C0PEP0D

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 834238)

C0PEP0D project – Deliverable 1.1
Data Management Plan

ABSTRACT :

This document presents the first version of the Data Management Plan (DMP) for the C0PEP0D project based on the ERC DMP template.

This DMP describes the data management life cycle for the data to be collected, processed and/or generated and is intended to be a living document.

As part of making research data findable, accessible, interoperable and re-usable (FAIR), this DMP includes information on:

- the handling of research data during and after the end of the project
- which 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.

Deliverable n° 1.1

Version N°0

Due date : 29 Feb 2020

Submission date: M2

Start date of the project: 1st Sept. 2019

Duration: 60 months

Deliverable leader: ECM

Dissemination level: Public

REFERENCES

Project Acronym	Project Number
C0PEP0D	834238

WP N°/title/leader	**WP1**	Management, dissemination and exploitation	ECM
Deliverable N°/title/leader	**D1.1**	Data Management Plan	ECM
Writer(s)	Pr. Eloy / ECM		
Organization	ECM		
Date of report	07/04/2022		

TABLE OF CONTENT

DATA MANAGEMENT PLAN

Update of the DMP record

Revision	Date	Description	Reviewer
0.0	19/09/2019	Initial outline	C. Eloy
1.9	07/04/2022	Mid-term scientific report	C. Eloy

Executive Summary

This document, D1.1 Data Management Plan (DMP) is a deliverable of the C0PEP0D project, which is funded by the European Research Council Executive Agency through ERC Programme under Grant Agreement N° 834238.

The objective of the C0PEP0D project is to decipher how planktonic copepods exploit hydrodynamic and chemical sensing to detect and track targets in turbulent flows.

Copepods are millimetric crustaceans that play a crucial role in marine ecosystems. They live in all seas and oceans and are thought to be the most abundant multi-cellular organism of the planet. Yet, copepods are blind. To detect preys, predators and mates, copepods use hydrodynamic and chemical sensing. How are they able to distinguish a meaningful signal in oceanic turbulence? Copepods being one of the greatest success-story of marine evolution, they likely evolved smart algorithms to process this sensing information. But today, these algorithms are poorly understood.

C0PEP0D aims at deciphering these algorithms by addressing three questions:
- **Q1: Mating.** How do male copepods follow the pheromone trail left by females?
- **Q2: Finding.** How do copepods use hydrodynamic signals to 'see'?
- **Q3: Feeding.** What are the best feeding strategies in turbulent flow?

C0PEP0D hypothesises that reinforcement learning can reverse-engineer copepod algorithms. To do so, we will build a virtual environment where copepods interact with a turbulent flow and learn. In this environment, copepods evolve generation after generation, with the goal of finding efficient neural networks mimicking the 1000-neuron brain of copepods.

The project will allow us to better understand the detecting and tracking skills of copepods, providing an inspiration for artificial sensors. By developing an evolutionary approach of reinforcement learning, we will also build a tool useful for biology and engineering.

1. Data management and responsibility

1.1 DMP Internal project Policy

The C0PEP0D project is engaged in the Open Research Data (ORD) pilot which aims at improving and maximising access to and re-using of research data generated by ERC projects. The C0PEP0D project takes into account the need to balance openness and protection of scientific information, commercialisation and Intellectual Property Rights (IPR), privacy concerns, security as well as data management and preservation questions.
The management of the project data/results requires decisions about the sharing of the data, the format/standard, the maintenance, the preservation, etc.
Thus the Data Management Plan (DMP) is a key element of good data management and is established to describe how the project will **collect**, **share** and **protect** the data produced during the project. As a living document, the DMP and its annexes can be updated over the lifetime of the project whenever necessary.

1.2 Data management responsible

The Data Management Responsible (DMR) in C0PEP0D project is the Principal Investigator (PI), represented by Pr. Christophe Eloy. The DMR is responsible for the respective local storage of the raw data and for the preparation of the metadata.

1.3 Data summary

Each data collection generated during the C0PEP0D project will be listed in Annex II table. Should this data collection be opened to public, a dedicated dataset document will be completed following the template in Annex III.

