D11.9 Data Management Plan

corosect.eu

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101016953

Author(s)/Organisation(s)	Dr. Alex Papadimitriou (CERTH)
Contributor(s)	Maja Zikic (FSH) Rico Mockel (UM) Laurens Nijs (KUL)
Work Package	WP11
Delivery Date (DoA)	30/06/2021
Actual Delivery Date	29/06/2021
Abstract:	This document aims to draft the CoRoSect Data Management Plan (DMP), which aims to describe the data management life cycle for the data to be collected, processed or generated by the project during its lifetime.

Document Revision History			
Date	**Version**	**Author/Contributor/ Reviewer**	**Summary of main changes**
10/2/2021	0.1	Dr. Alex Papadimitriou	Added content skeleton and indicative content
1/6/2021	0.2	Dr. Alex Papadimitriou	Added content
15/6/2021	0.3	Dr. Alex Papadimitriou	Added content from partner input
28/6/2021	0.4	Dr. Alex Papadimitriou	Integrated reviewer comments

Dissemination Level		
PU	Public	**X**
PP	Restricted to other programme participants (including the EC Services)	
RE	Restricted to a group specified by the consortium (including the EC Services)	
CO	Confidential, only for members of the consortium (including the EC)	

Funding Scheme: Innovation Action (IA) ● Topic: H2020-ICT-46-2020

Start date of project: 01 January, 2021 ● Duration: 36 months

CoRoSect Consortium			
Participant Number	**Participant organisation name**	**Short name**	**Country**
1	UNIVERSITEIT MAASTRICHT https://www.maastrichtuniversity.nl/	UM	NL
2	ETHNIKO KENTRO EREVNAS KAI TECHNOLOGIKIS ANAPTYXIS https://www.certh.gr/	CERTH	GR
3	HOCHSCHULE EMDEN/LEER https://www.hs-emden-leer.de/en/	HSEL	GER
4	LUONNONVARAKESKUS https://www.luke.fi/	LUKE	FIN
5	OULUN AMMATTIKORKEAKOULU OY - OULU UNIVERSITY OF APPLIED SCIENCES https://www.oamk.fi/fi/	OAMK	FIN
6	FUNDACION PARA LAS TECNOLOGIAS AUXILIARES DE LA AGRICULTURA http://www.fundaciontecnova.com/	TECNOVA	ES
7	KATHOLIEKE UNIVERSITEIT LEUVEN https://www.kuleuven.be/kuleuven/	KU LEUVEN	BEL
8	ATOS IT SOLUTIONS AND SERVICES IBERIA SL https://atos.net/en/	ATOS	ES
9	ROBOTNIK AUTOMATION SLL http://www.robotnik.es/	ROB	ES
10	AGVR BV www.agvegroup.com	AGVR	NL
11	NASEKOMO AD https://nasekomo.life/	NASEKOMO	BG
12	ENTOMOTECH SL http://entomotech.es/	ENTOMOTECH	ES
13	ENTOCYCLE LTD https://www.entocycle.com/	ENTOCYCLE	GB
14	SOCIETA AGRICOLA ITALIAN CRICKET FARM SRL https://www.italiancricketfarm.com/	ICF	IT
15	INVERTAPRO AS https://www.invertapro.com/	INVERTAPRO	NOR
16	FIELD LAB ROBOTICS BV https://www.fieldlabrobotics.com/	FLR	NL
17	FoodScale Hub https://foodscalehub.com/	FSH	RS
18	AgriFood Lithuania DIH https://www.agrifood.lt/	AFL	LT
19	CENTRO INTERNAZIONALE DI ALTISTUDI AGRONOMICI MEDITERRANEI http://www.iamb.it/	CIHEAM	IT

Table of Contents

List of tables

List of figures

List of Abbreviations and Acronyms	
CA	Consortium Agreement
DMP	Data Management Plan
IPR	Intellectual Property Rights
ORDP	Open Research Data Pilot
PSC	Project Steering Committee
UC	Use Case

Executive Summary

This deliverable is the first version of the CoRoSect plan for handling data and knowledge. The Data Management Plan (DMP) has been drafted in accordance to the regulations of the Pilot action on Open Access to Research Data of the Horizon 2020 programme (H2020). The objective of the DMP is to support the data management life cycle for all data that will be collected processed or generated by the project. It contains preliminary information about the data the project will generate, whether and how it will be exploited or made accessible for verification and re-use, and how it will be curated and preserved. CoRoSect declared its intention to participate in Open Research Data Pilot (ORDP), therefore this DMP has been prepared taking this fact into account. Actions taken in order to participate in ORD Pilot will be reported in subsequent versions of this deliverable.

To develop the present deliverable, an appropriate template was drafted based on the H2020 guidelines for the development of projects' DMP. This was circulated to all project partners so as to collect all relevant information concerning the datasets that are planned to be developed in the course of the project. On the basis of all partners' feedback, the preliminary data management plan of the project has been established, as described in the present deliverable.

Clearly, as the present deliverable has been drafted during the first project stages (M6), it can only reflect the intentions of the project partners toward developing the overall project's datasets. Thus, an updated version of the current deliverable could be submitted later in the project, even if not officially required, should this be judged to be of use to the consortium.

