Plan Overview

A Data Management Plan created using DMPonline.be

Title: WINGS.AI - Identification of flying insects through advanced acoustic design and artificial intelligence

Creator: Astrid Tempelaere

Affiliation: KU Leuven (KUL)

Funder: Fonds voor Wetenschappelijk Onderzoek - Research Foundation Flanders (FWO)

Template: FWO DMP (Flemish Standard DMP)

Project abstract:

In agriculture, insects are pivotal, with beneficial ones aiding pollination and pest insects harming crop yields. Balancing pest control while preserving beneficial insects is a challenge due to environmental concerns and regulations like the European Green Deal. Traditional pest monitoring is slow, costly, and limited in scope. Recent studies propose image analysis for larger insects, but smaller ones pose challenges. As an alternative, wingbeat analysis shows promise but faces practical obstacles. To address this, a holistic approach is needed for effective, environment-friendly pest control. This involves advancing technology for precise, real-time insect monitoring, considering various insect sizes and attraction rates, and meeting regulatory demands.

The primary objective of this project is to merge advanced acoustics with Deep Learning to establish a foundation for flying insect identification. Our research will concentrate on enhancing the signal-to-noise ratio of insect wingbeat signals in noisy environments and leveraging advanced Deep Learning techniques for signal analysis.

In Work Package 1, insect wingbeat signals will be captured in optimal conditions (anechoic chamber) using high-end microphones. This allows for accurately quantifying insect wingbeat behavior. In Work Package 2, the impact of noise sources on wingbeat signal measurements will be investigated under laboratory conditions and used for advanced acoustic design of a trap. Deep Learning will be investigated as an alternative for active noise canceling (WP3) and for efficient wingbeat classification (WP4). In WP5, a validation study will be carried out in collaboration with our user committee. Finally, management aspects of WINGS.AI together with the valorization and communication of results will constitute WP6.

ID: 211082

Start date: 01-10-2024

End date: 30-09-2028

Last modified: 16-01-2025

WINGS.AI - Identification of flying insects through advanced acoustic design and artificial intelligence Application DMP

Questionnaire

Describe the datatypes (surveys, sequences, manuscripts, objects …) the research will collect and/or generate and /or (re)use. (use up to 700 characters)

- optical wingbeat signals
- Acoustic wingbeat signals
- Acoustic environmental background signals
- DL models

Specify in which way the following provisions are in place in order to preserve the data during and at least 5 years after the end of the research? Motivate your answer. (use up to 700 characters)

1. Designation of responsible person: Niels Wouters
2. Storage capacity/repository
    - during the research: backups of MangGO, shared network storage, OneDrive folder
    - after the research: K-drive LVS (large volume storage)

What's the reason why you wish to deviate from the principle of preservation of data and of the minimum preservation term of 5 years? (max. 700 characters)

All data will be preserved for 10 years according to KU Leuven RDM policy.

Are there issues concerning research data indicated in the ethics questionnaire of this application form? Which specific security measures do those data require? (use up to 700 characters)

NA

Which other issues related to the data management are relevant to mention? (use up to 700 characters)

NA

WINGS.AI - Identification of flying insects through advanced acoustic design and artificial intelligence FWO DMP (Flemish Standard DMP)

1. Research Data Summary

List and describe all datasets or research materials that you plan to generate/collect or reuse during your research project. For each dataset or data type (observational, experimental etc.), provide a short name & description (sufficient for yourself to know what data it is about), indicate whether the data are newly generated/collected or reused, digital or physical, also indicate the type of the data (the kind of content), its technical format (file extension), and an estimate of the upper limit of the volume of the data.

				Only for digital data	Only for digital data	Only for digital data	Only for physical data
Dataset Name	Description	New or reused	Digital or Physical	Digital Data Type	Digital Data format	Digital data volume (MB/GB/TB)	Physical volume
		Please choose from the following options: • Generate New data • Reuse Existing data	*Please choose from the following options:* • Digital • Physical	*Please choose from the following options:* • Observational • Experimental • Compiled/aggregated data • Simulation data • Software • Other • NA	*Please choose from the following options:* • .por, .xml, .tab, .csv,.pdf, .txt, .rtf, .dwg, .gml, … • NA	*Please choose from the following options:* • <100MB • <1GB • <100GB • <1TB • <5TB • <10TB • <50TB • >50TB • NA	
Optical wingbeat signals		E	D	E	.wav	<100GB	
Acoustic wingbeat signals		N	D	E	.wav	<1TB	
Acoustic environmental background signals		N	D	E	.wav	<100 GB	
DL models		N	D	M	.pth	<100GB	

