.is
Reynir Smári Atlason (RSA)	CIRCULAR	reynir@circularsolutions.is
Melissa Tillotson	WAITROSE	melissa.tillotson@waitrose.co.uk
Ann-Kristin Schwarze	BIOZOON	Schwarze@biozoon.de
Charilaos Xiros	PROCESSUM	charilaos.xiros@processum.se
Pennanen Kyösti	VTT	Kyosti.Pennanen@vtt.fi
Haris Hondo	RISE	haris.hondo@ri.se

Document history

Document version #	**Date**	**Notes/Change**	**Status**
V. 0	03.02.2020	MSh prepared preliminary draft	draft
V. 0	06.03.2020	BS, GSt and MG reviewed preliminary draft	reviewed
V. 0	09.03.2020	RSA reviewed preliminary draft	reviewed
V. 1	18.03.2020	MSh compiled version V.1 based on reviews	reviewed
V.1	27.03.2020	BS, RJ, BJ and GSt reviewed version V.1	reviewed
V.2	30.03.2020	MSh compiled final version	Final
V2.1	06.04.2021	MSh updated final version with input from project partners regarding updated data sets (Revision at Month 18)	Updated Final

Table of content

1 Executive summary

NextGenProteins is participating in the H2020 Open Research Data Pilot. This voluntary participation entails three requirements:

(1) That NextGenProteins partner will deposit non-confidential collected research data in data repositories;
(2) That NextGenProteins project will take measures to enable third parties to access, mine, exploit, reproduce and disseminate this research data; and
(3) That NextGenProteins consortium will develop a Data Management Plan (DMP) detailing what kind of data the project is expected to generate, whether and how it will be exploited or made accessible for verification and reuse, and how it will be curated and preserved.

The NextGenProteins DMP is an open report describing the data life cycle, data summary, what and how the data is made FAIR, data sharing and IPR management, allocation of resources, data security and ethical aspects.

The current document represents the first version of the NextGenProteins DMP. The DMP will be a living document in which information will be revised through updates as the implementation of the project progresses and when significant changes occur in term of data or consortium policies and/or composition. The current Preliminary version of the DMP will be regularly revised during the lifespan of the project along the periodic assessment of the project (reviews in M18, M36 and M48).

2 Introduction

The objective of the Data Management Plan (DMP) is to formulate relevant aspects of making FAIR data– findable, accessible, interoperable and re-usable by encouraging sound data management as an essential part of research best practice.

The DMP defines the data management life cycle for the data to be collected, processed and/or generated in the frame of the NextGenProteins data management policy.

The Data Management Plan includes:

(1) the definition of the type of data that will be collected, processed and generated in the frame of the NextGenProteins project,
(2) whether and how the data will be shared and/or made accessible for verification and re-use, and
(3) how the data will be curated and preserved.

Participants of the NextgenProteins consortium must follow this DMP when managing NextGenProteins related data.

The data management is in accordance with European Commission (EC) guidelines and with the open access, FAIR and IPR data principles. In that sense, the NextGenProteins consortium might define certain datasets to remain closed for potential commercial exploitation according to the principle “as open as possible, as closed as necessary”.

The DMP reflects the current state of consortium agreements on data management and is consistent with exploitation and Intellectual Property Rights (IPR) requirements.

Note: The current document represents the first version of the NextGenProteins Data Management Plan. The DMP will be a living document in which information will be revised through updates as the implementation of the project progresses and when significant changes occur in term of data or consortium policies and/or composition. The current Preliminary version of the DMP will be regularly revised during the lifespan of the project along the periodic assessment of the project (reviews in M18, M36 and M48).

3 Data Life Cycle

The NextGenProteins Data Management Plan covers all the data lifecycle steps of the research data generated or collect in the project and is an important aspect to provide the project sustainability and security. Each dataset can be preserved in different way and may have different data access or usage policies. That is why it is of particular importance to keep track of the data lifecycle in NextGenProteins data management plan.

The lifecycle of the data is as follow:

- **Data Collection**: The first step is to collect/create data and to keep it in a workspace (backups are recommended);
- **Data Processing**: The second step is to identify, analyse and process data, always ensuring the quality of data. Before starting work with data, it is advisable to copy the raw data. The analysis of the research data may also require the collection of new data for the same or for other projects purposes;
- **Data Storage**: The data needs to be organized by specifying and choosing the file formats, its access policy, its metadata and must also be deposited in an online (and also local) repository. When the data is on a repository all the efforts need to be made to allow its long-term preservation;
- **Data Sharing**: After depositing the data in an online repository, it is are available to be accessed and discovered by other consortium members, and can then be used for other purposes (re-use).

