The PRIMA programme is an Art. 185 initiative supported and funded under Horizon 2020, the European Union's Framework Programme for Research and Innovation'

Deliverable title	**D7.6 VEGGIE-MED-CHEESES data management plan**
Deliverable Lead:	Università Politecnica delle Marche
Related Work Package:	**WP7 Multi-actor internal and external communication and technology transfer**
Related Task:	**T7.1 Establishment of a stakeholder-platform**
Author(s)	Lucia Aquilanti
Dissemination level	Public
Due Submission Date:	**31.10.2019**
Actual submission:	**31.10.2019**
Start date of project	01.05.2019
Duration	36 months (after project end extension: 48 months)
Abstract	A data management plan has been drafted and completed by Month 6, to manage VEGGIE-MED-CHEESES data overall collected from stakeholders as well as qualitative and quantitative analyses foreseen in WP2, 3, 4, 5, and 6. Such a data management plan has been drafted according to the H2020 Guidelines on Data management, including issues such as: "What types of data will the project generate/collect? What standards will be used? How will this data be exploited and/or shared/made accessible for verification and re-use? How will this data be curated and preserved?" This plan will be annually discussed and updated.

Versioning and Contribution History

Version	Date	Modified by	Modification reason
V1.0	15/10/2019	Lucia Aquilanti	Collection of the data description from each partner
V2.0	20/10/2019	Lucia Aquilanti	Completion of the draft of the deliverable
V3.0	22/10/2019	All partners	Final revision

Table of Contents

The aim of the VEGGIE-MED-CHEESES' Data Management Plan (DMP) is to ***identify the project's research data and to describe how to make them findable, accessible, interoperable and re-usable (FAIR*).** Following the H2020Data Management Plan Template, all partners involved in research's activities were asked to provide detailed information about the data generated during the entire project as reported in Section *1. Data Summary*.

In the Deliverable, an initial analysis on how the VEGGIE-MED-CHEESES Consortium intend to manage the amount of data produced in the VEGGIE-MED-CHEESES project is reported. The first checkpoint of the whole architecture of the DMP is the release of the first scientific publication that will be published within the VEGGIE-MED-CHEESES project: indeed, the data reported in this paper will be made available and interoperable to the larger typologies of stakeholders. To avoid issues related to PI rights and their access, as a first step in the strategy of development of DMP ***only data related to publications available to the public will be released***. In the VEGGIE-MED-CHEESES Project DMP is intended to be a living document in which information can be made available on a finer level of accuracy and details through updates as the implementation of the project progresses and when significant changes occur. Further update on Data Management will be provided in the mid-term reporting period.

1. Data Summary

What is the purpose of the data collection/generation and its relation to the objectives of the project?

For VEGGIE-MED-CHEESES data collection and integration is absolutely necessary. It is therefore of utmost importance that not only data is well generated but also well annotated using open standards and metadata as it is laid out in the following section. As VEGGIE-MED-CHEESES aims at uncovering potential biomarkers for demonstration of quality and/or authenticity of Mediterranean thistle-curdled cheeses, good documentation, data keeping and integration are necessary.

What types and formats of data will the project generate/collect?

We foresee that the following data sets will be collected and generated:

(i) morphological data related to spontaneous thistle populations (*Cynarahumilis*, *Onopordum tauricum*, *Onopordum platylepis* and *Cynara cardunculus*) growing in different Mediterranean areas (Spain, Italy, Tunisia).
(ii) genetic data related to distribution of SSR markers in the genomes of different thistle ecotypes within the wild thistle populations considered;
(iii) environmental data related to soil parameters, air temperature, rainfall, collected from the central Italian site and Tunisian one ?? designed for thistle cultivation
(iv) agronomic data related to sustainable thistle crops
(v) data related to dormancy of thistle seeds
(vi) chemical, biochemical and technological data related to the aqueous extracts obtained from both spontaneous and cultivated thistles
(vii) chemical and biochemical data related to the new proteases purified from the aqueous extracts from both spontaneous and cultivated thistles
(viii) physico-chemical, chemical, microbiological and sensory data related to thistle-curdled and control cheese prototypes
(ix) data on consumer preference for thistle-curdled and control cheese prototypes

These data sets will be produced by using specific equipment (e.g. photo-spectrometers, pH-meter, thermal cycler, electrophoresis apparatus, DNA sequencers, panel of sensory assessors etc.) and analytic techniques (e.g. Polymerase Chain Reaction – based approaches; next generation sequencing techniques; Bradford method for protein quantification, etc.). In addition, derived data from the original raw data sets will also be collected. This is important, as different analytical pipelines might yield different results or include ad-hoc data analysis parts. Therefore, specific care needs to be taken to document and archive these resources (including the analytic pipelines) as well.
Concerning formats, we assume to generate mainly documents (doc, docx, ppt, pptx, etc.), illustrations (png, jpeg, etc.), drawings (DWG, etc.) and raw data (xls, txt, etc).