1 Introduction

1.1. Overview

This document has the objective of drafting the CoRoSect Data Management Plan (DMP), which aims to describe the data management life cycle for the data to be collected, processed or generated by the project.

The European Commission has recognised that, apart from scientific publications, research data of a project is also important. Thus, there is a need for a structured approach to the use of research data within a research project. The CoRoSect DMP will outline how research data will be handled throughout the project and after its completion. It will describe what data will be collected and processed and how this data will be shared and/or made open, and how it will be preserved.

1.2. Structure of the deliverable

The deliverable is structured as follows:

Chapter 2 describes the general principles that will guide the drafting of the Data Management Plan.

Chapter 3 provides a description of the datasets that will be created and used within the CoRoSect project, their main attributes and characteristics, as well as the collection process.

Chapter 4 concludes the deliverable with a summary of main conclusions.

2 Data Management Plan – General Principles

2.1. Purpose of data collection

The fundamental aim of the system (and the great innovation it provides) will be to provide repetitive but also cognitively and physically demanding tasks, like transferring and handling of crates (de-stacking and stacking), monitoring of environmental conditions, larvae separation/detection, insect feeding, which require increased manual effort or continuous human supervision, with correspondingly automatic robotic-based procedures, as service in an I40-compliant Information-Communication Infrastructure. The overall goal is to create a collaboration environment, where humans and robots will harmoniously share and undertake at the same time different processing and manipulation tasks, targeting the application case of insect farming.

At this early stage of the project, a set of use cases (UCs) and scenarios have been defined, which will elicit the user requirements, the ethics/regulation/social acceptance and system requirements, model the dynamic assessment and define the system architecture and the system integration plan. A set of appropriate datasets needs to be identified early in the project, to evaluate the performance of the proposed technologies and to demonstrate the different UCs and scenarios. All these datasets have been defined after extensive discussions in order to provide the best possible coverage of user needs. However, introduction of another dataset is still possible through the lifetime of the project; in that case, a similar analysis will be performed and the DMP will be updated accordingly.

2.2. Participation in the Pilot on Open Research Data

CoRoSect participates in the Pilot on Open Research Data (ORD) launched by the European Commission along with the Horizon2020 programme. The consortium believes firmly in the concepts of open science, and the large potential benefits the European innovation and economy can draw from allowing reusing data at a larger scale. Therefore, an effort will be made to publish with open access a significant amount of data produced by the project. If certain datasets cannot be shared, a justification for opting out will be provided.

2.3. FAIR data

The elements of the FAIR Principles are related, but independent and separable. This principle defines how research outputs should be organised in order to adequately facilitate discovery, exchange and reuse by third-parties. Major funding bodies, including the European Commission, promote FAIR data to maximize the integrity and impact of their research investment[1].

[1] The Open Data Foundation http://www.odaf.org/

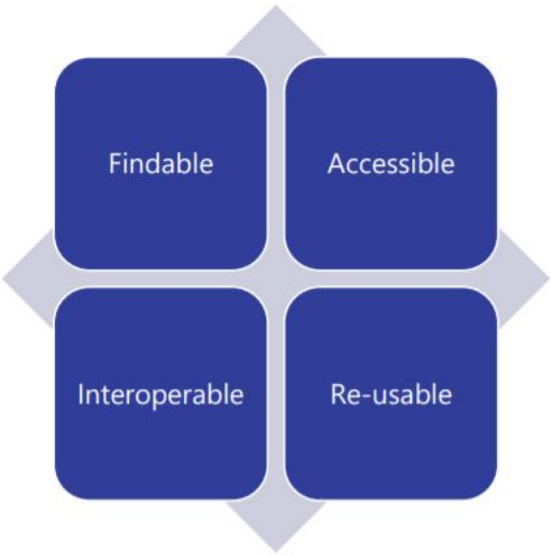

Figure 1 The FAIR Principle

Within CoRoSect, consortium members will respect the principles of FAIR data management, that is Findable, Accessible, Interoperable and Re-usable. To achieve this, a DMP should include information on:

- the handling of research data during and after the end of the project
- what data will be collected, processed and/or generated
- which methodology and standards will be applied
- whether data will be shared/made open access and
- how data will be curated and preserved (including after the end of the project).