Figure 1: Data Lifecycle (www.spirion.com)

4 Data Summary

4.1 Purpose of the data collection / generation

The objective of the NextGenProteins project is to optimise and validate, in an industrially relevant environment, the production of proteins from microalgae, single cells and insects and demonstrate their suitability as alternative sustainable sources in food and feed value chains.

The project will serve as a platform for industrial partners/entrepreneurs to take their innovations to the next level by turning them into relevant, credible products and thus, accelerate market-driven, customer- and consumer-responsive innovative EU alternative protein production. This will contribute to EU's food security and its goal of future proofing food and feed supply chains in a world faced with climate change, resource scarcity, increasing waste and aging population.

The methodology of NextGenProteins will focus on bioconversion of underutilized resources with the intention of expanding the sustainable European protein selection available for food and feed. Underutilized resources, e.g. plant-based biomass, often contain low or no amounts of protein; still they can be transformed through organism-based conversion processes that break them down, utilize their energy and convert into high quality proteins.

4.2 Types and formats of data the project will generate/collect

The NextGenProteins project will generate different datasets such as:

- Experiment data: characterisation of alternative proteins
- Sensory evaluation data of alternative proteins
- Environmental data: Production data – inputs/outputs, energy use (for LCA)
- Experiment data: application of alternative proteins
- Sensory evaluation data: application food products
- Consumer acceptance data: application food products
- Authorisations for animal experiments
- Animal testing data
- Sensory evaluation data
- Consumer insight data, Commercial data
- Stakeholder list - Personal data

The different datasets to be generated and or collected have different Access and Sharing Requirements. Therefore, special focus will be put on developing a platform for the creators of the data to communicate the desired use of the data. This may be different based on different companies sharing their data.

