HEARLIGHT

Deliverable D6.2
Data Management Plan

Due date of deliverable: M6
Actual submission date: M6

Start date of the project: 1st April, 2021 **Duration:** 48 months
Lead organisation name: *Institut Pasteur*
Revision: V1

Dissemination level	
Public - PU	**x**
Confidential, only for members of the consortium (including Commission Services) - CO	
Classified, as referred to in Commission Decision 2001/844/EC - CI	

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 964568

Data Management Plan – Hearlight

The Data Management Plan describes the data management life cycle for all datasets that will be collected, processed or generated in the project. It is a document describing what data will be collected, processed or generated and following what methodology and standards, whether and how this data will be shared and/or made open, and how it will be curated and preserved.

1.	**Information on the Data Management Plan**	
1.1.	Author(s) of the DMP	Brice Bathellier, brice.bathellier@pasteur.fr, Katharina Kopf, katharina.kopf@pasteur.fr
1.2.	Date of the first version of the DMP	October, 2021
1.3.	Current version of the DMP and date	V1, October, 2021
1.4.	Location of storage of the DMP	Initial and intermediate versions will be stored on an internal server and published on the project website. The final version will be stored on a data repository (to be determined).
2.	**Information on the project**	
2.1.	Name of the project funder and funding programme	The project is funded by the H2020 FET Open program
2.2.	Acronym of the project	HEARLIGHT
2.3.	Title of the project	Towards optogenetic cortical implants for the hearing impaired
2.4.	Project ID	964568
2.5.	Coordinator of the project	Brice Bathellier, brice.bathellier@pasteur.fr
2.6.	Organization and unit of the coordinator	Institut Pasteur
2.7.	Start date of the project	01/04/2021
2.8.	End date of the project	31/03/2025
3.	**Overview of the data**	
3.1.	What is the purpose of the data collection/generation?	We will acquire: 1/ neurometric measurements (electrophysiology and two-photon calcium imaging) in the mouse auditory cortex during acoustic and cortical implant stimulation 2/ psychometric measurements of sound discrimination in mice 3/ characterisation data for the stimulators developed in the projects
3.2.	How many dataset(s) will you generate during this project?	3 datasets
3.3.	What is the nature and format of generated/collected data?	Generated data will consist in: -.tiff for raw imaging data - binary file for electrophysiology and behavioral measurements - text (CSV), Matlab and Python files for pre-processed data - Excel sheets for characterization data
3.4.	Give the expected volume of generated data for this project	6.5 TB
3.5.	Will you also reuse existing data? If yes, specify their origin.	No old data will be used in this project. All deliverables from the project will come from original data acquired during this period.

4.	**Resources needed for data management**	
4.1.	What hardware resources do you need to manage your data?	Hardware resources such as external hard-drives and internal server space have been installed to ensure several safe back-ups.
4.2.	Who is in charge of data management during the research project?	The principal investigator of each task is responsible for: - data collection, processing and analysis - the generation of the metadata and documentation related to the data - data storage - data archiving and sharing
4.3.	What training or support do you think is necessary to help you manage your data?	No special training will be needed to curate and manage the data generated in this project.
4.4.	What budget do you have for managing your data? How do you intend to cover these costs?	The host institute and host lab have a dedicated budget to provide storage and maintenance of the data internally.
5.	**Legal and ethical aspects**	
5.1.	Does your project include personal data?	No
5.2.	Does your project include other data subject to a contractual, regulatory or legal obligation? If so, what type?	A patent application will be considered during the project. Some of the data cannot be disseminated before the patent application is filed.
6.	**Data management during the project**	
6.1.	What is the storage location of your data during the project?	Data are stored on a shared storage space provided by the partner Institutions. In addition, data are also stored in external hard drives as a backup.
6.2.	Do you use a file classification scheme to manage your data files? Briefly indicate how it is organized.	Yes, each unit created a file classification scheme at the beginning of the project. It is organized by data collection method (microscopy, electrophysiology, behavior, electronic device validation...) and then chronologically. Raw data and processed data are stored in different folders.
6.3.	What naming conventions do you use for your data? What rules do you use for clear versioning?	Processed and analyzed files will have the same suffix. Text documents follow date and version as well as authors initials throughout revision.
6.4.	What measures are in place to ensure the quality of the data?	In order to guarantee the quality of the data, several measures are implemented: - Independent repetition of experiments on biological (minimum of three repetitions on three different animals) - Standardization of data collection (all animals raised under the same conditions, temperature control, same stimulation conditions) - Regular review of data with PI

