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	Sebastian Haesler
FWO Project Number & Title	1272222NThe neural representation of self and others
Affiliation	NERF

Responsible: Cagatay Aydin

2. Data description	
Will you generate/collect new data and/or make use of existing data?	New data, Existing 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 Anatomical data: Microscopy images and anatomical reconstructions (.tif,jpeg,digital). Physical originals (brain sections) are saved as well. 100 GB maximum. Video and audio files Behavioral data (digital) including videos of breathing rate, pupil dilation, mouse position on the virtual track, stimulus delivery times, and choices. We will save raw data and their extracted information (.mov, .mat, tdms). 1TB. Electrophysiology data Electrophysiology data Simulation data Derived and compiled data Manuscripts When manuscripts are ready to submit, authors would consider depositing to open access preprint systems (e.g., https://www.biorxiv.org/) to get some feedbacks from the community. Algorithms and scripts

	During preprocessing, postprocessing of electrophysiological and behavior data. meta structures are going to be converted to python (.npy) and/or Matlab (.mat) file containers. Then with the help of custom-written scripts, these containers are going to be wrapped and plotted as manuscript-ready figures. All scripts are going to be deposited to open subversion control systems such as Github, and bitbucket. (e.g. https://github.com/nerf-common/chronic-neuropixels-protocol) Canonical data These datasets represent an important source of information for the laboratory of the PI (including future staff), for scientists, journalists and higher education teachers working in the field of system neuroscience, but also for non-profit organizations and industries active in the field of neuroscience.

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) (KU Leuven: 073/2019)
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?	An open-access subversion control system (from NERF) will be used to fly over the dataset and generate paper-ready figures from the uploaded metadata (e.g. https://github.com/nerf-common/chronic-neuropixels-protocol)

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 data will be made available in a python (.npy) or Matlab (.mat) format with an explicit name for each variable and a readme for a good comprehension of the dataset. The brain image will be uploaded with the metadata of the confocal microscope, reusable using the free software ImageJ. 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. Additional comments: 1. The behavioral and neural preprocessed data will be made available on a python (.npy) or matlab (.mat) format with an explicit name for each variable and a readme for a good comprehension of the dataset. 2. The raw data will be available on request and will be given with a readme ensuring a good understanding of the dataset. 3. The image of the brain will be upload with the metadata of the microscope, reusable using the free software ImageJ. 4. The code of the analysis will be put online on github as soon as the results will be made public. This code will be commented and a readme will garantee the good understanding for the users.

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. If applicable: Datasets collected in the context of clinical research, which fall under the scope of the Belgian Law of 7 May 2004, will be archived for 25 years, in agreement with UZ Leuven policy and the European Regulation 536/2014 on clinical trials of medicinal products for human use.
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. For the Raes lab: Data will be accessible to academic researchers via the European Genome-Phenome Archive. For industry, data access will be subject to specific contracts.
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 (like Excel). 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.