2.3.1. Findable

Data and metadata should be easily findable for humans as well as for computers. Crucial for data to be findable is whether "the data produced or used in the project is discoverable with metadata, identifiable and locatable by means of a standard identification mechanism."[2] To be findable any Data Object should be uniquely and persistently identifiable. The same Data Object should be persistent, the Data Object should contain basic machine actionable metadata that allows it to be distinguishable from other Data Objects. Identifiers for any concept used in data Objects should thus be Unique and Persistent. To make data findable following procedure should be considered:

- F1. (Meta)data are assigned a globally unique and persistent identifier
- F2. Data are described with rich metadata
- F3. Metadata clearly and explicitly include the identifier of the data they describe
- F4. (Meta)data are registered or indexed in a searchable resource[3]

In CoRoSect a variety of data sets is generated and collected. The data collection and processing workflow include discrete steps in which different versions of the data are stored. The original data collected from the data sources are stored and maintained in one step to ensure that the provenance

[2] See: https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf
[3] See: https://www.go-fair.org/fair-principles/

of the data can be traced back. Metadata for each data source are stored in a catalogue. After applying processing methods, a new version of the data is stored. Each set of data produced (dataset, deliverables, etc.) will be named in a uniform way and will include a table with a version control. The recommendations are as follows:

- Choose easily readable identifier names (short and meaningful);
- Do not use acronyms that are not widely accepted;
- Do not use abbreviations or contractions;
- Avoid Language-specific or non-alphanumeric characters;

More details on naming conventions are given in CoRoSect D12.7 Quality Assurance Plan.

2.3.2. Accessible

Accessibility requires that data can be always obtained by machines and humans upon appropriate authorisation through well-defined protocols. It should be determined what generated or used data will be made openly available. Where datasets cannot be shared or are restricted a clear explanation is required highlighting the legal or contractual reasons for the restriction of accessibility.

- A1. (Meta)data are retrievable by their identifier using a standardised communications protocol
    - A1.1 The protocol is open, free, and universally implementable
    - A1.2 The protocol allows for an authentication and authorisation procedure, where necessary
- A2. Metadata are accessible, even when the data are no longer available[4]

At the time of this first version of the Data Management Handling Plan information about which data can and will be published or made openly available is still imprecise. Data access will vary depending on the storage location. Starting with the use case data, measures will be taken to enable third parties to access, re-use, analyse, exploit, and disseminate the data (bound by the license specifications). Different access procedures will be implemented, enabling the export of an entire dataset as well as the provision of a querying interface for the retrieval of relevant subsets.

Access mechanisms will also be supported as much as possible by metadata enabling search engines and other automated processes to access the data using standard web mechanisms.

On partner's side, IT departments, quality assurance teams, systems administrators and data protection officers will guarantee the correct execution and give access only to staff involved in the project. Whereas the partners themselves follow the respective company's standard risk management to guarantee safety of data the procedure of a risk assessment formalization for the platform is in progress.

Project related datasets (deliverables, reports, code, DBs) will be stored in Coordinator's premises or Consortium agreed tools with proper access rights. Project administrative data is not intended to be publically shared or otherwise made available to third parties, access to the repository is granted by the coordinating partner to the partners actively involved in the development of the project.

Digital data will be released in machine-readable formats that supplement journal articles and presentations, sharing requirements are in conformity with open data models. Metadata of the datasets will be available in machine readable format following standards such as DCAT

[4] See: https://www.go-fair.org/fair-principles/

2.3.3. Interoperable

Interoperability refers to allowing data exchange and re-use between researchers, institutions, organisations, countries and other such parties. This element requires adherence 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. The metadata vocabularies, standards and methodologies must ensure interoperability.

- I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. (Meta)data use vocabularies that follow FAIR principles
- I3. (Meta)data include qualified references to other (meta)data[5]

2.3.4. Re-usable

Re-usability means that data can be easily utilised by third parties. The final goal of the European Commission's FAIR principles policy is to optimise the re-use of data. It is therefore important to determine how the data will be licensed to permit the widest re-use possible, when will data be made available for re-use, the time-frame of the intended reuse and data quality assurance processes. In cases where 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.

- R1. Meta(data) are richly described with a plurality of accurate and relevant attributes
- R1.1. (Meta)data are released with a clear and accessible data usage license
- R1.2. (Meta)data are associated with detailed provenance
- R1.3. (Meta)data meet domain-relevant community standards

Public deliverables will be openly accessible on the project website and as such are open for reuse by all interested researchers and institutions. The website is available at following URL: https://corosect.eu/.

All partners in the consortium are committed to engage in dissemination activities such as webinars, papers and presentations to promote the project and its findings within the scientific and the food safety community. The output will be made openly accessible for re-use whenever it is in alignment with the organizing institution. Regarding scientific outcomes, the consortium is committed to the Horizon2020 Open Access mandates and is planning to embrace all possible Open Access options known today. These include Gold Open Access, Green Open Access and self-archiving. The Consortium partners will therefore privilege Open Access journals or non-Open Access journals that support Green and Gold roads.

Administrative project data concerning management and finances is confidential within the consortium and not meant for re-use of third parties. Partners contact details, communications between partners and meeting minutes include personal data and privacy of correspondence and telecommunications have to be considered. What other platform and technical data will be made re-usable for research still has to be established. Given the very nature of the data exchanged as being business sensitive re-use will however have to be restricted for substantial parts to be in compliance with contractual commitments such as non-disclosure agreements necessary to get consent for the

[5]See: https://www.go-fair.org/fair-principles/

use of this data in the project, partners themselves can and will restrict re-usability. Plans for the re-use of data after the end of the project have still to be defined by the consortium

2.4. Interoperability

In a networked production system, all the system in the dynamic and open ecosystem should agree among themselves on the objectives to be supported in a mutual exchange of information and knowledge. Such decentralised structure of a production system needs to be designed according to meet the Industry 4.0 compliant communication structure, which ultimately enables cooperation between all the system and called as Interoperability.

