Project Number: 774571
Start Date of Project: 2017/11/01
Duration: 48 months

Type of document D7.2 – V1.0

Data Management Plan – Open Data Pilot

Dissemination level	PU
Submission Date	2018-04-30
Work Package	WP7
Task	T7.2
Type	Report
Version	1.0
Author	Emanuele Garone
Approved by	Andrea Gasparri, Riccardo Torlone

DISCLAIMER:
The sole responsibility for the content of this deliverable lies with the authors. It does not necessarily reflect the opinion of the European Union. Neither the REA nor the European Commission are responsible for any use that may be made of the information contained therein.

Executive Summary
This document describes the Data Management Plan of the project PANTHEON.

In particular this Deliverable describes the types and size of the data to be collected and gathered during the execution of the project; the procedures to be followed to ensure a FAIR (Findable, Accessible, Interoperable, Re-usable) access to data and the possible technical and legal/ethical issues.

Furthermore, this Deliverable describes the open-access approach to be followed for publishing scientific results.

Finally, this Deliverable is to be considered a live-document and is subject to regular updates, on a 12-months basis.

Table of Content

Abbreviations and Acronyms

AGRO	Agronomy Ontology
ASCII	American Standard Code for Information Interchange
GIS	Geographic Information System
GPS	Global Positioning System
IoT	Internet of Things
JSON	JavaScript Object Notation
LiDAR	Light detection and ranging
OBO	Open Biological and Biomedical Ontology
OWL	Web Ontology Language
RDF	Resource Description Framework
RTK	Real Tim Kinematics
SCADA	Supervisory Control and Data Acquisition
XML	Extensible Markup Language
WP	Work Package

1 Forewords

The project PANTHEON will offer to the scientific community both technical data to be used for further analyses and research and scientific publications.

2 Technical Data

2.1 Purpose of technical data collection/generation and its relation to project's objectives

The vision of project PANTHEON is to develop the agricultural equivalent of an industrial Supervisory Control And Data Acquisition (SCADA) system to be used for the precision farming of hazelnut orchards.

To do so PANTHEON will develop a system composed of fixed sensors (e.g. meteorological stations and soil moisture sensors) and ground and aerial robots that navigate the orchard to collect measurements using various kind of sensors (including high level imaging sensors such as LiDAR and multispectral cameras), achieving the resolution of the single tree.

The information will be sent to a central unit, which will store the data, process them, and extract synthetic indicators describing for each tree:

- water stress;
- presence of pests and diseases;
- geometry of the tree, including the possible presence and dimension of suckers;
- estimated number of nuts on the tree.

Based on these synthetic indicators, the system will elaborate a synoptic report for the agronomist in charge of the orchard, putting in evidence possible situations that may deserve attention, providing suggestions of intervention and, if requested, providing a historical view of the status of the plant and of the treatments already performed.

For some interventions, PANTHEON envisions the design and implementation of tailored algorithms based on these indicators to automatize farming operations such as the control of the irrigation level and suckers' elimination by robots.

The collection of data is pivotal to ensure the design and implementation of these techniques.
Briefly, primary goals of data collection can be summarized in the following two points:

1) Development, tuning, and validation of algorithms for the remote sensing of Hazelnut plantations. This includes the design of algorithms that will build the synthetic indicators on the basis of the data collected by the robots and by the fixed sensors. (WP4);
2) Development, tuning, and validation of the automatic feedback algorithms and of the expert system that will generate the synoptic reports (WP5 and WP6).

2.2 Origin of the Technical Data

All data generated by the sensors will be collected in the experimental hazelnut plantation "*Azienda Agricola Vignola*" which is located in the municipality of Caprarola, in the province of Viterbo, Italy. In particular the collected data will concern three specific plots of the plantation, highlighted in Fig. 1.