7.	**Data selection and long term preservation**	
7.1.	Are your data subject to preservation regulations? If yes, which ones?	No specific regulatory constraints
7.2.	Which datasets are of long-term value and should be preserved? What are the datasets to destroy?	All datasets will be preserved.
7.3.	On which platform or in which repository will the datasets be archived? Is this platform certified for long-term preservation and management?	After publication project, dataset 1 and 2 will be transferred to the ZENODO repository, which ensures sustainable archiving of the final research data. Dataset 3 may contain sensitive data. It cannot be made available on an external repository. It will therefore be preserved on Institutions' servers.
7.4.	Specify the formats chosen for archiving.	MAT, PDF, TIFF, PNG and CSV format files will be used as well as non-proprietary formats for electrophysiology and metadata.
7.5.	How long will the data be preserved?	Published datasets will be retained for the maximum duration allowed by the repository. For Zenodo, this corresponds to the lifetime of the host laboratory CERN, which currently has an experimental programme defined for the next 20 years at least. The raw data will be kept for an unlimited period of time as long as the space allocated within the partner institution is available.
7.6.	What is the expected volume of archived data?	8 Tb
7.7.	If a long-term preservation is needed, how do you intend to cover these costs?	The costs of long-term preservation will be covered by the Institut Pasteur and the other institutions.
8.	**Dataset #1**	
8.1.	**Data description**	
8.1.1.	ID and name of the dataset	1 Neurometric
8.1.2.	Who is the provider or producer of the data?	This dataset is generated by UNIBAS and IP.
8.1.3.	What are the nature and format of the data in this dataset?	Pictures in PNG and TIFF format. Movies of the imaging data in compressed .tiff Binary files for raw electrophysiological data. Matlab and Python files for imaging data.
8.1.4.	Describe in more detail the data in this dataset	This dataset includes movies or multiple 1D time series of brain activity during stimulation.
8.1.5.	Describe the method of data collection and/or generation	Data are generated by two-photon microscopy and multichannel electrophysiology and analysed by the Suite2P and KiloSort softwares. Matlab and Python will be used for data management, graph representations and data analysis.
8.1.6.	Describe your dataset with keywords	Two-photon microscopy, multichannel electrophysiology, auditory cortex, optogenetic, electrical stimulation
8.1.7.	Indicate the URL or the persistent identifier to access your dataset	TAB
8.1.8.	What is the expected volume of data in this dataset?	4 Tb

8.2.	**Making data openly accessible**	
8.2.1.	Will this dataset be freely accessible?	Pre-processed data will be made freely available at the time of pre-print publication. Raw data will be available upon request.
8.2.2.	Which repository did you chose to store the data of your dataset and make them accessible?	Data repository to be determined (Zenodo)
8.2.3.	Will this dataset be the subject of a patent application? If yes, this dataset has to be kept confidential.	No
8.2.4.	If this dataset has to be kept closed for other reasons, explain why.	No
8.2.5.	Specify how access to this dataset will be provided in case of restriction	The dataset is deposited on Zenodo but with restricted access. The access conditions are specified on Zenodo: the person who wishes to download the dataset must first explain how he/she intends to use it. Based on this justification, a decision will be made to grant or deny access.
8.2.6.	What software is necessary to read or access the data? Do you provide the documentation or the open source code of the software?	Matlab or Python
8.3.	**Making data findable**	
8.3.1.	Is this dataset identified by a persistent and unique identifier such as DOI (Digital Object Identifiers)? If not, describe how data and this dataset are identified.	Overall compiled into a dataset folder with generated DOI from the data repository.
8.3.2.	Which metadata standards do you use? If you don't use metadata standards, outline what type(s) of metadata will be created and how.	Metadata will follow the metadata standard or the guidelines of the repository chosen to store the data.
8.3.3.	Is this dataset described by keywords in order to make it easily findable?	Yes, this dataset will be described by keywords
8.3.4.	Do you provide a supplementary documentation in order to describe more precisely your data?	Yes, a file is available for each data to sum up the analyses that was performed.
8.4.	**Making data interoperable**	
8.4.1.	Are the data of this dataset technically interoperable?	Yes, we use only standard formats or Matlab and Python formats that are interoperable.
8.4.2.	If not, what methodologies will you apply to make your data interoperable?	N/A
8.4.3.	Specify whether you will be using standard vocabulary for all data types of your dataset, to allow inter-disciplinary interoperability. If not, will you provide mapping to more commonly used ontologies?	We do not use a specific ontology but we use a uniform vocabulary throughout the dataset. The vocabulary is based on commonly used terms in life sciences, specific terms are specified in the documentation associated with the dataset.