	Tecnical Information					Technical Description			Data Set - Access & Sharing					Archiving & Preservation				
	Partner Responsible	Data Set Name	Data Set reference	Data Set Description & Purpose	Category Type	Format	Size	Metadata & Standards	Access type	Access closed/restricted	Access requirements	Sharing type	Sharing unavailable	Data Collector	Quality Assurance Procedures	Archive & Preservation	Period of Time	Associated Costs
1.	RISE Processum	SCP cultivation properties and parameters	WP2	This data set will contain experimental results from SCP fermentation experiments. Certain parameters and properties will be monitored and optimized with regard to fish growth and health	Experiment data	Word, xls, Graphs	Up to 10 MB		Restricted	keep private/restricted untill publication or patent application		For Project purposes, only internal use		RISE	Procedures will follow RISE guidelines	RISE Processum archives		
2.	MATIS	FUNCTIONAL_PROPERTIES		This dataset contains results from assessment of functional properties directly in solution and in model systems to look at protein-protein interactions; alone, in mixture with each other and other food ingredients.	Experiment data: characterisation of alternative proteins	Excel, Word	1MB	Metadatafile with methodology and sample information	Restricted	Restricted within the consortium until agreed on publication	Excel, Word	Project use only until decided otherwise	Project use only until decided otherwise. Can contain sensitive information and/or good science results that could be shared in a peer reviewed article	MATIS	Trial protocol will ensure data quality. Data will be stored on a secured, IT specialists maintained internal server, in a NextGenProteins specific folder	Data will be stored on a secured, IT specialists maintained internal server, in a NextGenProteins specific folder	5 years, or until the data is made public or published in the scientific journals	Included in overhead
3.	biozoon	Characterisation report		Alternative protein characterisation will be performed and analysed in order to validate their functional properties (viscosity, solubility, sensory, etc.) within aqueous solution. Protein interactions with different hydrocolloids will be analysed.	Experiment data: characterisation of alternative proteins	Excel, Word	up to few GB	-	Private for company	Potential IPR	data saved on company internal server	only internal	potential IPR	BIOZOON Data protection manager and employees working in project	internal company server security will be applied	internal server	over 5 years	server specific cost (overheads)
4.	RISE	Raw material characterization		New protein raw materials provided by partners will be chacterized for important properties and functionality in order to increase understanding on how they can be used in food formulations. Properties that will be characterized include: water activity, colour, protein content, water content, protein solubility, foaming ability, gelling ability, emulsification ablity and stability, water holding capacity, oil holding capacity. Data will be collected by RISE in laboratory trials and will be stored and protected on RISE servers	Experimental data	Excel	1MB		Restricted	Data will be kept restricted due to possible publication in scientific journals later in the project	Data will be uploaded to appropriate channels to provide access to consortium members			RISE		Data will primarily be archived and preserved on RISE servers	Indefinetly	
5.	MATIS	SENSORY_PROPERTIES		This dataset contains results from evaluation of sensory properties (taste, odour and texture parameters) of the alternative proteins will be tested directly in suitable solutions and in model systems.	Organoleptic testing data of alternative proteins	Excel, Word	1MB	Metadatafile with methodology and sample information	Restricted	Restricted within the consortium until agreed on publication	Excel, Word	Project use only until decided otherwise. Rawdata only by partner.	Project use only until decided otherwise. Can contain sensitive information and/or good science results that could be shared in a peer reviewed article without sensitive information	MATIS	Trial protocol will ensure data quality. Raw Data will be stored on a secured, IT specialists maintained internal server, of partner. General data, non sensitive in a NextGenProteins specific folder	Raw Data will be stored on a secured, IT specialists maintained internal server, of partner. General data, non sensitive in a NextGenProteins specific folder	5 years, or until the data is made public or published in the scientific journals	Included in overhead
6.	CIRCULAR	RISE_LCI	CS_RISE_LCI	Contains the Life Cycle Inventory (LCI) to calculate the LCA for the protein production and understand the environmental impact. Data collection took place through google sheets. The provider of data was instructed, through a video tutorial, how to enter the data. The data is stored, until otherwise instructed, on the cloud provided by Google, administered by CIRCULAR.	Environmental data	Excel (Google sheets)	Approx 1 MB	No metadata is intentionally created	Private until otherwise decided	Sensitive Data: company may not want to disclose detail of their production processes	Microsoft Excel or Google Sheets	The dataset will not be available for sharing unless otherwise instructed from the providers of raw data, or a consensus is made within the project to share the dataset.	The sharing of the data may have heavy financial impact on the reporting companies and weaken their competitive status.	Collector: CIRCULAR Security: CIRCULAR (GOOGLE)	The data is stored in the Google cloud, within CIRCULAR's own area. No data is stored on USB's, personal computers or otherwise. However, when the LCA models will be structured, the data will be processed in personal computers.	Until otherwise instructed, the data will be stored on CIRCULAR's Google cloud platform.	5 years, or until the data is made public, or the data has been processed in an LCA and the results published in the scientific literature.	Subscription services to cloud platforms.
7.	CIRCULAR	MUTATEC_LCI	CS_MUTATEC_LCI															
8.	CIRCULAR	ENTOCUBE_LCI	CS_ENTOCUBE_LCI															
9.	CIRCULAR	ARBIOM_LCI	CS_ARBIOM_LCI															
