Data Management Plan - NERF

Project supervisors (from application round 2018 onwards) and fellows (from application round 2020 onwards) will, upon being awarded their project or fellowship, be invited to develop their answers to the data management related questions into a DMP. The FWO expects a **completed DMP no later than 6 months after the official start date** of the project or fellowship. The DMP should not be directly submitted to FWO, but to the research coordination office of the host institution.

At the end of the project, the **final version of the DMP** has to be added to the final report of the project; this should be submitted to FWO by the supervisor-spokesperson through FWO’s e-portal. This DMP may of course have been updated since its first version. The DMP is an element in the final evaluation of the project by the relevant expert panel. Both the DMP submitted within the first 6 months after the start date and the final DMP may use this template.

1. General Information	
Name applicant	Vincent Bonin
FWO Project Number & Title	12A5925N An Investigation of The Cell-Type Specific Rules of Connectivity in The Layer 2/3 Visual Cortical Network Through Anatomical Connections, Transcriptomics and Function
Affiliation	NERF

Responsible: Dylan Myers-Joseph, Vincent Bonin, Giuliano Maggi Olmedo

2. Data description	
Will you generate/collect new data and/or make use of existing data?	New data

Describe the origin, type and format of the data (per dataset) and its (estimated) volume If you **reuse** existing data, specify the **source** of these data. Distinguish data **types** (the kind of content) from data **formats** (the technical format).	Observational data Experimental data Digital images This project will generate several interlinked datasets that are essential for mapping the functional, anatomical, and molecular organisation of visual streams in the mouse cortex. The primary data types are derived from in vivo imaging experiments performed during behavioural tasks. These datasets are as follows: Two-photon calcium imaging recordings: - Format: .tif (ScanImage TIFF stack format) - Estimated size: ~60 GB/day per animal - Purpose: High-resolution recordings of neuronal activity in layer 2/3 of visual cortex during visual stimulation and task engagement. These data will be used to extract the receptive field properties and behavioural modulation of individual neurons, and to link them to their projection targets and transcriptomic profiles. Eye-camera recordings: - Format: .tif (frame-by-frame imaging) - Estimated size: ~50 GB/day per animal - Purpose: Monitoring the pupil location and size in real-time during the imaging sessions. These data are used to control for changes in arousal and gaze direction, which influence visual cortex activity. Face-camera recordings: - Format: .tif or video format (depending on implementation; stored frame-by-frame for synchronisation) - Estimated size: ~50 GB/day per animal (comparable to eye-camera) - Purpose: Recording facial expressions and upper-body movements. These data are used to extract motor and behavioural state variables (e.g., whisking, posture,

	orofacial movements) for inclusion in models of neuronal activity, enabling the mapping of non-visual behavioural modulation across visual cortical areas. Stimulus and behavioural event logs: - Format: .txt (timestamped event and stimulus metadata) - Estimated size: ~10 MB/session - Purpose: Precise logging of visual stimuli, trial events, and behavioural outcomes (e.g., wheel movements, rewards, etc). These metadata are critical for aligning external variables with neuronal activity during behavioural tasks. All datasets are time-synchronised using shared acquisition clocks and will be used together to model neural activity using statistical tools such as generalized linear models (GLMs). These integrated datasets are foundational for achieving the project's aims: to dissect the cell-type specific logic of information flow in visual cortex and uncover the link between function, connectivity, and molecular identity. Omics data In addition to the imaging and behavioural datasets, this project will generate transcriptomic datasets aimed at revealing the molecular identities of neurons involved in distinct visual streams. These datasets are: Single-nucleus RNA sequencing (snRNA-seq) datasets: - Format: Processed as .fastq files for raw reads, then converted to .h5ad (AnnData), .csv, or similar formats for expression matrices and metadata. - Estimated size: ~30–50 GB per experiment, depending on sequencing depth and number of nuclei (targeting ~20,000 nuclei per condition) - Purpose: To characterise the transcriptomic identity of neurons with defined projection patterns. Fluorescence-activated nuclear sorting (FANS) will be used to isolate nuclei from retrogradely labelled neurons in specific visual areas. These data will allow clustering of cells by gene expression and comparison of transcriptional signatures between feedforward and higher-order projection neurons, as well as