Will you re-use any existing data and how?
The project builds on existing data sets and relies on them. However, it is of course also important to include existing data sets on crude extracts on still unexplored thistles, new thistle crops and new thistle-curdled cheeses. Most data can simply be gathered from reference databases like the NCBI (research papers, literature reviews, DNA sequences, Protein sequences; sequence READ archives, etc.) and patents databases; these will be used as hints for the development/optimization of analytical procedures and production of thistle crops, aqueous extracts, thistle-curdled cheese prototypes);in case of protein sequences and 3D structures the Protein Data Bank RCSB – UniProt, Worldwide Protein Data Bank, etc..

What is the origin of the data?
Public data will be extracted as described in the previous paragraph. For VEGGIE-MED-CHEESES, specific data sets will be generated by the Consortium partners.
Briefly, most data will origin from the numerous activities carried out at UNIVPM by using specific equipment (e.g. photo-spectrometers, pH-meter, thermal cycler, electrophoresis apparatus, etc.) and analytic techniques (e.g. Polymerase Chain Reaction for DNA amplification; Bradford method for protein quantification, etc.)

(i) morphological characterization of spontaneous thistle populations (*Cynara humilis*, *Onopordum tauricum*, *Onopordum platylepis* and *Cynara cardunculus*) growing in different Mediterranean areas (Spain, Italy and Tunisia);
(ii) molecular characterization of SSR markers in different thistle ecotypes within the wild thistle populations considered;
(iii) environmental characterization (soil parameters, air temperature, rainfall) of the central Italian site designed for thistle cultivation and Tunisan one??
(iv) agronomic characterization of sustainable thistle crops
(v) thistle seeds germination testing
(vi) chemical, biochemical and technological characterization of the aqueous extracts obtained from both spontaneous and cultivated thistles
(vii) chemical and biochemical data related to the new proteases purified from the aqueous extracts from both spontaneous and cultivated thistles
(viii) physico-chemical, chemical, microbiological and sensory characterization of thistle-curdled and control cheese prototypes
(ix) consumer testing of thistle-curdled and control cheese prototypes

What is the expected size of the data?
From hundreds of megabytes to few gigabytes

To whom might it be useful ('data utility')?
Firstly, data produced within the VEGGIE-MED-CHEESES project will be of key importance for the whole VEGGIE-MED-CHEESES Consortium for internal use; secondly, they will be essential for production of both scientific documents

(research papers, abstracts at national/international conferences, patents, etc.) as well as educational material destined for a wide Platea including the supply dairy chain (farmers) and dairy industries.

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)?
The description of procedures to generate data is associated to a dataset (i.e. collection of data). At this stage of development of the VEGGIE-MED-CHEESE project, the specific typology and total number of variables in a single dataset table (see Data Summary) cannot be defined a-priori. The procedures for the identification of data are defined as follows:

- each dataset is first assigned to a unique identifier (ID);
- each dataset might also be associated to a Digital Object Identifier (DOI). The service is provided by the DOI (www.doi.org) community through a request to a local Registration Agency (RA). The use of DOIs is still under discussion by the VEGGIE-MED-CHEESES partners (→ to be eventually included in a further DMP version). The use of a DOI guarantees unique identification of the single dataset and the possibility of automatic data web retrieval.

What naming conventions do you follow?
Data variables will use standard names. This is e.g. the case for proteins, chemical compounds, minerals, vitamins, genes, etc. In the case of datasets, the dataset names will also encode the provenience.

Will search keywords be provided that optimize possibilities for re-use?
Keywords about the experiment and the consortium will be included as well as an abstract about the data, where useful.

Do you provide clear version numbers?
To maintain data integrity and to be able to re-analyze data, data sets will get version numbers and/or date

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.
One or more metadata files will be generated for each dataset. The metadata will be identified by the same unique ID of the related dataset, with a different suffix/extension. Each metadata file will be uploaded in a standardized format, depending on the dataset considered. Appropriate templates will be available for download to all VEGGIE-MED-CHEESES' partners in the Collaborative Platform (https://trello.com/b/BfZlqB3t). The metadata files will be stored in the cloud (DROPBOX) correlated to the VEGGIE-MED-CHEESES Collaborative Platform.

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.
By default all data sets from VEGGIE-MED-CHEESES will be shared and made openly available. This is however usually after a gratuity period allowing partners to importantly iterate through data and potentially clean up data in the process and to allow partners to exert their publishing and patenting rights prior to unlimited sharing.

How will the data be made accessible (e.g. by deposition in a repository)?
Data will be made available via the project specific website (https://veggiemedcheeses.com/). It will be ensured that data which can be stored in international specialized repositories (NCBI, Protein Data Bank RCSB –UniProt, Worldwide Protein Data Bank, etc.), will be stored and processed there as well.

What methods or software tools are needed to access the data?
No specialized software will be needed to access the data. Access will be possible via web interfaces once publicly available. For data processing after obtaining raw data, typical open source software can be used.

Is documentation about the software needed to access the data included?
As no software is needed, no documentation needs to be provided.