10.	CIRCULAR	ALGAENNOVATION_LCI	CS_ALGAENNOVATION_LCI															
11.	biozoon	Incorporation report		Application of alternative proteins will be performed and analysed in order to validate their performance and acceptance within texture modified model foods for elderly nutrition. Additionally, 3D printing feasibility will be analysed within this data set.	Experiment data:application of alternative proteins	Excel, Word	up to few GB	-	Private for company	Potential IPR	data saved on company internal server	only internal	potential IPR	BIOZOON Data protection manager and employees working in project	internal company server security will be applied	internal server	over 5 years	server specific cost (overheads)
12.	RISE	Product characterization (bakery products)		Properties of bakery products formulated including new protein sources will be analysed to order to understand how the new proteins are affecting the product properties. Data that will be collected during the project will include: microscopic structure, volume, water content, bake loss, porosity, appearance, colour, smell, taste, texture, processability of dough. Data will be collected by RISE in collaboration with Fazer during baking trials. Data will be stored and protected on RISE servers	Experimental data	Excel	1MB		Restricted	Data will be kept restricted due to possible publication in scientific journals later in the project. Additionaly,dataset could include data that is crutial for competitivness of industrial partner	Data will be uploaded to appropriate channels to provide access to consortium members			RISE		Data will primarily be archived and preserved on RISE servers	Indefinetly	
13.	RISE	Product characterization (emulsified products)		Properties of emulsified products formulated including new protein sources will be analysed to order to understand how the new proteins are affecting the product properties. Data that will be collected during the project will include: microscopic structure, water content, cooking loss, appearance, colour, smell, taste, texture, emulsion stability, processability of batter. Data will be collected by RISE in collaboration with Härryda Karlsson during emusification trials. Data will be stored and protected on RISE servers	Experimental data	Excel	1MB		Restricted	Data will be kept restricted due to possible publication in scientific journals later in the project. Additionaly,dataset could include data that is crutial for competitivness of industrial partner	Data will be uploaded to appropriate channels to provide access to consortium members			RISE		Data will primarily be archived and preserved on RISE servers	Indefinetly	
14.	MATIS	SENSORY_PROPERTIES_GK	Task 3.1	This dataset contains results from analysis of sensory properties of food products containing the alternative proteins.	Organoleptic testing data: application food products	Excel, Word	1MB	Statistical metadata with methodology and sample information	Restricted	Restricted within the consortium until agreed on publication	Excel, Word	Project use only until decided otherwise	Project use only until decided otherwise. Can contain sensitive information and/or good science results that could be shared in a peer reviewed article without sensitive information	MATIS	Trial protocol will ensure data quality. Raw Data will be stored on a secured, IT specialists maintained internal server, of responsible partner. General data, non sensitive in a NextGenProteins specific folder	Raw Data will be stored on a secured, IT specialists maintained internal server, of responsible partner. General data, non sensitive in a NextGenProteins specific folder	5 years, or until the data is made public or published in the scientific journals	Included in overhead.
15.	FAZER & RISE	SENSORY_PROPERTIES_FAZER	Task 3.2											RISE/FAZER Sweden				
16.	HARRYDA & RISE	SENSORY_PROPERTIES_HARRYDA	Task 3.3											RISE/Harryda Karlson				
17.	BIOZOON & TTZ	SENSORY_PROPERTIES_BIOZOON	Task 3.4											BIOZOON/TTZ				
18.	MATIS	Consumersurvey_CLT_OF F/ONLINE_GK	Task 3.1	This dataset contains results from consumer liking and attitudes of meals containing the alternative proteins. It will be registered by Matis in an excel file, kept on NextGenProteins specific folder on the Matis internal network and all data will be treated securely as described by the EU General Data Protection Regulation (GDPR) and stored appropriately	Non-personal identifiable data, Consumer acceptance data: application food products	Excel, Word	4MB	Statistical metadata with methodology and sample information	Restricted	Restricted within the consortium until agreed on publication	Excel, Word	Project use only until decided otherwise. Rawdata only by responsible partner.	Project use only until decided otherwise. Can contain sensitive information and/or good science results that could be shared in a peer reviewed article without sensitive information	MATIS	Trial protocol will ensure data quality. Raw Data will be stored on a secured, IT specialists maintained internal server, of responsible partner. General data, non sensitive in a NextGenProteins specific folder	Data will be stored on a secured, IT specialists maintained internal server, in a NextGenProteins specific folder. All data will be treated securely as described by the EU General Data Protection Regulation (GDPR) and stored appropriately	5 years, or until the data is made public or published in the scientific journals	Other cost is consumables re consumer testing (ca 1000 EUR)
19.	MATIS / FAZER / RISE	Consumersurvey_CLT_OF F/ONLINE_FAZER	Task 3.2															
20.	MATIS / HARRYDA / RISE	Consumersurvey_CLT_OF F/ONLINE_HARRYDA	Task 3.3															
21.	MATIS / BIOZOON / TTZ	Consumersurvey_CLT_OF F/ONLINE_BIOZOON	Task 3.4															
22.	Aquascot & Waitrose	Ethics requirements	WP 9 deliverable 9.3	Authorisations for animal experiments kept on file, procedures to ensure animal welfare and training certificates for personnel involved in animal experiments. While WR is not directly involved in the live fish experiments, Aquascot is, so data here would be gathered and stored.	Authorisations for animal experiments	Word	up to few GB		Private	For project use only	to be confirmed	For project use only	For project use only	Aquascot	to be defined	to be defined	to be defined	to be defined