1.4 Data nature, link with previous data and potential users

In the previous section “1.3 Data summary”, the C0PEP0D DMR lists the project’s data/results generated during the project and identifies which data will be opened. He also describes the link with previous data and potential users.
The basic rule is based on the fact that only Data needed to validate the results presented in scientific publications can be made accessible to third parties.
Research data linked to exploitable results, if any, will not be put into the open domain if they compromise its commercialisation prospects or have inadequate protection, which is a H2020 obligation.

2. FAIR Data

2.1 Making data findable

When a collection of data is ready to be published publicly, this data set will be archived on the Zenodo searchable data repository together with pertinent keywords and identified by means of a Digital Object Identifier, provided by Zenodo and linked to the published paper.
The source code of the virtual environment will be made freely available on an open-source platform (GitHub).
As part of the attached documentation, the file naming convention will be specified on a case-by-case basis. In case of successive versions of a given dataset, version numbers will be used. Where relevant, the databases will be linked to metadata such as movies or sound recordings.

2.2 Making data openly accessible

By default, all scientific publications will be made publicly available with due respect of the Green / Gold access regulations applied by each scientific publisher.
Should a publisher imposes an embargo longer than 6-months under the Green access, and in order to comply with Art. 29.2 of the ERC Grant Agreement provisions, a Gold access will be selected.
All scientific publications will be made freely accessible through the project web site and the open access online repositories ArXiv/BioRxiv and HAL. The databases that will be selected to constitute the project validation benchmarks will be archived on the Zenodo platform, and linked from the C0PEP0D project website. Ascii-readable file formats will be preferred for small datasets, and binary encoding will be implemented for large datasets, using freely available standard formats (e.g. the CFD Generic Notation System) for which the source and compiled import libraries are freely accessible. In the latter case, the structure of the binary records (headers) will be documented as part of the dataset. The DMR will examine the suitability of the datasets produced by the project for public dissemination, as well as their proper archival and documentation.

2.3 Making data interoperable

The interoperability of the published datasets will be enforced by the adoption of the RDA DMP Common Standard. This shared vocabulary will be adopted for the definition of the datasets, including variable names and units (See Annex III).

2.4 Increase data re-use

Data from public databases are open access and used a common creative licence CC-BY-NC-SA (Attribution-NonCommercial-ShareAlike 4.0 international)
With the impulsion of C0PEP0D project, the open access databases can be used by other laboratories and industrials to made comparison with other experimentations.
Methods developed and physical analysis become references to other test cases and improve the knowledge of the community.

3. Allocation of resources

Costs related to the open access and data strategy:

- Data storage in IRPHE data repositories: Included in structural operating cost.
- Data archiving with Zenodo or GitHub repositories: Free of charge.

DMR during the project:
The PI is responsible for the establishment, the updates during the lifetime of the project and the respect of the Data Management Plan. The relevant experimental data and the generated data from numerical simulations during the C0PEP0D project will be made available to the Team Members within the frame of the IPR protection principles and the present Data Management Plan.

4. Data security

Long-term preservation: Using Zenodo and GitHub data repositories.
Data Transfer: Using Zenodo and GitHub web platforms
Intellectual property: All data set contains are attached to a common creative licence.

5. Ethical aspects

The data generated by the C0PEP0D project is not subject to ethical issues.

6. Other

No other procedure for data management.

Annex I

Listing of datasets generated by the C0PEP0D project.
The grey parts will be filled when they are known.

Dataset Name	Nature of the Data	Format	Size	Purpose	Confidentiality level	dataset_id
1. sheld0n	software	C++ and Python	26 MB	A code that enables complex active particle advection in flows.	Public	10.5281/zenodo.6420863
2. otto	software and deep learning models	Python and tensorflow models	104 MB	Python package to learn, evaluate and visualize strategies for odor-based searches.	Public	10.5281/zenodo.6420861
3. dataset for otto	dataset	csv	69 kB	Dataset containing the results presented in "Searching for a source without gradients: how good is infotaxis and how to beat it".	Public	10.5281/zenodo.6125392
4. XX	Nature	Format	X GB	Purpose	Public / Confidential	dataset_id

- **Nature of the data**: experimental data, numerical data, documentation, software code, hardware, etc.
- **Format**: can be .pdf / .csv / .txt, etc.
- **Size**: expected size of the data
- **Purpose / objective**: purpose of the dataset and its relation to the objectives of the project.
- **Confidentiality level**:

  * “Public”: when data needed for the verification of results published in scientific journals can be made accessible to third parties.