	among converging streams. Spatial transcriptomics datasets (MERSCOPE platform): - Format: Multichannel fluorescence images (.tiff) of tissue slices, with accompanying gene expression data stored as spatial matrices (.csv, .json, or proprietary Vizgen formats) - Estimated size: ~100–200 GB per brain region imaged, depending on number of genes and resolution - Purpose: To spatially map the expression of a panel of several hundred genes within the visual cortex. This dataset will be registered to functional imaging data to determine whether neurons' gene expression profiles correspond with their receptive field properties and behavioural modulation. These data will also be used to test whether transcriptomic cell types are anatomically or functionally segregated and to refine existing transcriptomic classifications of layer 2/3 excitatory neurons. Both datasets will be analysed in combination with functional and anatomical information to investigate whether transcriptomic identity is aligned with specific visual streams, projection targets, or behavioural roles. They represent a key pillar of the integrative multimodal approach at the heart of the project. Simulation data Derived and compiled data Research documentation

	All documentation associated with the project will be maintained as .pdf files or in markdown format. Manuscripts Manuscripts generated from the project and all figures associated with them will be in .pdf format - and will have a DOI associated. The code used to generate the figures and analyse the data for those manuscripts will be in Github repositories written in python. Canonical data These datasets represent an important source of information for the laboratory of the PI (including future staff), for scientists, educators, and science communicators working in the fields of systems neuroscience, visual processing, and molecular neurobiology, but also for non-profit organizations and research consortia focused on brain mapping and neurodevelopmental disorders. By combining in vivo neuronal activity, behavioural data, anatomical connectivity, and gene expression at single-cell resolution, this dataset will provide a rich multimodal resource for studying how information flows through the brain—and how distinct neuronal types contribute to perception and behaviour.

3. Ethical and legal issues	
Will you use personal data? If so, shortly describe the kind of personal data you will use AND add the reference to your file in your host institution's privacy register. In case your host institution does not (yet) have a privacy register, a reference is not yet required of course; please add the reference once the privacy register is in place in your host institution.	No
Are there any ethical issues concerning the creation and/or use of the data (e.g. experiments on humans or animals, dual use)? If so, add the reference to the formal approval by the relevant ethical review committee(s).	Yes Refer to ethical committee approval: - for work with laboratory animals: Ethical Committee Animal Experimentation (ECD) P144/2020 Specific examples: -Genetically modified organisms: animals are housed in facilities of the Laboratory Animal Center of KU Leuven or in the facilities at NERF, which applies Standard Operation Procedures concerning housing, feeding, health monitoring to assure consistent care in accordance with European and national regulations and guidelines. Animal administrative, husbandry and animal welfare data are sensitive data and are stored in the LAIS database according to security procedure of KU Leuven.
Does your work possibly result in research data with potential for tech transfer and valorisation? Will IP restrictions be claimed for the data you created? If so, for what data and which restrictions will be asserted?	No

Do existing 3rd party agreements restrict dissemination or exploitation of the data you (re)use? If so, to what data do they relate and what restrictions are in place?	No

4. Documentation and metadata	
What documentation will be provided to enable understanding and reuse of the data collected/generated in this project?	All of the code used to analyse the data will be documented and uploaded to an openly accessible Github repository, with clear examples of how to interface with the data. For computations ran in central NERF computing facilities (nerfcluster-fs), a json file is automatically generated for certain type of jobs as spikesorting, which contains metadata. This contains information as input file, location of output file, computation-date and parameters used along the calculation. Furthermore, in case of conda environments, an yml file is automatically generated which contains python packages and their versions. These files are saved in a directory choosen by an user, and automatically copied to a directory managed by the admin of the system. These metadata files augment the manner to reproduce results and a posrteriori understanding of data production.