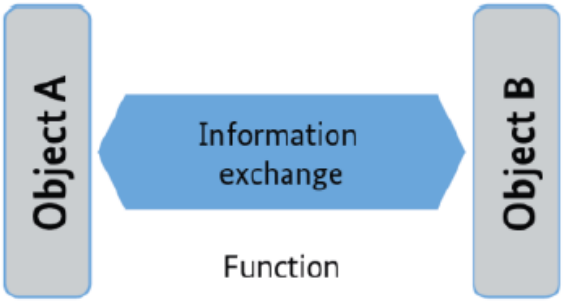

Figure 2: Information Exchanged in a function

Interoperability is the ability of different objects, can be from the same vendor or from different vendors for exchanging information and use that information for the right cooperation. In other words, two or more systems are interoperable if they can perform cooperatively a specific function using the information that is exchanged. Industry 4.0 defines the suitable data structures for the exchange of data and their defines meaning. This exchange of data in the standardized form with its defined meaning is called as semantic interoperability. The Asset Administration Shell is a standard that will provide a model for the interoperability of the assets [1].

Complex information from various subsystems of a machine or an asset must be provided, processed, and interpreted. The interactions between the Administration Shells of the contributing subsystems component orchestrate the Industry 4.0-system to execute the value chains of an ecosystem. For this purpose, the administration shells require a common language to accomplish a high degree of interoperability in the storage, exchange, and processing of data in such a technical system at the information level [2].

2.5. Security

The datasets foreseen to be collected through CoRoSect are of high value and may contain sensitive personal data. Special care should be taken to prevent such datasets to leak or become hacked. This is another key aspect of CoRoSect data management, and all data repositories used by the project will include effective protection.

A holistic security approach will be followed, in order to protect the pillars of information security (confidentiality, integrity, availability). The security approach will consist of a methodical assessment of security risks followed by their impact analysis. This analysis will be performed on the personal information and data processed by the proposed system, their flows and any risk associated to their processing. Security measures will include the implementation of PAKE protocols, such as the SRP protocol, and protection about bots such as captcha technologies. Moreover, the pilot sites shall apply

monitored and controlled procedures related to the data collection, their integrity and protection. The data protection and privacy of personal information will include protective measures against infiltration as well as physical protection of core parts of the systems and access control measures.

2.6. Ethical aspects and personal data protection

Activities within CoRoSect will involve the participation of animals and in particular insects for related data collection purposes. CoRoSect activities will involve human participants for various human activity – related data collection purposes. Therefore, it is clear that in some cases personal data will have to be collected.

Such data will be protected in accordance with the EU's Data Protection laws, where the main legislation is the 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, also known as the 'General Data Protection Regulation' (GDPR), which entered into force on 25 May 2018.

Further information on how personal data collection and handling should be approached in the CoRoSect project are provided in the deliverables of WP1 "Ethical, Legal and Social Implications (ELSI) of human-robot collaboration in industrial automation". The CoRoSect consortium, guided by the responsible partner KUL, will investigate if there are further issues that can have impact on data sharing.

2.6.1. Transfer to third countries

Any transfer of personal data to recipients in a third country or international organisation is prohibited unless they fully comply with the conditions set out in Chapter V of the GDPR. (Art 44- 50) The scope of the GDPR is to protect natural person's data. Although the GDPR recognises the necessities of international trade and international cooperation and the hereby increased flow of data, the core of the GDPR shall not be undermined by transferring data to and from countries outside the Union[6]. Thus the transfer of personal data to a third country or to an international organisation, known as 'data export',[7] is not allowed unless the jurisdiction in which the recipient is located is deemed to provide an adequate level of data protection, the data exporter puts in place appropriate safeguards or a derogation or exemption applies.

A third country is a country other than the EU member states and the three additional EEA countries (Norway, Iceland, and Liechtenstein) which have adopted a national law implementing the General Data Protection Regulation (GDPR). The CoRoSect consortium includes a partner from a non-EU country, FSH (Serbia). Thus any transfer of personal data to and from Jordan or Egypt are subject to the restrictions set out in the GDPR.

In late 2018, Serbia updated its data protection law to better align with the EU General Data Protection Regulation. Serbia enacted a new Data Protection Law on 9 November 2018 (published in the Official Gazette of the Republic of Serbia, no. 87/2018) ("DP Law"). Although the DP Law entered into force 21 November 2018, its effective date was postponed until 21 August 2019 (except for the maintenance of the Central Register of Personal Databases which has already been terminated). Its content is largely harmonized with the GDPR and it is now fully effective as of 21 August 2019.

[6] Recital 101 GDPR

[7] Gawronski (ed.) (2019): Guide to the GDPR, p. 99.

The crucial element in all data transfer to third countries is that the controller must ensure by other means that the personal data of the data subject will be sufficiently protected by the recipient.

