FWO DMP Template - Flemish Standard Data Management Plan

Version KU Leuven

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 submitted to FWO but to the research co-ordination office of the host institute; FWO may request the DMP in a random check.

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.

The DMP template used by the Research Foundation Flanders (FWO) corresponds with the Flemish Standard Data Management Plan. This Flemish Standard DMP was developed by the Flemish Research Data Network (FRDN) Task Force DMP which comprises representatives of all Flemish funders and research institutions. This is a standardized DMP template based on the previous FWO template that contains the core requirements for data management planning. To increase understanding and facilitate completion of the DMP, a standardized **glossary** of definitions and abbreviations is available via the following link.

1. General Project Information	
Name Grant Holder & ORCID	Name Grant Holder: Philippe Lemey ORCID: 0000-0003-2826-5353
Contributor name(s) (+ ORCID) & roles	Contributor name: Liana Kafetzopoulou ORCID: 0000-0003-4531-1374
Project number [1] & title	Project number: G005323N Project Title: Intra-host evolution of Lassa virus during acute infection and virus persistence
Funder(s) GrantID [2]	FWO GrantID: G005323N
Affiliation(s)	☒ KU Leuven ☐ Universiteit Antwerpen ☐ Universiteit Gent ☐ Universiteit Hasselt ☐ Vrije Universiteit Brussel ☐ Other: ROR identifier KU Leuven: 05f950310

[1] "Project number" refers to the institutional project number. This question is optional. Applicants can only provide one project number.

[2] Funder(s) GrantID refers to the number of the DMP at the funder(s), here one can specify multiple GrantIDs if multiple funding sources were used.

Please provide a short project description	Every year, hundreds of thousands of people become infected with Lassa virus (LASV). It is endemic in West Africa, with Nigeria being the most affected country. Transmission to humans occurs primarily through contact with infected rodents of the species Mastomys natalensis. About 80% of infections are asymptomatic. A small proportion of LASV infections lead to an acute viral haemorrhagic disease, Lassa fever, with a case fatality rate of 20%. There are neither effective therapies nor vaccines. Little is known about pathogenesis factors. After recovery from acute illness, the virus persists in various body fluids. We hypothesise that changes in the viral genome during the course of infection play a role in the pathophysiology of acute disease and viral persistence, an aspect that has not yet been explored. Therefore, in this project we aim to elucidate viral evolution within the host and examine a possible role in disease progression and persistence. To do this, we will use a unique collection of patient samples from two longitudinal studies in Nigeria in which Lassa fever patients were followed both during hospitalisation and months to years after discharge. In the latter study, significant viral persistence was found, particularly in semen samples. We will develop experimental protocols for sensitive and detailed description of viral sequences in all sample types. These protocols will then be used to deep-sequence the virus at different time points in the course of the disease; both in patients who died from the disease and in survivors in whom the virus persisted to varying degrees. Using these data, we aim to reconstruct the dynamics of the genetic diversity of LASV in different body fluids and relate it to factors of acute disease or persistence.

2. 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 [3].

				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
Sample metadata	Pseudo-anonymised metadata relating to samples collected from patients including: age, sex, results diagnostic test Lassa, country, state, city, diagnostic test location, date seen at health facility, date of onset of disease, date of specimen collection, symptoms, disease outcome	☐ Generate new data ☒ Reuse existing data	☒ Digital ☐ Physical	☐ Audiovisual ☐ Images ☐ Sound ☒ Numerical ☒ Textual ☐ Model ☐ Software ☐ Other:	Excel files and mySQL database	☐ < 1 GB ☒ < 100 GB ☐ < 1 TB ☐ < 5 TB ☐ > 5 TB ☐ NA	NA
Laboratory procedure metadata	Data relating to samples processed in the laboratory, all necessary information relating to the laboratory procedures performed	☒ Generate new data ☐ Reuse existing data	☒ Digital ☐ Physical	☐ Audiovisual ☐ Images ☐ Sound ☒ Numerical ☒ Textual ☐ Model	Excel files and mySQL database	☐ < 1 GB ☒ < 100 GB ☐ < 1 TB ☐ < 5 TB ☐ > 5 TB ☐ NA	NA

[3] Add rows for each dataset you want to describe.