Will a metadata standard be used? If so, describe in detail which standard will be used. If not, state in detail which metadata will be created to make the data easy/easier to find and reuse.	Yes The metadata files automatically generated for computations executed at the nerfcluster-fs, are placed in a folder which is indexed by the job-scheduler (SLURM), which is accompanied by the execution date. Therefore, the manner to search for metadata is slurmID-DD.MM.YYYY. This meta data is reachable by the user at http://nerfcluster-fs:8080/singlejob , which is browsable from the nerf-network. For computations ran in central NERF computing facilities (nerfcluster-fs), a json file is automatically generated for certain type of jobs as spikesorting, which contains metadata. This contains information as input file, location of output file, computation-date and parameters used along the calculation. Furthermore, in case of conda environments, an yml file is automatically generated which contains python packages and their versions. These metadata files are placed in a folder which is indexed by the job-scheduler (SLURM), which is accompanied by the execution date. Therefore, the manner to search for metadata is slurmID-DD.MM.YYYY. This meta data is reachable by user at http://nerfcluster-fs:8080/singlejob , which is browsable from the nerf-network.

5. Data storage & backup during the FWO project	
Where will the data be stored?	AT NERF, we have a data storage system with 327 TB capacity. The system has multiple layers of protection to ensure long-term data retention. First, the system is distributed into two datacenter sites which act as a mirror. Second, snapshots of the data are taken regularly that allow recovering from accidental corruption or deletion of data. Third, the system screens the data on a regular basis to avoid data corruption due to bit-rot. Lastly, the disks in each server are configured using RAIDZ2 (a type of ZFS). NERF storage system for actual data, it is composed of two subsystems, one is a dedicated archive system and the another one is for Work In Progress (WIP). The former is an object storage system, for which NERF has an active maintenance contract with the provider (Cloudian), while the latter is a filesystem based on openZFS. The archive system serves to store data that needs to be kept long term (years) due to legal requirements or for a later analysis. This system is so-called “nerfhf01”. The WIP system offers a high throughput which suits best for highly demanding daily IO operations. This system is so-called “nerffs13”

How will the data be backed up?	The storage system at NERF has multiple layers of protection to ensure long-term data retention. The WIP server (nerffs13) has a "twin" server located in a different data center which acts as a mirror of the former. This provides data backup in case of full failure of the nerffs13, whether caused for severe hardware issues or in case the entire data center is compromised. Furthermore, snapshots of the data are taken regularly that allow recovering from accidental corruption or deletion of data, which in combination with a RAIDZ2 (zfs-raid) configuration provides a strong data redundancy per server. Lastly, the system screens the data on a regular basis to avoid data corruption due to bit-rot. The archive system (nerfhf01) is a redundant system on itself, this is composed of several nodes distributed in multiple data centers. The nerfhf01 technology allows to have one entire node down and data is not compromised.

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 STORAGE: the archive and WIP systems have roughly 500TB of capacity each, until 2021. These can be expanded upon necessity and considering technical specs. This task is managed by the admin of the system, who also performs the upgrades and provides data storage monitoring and reporting BACKUP: the archive system, up-to-date, comprises 13 nodes (servers) distributed across three server rooms located in different buildings. This yields a full data redundancy for long term storage data, which is protected even if more than one entire node fails. In the case of the WIP system, this is composed of 2 servers placed in different buildings. One of the servers is the one that users connect to, while the another one has a copy of all the data which is nightly updated, yielding a strong data redundancy. Moreover, each of these servers are in in a RAID-Z2 configuration, meaning that if until two disks fail at the same time, data integrity is preserved.
What are the expected costs for data storage and backup during the project? How will these costs be covered? Although FWO has no earmarked budget at its disposal to support correct research data management, FWO allows for part of **the allocated project budget** to be used to cover the cost incurred.	Based on the last two years expenses and data storage forecast, NERF costs for the storage system comprises the hardware itself, and license and maintenance costs. The former amounts to 45000€ per year and the latter to 15000€ per year. These costs are covered by the NERF central budget.

Data security: how will you ensure that the data are securely stored and not accessed or modified by unauthorized persons?	NERF servers are in imec campus at Leuven. Thus, we have strong network protection as provided by imec firewalls. Moreover, imec provides a dedicated VLAN for NERF, meaning that only registered devices can access to the NERF network from the imec campus. For users outside of the imec campus, a Cisco AnyConnect VPN can be used to access to the NERF network. The VPN login authorization is setup by two factors authentification for each user. This VPN is provided and maintained by imec. In addition to that network security, the access to our storage servers from user computers is via SMB protocol. Therefore, each research group at NERF has their own “SMB accounts” as setup in the storage server by the system admin. Consequently, whether a device in-imec-campust or out-imec-campus attempts to access to the NERF network and thereafter to NERF storage servers, security layers on the network side and server accounts have to be passed first. This strongly reduces the likelihood that unauthorized persons access to NERF data.