	Tecnical Information					Technical Description			Data Set - Access & Sharing					Archiving & Preservation				
	Partner Responsible	Data Set Name	Data Set reference	Data Set Description & Purpose	Category Type	Format	Size	Metadata & Standards	Access type	Access closed/restricted	Access requirements	Sharing type	Sharing unavailable	Data Collector	Quality Assurance Procedures	Archive & Preservation	Period of Time	Associated Costs
23.	MATIS	SALMON_LAB_TRIAL_ALGAENNOVATION		This dataset contains results from the salmon lab trial on growth, welfare, digestibility, feed intake, mortality and other trial parameters and conditions. It will be registered by Matis in an excel file, kept on NextGenProteins specific folder on the Matis internal network	Animal testing data	Excel, Word	1MB	No metadata	Restricted	Restricted within the consortium until agreed on publication	Excel, Word	Project use only until decided otherwise	Project use only until decided otherwise. Can contain sensitive information and/or good science results that could be shared in a peer reviewed article	MATIS	Trial protocol will ensure data quality. Data will be stored on a secured, IT specialists maintained internal server, in a NextGenProteins specific folder	Data will be stored on a secured, IT specialists maintained internal server, in a NextGenProteins specific folder	5 years, or until the data is made public or published in the scientific journals	Included in overhead
24.	MATIS	SALMON_LAB_TRIAL_MUTATEC																
25.	MATIS	SALMON_LAB_TRIAL_ARBIOM																
26.	MATIS	SALMON_LAB_TRIAL_PROCESSUM																
27.	MATIS	SALMON_FIELD_TRIAL_ALGAENNOVATION		This dataset contains results from the salmon field trial on growth and other trial parameters and conditions. It will be registered by MOWI, kept on MOWIS internal network	Animal testing data	Excel, Word	1MB	No metadata	Restricted	Restricted within the consortium until agreed on publication	Excel, Word	Project use only until decided otherwise	Project use only until decided otherwise. Can contain sensitive information and/or good science results that could be shared in a peer reviewed article	MOWI	Trial protocol will ensure data quality. Data will be stored on a secured internal server	Data will be stored on a secured internal server	5 years, or until the data is made public or published in the scientific journals	Included in overhead
28.	MATIS	SALMON_FIELD_TRIAL_MUTATEC																
29.	MATIS	SALMON_FIELD_TRIAL_ARBIOM																
30.	Aquascot & Waitrose	Quality (Organoleptic) tests	WP 4	Quality of harvested salmon will be evaluated both by Aquascot and third party contracted by waitrose	Organoleptic testing data	to be confirmed			Private	For project use only	to be confirmed	For project use only	For project use only	Aquascot and third party contracted by waitrose	to be defined	to be defined	to be defined	to be defined
31.	Waitrose	Waitrose consumer insight and data in the market	WP5	Waitrose might gather consumer insight data and use it for the purposes of NEXTGENProteins in WP5.	Consumer insight data, Commerical data	to be confirmed			Private	For project use only. There is no need to make consumer insight data public. It wouldn't make commercial sense to publish consumer data on alternative proteins.	to be confirmed	For project use only	For project use only	Waitrose	to be defined	to be defined	to be defined	to be defined
32.	MATIS	FocusGroups_T5.1.1	Task 5.1	This dataset contains results from consumer focusgroups about alternative proteins and their use. It will be registered by Matis in an excel file, kept on NextGenProteins specific folder on the Matis internal network	Non-personal identifiable data, Consumer insight data	Non-personal identifiable data, sound recording converted to Word, Excel	10MB	Metadata with methodology and description of criteria	Restricted	Restricted within the consortium until agreed on publication	Excel, Word	Project use only until decided otherwise	Project use only until decided otherwise. Can contain sensitive information and/or good science results that could be shared in a peer reviewed article	MATIS	Trial protocol and manuscript will ensure data quality. Data will be stored on a secured, IT specialists maintained internal server, in a NextGenProteins specific folder	Data will be stored on a secured, IT specialists maintained internal server, in a NextGenProteins specific folder. All data will be treated securely as described by the EU General Data Protection Regulation (GDPR) and stored appropriately	5 years, or until the data is made public or published in the scientific journals	Included in overhead. Other cost is consumables re focus grous (ca 1500 EUR)