As an entity operating in a European environment and often handling EU data, which falls under the GDPR, FSH continues to be vigilant when it comes to respecting and following the GDPR. A dedicated GDPR responsible person, officially Data Protection Officer (Ms. Maja Zikic), has been assigned in order to keep up with the latest developments in the field and implement them in FSH daily operations. As a result, FSH is fully compliant with both the Serbian "DP Law" as well as the GDPR.

FSHs activities in the CoRoSect project involve a number of key stakeholders. This is achieved through FSH dissemination, communication, and community building activities. Their first contact point with CoRoSect is via CoRoSect.eu website:

- Contact form on the CoRoSect.eu (available at: https://corosect.eu/contact/)
- Newsletter subscription box on CoRoSect.eu (available at: https://corosect.eu/)

The consent is obtained each time our stakeholders try to reach out to us via the CoRoSect.eu Contact page (available at: https://corosect.eu/contact/). Before they click the "SEND" button, they must confirm that they have read and understood the CoRoSect Privacy Policy (available at: https://corosect.eu/privacy-policy/).

3 Data Summary

In this chapter, the datasets that are planned to be captured by the CoRoSect partners are described. The description of the datasets complies with the EC guidelines of the dataset aspects that should be reported in DMPs of the H2020 projects[8].

The descriptions of the datasets, in this initial version of the deliverable, may not include information about the exact storage of the datasets, the data erasure policy of personal data or anonymisation of personal data and the specific technical and organisational measures of partners to secure personal data when carrying out the pilots. Such information will be made available once it becomes clearer and will be included in subsequent versions of the deliverable.

3.1. Summary of the foreseen CoRoSect datasets

A summary of the planned datasets and a short description of each one is presented in the table below.

Dataset name	Short Description
DS1.CERTH Object Recognition Dataset	A large-scale dataset of images and associated annotations will be collected depicting insects of various life-cycle stages.
DS2.CERTH Object Manipulation Tasks	A set of videos/images associated with fine manipulation tasks currently performed by human workers at end-user sites.
DS3.UM Manipulation Task Demonstrations	A set of videos/images associated with fine manipulation tasks currently performed by human workers at end-user sites, to be used as training and demonstrations trajectories by learning from demonstration algorithms. The recordings will contain human workers performing atomic tasks at various stages of insect rearing

[8] http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

	processes, which have been previously identified as use-cases for the fine manipulation robot.
DS4.FSH CoRoSect.eu Newsletter	This dataset is collected through the CoRoSect website (available at: corosect.eu). It contains name and last name, professional email address and company/ organization name of key stakeholders interested in CoRoSect activities and results.
DS5.ICF Breeding Processes	The data are a series of images and videos illustrating ICF's breeding processes.
DS6.ICF Cricket Production Processes	Information on ICF's cricket production process. Graphs, data and tables and findings relating to the biology of the insect.

Table 1: List of foreseen datasets to be collected within CoRoSect.

3.2. DS1.CERTH – Object Recognition Dataset

General description

A large-scale dataset of images and associated annotations will be collected depicting insects of various life-cycle stages.

DS1.CERTH – Object Recognition Dataset

Dataset description

Data Origin
The dataset will include a collection of RGB-D images of insects captured from various viewpoints using an appropriate depth sensor (e.g., RealSense) or other sensing/scanning devices. Specifically, the capturing process will take place at the end users' establishment over an estimated period of two (2) days.

Nature and scale of data
The data will consist of images depicting real insects at various life-cycle stages for evaluation. The target is that the dataset will contain the breeding process of at least 10 instances of each insect type.

To whom could the dataset be useful
The dataset will be valuable for benchmarking algorithms for (small) object recognition, anomaly detection, among others.

Related scientific publication(s)
The dataset will accompany our research results in the fields of (small) object recognition and anomaly detection.

Indicative existing similar data sets
There are several public datasets containing RGB-D images of objects aimed at object recognition.
The UW dataset (http://www.cs.washington.edu/rgbd-dataset/)
The Berkley's B3DO dataset (http://kinectdata.com/)
The Berkley's BigBird dataset (http://rll.berkeley.edu/bigbird/).

Personal data

Do you collect or process personal data in the context of CoRoSect?

☐ **Yes** — Please continue with next question

☒ **No** — You can leave this section open

Personal data collection

Actor	Type of data[9]	If generated, is it shared with third parties?	If collected, from which second/third party do the data originate from?

Purpose of data collection
N/A.

Legal basis for processing data
We can rely on the following legal basis for processing and collecting personal data (check/color the box):

[9] Please indicate if the data is, in your view, personal or technical and why this data is generated or collected.

- ☐ Consent data subject
- ☐ Performance of contract
- ☐ Legal obligation
- ☐ Vital interest of data subject
- ☐ Task carried out in the public interest or in the exercise of official authority
- ☐ Legitimate interest of the controller

Explain your choice: …

Data transfers to non-EU countries
N/A

Security of data
N/A

Metadata and annotation

Indicative metadata include a) foreground-background masks for training images, b) camera calibration information, c) camera pose matrix for each viewpoint, d) object identifier, description category label and 3D pose annotation. The metadata will be in a format that may be easily parsed with open-source software.