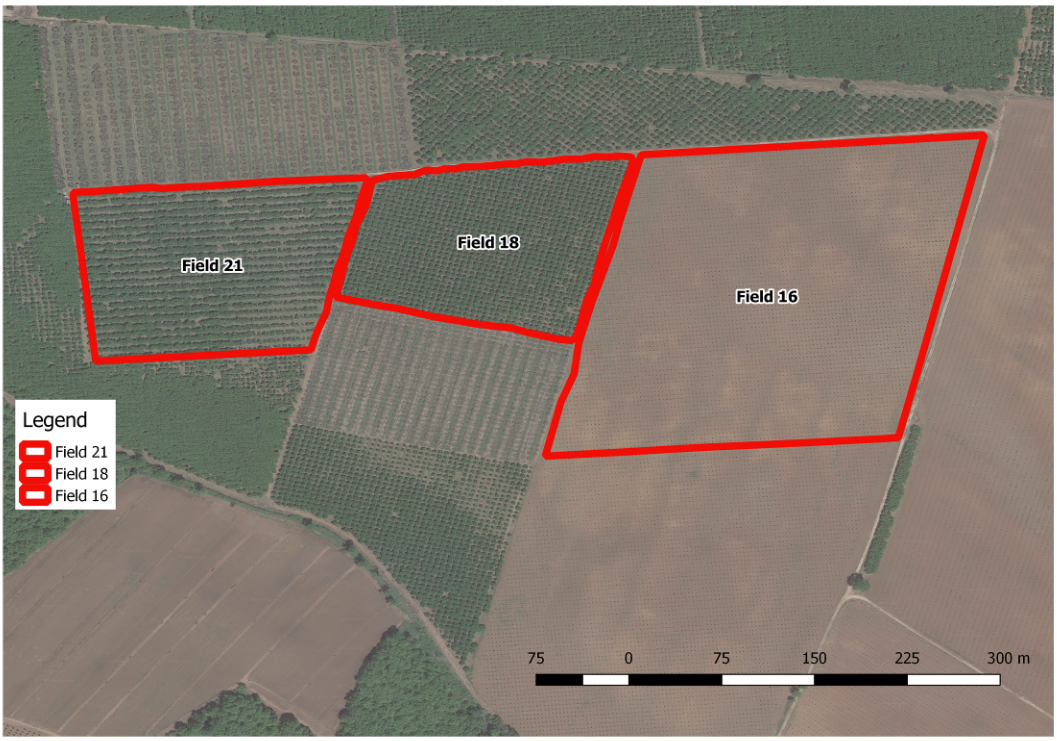

Fig. 1: Fields for the Pantheon project data collection activity.

The current plan foresees both the collection of general data concerning the entire areas (e.g. aerial images, soil analysis, weather conditions data, etc.) and the continuous collection of data on a selected subset of trees over the four years of the project.

At the current stage, we foresee that a total of ca. 48 trees will be selected to collect different kind of measurements over the four years of the project PANTHEON. In particular, they will be organized as follows:

- *Water stress:* ca. 10 trees selected in field 18 and ca. 10 trees selected in field 16;
- *Sucker detection and control:* ca. 6 trees in field 18;
- *Fruit detection:* ca. 6 of the trees selected in field 16;
- *Tree geometry reconstruction*: ca. 6 trees selected in field 16;
- *Pest and disease detection*: ca. 10 trees selected in field 21

The selected trees will be continuously monitored manually by PANTHEON agronomists tentatively every ten days and autonomously by the ground and aerial robots tentatively once a month. Full details concerning the procedures for the trees selection will be part of Deliverable D2.3 "Real-world (1:1 scale) hazelnut orchard for final demo".

The data collected by the robots will be stored in a database. This data-set will be used (especially in the first part of the project) to develop, train, tune and validate the automatic analysis algorithms, while the data collected manually by the agronomists will be used as *ground truth* for benchmarking. Furthermore, this dataset will be used (mostly in the second part of the project) to validate the effectiveness of the expert system to identify needs and proposed the right corrective action (WP5)

A more detailed overview of the data that will be collected is reported in the next subsection.

2.3 Types and formats of technical data will the project generate/collect

As explained, different types of data will be collected/generated during the project. They will contain evaluations and measurements performed with various techniques and sensors both on single trees and on entire areas.

In principle, the collected data can be divided in the following classes:

1) General information on the orchard: descriptive information of each area including:

a. number of trees,
b. agronomic age and history,
c. type of irrigation;
d. composition of the soil in each area,
e. ID and geo-localization of each tree in the orchard,
f. altimetric characterization of each point of the orchard,
g. geo-localization of the irrigation installation.

Data will be provided in the following formats:

a) A ***.json*** file with a synthetic description of each area, its history, and including the ID of each tree, its age, an indication of the cultivar and its geo-localization.
b) A more complete standard GIS format (e.g. the Geography Markup Language) containing the map of the orchard with all the relevant information (trees ID and position, irrigation lines, altimetry).