				☐ Software ☐ Other:			
Sequencing reads	Sequencing reads output from Illumina deep sequencing	☒ Generate new data ☐ Reuse existing data	☒ Digital ☒ Physical	☐ Audiovisual ☐ Images ☐ Sound ☐ Numerical ☒ Textual ☐ Model ☐ Software ☐ Other:	Fastq files	☐ < 1 GB ☐ < 100 GB ☐ < 1 TB ☒ < 5 TB ☐ > 5 TB ☐ NA	NA
Sequencing data analysis and results	Sequencing data analysis and result output files including: fasta, fastq, csv, vcf	☒ Generate new data ☐ Reuse existing data	☒ Digital ☐ Physical	☐ Audiovisual ☐ Images ☐ Sound ☒ Numerical ☒ Textual ☐ Model ☐ Software ☐ Other:	Text files (including following formats: fasta, fastq, csv, vcf)	☐ < 1 GB ☐ < 100 GB ☐ < 1 TB ☒ < 5 TB ☐ > 5 TB ☐ NA	NA
Custom made scripts	Portfolio of custom-made scripts for the processing and analysis of the metadata and results generated as part of the project	☒ Generate new data ☐ Reuse existing data	☒ Digital ☐ Physical	☐ Audiovisual ☐ Images ☐ Sound ☐ Numerical ☒ Textual ☐ Model ☐ Software ☐ Other:	Text files	☐ < 1 GB ☒ < 100 GB ☐ < 1 TB ☐ < 5 TB ☐ > 5 TB ☐ NA	NA
Data on phylogenetic inference	Phylogenetic inference of the sequences obtained and result output files	☒ Generate new data ☐ Reuse	☒ Digital ☐ Physical	☐ Audiovisual ☐ Images ☐ Sound	Text files and nexus/newick tree files	☐ < 1 GB ☐ < 100 GB ☐ < 1 TB	NA

		existing data		☒ Numerical ☒ Textual ☐ Model ☐ Software ☐ Other:		☐ < 5 TB ☐ > 5 TB ☐ NA	

GUIDANCE:
The data description forms the basis of your entire DMP, so make sure it is detailed and complete. It includes digital and physical data and encompasses the whole spectrum ranging from raw data to processed and analysed data including analysis scripts and code. Physical data are all materials that need proper management because they are valuable, difficult to replace and/or ethical issues are associated. Materials that are not considered data in an RDM context include your own manuscripts, theses and presentations; documentation is an integral part of your datasets and should described under documentation/metadata.
RDM Guidance on data

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.	The project re-uses existing data which are partly published and partly unpublished or available online. Sample metadata will be used from three pre-existing studies: (1) "Leftover samples" protocol (access to the initial diagnostic sample) (2) the "Pathogenesis study" and (3) the "Follow-up study". Published data exists under the following DOI: 1. Metagenomic sequencing at the epicenter of the Nigeria 2018 Lassa fever outbreak: 10.1126/science.aau9343 2. Virus persistence after recovery from acute Lassa fever in Nigeria: a 2-year interim analysis of a prospective longitudinal cohort study: 10.1016/S2666-5247(21)00178-6
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, refer to specific datasets or data types when appropriate and provide the relevant ethical approval number.	☒ Yes, human subject data; provide SMEC or EC approval number: S67612 ☐ Yes, animal data; provide ECD reference number: ☐ Yes, dual use; provide approval number: ☐ No Additional information: S67612

Will you process personal data[4]? If so, please refer to specific datasets or data types when appropriate and provide the KU Leuven or UZ Leuven privacy register number (G or S number).	☒ Yes (provide PRET G-number or EC S-number below) ☐ No Additional information: S67612
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 ☒ No If yes, please comment:
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 to what data they relate and what restrictions are in place.	☐ Yes ☒ No If yes, please explain:
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 to what data they relate and which restrictions will be asserted.	☐ Yes ☒ No If yes, please explain:

3. Documentation and Metadata

[4] See Glossary Flemish Standard Data Management Plan

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). *RDM guidance on documentation and metadata*.	Sample metadata will be combined in a relational database allowing for fast and efficient information interrogation. All laboratory procedures, all bioinformatics analysis and testing and any relating data is heavily documented; an electronic lab book system has been created and implemented for tracking all processes conducted. All scripts generated will be commented/documented and will include README files. Log files will be generated during data analysis with detailed information on the analysis process performed on each run and sample analysed. All public databases used within our analysis (eg. Genome reference database) will contain download dates and/or version control Version control is used across all scripts.
Will a metadata standard be used to make it easier to **find and reuse the data**? If so, please specify which metadata standard will be used. If not, please specify which metadata will be created to make the data easier to find and reuse. *REPOSITORIES COULD ASK TO DELIVER METADATA IN A CERTAIN FORMAT, WITH SPECIFIED ONTOLOGIES AND VOCABULARIES, I.E. STANDARD LISTS WITH UNIQUE IDENTIFIERS.*	☒ Yes ☐ No If yes, please specify (where appropriate per dataset or data type) which metadata standard will be used: A data schema was internally designed to ensure Findable, Accessible, Interoperable and Reusable (FAIR) metadata. All metadata is stored within a relational (MySQL) database. Within the database, data is structured with unique identifiers containing specific attributes with data type declaration and is therefore machine-readable. If no, please specify (where appropriate per dataset or data type) which metadata will be created:

4. Data Storage & Back-up during the Research Project

Where will the data be stored? *Consult the interactive KU Leuven storage guide to find the most suitable storage solution for your data.*	☐ Shared network drive (J-drive) ☐ Personal network drive (I-drive) ☐ OneDrive (KU Leuven) ☐ Sharepoint online ☐ Sharepoint on-premis ☒ Large Volume Storage ☐ Digital Vault ☒ Other: KUL hosted MySQL Server
How will the data be backed up? *WHAT STORAGE AND BACKUP PROCEDURES WILL BE IN PLACE TO PREVENT DATA LOSS?*	☒ Standard back-up provided by KU Leuven ICTS for my storage solution. ☐ Personal back-ups I make (specify) ☐ Other (specify)
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 ☐ No If no, please specify:
How will you ensure that the data are securely stored and not accessed or modified by unauthorized persons? *CLEARLY DESCRIBE THE MEASURES (IN TERMS OF PHYSICAL SECURITY, NETWORK SECURITY, AND SECURITY OF COMPUTER SYSTEMS AND FILES) THAT WILL BE TAKEN TO ENSURE THAT STORED AND TRANSFERRED DATA ARE SAFE.* *Guidance on security for research data*	All data is stored on locations where only authorized persons can access. No unauthorized persons can access the data. All data and files generated as part of this project will be stored on KU Leuven servers using the network drives or SharePoint. Additionally, all researchers actively working on the project follow a clean desk policy. Secure solutions (Belnet Filesender) for sharing data with persons outside KU Leuven will be used when data and files need to be shared between the project partners.

What are the expected costs for data storage and backup during the research project? How will these costs be covered?	L drive: €569,2/year KUL hosted MySQL Server: € 91,46/year **Total: € 660.66/year** These costs have been anticipated and will be covered from project funding that is already available.

5. 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...). *Guidance on data preservation*	☒ All data will be preserved for 10 years according to KU Leuven RDM policy ☐ All data will be preserved for 25 years according to CTC recommendations for clinical trials with medicinal products for human use and for clinical experiments on humans ☐ Certain data cannot be kept for 10 years (explain)
Where will these data be archived (stored and curated for the long-term)? *Dedicated data repositories are often the best place to preserve your data. Data not suitable for preservation in a repository can be stored using a KU Leuven storage solution, consult the interactive KU Leuven storage guide.*	☐ KU Leuven RDR ☒ Large Volume Storage (longterm for large volumes) ☐ Shared network drive (J-drive) ☐ Other (specifiy):

What are the expected costs for data preservation during the expected retention period ? How will these costs be covered?	These involve the same costs as during the project duration L drive: €569,2/year KUL hosted MySQL Server: € 91,46/year **Total: € 660.66/year** These costs have been anticipated and will be covered from funds that are available to the PI.

