A Collaborative Paradigm for Human Workers and Multi-Robot Teams in Precision Agriculture Systems (CANOPIES)

This project is funded by the European Union G.A No 101016906

Project Number: 101016906
Start Date of Project: 2021/01/01
Duration: 48 months

Type of document D10.2 – V1.0

Data Management Plan

Dissemination level	Public
Submission Date	2021-06-29
Work Package	WP10
Task	T10.2
Type	ORDP: Open Research Data Pilot
Version	1.0
Author	Alberto Sanfeliu and Ana Puig-Pey (UPC)
Approved by	Consortium

DISCLAIMER:
The sole responsibility for the content of this deliverable rests with the authors. It does not necessarily reflect the opinion of the European Union. Neither the Directorate-General for Communications Networks, Content and Technology, Artificial Intelligence and Digital Industry nor the European Commission are responsible for any use that may be made of the information

Executive Summary

This document provides indications of the type of data of the CANOPIES Project that will be collected, how the data will be preserved and what are the adopted sharing policies towards making these data readily available to the research community.

The data generated in CANOPIES project will be used in the development of the scientific research of the project and in the dissemination of the results. Several datasets will be created through the period of the project and each one of them will be structured following the FAIR data management guidelines for the H2020 Program by the EC.

In the CANOPIES project, we are considering to create four set of datasets: (1) one that will be used for design human-robot safe interaction strategy in the context of precision agriculture; (2) the second one that will be used for helping in the design of the navigation algorithms; (3) the third one that will be used for designing classifiers for vineyard grape cluster detection and tree branches detection; (4) and the fourth one that will be used for designing navigation and human-robot interaction algorithms, using virtual reality sensors.

Table of Content

Abbreviations and Acronyms

DoA	Description of Action
DMP	Data Management Plan
DOI	Digital Object Identifier
EC	European Commission
FAIR	Findable, Accessible, Interoperable and Reusable
WP	Work Package

1 Data summary

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

The data generated in CANOPIES project will be used in the development of the scientific research of the project and in the dissemination of the results as scientific articles, conferences and peer-reviewed publications, and the re-use and verification of these, making possible innovative proposals, from solution designs to prototypes, and industrial products. The data is expected to include publications, internal reports, code, experimental telemetry, models, experimental/simulation videos and learning datasets.

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

Several datasets will be created through the period of the project and each one of them will be structured in a different way, although all of them will be captured and recorded using ROS (Robotic Operating System) and the general recording file will be ROS bag file.

The first set of databases has the purpose to design human-robot safe interaction strategy in the context of precision agriculture. These data will be useful to: (1) assess safe and dangerous situations and have the robot behave accordingly; (2) design human-like robot manipulation strategies in order to have the robot perform agronomic operations in a manner similar to its human counterpart; and (3) design force control strategy to have the robot physically interact with the human operator. At this stage of the development of the project, the specific typology and total number of variables of this dataset cannot be defined a-priori. However, data are numeric and mostly consist in RGB-D sensorial information made of RGB images and Depth data, and of interaction force/torque profiles. Data will be in the form of ROS bag data.

The second set of datasets has the purpose of helping in the design of the navigation algorithms. They will serve for developing: (1) navigation algorithms for the robotic platforms; and (2) estimation, coordination and optimization algorithms for moving the robotic platforms within the vineyard. For the navigation algorithms the data that we expect to collect will be mostly related to the following sensors: GPS, IMU, LIDARs. For the estimation, coordination, and optimization algorithms the data that we expect to collect/generate will be mostly related to the (internal) state of the robots, e.g., pose, battery status etc. These activities will be carried out by resorting to the ROS middleware and therefore the format of the data will be the dictated by the kind of ROS messages used.

The third set of datasets has the purpose of designing classifiers for vineyard grape cluster detection and tree branches detection. They will serve for developing: (1) vineyard grape cluster detection for harvesting; (2) tree branches detection for pruning; and (3) the quality of the grapes. These data will be captured in real time and real scenarios and it will use the following sensors: RGB-D images, LiDAR depth data, GPS RTK, possible some multi-spectral imaging of the table grape bunches and a portable refractometer. The last one will be used to identify the vineyard tree in the field. Data collected will be in the form of ROS bag data.

The fourth set of datasets has the purpose of designing navigation and human-robot interaction algorithms, using virtual reality sensors. These data are collected to be the benchmark for comparison with the data generated by the simulation. In order to establish a simulation with conditions similar to the real environment, the project needs to collect data from different navigation and agricultural sensors used in the robots. These data will be used as a reference when comparing with the sensor data generated during the simulation and transmitted virtually to the robots for movement, detection and perform tasks inside the simulation. As well topographic data of test plots, acceleration data when the robot is moving, GPS data for location and IMU (Inertial Measurement Unit). These data will be used as a reference to correlate the physics of movement of the robot in the simulation. Data collected will be in the form of ROS bag data.

Will you re-use any existing data and how?

No existing data is expected to be re-used for most of the activities.