***2) Agronomic Data collected manually*:** results of agronomical evaluations performed by PANTHEON agronomists on the selected trees. This includes:

a. the evaluation of the phenology,
b. the evaluation of the biometric variables,
c. the detection of pests and diseases,
d. the evaluation of suckers.

Further data that will be collected manually concerns the yearly hazelnut yield of each plant under observation. It is expected that all the information will be collected using standardized protocols. Details concerning the protocols to be used will be part of Deliverable D2.3 “Real-world (1:1 scale) hazelnut orchard for final demo”. The data will be stored in tables using Excel ***.xlsx*** files.

***3) Raw Remote Sensing Data collected by the robots*:** data collected by the various sensors mounted on the ground and aerial robots of the project. More specifically it will consist of:

a. images captured with RGB,
b. images captured with Multispectral and Thermal Cameras,

c. 3D measurements captured with Lidar,
d. Data relative to their triggering (RTK-GPS position, date and time, orientation of the gimbal, orientation and speed of the robot).

More specifically the data collected by the Unmanned Aerial Robot will be

a. Sony a5100: ***.raw*** RGB images,
b. Tetracam MCAWL ***.raw*** multispectral images,
c. Teax ThermalCapture 2.0 ***.raw*** themal images.

Each of these images will be associated with a JSON object containing the description of the data, date and time of the capture, GPS positioning of the image, and all the data concerning the telemetry of the UAV and the position of the gimbal at the time of the trigger. The JSON objects will be collected in a ***.json*** file.

The data collected by the Ground Robot mostly consists of the three main sensors

a. Faro Focus S70 (laser scanner) ***.fls*** files containing the 3D point cloud
b. Sony a5100: ***.raw*** RGB Images
c. MicaSense RedEdge-M ***.raw*** *multispectral images*

Each of these images will be associated with a JSON object containing the description of the data, date and time of the capture, GPS positioning of the image, and all the data concerning the telemetry of the ground robot and the position of the gimbals at the time of the trigger. The JSON objects will be collected in a ***.json*** file. It is also foreseen to store the data of the extra navigation sensors (e.g. the navigation lidar) in ***.raw*** for comparison purposes**.**

4) Elaborated Remote Sensing Data**:** processed data computed starting from the raw remote sensing data. These data include both data resulting from pre-processing (filtering, homogenization, etc.) and real derived data, such as: orthophotos of the orchards and of some of its parts, graph representation of the hazelnut tree structure, water stress maps, indicators on the presence of suckers, estimation of the state of health of the plants. At the current stage the format of these data has not been defined yet, however, whenever possible standard **XML** or **JSON** formats will be used.

5) Measurements collected by the fixed IoT infrastructure: measurements collected on the field 24/7 by the fixed Internet of Things (IoT) infrastructure composed of a weather station and moisture sensors placed in different parts of the orchard. These data will be collected as ASCII files and possibly converted to Excel ***.xlsx*** files.

6) History of the plants: It represent the history of all the treatments sustained by the plants. This will be recorded in an Excel ***.xlsx*** files.

At current state, we expect that all the data will be collected in a NoSQL database for easy queries and all the generated files will have an associated JSON object containing all relevant information.

2.4 Re-use of existing data

No re-use of any existing data is foreseen at the present stage

2.5 Expected size of the data

The expected total size of the generated data mostly depends on the remote sensing activities and their subsequent analysis. Diversely, all the other information that will be collected during the entire duration of the project (information on the orchard, manual sampling and sampling from the infrastructure) will amount to less than 200 MB.
Roughly the data gathered by the remote sensing-based activities (in particular the .raw and pre-processed images) will represent ~95% of the whole technical data managed during the project.

2.5.1 UAV data

For what concerns the remote sensing performed through the UAV (water stress and pest and disease detection) the raw file size for each capture is about

1) **28 MB** for the Sony a5100 RGB camera;
2) **15 MB** for the Tetracam MCAW multispectral camera;
3) **0.8 MB** for the Teax ThermalCapture 2.0 thermal camera.

Which amount to approximately **44 MB** per capture. For each day of measurement, we assume approximatively 2000 captures, for a total of ca. **90 GB/day.** At this point, by assuming a minimum of 7 measurements per year (full details about the calendar of automated sampling activities will be part of Deliverable D2.3 "Real-world (1:1 scale) hazelnut orchard for final demo"), a total of ca. **0.63 TB/year** of raw image data from the UAV is reached**,** which will result in ca. **2.5 TB** of **raw image data** from the UAV in the entire duration of the project.