33.	VTT	Focus group data for T5.1.1 (Finland and Italy)		Qualitative data in a format of recorded online discussions. The data is related to NextGenProteins objective to understand consumers' attitudes and beliefs towards the NextGenProteins ingredients. The data is collected through online discussions and stored in VTT protected cloud service as well as on personal computers of the responsible researchers. The data will be destroyed after the time period described in Finnish ethical principles of research within social sciences.	Consumer insight data	Sound recording converted to Word	Approx. 167 MB		Raw data is restricted to VTT responsible researchers. Anonymised data is restricted to NextGenProteins partners. Processed and anonymised data is public in a form of publications.	Data might contain information, which reveals the identity of study participants. Data can be used to generate insights which the NextGenProtein partners have priority access.	Recordings: any software which opens .WMA files. Transcriptions: MS Word	Project use only until decided otherwise	Raw data might contain information, which reveals the identity of study participants.	VTT	Trial protocol and manuscript will ensure data quality. Data will be stored on a secured, IT specialists maintained internal server, in a NextGenProteins specific folder	Data will be archived in VTT protected cloud service.	5 years, or until the data is made public or published in the scientific journals	Included in overhead.
34.	VTT	Online survey data for T5.1.1 (Finland and two other countries specified later)		Fully anonymised numerical data. The data is related to understand consumers' attitudes and beliefs towards NextGenProteins ingredients. The data is collected through online survey and stored in VTT protected cloud service as well as on personal computers of the responsible researchers. The data will be destroyed after the time period described in Finnish ethical principles of research within social sciences.	Consumer insight data	SPSS, Excel	Approx. 2MB		Raw data is restricted to VTT responsible researchers and other NextGenProteins partners. Processed data is public in a form of publications.	Data can be used to generate insights which the NextGenProtein partners have priority access.	IBM SPSS	Project use only until decided otherwise	Data can be used to generate insights which the NextGenProtein partners have priority access	Data collection: Bilendi. Preservation and security: VTT	The survey instrument will ensure data quality. Data will be stored on a secured, IT specialists maintained internal server, in a NextGenProteins specific folder	Data will be archived in VTT protected cloud service.	Data will be stored according to the time frame given in Finnish ethical principles of research within social sciences	Included in overhead.
35.	VTT	Stakeholder interview data (Finland)	Task 5.1	Qualitative data in a format of recorded online discussions. The data is related to NextGenProteins objective to understand stakeholder attitudes and beliefs towards the NextGenProteins and theoir production technologies. The data is collected through online discussions and stored in VTT protected cloud service. The data will be destroyed after the time period described in Finnish ethical principles of research within social sciences.	Stakeholder insight data	MS Stream	Appr. 2 MB		Raw data is restricted to VTT responsible researchers.	Anonymised data is restricted to NextGenProteins partners. Processed and anonymised data is public in a form of publications.	MS Office			VTT		Data will be archived in VTT protected cloud service.	Data will be stored according to the time frame given in Finnish ethical principles of research within social sciences	
36.	TTZ	Stakeholder list	WP 7	TTZ gathered information about stakeholders to put into a database of stakeholders – Partners involved: all NextGenProteins partners	Stakeholder list: Personal data	Excel	up to few GB		Private until otherwise decided	Personal data will not be disclosed. Potentially mapping of actors to be shared.	to be confirmed	For project use only until otherwise decided	For project use only until otherwise decided	TTZ	to be defined	Data will be stored on a secured, IT specialists maintained internal server, in a NextGenProteins specific folder	5 years	server specific costs included in overheads
37.	VTT	RRI Thinking tool PDF reports	WP7	VTT gathered PDF documents of each WP, which answered the Societal Readiness Thinking Tool questions in https://newhorizon.eu/thinking-tool/. These pdf documents show how RRI aspects are considered in the project.	Qualitative answers	PDF	tens of kB		Open	Open	PDF	Open, public	Open, public	VTT		Data will be archived in VTT protected cloud service.	Data will be stored according to the time frame given in Finnish ethical principles of research within social sciences	