However, there are datasets closely related to the detection of wine grapes, but they do not provide any sugar content labels. An example of such data is the Embrapa Wine Grape Instance Segmentation Dataset – Embrapa (WGISD). This kind of data could be used to train initial detectors and trackers, or to provide initial data to perform transfer learning when more specific data will be collected.

What is the origin of the data?

The data will be generated in the experimental campaigns and in the experiments made in the research labs, during the project’s lifespan, including: partner’s pre-existing data, data from the scientific literature and real-world measurement data. The origin is observational and sensorial, including standard RGB-D cameras, multi-spectral sensors, 3D LiDAR sensors and force/torque sensors. Data will also be originated from navigation and agronomic perception sensors. GPS and IMU data will be generated when testing the robot's movement.

What is the expected size of the data?

In general, the expected size of the data is not currently known. Some of the datasets could be huge, for example the agronomic detection of clusters can require many high-resolution images and videos across many days per seasons.

To whom might it be useful ('data utility')?

Data will be useful for the purpose of algorithms’ design and, then, for the members of the CANOPIES Consortium who are involved in the design and testing of HRI strategies. By making these data available, the larger research community would be able to re-use and verify it in the following topics: robotics, human-robot collaboration and agricultural robotics. Moreover, it is also expected that the industrial partners use it for the future automatization of their agricultural activities.

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)?

Some of the data will be identified by a unique (randomly generated) identifier within the consortium repository that is being used within the CANOPIES project. The data to be uploaded will be assigned an unique URL or DOI, making it uniquely identifiable and, thus, traceable and referenceable. The implementation of the data description depends on the typology of datum considered: for each acquisition a brief text description is required in which relevant and anonymous information concerning the experiment are reported.

Other data is not intended to be used by third parties. It is solely used for the development of the simulation system. In that case, it will not be implemented an identification mechanism for the data.

What naming conventions do you follow?

In order to be able to distinguish and easily identify some of the collected data sets, each dataset will be assigned with an unique name. All data files produced will include the term "CANOPIES", followed by the date of the experiment (YYYY-MM-DD), a Unique Random Identifier (URI) and a short title:

CANOPIES_[YYYY-MM-DD][URI]_[Short Title].[extension]

Will search keywords be provided that optimize possibilities for re-use?

For some datasets, search keywords will be provided to facilitate re-use. A pre-defined set of keywords will be defined and shared with the Consortium to guarantee consistency.

Do you provide clear version numbers?

We will not provide any versioning mechanisms for datasets.

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.

For some datasets, custom descriptive metadata will be associated with each data set to precisely describe the content of them. A common xml format will be defined to aggregate the relevant information of the data set, e.g., date of the experiment, place of the experiment and involved sensors.

For others datasets, in compliance with the purpose of the data, we do not intend to use any metadata outside those being part of the ROS standard messages.

2.2 Making data openly accessible

Which data produced and/or used in the project will be made openly available as the default? If certain datasets cannot be shared (or need to be shared under restrictions), explain why, clearly separating legal and contractual reasons from voluntary restrictions. Note *that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if relevant provisions are made in the consortium agreement and are in line with the reasons for opting out.*

Some of the data collected is for the design and development of the project algorithms. Once these tasks will be completed, some of the datasets will be opened to the general public.

How will the data be made accessible (e.g., by deposition in a repository)?

Some data will be made available by offering access to a public portion of the consortium repository that is being used within the CANOPIES project.

What methods or software tools are needed to access the data?

Any web browser will be enough to download the data from the repository. Other software tools may be used, but there will be specified through the progression of the project.

Is documentation about the software needed to access the data included?

No documentation is required.

Is it possible to include the relevant software (e.g., in open-source code)?

The code will mainly consist in MATLAB, python, C and C++, and ROS packages. There will be a Readme.txt file that will describe where is the code.

Where will the data and associated metadata, documentation and code be deposited? Preference should be given to certified repositories which support open access where possible.

Some of the data with associated metadata and documentation will be deposited on Microsoft Sharepoint, which is the Consortium data repository that is being used within the CANOPIES project. The project partners could also use their open access repositories (for example https://upcommons.upc.edu/ and DiVA (the Swedish open-access repository of publications)), etc.

MS SharePoint is a web-based collaborative platform of Microsoft that can be used through a web explorer like Microsoft Edge, Internet Explorer, Chrome or Firefox, and can be used as a secure place to store, organize, share, and access data from any device. In this repository, the shared library "project-canopies" adopted as private repository for the CANOPIES project will inherit all the features and flexibility of the SharePoint permission system. Details can be found on the MS SharePoint documentation library available at https://docs.microsoft.com/en-us/sharepoint/. Data will be made available by offering access to a public portion of the Consortium data repository.

Have you explored appropriate arrangements with the identified repository?

The data in MS SharePoint repository that will be used for large amount of data, e.g., navigation data collection campaign, will be hosted by Roma Tre University. While full complete access will be provided to the Consortium, a public section will be created for sharing the data with the scientific community at large not directly involved within the CANOPIES project.

If there are restrictions on use, how will access be provided?

As far as data is concerned, access will be regulated through the MS Sharepoint login access policy for the private part of the repository. For the public part, an open access with reading only permissions will be guaranteed to whoever is interested in using the data.