If you reuse existing data, please specify the source, preferably by using a persistent identifier (e.g. DOI, Handle, URL etc.) per dataset or data type:

Data will be collected in the project and combined with public datasets, for instance:
Mosquito Wingbeat Recordings: https://www.kaggle.com/datasets/potamitis/wingbeats (DOI: 10.1016/j.compag.2020.105849)

Are there any ethical issues concerning the creation and/or use of the data (e.g. experiments on humans or animals, dual use)? Describe these issues in the comment section. Please refer to specific datasets or data types when appropriate.

- No

Will you process personal data? If so, briefly describe the kind of personal data you will use in the comment section. Please refer

to specific datasets or data types when appropriate.

- No

Does your work have potential for commercial valorization (e.g. tech transfer, for example spin-offs, commercial exploitation, ...)? If so, please comment per dataset or data type where appropriate.

- Yes

All the new data stated on the data list will be potentially used as a basis for establishing detection model for enhancing integrated pest management in the agrifood industry. Hardware and software companies developing microphone sensors, monitoring tools, or high-end AI solutions could benefit from all the data generated in this project.

Do existing 3rd party agreements restrict exploitation or dissemination of the data you (re)use (e.g. Material/Data transfer agreements/ research collaboration agreements)? If so, please explain in the comment section to what data they relate and what restrictions are in place.

- No

Are there any other legal issues, such as intellectual property rights and ownership, to be managed related to the data you (re)use? If so, please explain in the comment section to what data they relate and which restrictions will be asserted.

- No

2. Documentation and Metadata

Clearly describe what approach will be followed to capture the accompanying information necessary to keep data understandable and usable, for yourself and others, now and in the future (e.g., in terms of documentation levels and types required, procedures used, Electronic Lab Notebooks, README.txt files, Codebook.tsv etc. where this information is recorded).

During WINGSAI, large optic and acoustic wingbeat signals will be generated, along with reference labels for each object. Data will come in following size and formats:

- Optics data (WAV format), (10-20 kB each). Total expected size ~50 GB
- Acoustics data (WAV format), (600-1000 kB each). Total expected size ~1TB

Data processing toolboxes will be written to be compatible with Python (building on backends in C++, Matlab, etc.). Presentation files and manuscripts (reports, conference papers, and journal articles) will be produced using MS Office and LaTeX.
A dedicated storage share is available on which each PhD student stored data files, with corresponding metadata in a cloud-based data management system. For the large volume of original optics and acoustics data during and minimally 5 years after the project, we will rent large volume storage of KU Leuven. Costs will be covered by the project consumables budget. For daily work, a separate folder share is provided for the research with automatic back-up. Work-related files of the researcher will be on a OneDrive cloud folder (and GitLab cloud for code) that is daily backed up to a KU Leuven file server, where it is shared with promotors. After conclusion of the PhD, files are transferred to a KU Leuven archive. For the active data management during the project, we plan to use the new ManGO platform, based on iRODS (open-source software), provided by KU Leuven-ICTS.

Will a metadata standard be used to make it easier to find and reuse the data? If so, please specify (where appropriate per dataset or data type) which metadata standard will be used. If not, please specify (where appropriate per dataset or data type) which metadata will be created to make the data easier to find and reuse.

- Yes

Metadata will be generated following the standards used in 'ManGO', the new KU Leuven system for data management. Metadata schemas will be created in collaboration with project partners, to make data reusable and findable within the ManGO platform.

3. Data storage & back-up during the research project

Where will the data be stored?

- ManGO
- Sharepoint (Teams folder)

The data are stored on the Sharepoint (Teams folder) and on ManGO (iRODS system provided by the Research Data Management and ICTS teams of KU Leuven).

How will the data be backed up?

Standard back-up provided by KU Leuven ICTS for my storage solution

Is there currently sufficient storage & backup capacity during the project? If yes, specify concisely.
If no or insufficient storage or backup capacities are available, then explain how this will be taken care of.

- Yes

If needed, more storage is provided by KU Leuven for our ManGO upon request.

How will you ensure that the data are securely stored and not accessed or modified by unauthorized persons?