Open access to data

Access type
Open access.

How will the data be made accessible
The data will be stored on CERTH infrastructure and will be made available to everyone who requests it via an online form.

What methods or software tools are needed to access the data
The dataset will be designed to allow easy reuse with commonly available tools and software libraries, i.e., any commercially available video and image player could be used to view the collected dataset.

Where will the data and associated metadata, documentation and code be deposited
The dataset will be accommodated at CERTH's own private and internal data storage server.

Data re-use

Data preservation period
The dataset will be preserved by CERTH at least until the end of the project.

Embargo periods
Some datasets maybe available only after a corresponding paper is accepted and published.

Allocation of resources

Indicative associated costs for data archiving and preservation
Probably two dedicated hard disk drives will be allocated for the dataset; one dedicated to the public part and one to the private. No costs are currently foreseen regarding its preservation.

Indicative plan for covering the above costs
Small one-time costs covered by CoRoSect.

Partners activities and responsibilities

Partner Owner / Data Collector
CERTH

Partner in charge of the data analysis
CERTH

Partner in charge of the data storage CERTH **WPs and Tasks** The data is going to be collected within activities of WP5 and are planned to be mainly used in the research efforts of the same WP.

Table 2: DS1.CERTH - Object Recognition Dataset description.

3.3. DS2.CERTH – Object Manipulation Tasks

General description

A set of videos associated with fine manipulation tasks currently performed by human workers at end-user sites.

DS2.CERTH – Object Manipulation Tasks
Dataset description
Data Origin The dataset will include a collection of RGB-D videos of human workers captured from various viewpoints using an appropriate depth sensor (e.g., RealSense) or other sensing/scanning devices, performing certain manipulation tasks that have yet to be defined. Specifically, the capturing process will take place at the end users' establishment over an estimated period of two (2) days. **Nature and scale of data** The data will consist of videos depicting object handling tasks for evaluation. The target is that the dataset will contain the execution of at least 3 manipulation tasks performed by human workers during insect manipulation. **To whom could the dataset be useful** The dataset will be valuable for benchmarking algorithms for Active Learning, Immitation Learning and Learning from user input. **Related scientific publication(s)** The dataset will accompany our research results in the fields of Active Learning, Immitation Learning and Learning from user input. **Indicative existing similar data sets** There are several public datasets containing RGB-D images of object manipulation tasks aimed at Immitation learning and learning from human input: https://sites.google.com/site/brainrobotdata/home/multiview-pouring http://www.ycbbenchmarks.org/ https://www.robonet.wiki/
Personal data
Do you collect or process personal data in the context of CoRoSect? ☐ **Yes** — Please continue with next question ☒ **No** — You can leave this section open **Personal data collection** Actor | Type of data[10] | If generated, is it shared with third parties? | If collected, from which second/third party do the data originate from? | | | | **Purpose of data collection** N/A **Legal basis for processing data**

[10] Please indicate if the data is, in your view, personal or technical and why this data is generated or collected.

We can rely on the following legal basis for processing and collecting personal data (check/color the box):

- ☐ Consent data subject
- ☐ Performance of contract
- ☐ Legal obligation
- ☐ Vital interest of data subject
- ☐ Task carried out in the public interest or in the exercise of official authority
- ☐ Legitimate interest of the controller

Explain your choice: …

Data transfers to non-EU countries
N/A

Security of data
N/A

Metadata and annotation

Indicative metadata include a) foreground-background masks for training images, b) camera calibration information, c) camera pose matrix for each viewpoint, d) object identifier, description category label and 3D pose annotation. The metadata will be in a format that may be easily parsed with open-source software.

Open access to data

Access type
Open access.

How will the data be made accessible
The data will be stored on CERTH infrastructure and will be made available to everyone who requests it via an online form.

What methods or software tools are needed to access the data
The dataset will be designed to allow easy reuse with commonly available tools and software libraries, i.e., any commercially available video and image player could be used to view the collected dataset.

Where will the data and associated metadata, documentation and code be deposited
The dataset will be accommodated at CERTH's own private and internal data storage server.

Data re-use

Data preservation period
The dataset will be preserved by CERTH at least until the end of the project.

Embargo periods
Some datasets maybe available only after the corresponding paper is accepted and published.

Allocation of resources

Indicative associated costs for data archiving and preservation
Probably two dedicated hard disk drives will be allocated for the dataset; one dedicated to the public part and one to the private. No costs are currently foreseen regarding its preservation.

Indicative plan for covering the above costs
Small one-time costs covered by CoRoSect.

Partners activities and responsibilities

Partner Owner / Data Collector
CERTH

Partner in charge of the data analysis
CERTH

Partner in charge of the data storage

CERTH **WPs and Tasks** The data are going to be collected within activities of WP8 and are planned to be mainly used in the research efforts of the same WP.

Table 3: DS2.CERTH – Object Manipulation Tasks Dataset description

3.4. DS3.UM - Manipulation Task Demonstrations

General description

A set of videos/images associated with fine manipulation tasks currently performed by human workers at end-user sites, to be used as training and demonstrations trajectories by learning from demonstration algorithms. The recordings will contain human workers performing atomic tasks at various stages of insect rearing processes, which have been previously identified as use-cases for the fine manipulation robot.