  * “Confidential”: when data associated with results may have potential for commercial or industrial protection and thus will not be made accessible to a third party

Annex II

Table linking C0PEP0D project scientific publications with open datasets.

The grey parts will be filled when they are known.

Publication Title	**arXiv ID**	**dataset ID**
How memory architecture affects learning in a simple POMDP: the two-hypothesis testing problem	2106.08849	DOI
Surfing on turbulence	2110.10409	10.5281/zenodo.6420863
Unsteady and inertial dynamics of an active particle in a fluid	2105.01408	DOI
Searching for a source without gradients: how good is infotaxis and how to beat it	2105.01408	10.5281/zenodo.6420861 10.5281/zenodo.6125392

Annex III

I) In order to help the identification of the datasets opened to public as part of the C0PEP0D project, those latters will be linked to the following Metadata :

Metadata type	Name	Type	Data
DMP	title	String	C0PEP0D Data Management Plan
DMP	language	String	English
DMP	dmp_id	TypedIdentifier	10.5281/zenodo.3497406
DMP	created	Date	19/09/2019
DMP	modified	Date	
DMP	Ethical_Issues_exist	Boolean	No
Project	title :	String	C0PEP0D
Project	description :	String	The objective of the C0PEP0D project is to decipher how planktonic copepods exploit hydrodynamic and chemical sensing to detect and track targets in turbulent flows
Project	projectStart :	Date	01/09/2019
Project	projectEnd :	Date	31/08/2024
Funding	grantID	TypedIdentifier	834238
Funding	funderID	String	ERC-2018-ADG
Contact	name	String	Christophe Eloy
Contact	mbox	String	christophe.eloy@centrale-marseille.fr
Contact	contact_id	TypedIdentifier	0000-0003-4114-7263

II) The dataset template below will be completed for each dataset that will be open to public, once the data are generated.

The grey parts will be filled when they are known.

dataset type	Name	Type	Data
Dataset	title	String	sheld0n
Dataset	description	String	A code that enables complex active particle advection in flows.
Dataset	type	<dict>	software
Dataset	keyword	String	active particles, turbulence, plankton
Dataset	dataset_id	TypedIdentifier	10.5281/zenodo.6420863
Dataset	personal_data	Boolean	no
Dataset	sensitive_data	Boolean	no
Distribution	title	String	Surfing on turbulence
Distribution	description	String	Many planktonic organisms are motile and perceive their environment with flow sensors. Can they use flow sensing to travel faster in turbulence? To address this question, we consider plankters swimming at constant speed, whose goal is to move upwards. We propose an analytical behavior that allows plankters to choose a swimming direction according to the local flow gradients. We show numerically that such plankters can "surf" on turbulence and reach net vertical speeds up to twice their swimming speed. This physics-based model suggests that planktonic organisms can exploit turbulence features for navigation.
Distribution	format		.pdf
Distribution	access_url	url	https://doi.org/10.48550/arXiv.2110.1040
Distribution	data_access	<open/closed/shared>	open
Distribution	available_till	Date	n/a
Host	title	String	Zenodo
Licence	title	String	MIT
Licence	license_ref	url	https://github.com/C0PEP0D/sheld0n/blob/master/LICENSE
Licence	start_date	Date	2 nov 2021