Is there a need for a data access committee?

No need for a data access committee is foreseen. Data will be completely anonymous and not containing specific biometric or medical data; moreover, access will be restricted to the project members.

Are there well described conditions for access (i.e., a machine-readable license)?

Conditions for access will be based on a username and password system both on MS SharePoint and github platforms. These are standard platforms with clear access documentation already available.

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

As far as the private part the data repository hosted by Roma Tre University is concerned, the identification will be carried out using the standard Microsoft Sharepoint login access policy. Further information concerning the identification can be found in the Microsoft Sharepoint documentation.

As far as the public part of the data repository hosted by Roma Tre University is concerned, no identification will be requested for making the data available to the scientific community. Making data interoperable

Are the data produced in the project interoperable, that is allowing data exchange and re-use between researchers, institutions, organisations, countries, etc. (i.e., adhering to standards for formats, as much as possible compliant with available (open) software applications, and in particular facilitating re-combinations with different datasets from different origins)?

Data collection campaigns will mostly consist in ROS bag files composed of raw sensor data collection. Indeed, ROS is a very well-known and widely used middleware used by the scientific community and this will naturally maximize project interoperability, in terms of data exchange and re-use between researchers, institutions, organizations, countries.

What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?

For ROS datasets, standard ROS messages which are well-known by the research community will be mostly employed for the navigation data set. The GNSS Software Defined Receiver Metadata Standard, which defines XML formats for Global Navigation Satellite Systems, will be taken into account as a reference for defining the metadata vocabulary.

Will you be using standard vocabularies for all data types present in your data set, to allow inter-disciplinary interoperability?

N.A

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?

N.A

2.3 Increase data re-use (through clarifying licenses)

How will the data be licensed to permit the widest re-use possible?

Some data will be licensed under the GNU license.

Other datasets will not be licensed for any re-use and their use will be limited to the duration of the project only.

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.

The data will be available to the CANOPIES members for the design of the algorithms. These data will be open to general public, once they have been published or patented and the members of the Consortium agree in making them public.

Other datasets will not be made available for re-use.

Are the data produced and/or used in the project useable by third parties, after the end of the project? If the re-use of some data is restricted, explain why.

No restriction is expected for the data.

However, it could be that some data will not be open to the public, because, for example a company wants to obtain a patent or wants to commercialize the dataset.

How long is it intended that the data remains re-usable?

In general, the data will remain re-usable up to the end of the project. No guarantee will be given after that period, but we plan to make the data available also afterwards, if the conditions allow for it.

As far as the re-usability is concerned, being some of the data collected in ROS format, it will be re-usable as long as the ROS tools to access bags file will remain as for the current version of the ROS system in use, i.e., ROS Noetic.

Are data quality assurance processes described?

The quality of the dataset is guaranteed by the platform functioning. No data quality assurance process is expected for the collected data.

3 Allocation of resources

What are the costs for making data FAIR in your project?

Data will be stored by resorting to internal hardware and software resources offered free of charge by Roma Tre University.

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).

It is part of the project budget.

Who will be responsible for data management in your project?

The project coordinator - Roma Tre University - has the ultimate responsibility for the data management in the project. He will be responsible of the management of the repository, while each Partner of the Consortium will be responsible of the management of the data collection campaigns carried out for the scientific activities they are leading.

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)?

Regarding the question of long-term data preservation, no specific arrangements has been done in the consortium yet.

4 Data security

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

Data security will be guaranteed through the access functionalities of Microsoft Sharepoint. Sign-in will be required to access the data repository. Public data will not require any authentication to boost data sharing in the scientific community but will be accessible only with reading permissions. Only partners of the Consortium will be enabled to edit the public data.

Data recovery will be provided through the recycle bin functionality of Microsoft Sharepoint. This allows to store and possibly recover deleted contents. The Microsoft Sharepoint version control feature will be also exploited for libraries of interest, guaranteeing the possibility to keep track of all changes made on the libraries.

No transfer of sensitive data is present for the data collection campaigns required for the development of the navigation algorithms.

Is the data safely stored in certified repositories for long term preservation and curation?

All the repositories chosen ensure long term preservation and curation.

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 DoA.

All the activities carried out under the CANOPIES project comply with ethical principles and relevant national, EU and international legislation, as the Charter of Fundamental Rights of the European Union and the European Convention on Human Rights.

In this regard, collected data that contain LIDAR representation of the project personnel, is inherently anonymized not being possible to identify a person from their lidar representation. Should any issue arise, proper actions will be taken to anonymize and de-identify the subjects by altering relevant physical aspects.

Some of the data contain RGB-D images and 3D lidar representation of some of the project personnel. In general, individual names and organizations as well as other relevant information will not be registered. Sensitive issues will be carefully avoided and, if any personal information is content in the data, it will be ensured that they cannot be directly associated to individuals. In the specific case, data will be anonymized and de-identified whenever required by removing face details and other relevant physical aspects.

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

The tasks that involve humans, regarding personal data from questionnaires, require informed consent and will be aligned with Article 34 of the Grant Agreement.

6 Other issues

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

No