8.5.	**Increase data reuse**	
8.5.1.	At the end of the project, can the data of this dataset be reused by third parties? If reuse is restricted, explain why.	Once published, this dataset may be reused by the scientific community, with the restriction that the data may not be used for commercial purposes (a CC-BY-NC license will be associated with the dataset).
8.5.2.	What license will be assigned to your dataset to permit the widest reuse possible?	CC-BY-NC license (https://creativecommons.org/licenses/by-nc/4.0/)
8.5.3.	When will the dataset be available for reuse? If applicable, specify why and for what period an embargo is needed.	Data will be available for reuse immediately after manuscript publication.
8.5.4.	Specify how long the dataset will remain reusable	Data will be stored during the maximum allowed time by the Institut Pasteur storage allowance capacity and the repository chosen to publish the data. Data will also be stored in external hard disks in the laboratory.
8.6.	**Data security**	
8.6.1.	Does this dataset have to remain confidential during your project? If so, can you specify to whom it can be made accessible?	Until publication, raw data and dataset will be accessible only to lab members.
8.6.2.	During the project (before storage of data in a repository), is the dataset safely stored?	During the research project, data will be stored in secure internal Pasteur or Uni Basel servers. In addition, data will be also stored in external hard-drives or computer hard drives as a second backup.
8.6.3.	Has the data repository chosen to store the dataset after the project implemented a security policy regarding its information system?	As this dataset is not confidential, it does not need to be protected in a secure repository It will be kept on a public repository after the end of the project.
8.6.4.	What security measures are in place for data collection and exchange?	Before publication, data is shared using a secure software. (e.g. Syncplicity)

9.	**Dataset #2**	
9.1.	**Data description**	
9.1.1.	ID and name of the dataset	2 Psychometry
9.1.2.	Who is the provider or producer of the data?	This dataset is generated by UNIBAS and IP.
9.1.3.	What are the nature and format of the data in this dataset?	Binary files for raw data. Movies of behaving mice. Matlab and Python files for preprocessed data.
9.1.4.	Describe in more detail the data in this dataset	This dataset includes behavioural responses and parameters.
9.1.5.	Describe the method of data collection and/or generation	Data are generated by behavioural tasks in animals run by costum-made setups and softwares. Matlab and Python will be used for data management, graph representations and data analysis.
9.1.6.	Describe your dataset with keywords	Behavior, auditory cortex, optogenetic, electrical stimulation