5 FAIR data

NextGenProteins consortium is aware of the directives for Open Access of Publications and Research Data in the H2020 projects and in taking part in the Open Research Data Pilot. Making research data findable, accessible, interoperable and re-usable (FAIR) is an integral part of the process of open science and research.

Making research FAIR data enable both scientific research and society to leverage the benefits of such data and make a significant contribution to economic growth. The FAIR data principles presented below are particularly helpful and allows:

- support to knowledge discovery and innovation;
- data and knowledge integration support;
- sharing and data re-use;
- support data and metadata to be machine-readable;
- data discoveries through the harvest and analysis of multiple datasets.

When managing research data, these principles should be heeded to ensure that the NextGenProteins research data will be shared in a way that enables and enhances re-use, by humans and machines.

5.1 Making data findable, including provisions for metadata

NextGenProteins produced and/or used research data must be easily discoverable with metadata, identifiable and locatable by means of standard identification mechanism in order to be used and re-used. Both data and metadata should be easy to find for humans and computers and therefore the use of machine-readable metadata is essential for datasets to be automatically findable.

5.1.1 File Identification

It is important to provide from the beginning of the NextGenProteins project a correct identification of files where the same structure must be used in both the active data and backup data. For the identification of data files, it is recommended to use a descriptive name to reflect the contents of the file and if it is necessary to use the date, it should be specified in a standard format (e.g. “YYYYMMDD”). We propose following file name convention for NextGenProteins data:

Example: “**20200330_NGP_TTZ_water_temperature_reactor1_v1.0.xlsx**”

- Date (in this example: “20200330”)
- Prefix to specify NextGenProteins data (in this example: “NGP”)
- Partner name (in this example: ”TTZ”)
- Intuitive title (in this example: “water_temperature_reactor1”)
- For each new version, specify the version number (in this example: “v1.0”)
- The file format (in this example: “.xlsx”)

5.1.2 File Metadata

It is important to provide metadata with the data file because it provides information that describe the data and context. Metadata should be enough by itself if data is discovered to understand the data.

Each NextGenProteins data should provide the following metadata:

- Research Data Name: Title of the research data, easy to search
- Data Type
- Responsible Partner
- Description and purpose: description should include the procedures followed to obtain those results and his purpose and benefits
- Version
- Keywords
- Access: specify the rights to access the NextGenProteins data

5.1.3 Data Storage

Making NextGenProteins research data easily findable and identifiable by consortium partners is fundamental. For that it requires active backups and to deposit the research data in a data repository. For this purpose, we deployed a "partner-only" storage area accessible by all consortium partners in the website of NextGenProteins, in which we intend to keep all of our data. The Coordinator is responsible for data safety at the "partner-only" storage area.

However, regarding Open Research Data Pilot, we need to keep our data visible and findable and are therefore considering using Zenodo for our open data (https://zenodo.org/) at later stage of the project.

Each partner takes care of the local data storage. Source data should be always kept separate from the on-going work or final data.

The NextGenProteins research data need to be stored during the period of the project and must be preserved for at least 5 years after the end of NextGenProteins project.

- research data and metadata must be given a unique, persistent, global identifier
- research data should be described with well-founded metadata
- research data and metadata must be registered/indexed in a searchable resource (repository)
- Metadata shall specify the identifier for the NextGenProteins research data
- Naming must provide project name NextGenProteins and should use a descriptive name, date and clear version number
- Search keywords should be provided to optimize possibilities for re-use

5.2 Making data openly accessible

All data produced and/or used in the project must be made openly available as the default. However, data do not need to be open if there are good reasons such as privacy concerns,

patent issues or commercial interests. However, there must be transparency in the conditions of access and re-use.

The data must be made accessible by deposition in a certified repository, which support open access. Appropriate arrangements need to be explored once the repository is identified concerning potential restriction on use, methods or software tools needed to access data, the need for a data access committee, description of conditions for access (i.e. a machine-readable license) and how the identity of the person accessing the data should be ascertained.

- research data and metadata need to be recoverable through their identifier using a standardized communication protocol that needs to be open and free allowing, if required, authentication and authorization procedures
- Metadata must be accessible, even when NextGenProteins data is no longer available

5.3 Making data interoperable

The data produced in the NextGenProteins project must be interoperable to allow data exchange and re-use between researchers, institutions, organisations, countries, etc. To be interoperable, research data and metadata need to adhere to standards for formats, and as much as possible compliant with available (open) software applications, and in particular facilitating re-combinations with different datasets from different origins.

- research data and metadata must use a formal, accessible, shared, and widely applicable language for knowledge representation
- research and metadata should have vocabularies that follow the FAIR data principles and if necessary generate project specific ontologies or vocabularies, as well as related mappings
- research data and metadata should include qualified references to other research data or metadata
- data must be recorded using digital and user-friendly format since the choice of an accessible format allows their preservation, access and share with third parties
- When choosing a file format, NextGenProteins project will prefer format that are non-proprietary, have no encryption, are uncompressed, are open and documented by the community, have common character encoding and are adapted for the data type
- NextGenProteins will prefer the most used formats, according to the European Data Portal: CSV, TXT, HTML, JSON, PDF, XLS and XML

5.4 Increase data re-use (through clarifying licences)

Optimize data re-use is the basic purpose of the FAIR data principles. NextGenProteins research data should therefore maintain its initial richness and must be clearly described.