6. Data preservation after the end of the FWO project FWO expects that data generated during the project are retained for a period of minimally 5 years after the end of the project, in as far as legal and contractual agreements allow.	
Which data will be retained for the expected 5 year period after the end of the project? In case only a selection of the data can/will be preserved, clearly state the reasons for this (legal or contractual restrictions, physical preservation issues, ...).	The minimum preservation term of 5 years after the end of the project will be applied to all datasets. All datasets will be stored on the university's central servers with automatic back-up procedures for at least 5 years, conform the KU Leuven RDM policy.
Where will these data be archived (= stored for the long term)?	Data that needs long term storage is stored in nerfhf01.
What are the expected costs for data preservation during these 5 years? How will the costs be covered? Although FWO has no earmarked budget at its disposal to support correct research data management, FWO allows for part of **the allocated project budget** to be used to cover the cost incurred.	That amounts roughly to 60000€ per year, and will be covered by the NERF central budget.

7. Data sharing and reuse	
Are there any factors restricting or preventing the sharing of (some of) the data (e.g. as defined in an agreement with a 3rd party, legal restrictions)?	No
Which data will be made available after the end of the project?	Participants to the present project are committed to publish research results to communicate them to peers and to a wide audience. All research outputs supporting publications will be made openly accessible. Depending on their nature, some data may be made available prior to publication, either on an individual basis to interested researchers and/or potential new collaborators, or publicly via repositories (e.g. negative data). We aim at communicating our results in top journals that require full disclosure upon publication of all included data, either in the main text, in supplementary material or in a data repository if requested by the journal and following deposit advice given by the journal. Depending on the journal, accessibility restrictions may apply. Biological material will be distributed to other parties if requested
Where/how will the data be made available for reuse?	In an Open Access repository, Upon request by mail
When will the data be made available?	Upon publication of the research results

Who will be able to access the data and under what conditions?	Whenever possible, datasets and the appropriate metadata will be made publicly available through repositories that support FAIR data sharing. As detailed above, metadata will contain sufficient information to support data interpretation and reuse, and will be conform to community norms. These repositories clearly describe their conditions of use (typically under a Creative Commons CC0 1.0 Universal (CC0 1.0) Public Domain Dedication, a Creative Commons Attribution (CC-BY) or an ODC Public Domain Dedication and Licence, with a material transfer agreement when applicable). Interested parties will thereby be allowed to access data directly, and they will give credit to the authors for the data used by citing the corresponding DOI. For data shared directly by the PI, a material transfer agreement (and a non-disclosure agreement if applicable) will be concluded with the beneficiaries in order to clearly describe the types of reuse that are permitted.
What are the expected costs for data sharing? How will these costs be covered? Although FWO has no earmarked budget at its disposal to support correct research data management, FWO allows for part of **the allocated project budget** to be used to cover the cost incurred.	It is the intention to minimize data management costs by implementing standard procedures e.g. for metadata collection and file storage and organization from the start of the project, and by using free-to-use data repositories and dissemination facilities whenever possible. Data management costs will be covered by the laboratory budget. A budget for publication costs has been requested in this project.

8. Responsibilities	
Who will be responsible for the data documentation & metadata?	Metadata will be documented by the research and technical staff at the time of data collection and analysis, by taking careful notes in dedicated files (.csv, .md and .pdf). In addition, as indicated in the section 11, jobs launched in the nerfcluster will automatically create metadata, where the responsable is the system administrator of NERF.
Who will be responsible for data storage & back up during the project?	As long as the data is in the central storage system of NERF, the responsable is the system administrator of NERF.
Who will be responsible for ensuring data preservation and sharing?	The researchers involved and the PI.
Who bears the end responsibility for updating & implementing this DMP? Default response: The PI bears the overall responsibility for updating & implementing this DMP	The PI is ultimately responsable for all data management during and after data collection, including implementation ad updating DMP.