9.1.7.	Indicate the URL or the persistent identifier to access your dataset	TAB
9.1.8.	What is the expected volume of data in this dataset?	2 Tb
9.2.	**Making data openly accessible**	
9.2.1.	Will this dataset be freely accessible?	Pre-processed data will be made freely available at the time of pre-print publication. Raw data will be available upon request.
9.2.2.	Which repository did you chose to store the data of your dataset and make them accessible?	Data repository to be determined (Zenodo)
9.2.3.	Will this dataset be the subject of a patent application? If yes, this dataset has to be kept confidential.	No
9.2.4.	If this dataset has to be kept closed for other reasons, explain why.	No
9.2.5.	Specify how access to this dataset will be provided in case of restriction	The dataset is deposited on Zenodo but with restricted access. The access conditions are specified on Zenodo: the person who wishes to download the dataset must first explain how he/she intends to use it. Based on this justification, a decision will be made to grant or deny access.
9.2.6.	What software is necessary to read or access the data? Do you provide the documentation or the open source code of the software?	Matlab or Python
9.3.	**Making data findable**	
9.3.1.	Is this dataset identified by a persistent and unique identifier such as DOI (Digital Object Identifiers)? If not, describe how data and this dataset are identified.	Overall compiled into a dataset folder with generated DOI from the data repository.
9.3.2.	Which metadata standards do you use? If you don't use metadata standards, outline what type(s) of metadata will be created and how.	Metadata will follow the metadata standard or the guidelines of the repository chosen to store the data.
9.3.3.	Is this dataset described by keywords in order to make it easily findable?	Yes, this dataset will be described by keywords.
9.3.4.	Do you provide a supplementary documentation in order to describe more precisely your data?	Yes, a file is available for each data to sum up the analyses that was performed.
9.4.	**Making data interoperable**	
9.4.1.	Are the data of this dataset technically interoperable?	Yes, we use only standard formats or Matlab and Python formats that are interoperable.
9.4.2.	If not, what methodologies will you apply to make your data interoperable?	N/A

9.4.3.	Specify whether you will be using standard vocabulary for all data types of your dataset, to allow inter-disciplinary interoperability. If not, will you provide mapping to more commonly used ontologies?	We do not use a specific ontology but we use a uniform vocabulary throughout the dataset. The vocabulary is based on commonly used terms in life sciences, specific terms are specified in the documentation associated with the dataset.
9.5.	**Increase data reuse**	
9.5.1.	At the end of the project, can the data of this dataset be reused by third parties? If reuse is restricted, explain why.	Once published, this dataset may be reused by the scientific community, with the restriction that the data may not be used for commercial purposes (a CC-BY-NC license will be associated with the dataset).
9.5.2.	What license will be assigned to your dataset to permit the widest reuse possible?	CC-BY-NC license (https://creativecommons.org/licenses/by-nc/4.0/)
9.5.3.	When will the dataset be available for reuse? If applicable, specify why and for what period an embargo is needed.	Data will be available for reuse immediately after manuscript publication.
9.5.4.	Specify how long the dataset will remain reusable	Data will be stored during the maximum allowed by the Institut Pasteur and University of Basel storage allowance capacity and the repository chosen to publish the data. Data will also be stored in external hard disks in the laboratory.
9.6.	**Data security**	
9.6.1.	Does this dataset have to remain confidential during your project? If so, can you specify to whom it can be made accessible?	Until publication, raw data and dataset will be accessible only to lab members.
9.6.2.	During the project (before storage of data in a repository), is the dataset safely stored?	During the research project, data will be stored in secure internal Pasteur and Uni Basel servers. In addition, data will be also stored in double external hard-drives, and my work computer under my user account.
9.6.3.	Has the data repository chosen to store the dataset after the project implemented a security policy regarding its information system?	As this dataset is not confidential, it does not need to be protected in a secure repository It will be kept on a public repository after the end of the project.
9.6.4.	What security measures are in place for data collection and exchange?	Before publication, data is shared using a secure software. (e.g. Syncplicity)

10.	**Dataset #3**	
10.1.	**Data description**	
10.1.1.	ID and name of the dataset	3, Device characterisation
10.1.2.	Who is the provider or producer of the data?	This dataset is generated by NTNU, UoS, Novagan, EMSE
10.1.3.	What are the nature and format of the data in this dataset?	Text (CSV), Matlab