- data and metadata must have a plurality of precise and relevant attributes;
- data and metadata must be released with a clear and accessible data usage license;
- data and metadata need to be associated with their origin;
- data and metadata must be aligned with the community standards and relevant to their domain;

- NextGenProteins should define for how long it is intended that the data remains re-usable;
- NextGenProteins should define and describe data quality assurance processes

6 Data sharing and IPR Management

In NextGenProteins project, research data will be owned by the one who generated them, and the consortium intends to provide as early as possible publicly available and easily discoverable data. This approach aims to maximize NextGenProteins visibility, the exploration of its results, the long-term impact and allowance to other researchers to use them.

Interoperability and data re-use must be provided following the FAIR data principles. The research data should be shared in an easy and transparent way to ensure that it can be understood and accessed by other researchers, institutions and organisations, along with the metadata and available documentation.

However, in NextGenProteins project, there will be a usage of very sensitive data coming from previous consortium members experiments or from their partners. This highly sensitive data will not be shared, even within the consortium. If this data is of particular importance for a member of NextGenProteins consortium, we will determine case by case how to share the data (anonymized, mean value, etc...).

In a project like NextGenProteins, it is fundamental to carry out property rights of the data used and generated during the project. Since the consortium will produce research data, publications and underlying data, all the Intellectual Property Rights (IPR) must be safeguarded using explicit licenses to make them openly accessible:

- Publications – The copyrights must be safeguarded and appropriate licenses to publications should be granted. Creative Commons licenses are recommended since offer useful licensing solutions in order to provide Open Access to third parties;
- Research data - To make research data more openly accessible, explicit licenses such as Creative Commons Attribution 4.0 (CC BY) or Creative Commons CCZero (CC0) should be attached to the deposited research data in the repository. In order to help the selection of the license for the research data, it is recommended to be used the EUDAT B2SHARE tool that includes an integrated license wizard to facilitates the selection.

There are two exemptions from the requirement to upload research data:

1) The research data is confidential and cannot be made available (e.g. due to it containing personal information, commercially sensitive data, etc.). However, it is still needed to fill in the data management plan form with as much detail as possible, and also explain why the data cannot be made available.
2) The datasets consist of pre-made datasets already available online. If so, it is possible to simply provide a link to where the original datasets can be found in your data

management plan form. It is still needed to fill in all the information required in form, though.

7 Allocation of resources

Costs related to open access to research data are eligible as part of the Horizon 2020 grant. In that sense, the costs for making FAIR data in the NextGenProteins project have been budgeted as such and will be covered by the grant.

The resources for long-term preservation will be discussed and decided (costs and potential value, who decides and how what data will be kept and for how long) along the NextGenProteins project.

8 Data security

When collecting/creating research data, it is recommended to follow good practices since it may come from different origins and can have many forms. The most common are text, numeric, audio, code, pictures and videos. It is recommended to keep the collected data in a workspace and to make backups.

All necessary steps must be taken to prevent both publications and research data from being leaked or hacked in order not to damage the NextGenProteins project plan as well as the opportunities and individual partner plan.

9 Ethical aspects

In the frame of the project, no sensitive personal data such as genetic, biometric, health data, or personal data revealing racial/ethnic origin, political opinions, religious or ideological convictions or trade union membership will be retrieved.

Specifically, in activities linked with stakeholders and consumers engagement, personal data might be collected, processed and stored. However, only the personal data required for the evaluation of the research will be collected in the form of coded samples.

All partners that process personal data in the NextGenProteins project will comply with the GDPR and the H2020 ethics standards. These standards are described in deliverable D9.2

10 Conclusions

The DMP details what kind of data the project is expected to generate, whether and how these will be exploited or made accessible for verification and reuse, and how they will be curated and preserved.

The project coordinator MATIS is currently preparing a graphical guide that shows which data to upload and how. It will be attached to next update of DMP in M18, and communicated to the consortium as soon as available.

11 References

EUROPEAN COMMISSION, D. (2016, July 26). Guidelines on FAIR Data Management in Horizon 2020 - Version 3.0.

FORCE11. (2011-2017). FORCE11 The Future of Research Communications and e-Scholarship. Retrieved 2019, from https://www.force11.org/group/fairgroup/fairprinciples

REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (GDPR). (n.d.).

Talwar Thakore & Associates. (2018). Data Protected - India. Retrieved 2019, from https://www.linklaters.com/de-de/insights/data-protected/data-protected---india