To receive the final multispectral orthoimages, which are needed to calculate the spectral indices, a post processing is required. In the **first testing phase** (year 1-2) and in the **development phase** (year 3) intermediate files are generated to evaluate the correctness of the results and to further develop the algorithms. More and more of these files can be deleted with progressive development of the project.

Based on the current design of the processing chain we assume to generate each measurement day post processed data in a magnitude of

- **390 GB** in the testing phase;
- **35 GB** in the development phase;
- **30 GB** for the final product.

Assuming 7 measurement days per year this results in a data volume of about

- **2.8 TB/year** in the testing phase;
- **0.3 TB/year** in the development phase and for the final product.

So about **6.2 TB post processed UAV remote sensing data** will be generated during the entire duration of the project.

2.5.2 UGV data

To perform the remote sensing activities through the UGV (tree geometry reconstruction, suckers detection and fruit detection) we plan to capture each tree by 4 Lidar scans and by 16 photo shoots per camera. Based on the sensor characteristics, it is foreseen that for each tree and day of measure raw sensor files are generated with a volume of at most

1) **0.25 GB** for the Faro Focus S70 laser scanner (.fls);
2) **0.45 GB** for the Sony a5100 RGB camera (.raw);
3) **0.05 GB** for the MicaSense RedEdge-M multispectral camera (.raw).

For the UGV the amount of data depends on the specific operation and on the phase of development of the project. 60 trees measured with all sensors and 12 trees measured with Lidar only (full details about the calendar of automated sampling activities will be part of Deliverable D2.3 "Real-world (1:1 scale) hazelnut orchard for final demo"), it is possible to estimate the total amount of data generated every year. For the various activities we foresee that every year we will measure

- **60 trees** scanned by Lidar;
- **48 trees** captured by the cameras.

So, each **year** we will generate approximately **39 GB raw UGV sensor data** in the field. A data volume of **9 GB/day** is not exceeded.

To receive the multispectral point clouds and image data used for further analyzes, the raw data has to be post processed. In the **first testing phase** (year 1-2) it is important to store more data (including more .raw and intermediate formats data) to evaluate the correctness of all intermediate processing steps. Based on the current design of the processing chain, for each tree post processed data is generated with a data volume of approximatively

- **2.7 GB MB** for the laser scanner;
- **8.1 GB MB** for the RGB and multispectral cameras.

For the **development phase** (year 3) and the **final product** (year 4) most intermediate and temporary files can be deleted, and the amount of post processed data for each tree will decrease to

- **0.75 GB** for the laser scanner;
- **3.6 GB** for the RGB and multispectral cameras.

Based on the planned data acquisition design we will generate approximately

- **550 GB/year** for in the testing phase;
- **260 GB/year** for the development phase and final product.

So, we will generate approximately **1.6 TB post processed UGV remote sensing data** during the entire duration of the project.

2.5.3 Total data volume

We estimate that approximatively **1.7 TB** will be generated during the entire duration of the project coming from the main sensors of the ground robots and **8.7 TB** coming from the sensors of the UAV. Considering all the data acquired from all the various sources it is reasonable **to estimate the total amount of data that will be generated in the order of 10-15 TB.**

2.6 Third parties possibly interested in the data

The consortium believes that the third party possibly interested in the data are mostly research group on remote sensing that may want to reuse the collected data to test and validate new algorithms and research groups interested on hazelnut plantation that may be interested in validating current best practices or formulating new paradigms for orchards management.

3 FAIR data

3.1 Making data findable, including provisions for metadata

3.1.1 Name Convention and Provision of Metadata

All data will be stored following the following name convention:

TypeodData-CalendarDay-SequentialNumber.extension

where:

- **Type of data:** represent a code of the type of data composed of four capital letters. The meaning of each code will be developed during the project.
- **Calendar Day:** follows the convention YYYY.MM.DD
- **Sequential Number:** is the progressive number for that specific kind of data generated in that day
- **Extension:** is the one proper for that type of data

This naming allows to easily find and order the data for type, date and sequence, for instance *UAV1-2018.08.03-1.raw* represent the first capture from the first sensor of the UAV on the 3rd of August 2018 whose format is a .raw.

Together with each generated file, there will be always be an associated JSON object that will be stored in a ***.json*** file containing the relevant metadata and extra information that might be needed.