DS3.UM - Manipulation Task Demonstrations
Dataset description
Data Origin The dataset will be captured using cameras from multiple viewpoints at the end-users' establishments. Recordings may take multiple days depending on potential fine manipulation tasks. At least RGB video data would be required for image analysis and pose detection approaches, with added benefit if depth information is already included in data (RGB-D). **Nature and scale of data** The data will be video data depicting specific atomic tasks performed at the end users' establishments at various stages of the industrial processes. The dataset would contain at least 10 instances of each atomic task that can be recorded, as well as labelling of the ground truth task being depicted. The length of the videos would be entirely dependent on the task depicted. **To whom could the dataset be useful** The dataset will be valuable for learning from demonstration approaches to training robotic manipulators, as well as machine learning solutions to related use-cases. **Related scientific publication(s)** The dataset will accompany research results in the field of fine insect manipulation and learning from demonstration. **Indicative existing similar data sets** Multiple Interactions Made Easy (MIME) dataset https://sites.google.com/view/mimedataset
Metadata and annotation
Indicative metadata include a) ground truth task description and label, b) camera calibration information, c) camera pose matrix for each viewpoint. The metadata will be in a format that may be easily parsed with open-source software.
Open access to data
Access type Restricted access. Access limited to partners of the CoRoSect consortium. **How will the data be made accessible** Data will be shared with other CoRoSect partners. **What methods or software tools are needed to access the data** The dataset will be designed to allow easy reuse with commonly available tools and software libraries, i.e., any commercially available video and image player could be used to view the collected dataset. **Where will the data and associated metadata, documentation and code be deposited** The dataset will be accommodated at UM on an internal storage server/device.
Data re-use

Data preservation period
The dataset will be preserved by UM at least until the end of the project.

Embargo periods
Some datasets may be available only after the corresponding paper is accepted and published.

Allocation of resources

Indicative associated costs for data archiving and preservation
No costs are currently foreseen regarding its preservation.

Indicative plan for covering the above costs
Small one-time costs covered by CoRoSect.

Partners activities and responsibilities

Partner Owner / Data Collector
UM

Partner in charge of the data analysis
UM

Partner in charge of the data storage
UM

WPs and Tasks
The data are useful for WP6, WP7 and WP8 and are planned to be mainly used in the research efforts of the corresponding work packages.

Table 4: DS3.UM - Manipulation Task Demonstrations dataset description.

3.5. DS4.FSH – CoRoSect.eu Newsletter

General description

This dataset is collected through the CoRoSect website (available at: corosect.eu). It contains name and last name, professional email address and company/ organization name of key stakeholders interested in CoRoSect activities and results.

DS4.FSH – CoRoSect.eu Newsletter
Dataset description
Data Origin Data is collected through a Newsletter subscription box on the CoRoSect website (available at: corosect.eu) The dataset contains name and last name, professional email address and company/ organization of CoRoSect.eu newsletter subscribers. **Nature and scale of data** The data is mainly in the text format. The dataset is small (approximately 50 MB) **To whom could the dataset be useful** This dataset will be used for the purposes of dissemination, communication and exploitation activities targeting robotics/ AI related stakeholders, target groups, and future collaborators. **Related scientific publication(s)** Not applicable. **Indicative existing similar data sets** Not applicable.
Metadata and annotation
Not applicable.
Open access to data
Access type Restricted – the dataset is available only to CoRoSect Dissemination, Communication and Exploitation Task Force (FSH) which is managing the project's mailing list. **How will the data be made accessible** The dataset contains personal information. Therefore, the dataset will not be accessible to people/ organizations outside the consortium. In fact, this dataset is directlz accessible only to the CoRoSect Dissemination and Communication Manager – FSH. In case one of the CoRoSect partners wishes to make an announcement or publish a relevant update through the CoRoSect mailing list, they have to reach out to FSH first. The communication workflow is described in detail in D11.1. **What methods or software tools are needed to access the data** To access the data, we are using Mailchimp, email marketing automation platform. Data can be exported in a .csv or .xlxs format. **Where will the data and associated metadata, documentation and code be deposited** Data will be deposited in the CoRoSect Teams channel.
Data re-use
Data preservation period Until the end of the project (December 2023). Data needed to answer to potential audits by the European Commission services may be kept for up to 5 years after the end of the Project. **Embargo periods** Not applicable.
Allocation of resources
Indicative associated costs for data archiving and preservation None. **Indicative plan for covering the above costs** Not applicable.

Partners activities and responsibilities
Partner Owner / Data Collector FSH **Partner in charge of the data analysis** FSH **Partner in charge of the data storage** FSH **WPs and Tasks** The data collected in this dataset will be used in the scope of WP11 – Dissemination, Communication and Exploitation (led by FSH), and will be used primarily in the scope of T11.1 – Dissemination and Communication Plan (FSH), T11.2 – Communication and Dissemination Activities (FSH), T11.3 – Community Building (AFL) for the purposes of reaching out to key stakeholders, target groups and potential collaborators and informing them about CoRoSect use cases, events/ activities, relevant updates from our ecosystem, and more.