10.1.4.	Describe in more detail the data in this dataset	Binary files for raw data. Spectral measurements, current-voltage measurements, stability data, e.g. intensity versus. time
10.1.5.	Describe the method of data collection and/or generation	Data are generated by custom-made measurement systems and software. Matlab will be used for data management, graph representations and data analysis.
10.1.6.	Describe your dataset with keywords	Impedance spectroscopy, photolithography, optical characterization, current-voltage characterization
10.1.7.	Indicate the URL or the persistent identifier to access your dataset	TAB
10.1.10.	What is the expected volume of data in this dataset?	2 Tb
10.2.	**Making data openly accessible**	
10.2.1.	Will this dataset be freely accessible?	Pre-processed data will be made freely available at the time of pre-print publication. Raw data will be available upon request.
10.2.2.	Which repository did you chose to store the data of your dataset and make them accessible?	Zenodo
10.2.3.	Will this dataset be the subject of a patent application? If yes, this dataset has to be kept confidential.	No
10.2.4.	If this dataset has to be kept closed for other reasons, explain why.	No
10.2.5.	Specify how access to this dataset will be provided in case of restriction	The dataset is deposited on Zenodo but with restricted access. The access conditions are specified on Zenodo: the person who wishes to download the dataset must first explain how he/she intends to use it. Based on this justification, a decision will be made to grant or deny access.
10.2.6.	What software is necessary to read or access the data? Do you provide the documentation or the open source code of the software?	Raw text files (CSV), Matlab or Python
10.3.	**Making data findable**	
10.3.1.	Is this dataset identified by a persistent and unique identifier such as DOI (Digital Object Identifiers)? If not, describe how data and this dataset are identified.	Overall compiled into a dataset folder with generated DOI from the data repository.
10.3.2.	Which metadata standards do you use? If you don't use metadata standards, outline what type(s) of metadata will be created and how.	Metadata will follow the metadata standard or the guidelines of the repository chosen to store the data.
10.3.3.	Is this dataset described by keywords in order to make it easily findable?	Yes, this dataset will be described by keywords
10.3.4.	Do you provide a supplementary documentation in order to describe more precisely your data?	Yes, a file is available for each data to sum up the analyses that was performed.

10.4.	**Making data interoperable**	
10.4.1.	Are the data of this dataset technically interoperable?	Yes, we use only standard text formats or Matlab and Python formats that are interoperable.
10.4.2.	If not, what methodologies will you apply to make your data interoperable?	N/A
10.4.3.	Specify whether you will be using standard vocabulary for all data types of your dataset, to allow inter-disciplinary interoperability. If not, will you provide mapping to more commonly used ontologies?	We do not use a specific ontology, but we use a uniform vocabulary throughout the dataset. The vocabulary is based on commonly used terms in engineering and life sciences, specific terms are specified in the documentation associated with the dataset.
10.5.	**Increase data reuse**	
10.5.1.	At the end of the project, can the data of this dataset be reused by third parties? If reuse is restricted, explain why.	Once published, this dataset may be reused by the scientific community, with the restriction that the data may not be used for commercial purposes (a CC-BY-NC license will be associated with the dataset).
10.5.2.	What license will be assigned to your dataset to permit the widest reuse possible?	CC-BY-NC license (https://creativecommons.org/licenses/by-nc/4.0/)
10.5.3.	When will the dataset be available for reuse? If applicable, specify why and for what period an embargo is needed.	Data will be available for reuse immediately after manuscript publication.
10.5.4.	Specify how long the dataset will remain reusable	Data will be stored during the maximum allowed by our storage allowance capacity and the repository chosen to publish the data. Data will also be stored in external hard disks in the laboratory.
10.6.	**Data security**	
10.6.1.	Does this dataset have to remain confidential during your project? If so, can you specify to whom it can be made accessible?	Until publication, raw data and dataset will be accessible only to lab members.
10.6.2.	During the project (before storage of data in a repository), is the dataset safely stored?	During the research project, data will be stored in secure internal servers. In addition, data will be also stored in double external hard-drives, and my work computer under my user account.
10.6.3.	Has the data repository chosen to store the dataset after the project implemented a security policy regarding its information system?	As this dataset is not confidential, it does not need to be protected in a secure repository It will be kept on a public repository after the end of the project.
10.6.4.	What security measures are in place for data collection and exchange?	Before publication, data is shared using a secure software. (e.g. Syncplicity)