3.1.2 Structure of the metadata (including keywords and version numbers)

Each generated file will have an accompanying JSON object that will be stored in a ***.json*** file which will be structured to include the following information

- **General information on the data:** it contains metadata such as the name-file including the data and its key, a description of the nature of the data (including versioning), keywords for easy searchability, and indication on the license under which the data are distributed.

- **Accessibility information:** it contains information on how to read the data. It includes the format of the file (with possible versions), when relevant indication on the way the data is structured (e.g. convention for tables), and suggestions on the software to open the data (including an URL to the software producer, when available).

- **Service information:** Contain the extra information on the data acquisition. It will always contain the timestamp of the acquisition, and the GPS coordinates of the acquisition, together with any other information that can be useful for the elaboration of the data.

A tentative structure of a possible .json describing data is reported hereafter

```
{
  "generalInfo" : {
    "filename" : "TypeodData.extension",
    "key" : " CalendarDay-SequentialNumber",
    "description" : "Here a description of the file and its content",
    "keywords" : [ "Keyword1", " Keyword2", " Keyword3"],
    "copyrightOwner" : "H2020 EU Project PANTHEON, www.project-pantheon.eu"
    "copyrightLicense" : "Type of licence with which data are released"
},
  "dataInfo" : {
    "formatFile" : "format file",
    "structure" : "Possible description of the information file",
    "supportSoftware" : "Name of the software to open the data",
    "urlSoftware" : "if available, URL to a software to open the data"
},
    "serviceInfo" : {
    "timeStamp" : "Timestamp in Unix Epoch format",
    "gps" : ["Latitude","Longitude","Altitude"],
      ….
}
```

3.2 Making data openly accessible

3.2.1 Default Open Access, Exceptions and Temporary Embargos

In line of principle, it is intention of the consortium to make all the collected data publicly available by default at the end of the project, so that they can be re-used by the project partners and by third parties.

Exceptions to this general principle will be made on the basis of:

- Possible well-motivated objections raised by either one of the partner or by the owner of the hazelnut orchard *"Azienda Agricola Vignola"* concerning the disclosure of sensitive information that might jeopardize the economical exploitation of the results of the project or legitimate economical/privacy interests of the involved organizations. The pertinence of the objections must be approved by the consortium boards.

- Technical difficulties in publicly sharing the data due to the size of the database and the associated bandwidth requirements. Should this be the case, a representative sample of the data will be selected and will be made publicly available on the internet without any access restriction. The consortium will grant access to the entire dataset upon request.

Furthermore, any consortium partner may request a temporary embargo on any specific subset of the data up to the time that scientific publication, patents, or products based on those data are published.

The means to make the data publicly available will be detailed in Section 3.2.3.

3.2.2 Software to access the data

As already detailed in Section 2, the technical data generated is either raw data from the various sensors (and that as such follow the specifications of the sensors manufacturer) or processed data provided in the most common storage formats.

In the **JSON** object accompanying every generated data file, it is foreseen a field which describes the type of the data, its internal structure (when relevant), and a suggestion on the software to be used. The JSON object will also contain a link to the suggested software to access the data. Whenever possible, link to downloadable open source software will be provided.

3.2.3 Repository and Access to the Data

All data will be stored in a NoSQL database (the same that will be used within the central unit for the project). The database will run on the main workstation of the project, installed at the University of Roma Tre.

To make the data accessible, a webpage connected to the project webpage will be created as a front-end to the NoSQL database. The page will describe the content of the database, and the instructions for accessing it.

The possibility to also upload the material on a public repository for research data sharing (e.g. https://zenodo.org/) will be evaluated. However, at the current stage this solution seems non-practicable given the very large size of the generated database. A possible solution could be to select a representative subset of the data (e.g. all the measurements concerning a very small number of trees) to be uploaded on a standard repository for research data sharing and clearly putting a disclaimer that a large dataset is accessible at the project website upon request.

The access to the database will be through a login and password. The login and password obtainable through the front-end will require the Name, Last Name and institutional email registration. The user will have read-only privileges to the data and he/she will not have the access to restricted data or embargoed data. Access to restricted or embargoed data will be possibly granted upon motivated request to the Consortium. The personal data of the registered users (name, last name, and email) will be accessible only to the system administrator.