Is it possible to include the relevant software (e.g. in open source code)?
As stated above, software is only needed AFTER data has been obtained by a user in order to process and/or analyze the data. Here we use publicly available open source certified software.

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.
As noted above specialized repositories like NCBI, Protein Data Bank RCSB – UniProt, Worldwide Protein Data Bank, etcare very likely the most common ones.

Have you explored appropriate arrangements with the identified repository?
The submission is for free and it is the goal (at least of NCBI, Protein Data Bank RCSB – UniProt, Worldwide Protein Data Bank) to obtain as much data as possible. Therefore, arrangements are neither necessary nor useful.

If there are restrictions on use, how will access be provided?
There are no restrictions.

Is there a need for a data access committee?
Consequently, there is no need for a committee.

Are there well described conditions for access (i.e. a machine readable license)?
Yes where possible e.g. Creative Commons Rights Expression Language(CC REL)will be used.

How will the identity of the person accessing the data be ascertained?
In case data are only shared within the VEGGIE-MED-CHEESES Consortium due to data cleanup tasks and/or publication preparations, a user specific log in is necessary. In case data are publicly available, these data have to be available without siphoning out personal information. I.e. access is anonymous.

2.3. Making data interoperable

Are the data produced in the project interoperable, that is allowing data exchange and re-use between researchers, institutions, organizations, 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)?
At all times, data will be stored in common and openly defined formats. By default no proprietary formats will be used.

What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?
As mentioned above, we foresee to use e.g. NCBI for DNA data, Protein Data Bank RCSB – UniProt and Worldwide Protein Data Bankfor proteases data, but will also be relying on specific standard operating procedure (SOP) established in the VEGGIE-MED-CHEESES project. The latter will thus allow integrating data across projects and safeguards reusing established and tested protocols. Additionally, we will use ontology terms to enrich the data sets relying on free and open ontologies.

Will you be using standard vocabularies for all data types present in your data set, to allow inter-disciplinary interoperability?
Indeed, common and open biomedical and biological ontologies as well as plant ontology will be used.

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?
Common and OPEN ontologies will be used thus this question does not apply.

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

How will the data be licensed to permit the widest re-use possible?
Open licenses will be used, such as CC.

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.
In general, due time will be given to the VEGGIE-MED-CHEESES Consortium partners to exploit the data for publication and/or PI issues first. This is not least due to the fact that data will often see an increase in quality by additional cleaning up for publications. All VEGGIE-MED-CHEESES Consortium partners will be encouraged to make data available prior to publication under pre-publication agreements such as those started in Fort Lauderdale and set forth by the Toronto International Data Release Workshop.

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.
There will be no restrictions once data is made public.

How long is it intended that the data remains re-usable?
Data will be made available for many years and potentially indefinitely after the end of the project.

Are data quality assurance processes described?
Data will be analyzed using automatic procedures as well as by manual curation.

3. Allocation of resources

What are the costs for making data FAIR in your project?
The costs comprise data curation, setup of databases and long term sustenance and storage including electricity. The costs will amount to approximately 5.000 € of which most can be covered by eligible costs that will be claimed.

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).
The cost is covered in WP7 –Multi-actor internal and external communication and technology transfer.

Who will be responsible for data management in your project?
Partner 1 – D3A-UNIVPM

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)?
The partner D3A-UNIVPM decides on preservation. The partner has pledged long term (potentially for decades) support based on own costs; e.g. DNA data will also be available through NCBIwhereas protease sequences and 3D structures will be available through Protein Data Bank RCSB – UniProt and Worldwide Protein Data Bank (so data will be available from redundant resources).

4. Data security

What provisions are in place for data security (including data recovery as well as secure storage and transfer of sensitive data)?
Once data is transferred to the VEGGIE-MED-CHEESES database, data security standards will be imposed. Data will be stored and shared in the private VEGGIE-MED-CHEESES' Collaborative Platform (https://trello.com/b/BfZlqB3t) with restricted access (username + password) to authorized users. As an initial step, only the VEGGIE-MED-CHEESES Consortium Partners will have access to the cloud storage where dataset and metadata are filed.

Is the data safely stored in certified repositories for long term preservation and curation?
Nucleotide datawill be also made available upon publication via the standards NCBI whereas protease data will be made available upon publication in the Protein Data Bank RCSB – UniProt and Worldwide Protein Data Bank.

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).
At the moment, we do not foresee ethical or legal issues with data sharing. The VEGGIE-MED-CHEESES Consortium makes its best efforts to ensure data sharing, at the very latest after exploitation through papers and patents.

Is informed consent for data sharing and long term preservation included in questionnaires dealing with personal data?
The only personal data that will potentially be stored is the submitter name and affiliation in the metadata. As lengthy questionnaires tend to stifle careful answering and deposition, this will be highlighted again to the submitters and they can opt out in which case only their institution will be mentioned. This is however a very unlikely case, as data evaluation will be published in scientific journals anyway providing the names of the authors.

7. Other issues

Do you make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones?
Not applie