Table 5: DS4.FSH – CoRoSect.eu Newsletter dataset description.

3.6. DS5.ICF – Breeding Processes

General description

The data are a series of images and videos illustrating ICF's breeding processes.

DS5.ICF – Breeding Processes
Dataset description
Data Origin Data from ICF **Nature and scale of data** The data are in video and image formats (JPG, NPG) and in PSD format. The storage of all data has an estimated size of 2 GB. **To whom could the dataset be useful** Useful to the whole consortium for understanding our production processes, except for other end-users **Related scientific publication(s)** No **Indicative existing similar data sets** No
Metadata and annotation
No
Open access to data
Access type No data provided by us may be disclosed outside CoRoSect. **How will the data be made accessible** The data will be uploaded on the dedicated MS teams platform or alternatively sent by e-mail to the interested parties, if and only if they belong to the CoRoSect project. **What methods or software tools are needed to access the data** General ICF data are stored on Drop-Boxes to which only a few appointed ICF employees have access. Accounting data and financial data in general are stored on a platform in communication with the Italian Revenue Agency; these are stored and protected for 5 years from their issue. **Where will the data and associated metadata, documentation and code be deposited** ICF will explore data repository options and decide on the appropriate repository after the data collection process has been completed to ensure updated options.
Data re-use
Data preservation period ICF stores its data for 5 years **Embargo periods** ICF stores its data for 5 years
Allocation of resources
Indicative associated costs for data archiving and preservation 500,00 euros/year **Indicative plan for covering the above costs** - ICF will uptake the costs of preservation

Partners activities and responsibilities
Partner Owner / Data Collector ICF **Partner in charge of the data analysis** ICF **Partner in charge of the data storage** ICF **WPs and Tasks** WP 02

Table 6: DS5.ICF – Breeding Processes dataset description.

3.7. DS6.ICF - Cricket Production Processes

General description

Information on ICF's cricket production process. Graphs, data and tables and findings relating to the biology of the insect.

DS6.ICF - Cricket Production Processes
Dataset description
Data Origin Data from ICF **Nature and scale of data** XXX **To whom could the dataset be useful** Useful to the whole consortium for understanding our production processes, except for other end-users **Related scientific publication(s)** - **Indicative existing similar data sets** -
Metadata and annotation
Open access to data
Access type No data provided by us may be disclosed outside CoRoSect. **How will the data be made accessible** The data will be uploaded on the dedicated MS teams platform or alternatively sent by e-mail to the interested parties, if and only if they belong to the CoRoSect project. **What methods or software tools are needed to access the data** General ICF data are stored on Drop-Boxes to which only a few appointed ICF employees have access. Accounting data and financial data in general are stored on a platform in communication with the Italian Revenue Agency; these are stored and protected for 5 years from their issue. **Where will the data and associated metadata, documentation and code be deposited** ICF will explore data repository options and decide on the appropriate repository after the data collection process has been completed to ensure updated options.
Data re-use
Data preservation period for 5 years **Embargo periods** for 5 years
Allocation of resources
Indicative associated costs for data archiving and preservation -500,00 euros/year

Indicative plan for covering the above costs - ICF will uptake the costs of preservation
Partners activities and responsibilities
Partner Owner / Data Collector ICF **Partner in charge of the data analysis** ICF **Partner in charge of the data storage** ICF **WPs and Tasks** WP02-WP03

Table 7: DS6.ICF - Cricket Production Processes.

4 Conclusions

A first version of the project's Data Management Plan (DMP) has been prepared and presented in this deliverable. However, this is a "live" document, which will be constantly updated during the project, fully addressing the lifecycle and public availability of generated research data as well as any other type of data that is gathered within the scope of the project.

So far, six datasets have been identified and will be made open at the early stages of the project and will be offered to the Open Research Data Pilot. The DMP will detail what data the project will generate, whether and how it will be exploited or made accessible for verification and re-use, and how it will be curated & preserved.

References

1. Birgit Boss, Somayeh Malakuti, Shi-Wan Lin, Thomas Usländer, Erich Clauer, Michael Hoffmeister, Ljiljana Stojanovic, Digital Twin and Asset Administration Shell Concepts and Application in the Industrial Internet and Industrie 4.0, Available: (https://www.plattform-i40.de/PI40/Redaktion/EN/Downloads/Publikation/Digital-Twin-and-Asset-Administration-Shell-Concepts.pdf?__blob=publicationFile&v=7). Accessed on 15th October,2020
2. Bundesministerium für Wirtschaft und Energie (BMWi), Positionspapier – Interoperabilität: Interoperabilität – Unsere Vision für Industrie 4.0: Maschinen sprechen in vernetzten digitalen Ökosystemen interoperabel miteinander. [Online]. Available: https://www.plattform-i40.de/PI40/Redaktion/DE/Downloads/Publikation/Positionspapier-Interoperabilität.pdf?__blob=publicationFile&v=3## (accessed: Jan. 14,2020).

CoRoSect

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101016953