The data is kept in the KU Leuven Sharepoint (Teams folder), with access restricted to the account holders and granted to the IT department only in specific situations. As for ManGO, only a select group of users are allowed to view or change the data, but they must first be added as members of the group by one of the ManGO managers before they can do so. This is done by sending the users an invitation to join the group.

What are the expected costs for data storage and backup during the research project? How will these costs be covered?

For ManGO, there is a fee of €35 per terabyte annually. It's anticipated that approximately 10 terabytes (TB) will be generated throughout this project, amounting to an annual storage expense of €350. These costs will be funded by the project funds of the associated partners participating in the project.

4. Data preservation after the end of the research project

Which data will be retained for at least five years (or longer, in agreement with other retention policies that are applicable) after the end of the project? In case some data cannot be preserved, clearly state the reasons for this (e.g. legal or contractual restrictions, storage/budget issues, institutional policies...).

All data will be preserved for 10 years according to KU Leuven RDM policy

Where will these data be archived (stored and curated for the long-term)?

- Other (specify below)
- Large Volume Storage (longterm for large volumes)

The data will be stored on K-drive or LVS (both at € 100,86 / TB / year).

What are the expected costs for data preservation during the expected retention period? How will these costs be covered?

For ManGO, there is a fee of €35 per terabyte annually. It's anticipated that approximately 10 terabytes (TB) will be generated throughout this
project, amounting to an annual storage expense of €350 or a total of €3,500 for a retention period of 10 years. These costs will be funded by
the project funds of the associated partners participating in the project.

5. Data sharing and reuse

Will the data (or part of the data) be made available for reuse after/during the project? In the comment section please explain per dataset or data type which data will be made available.

- Yes, in a restricted access repository (after approval, institutional access only, …)

It is our intention to make the publicly available for reuse some time after completion of the project, once they have been properly valorized.

If access is restricted, please specify who will be able to access the data and under what conditions.

During the embargo period, the data will only be accessible for members of the research group involved in the project.
Afterwards, the data will be made available to other researchers for non-commercial use.

Are there any factors that restrict or prevent the sharing of (some of) the data (e.g. as defined in an agreement with a 3rd party, legal restrictions)? Please explain in the comment section per dataset or data type where appropriate.

- Yes, Intellectual Property Rights

The access may be restricted for some time (embargo) to protect the KU Leuven IP rights.

Where will the data be made available? If already known, please provide a repository per dataset or data type.

No decision has been made with respect to this.

When will the data be made available?

- Other (specify below)
- Upon publication of research results

Data that is used in potential publications will be available upon publication of research results.

Data that might be valorized will not be available until they are properly valorized.

Which data usage licenses are you going to provide? If none, please explain why.

- CC-BY 4.0 (data)
- Data Transfer Agreement (restricted data)
- MIT licence (code)
- GNU GPL-3.0 (code)

Not decided yet. For code we consider both the MIT licence or the GNU GPL-3.0 license. During the embargo period, we will have to work
with Data Transfer Agreements. Afterwards, we could move to CC-BY 4.0.

Do you intend to add a PID/DOI/accession number to your dataset(s)? If already available, you have the option to provide it in the comment section.

- Yes

Yes, a PID will be added upon deposit in a data repository

What are the expected costs for data sharing? How will these costs be covered?

Sharing data without incurring extra costs is expected as the data will be shared directly from their host locations. Expenses related to sharing will be assumed by the relevant research groups or the chosen repository, depending on the specific sharing arrangement selected.

6. Responsibilities

Who will manage data documentation and metadata during the research project?

The responsibility for managing data documentation and metadata falls upon the PhD researchers working on the project, with guidance from their respective promoters and data management representatives: Niels Wouters at MeBioS.

Who will manage data storage and backup during the research project?

The responsibility for managing data storage and backup falls upon the PhD researchers working on the project, with guidance from their respective promoters and data management representatives: Niels Wouters at MeBioS. The backups of ManGO, the shared network storage and the settings of the OneDrive folder are managed by SET-IT and RDM teams of KU Leuven.

Who will manage data preservation and sharing?

The responsibility for managing data preservation and sharing falls upon the PhD researchers working on the project, with guidance from their respective promoters and data management representatives: Niels Wouters at MeBioS.

Who will update and implement this DMP?

The responsibility for updating and implementing this DMP falls upon the data management representatives: Astrid Tempelaere at MeBioS.