6. Data Sharing and Reuse	
Will the data (or part of the data) be made available for reuse after/during the project? Please explain per dataset or data type which data will be made available. *NOTE THAT 'AVAILABLE' DOES NOT NECESSARILY MEAN THAT THE DATA SET BECOMES OPENLY AVAILABLE, CONDITIONS FOR ACCESS AND USE MAY APPLY. AVAILABILITY IN THIS QUESTION THUS ENTAILS BOTH OPEN & RESTRICTED ACCESS. FOR MORE INFORMATION: HTTPS://WIKI.SURFNET.NL/DISPLAY/STANDARDS/INFO-EU-REPO/#INFOEUREPO-ACCESSRIGHTS*	☒ Yes, as open data ☐ Yes, as embargoed data (temporary restriction) ☒ Yes, as restricted data (upon approval, or institutional access only) ☒ No (closed access) ☐ Other, please specify: Sample metadata: closed access Laboratory procedure metadata: restricted access Sequencing reads: restricted access Sequencing data analysis and results: open data Custom made scripts: open data Data on phylogenetic inference: open data

If access is restricted, please specify who will be able to access the data and under what conditions.	Laboratory procedure metadata: institutional access Sequencing reads: as read data could contain human reads this is restricted. However, upon request we can map and share viral reads only.
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 per dataset or data type where appropriate.	☐ Yes, privacy aspects ☐ Yes, intellectual property rights ☐ Yes, ethical aspects ☐ Yes, aspects of dual use ☐ Yes, other ☒ No If yes, please specify:
Where will the data be made available? If already known, please provide a repository per dataset or data type.	☐ KU Leuven RDR ☒ Other data repository (specify): NCBI ☒ Other (specify) Sequencing data analysis and results: NCBI Custom made scripts: Github Data on phylogenetic inference: Github or as manuscript supplementary information
When will the data be made available?	☒ Upon publication of research results ☐ Specific date (specify) ☐ Other (specify)

Which data usage licenses are you going to provide? If none, please explain why. *A DATA USAGE LICENSE INDICATES WHETHER THE DATA CAN BE REUSED OR NOT AND UNDER WHAT CONDITIONS. IF NO LICENCE IS GRANTED, THE DATA ARE IN A GREY ZONE AND CANNOT BE LEGALLY REUSED. DO NOTE THAT YOU MAY ONLY RELEASE DATA UNDER A LICENCE CHOSEN BY YOURSELF IF IT DOES NOT ALREADY FALL UNDER ANOTHER LICENCE THAT MIGHT PROHIBIT THAT.* *Check the RDR guidance on licences for data and software sources code or consult the License selector tool to help you choose.*	☐ CC-BY 4.0 (data) ☐ Data Transfer Agreement (restricted data) ☒ MIT licence (code) ☐ GNU GPL-3.0 (code) ☐ Other (specify)
Do you intend to add a PID/DOI/accession number to your dataset(s)? If already available, please provide it here. *INDICATE WHETHER YOU INTEND TO ADD A PERSISTENT AND UNIQUE IDENTIFIER IN ORDER TO IDENTIFY AND RETRIEVE THE DATA.*	☒ Yes, a PID will be added upon deposit in a data repository ☐ My dataset already has a PID ☐ No
What are the expected costs for data sharing? How will these costs be covered?	No costs are expected for data sharing on locations where the data and scripts will be made publicly available (NCBI, Github). Any publications costs associated will be covered by project funding already available.

7. Responsibilities	
Who will manage data documentation and metadata during the research project?	Ine Boonen, Liana Kafetzopoulou
Who will manage data storage and backup during the research project?	Ine Boonen, Liana Kafetzopoulou
Who will manage data preservation and sharing?	Ine Boonen, Liana Kafetzopoulou
Who will update and implement this DMP?	Liana Kafetzopoulou, Philippe Lemey