3.2.4 Licenses

The data will be released under **Creative Common Attribution-NonCommercial-ShareAlike** licence, for details on this licence please refer to https://creativecommons.org/licenses/by-nc-sa/3.0/legalcode. The information on the licenses will be reported in each **JSON** description as well as on the front page of the repository

Figure 2 – The data will be released under Creative Common Attribution-NonCommercial-ShareAlike License

3.3 Making data interoperable

Since the developed data will be stored in the most common formats, it is reasonable to expect that data could be re-used with a good level of interoperability. The use of the ***.json*** auxiliary file to explicit the data types, and possible internal structure of the date will facilitate the interoperability. Furthermore, as the data will be collected in a NoSQL database, access to the elaborated data (and possible conversion to specific reporting formats) will be easily achieved.

To make our data interoperable with other agricultural-related databases and support inter-disciplinary interoperability we will use metadata vocabularies (based on RDFS) and standard ontologies (based on OWL) for agronomists, such as, AGRO (the AGRonomy Ontology)[1], developed by The Open Biological and Biomedical Ontology (OBO) Foundry.

3.4 Increase data re-use (through clarifying licences)

3.4.1 Licensing to increase re-use

The data will be publicly released under **Creative Common Attribution-NonCommercial-ShareAlike** licence. The information on the licenses will be reported in each ***.json*** description as well as on the front page of the repository.

Summarizing from the **Creative Common** website (https://creativecommons.org/licenses/by-nc-sa/3.0/) this license allows to freely:

- **Share** – Copy and redistribute the data in any medium or format
- **Adapt** – Remix, transform and build upon the data

Under the following conditions:

- **Attribution** — The user must give appropriate credit to the licensor, provide a link to the license, and indicate if changes were made. The user may do so in any reasonable manner, but not in any way that suggests the licensor endorses the user or the use of the data.

- **NonCommercial** — The user may not use the material for commercial purposes. The PANTHEON consortium pledge to not consider publication of scientific papers on peer-reviewed journal a commercial purpose.

- **ShareAlike** — If the user remix, transform, or build upon the data, he must distribute his contributions under the same license as the original.

3.4.2 Availability of the data

The consortium will ensure the public access to the generated database starting from the beginning of the fourth year of the project, taking into account the possible exceptions highlighted in Section

[1] http://www.obofoundry.org/ontology/agro.html

3.1.1. The consortium will ensure the internet availability of the database at least 2 years after the end of the project.

3.4.3 Description of the data quality assurance process

The consortium will comply with high standard of data collection. Full details concerning the methods for data collection and protocols will be part of the Deliverable D2.3 “Real-world (1:1 scale) hazelnut orchard for final demo”.

4 Data security

Data will be stored in a server which will be physically located at Roma Tre University and protected by a firewall. In particular, the server will be a cluster of Standard Linux-based workstation equipped with the latest versions of open-source security tools. Regarding data reliability and fault-tolerance, data will we replicated in the local server. In addition, whenever possible, the other partners of the consortium will keep copies of the data sets to ensure some redundancy against possible failures.

5 Scientific Publications

All scientific outcomes will be provided in open access mode. In particular, the 'green' open access model will be used. Every scientific outcome generated in the project will be self-archived in three locations: on the project website, on arXiv, and on Researchgate to ensure maximal visibility. The researchers will be instructed to publish only in journal and conferences ensuring self-archiving (green publishers). Exceptions to this policy must be authorized by the Project Management Committee. The authorization to publish on journal/conference not ensuring self-archiving will be granted only if motivated by reasons of opportunity.

6 Ethical aspects

No ethical aspects concerning data sharing is expected. If any should raise (e.g. images capturing neighboring fields or unexpected people passing by), proper actions will be taken, e.g., data removal.

At the current stage is foreseen that the database will not contain any personal information except:

- Progressive ID of the Agronomic Experts (for agronomical evaluation). As described in the Ethics deliverable D8.1 the real identity behind the evaluation number will be known only to the leader of WP5, who will store it in a nondigital register for his eyes only together with the copies of the informed consent that the expert will sign (for a fac-simile of the informed consent please refer to deliverable D8.1)

- Authors or of the data. The author of the data will be given the possibility to appear in the database with his real name of with a standardized nickname. In both cases he will sign an informed consent that will be kept by the data management responsible

- Name, Last Name and Email of each user of the database. This information will be restricted (only the system administration will have access it). All people signing up in the repository will have to agree on an informed consent form on the use of personal data complying with the Italian legislation.