dataset type	Name	Type	Data
Dataset	title	String	otto
Dataset	description	String	Python package to learn, evaluate and visualize strategies for odor-based searches.
Dataset	type	<dict>	software
Dataset	keyword	String	Infotaxis, odor tracking, deep reinforcement learning
Dataset	dataset_id	TypedIdentifier	10.5281/zenodo.6420861
Dataset	personal_data	Boolean	no
Dataset	sensitive_data	Boolean	no
Distribution	title	String	**Searching for a source without gradients: how good is infotaxis and how to beat it**
Distribution	description	String	Infotaxis is a popular search algorithm designed to track a source of odor in a turbulent environment using information provided by odor detections. To exemplify its capabilities, the source-tracking task was framed as a partially observable Markov decision process consisting in finding, as fast as possible, a stationary target hidden in a 2D grid using stochastic partial observations of the target location. Here we provide an extended review of infotaxis, together with a toolkit for devising better strategies. We first characterize the performance of infotaxis in domains from 1D to 4D. Our results show that, while being suboptimal, infotaxis is reliable (the probability of not reaching the source approaches zero), efficient (the mean search time scales as expected for the optimal strategy), and safe (the tail of the distribution of search times decays faster than any power law, though subexponentially). We then present three possible ways of beating infotaxis, all inspired by methods used in artificial intelligence: tree search, heuristic approximation of the value function, and deep reinforcement learning. The latter is able to find, without any prior human knowledge, the (near) optimal strategy. Altogether, our results provide evidence that the margin of improvement of infotaxis toward the optimal strategy gets smaller as the dimensionality increases.
Distribution	format		pdf
Distribution	access_url	url	https://doi.org/10.48550/arXiv.2112.10861
Distribution	data_access	<open/closed/shared>	open
Distribution	available_till	Date	n/a
Host	title	String	Zenodo
Licence	title	String	MIT
Licence	license_ref	url	https://github.com/C0PEP0D/otto/blob/main/LICENSE
Licence	start_date	Date	7 April 2022

dataset type	Name	Type	Data

Dataset	title	String	Dataset for otto
Dataset	description	String	Dataset containing the results presented in "Searching for a source without gradients: how good is infotaxis and how to beat it".
Dataset	type	<dict>	dataset
Dataset	keyword	String	Natural Laminar Flow, Reduced Friction Drag
Dataset	dataset_id	TypedIdentifier	10.5281/zenodo.6125392
Dataset	personal_data	Boolean	no
Dataset	sensitive_data	Boolean	no

Distribution	title	String	**Searching for a source without gradients: how good is infotaxis and how to beat it**
Distribution	description	String	Infotaxis is a popular search algorithm designed to track a source of odor in a turbulent environment using information provided by odor detections. To exemplify its capabilities, the source-tracking task was framed as a partially observable Markov decision process consisting in finding, as fast as possible, a stationary target hidden in a 2D grid using stochastic partial observations of the target location. Here we provide an extended review of infotaxis, together with a toolkit for devising better strategies. We first characterize the performance of infotaxis in domains from 1D to 4D. Our results show that, while being suboptimal, infotaxis is reliable (the probability of not reaching the source approaches zero), efficient (the mean search time scales as expected for the optimal strategy), and safe (the tail of the distribution of search times decays faster than any power law, though subexponentially). We then present three possible ways of beating infotaxis, all inspired by methods used in artificial intelligence: tree search, heuristic approximation of the value function, and deep reinforcement learning. The latter is able to find, without any prior human knowledge, the (near) optimal strategy. Altogether, our results provide evidence that the margin of improvement of infotaxis toward the optimal strategy gets smaller as the dimensionality increases.
Distribution	format		pdf
Distribution	access_url	url	https://doi.org/10.48550/arXiv.2112.10861
Distribution	data_access	<open/closed/shared>	open
Distribution	available_till	Date	n/a

Host	title	String	Zenodo

Licence	title	String	CC Attribution 4.0 International
Licence	license_ref	url	https://creativecommons.org

Licence	start_date	Date	17 Feb. 2022

dataset type	Name	Type	Data
Dataset	title	String	Title of the dataset
Dataset	description	String	Description of the dataset
Dataset	type	<dict>	text, numbers, images, 3Dmodels, software, audio files, video files, reports, …
Dataset	keyword	String	Natural Laminar Flow, Reduced Friction Drag
Dataset	dataset_id	TypedIdentifier	DOI number provided by Zenodo
Dataset	personal_data	Boolean	no
Dataset	sensitive_data	Boolean	yes/no
Distribution	title	String	Title of the publication
Distribution	description	String	Executive summary of the publication
Distribution	format		• .txt • .ascii, .xlsx, • .jpg,.tif, .ps, .png • .DWG, .DXF, .DGN, .STL ,.3DS, • .f, .c, • .WAV, .AIF, .MP3, .MID • .MPG, .MOV, .WMV, .RM • .doc, .eps, .ps, .pdf
Distribution	access_url	url	url
Distribution	data_access	<open/closed/shared>	open / closed / shared
Distribution	available_till	Date	End date
Host	title	String	Zenodo
Licence	title	String	CC BY SA NC
Licence	license_ref	url	https://creativecommons.org
Licence	start_date	Date	Date of publication