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	Nathalie Veyt , ORCID ID: 0000-0002-6472-1822
Contributor name(s) (+ ORCID) & roles	Principal investigator: prof. dr. Wim Wuyts , ORCID ID: 0000-0001-9648-3497
Project number [1] & title	Genetic risk stratification in familial idiopathic pulmonary fibrosis to refine the multimodal screening program for relatives
Funder(s) GrantID [2]	Fonds voor Wetenschappelijk Onderzoek – Research Foundation Flanders (FWO): 1SHA424N
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	Idiopathic pulmonary fibrosis (IPF) is the most common fibrotic interstitial lung disease. The disease is rapidly progressive and unfortunately the diagnosis is often made in an advanced disease stage, which makes the prognosis very poor. So far, only two antifibrotic drugs (pirfenidone and nintedanib) can slow down disease progression. As these drugs are equally effective in an early compared to advanced disease stage, early diagnosis is crucial. The increased occurrence of IPF within families (familial IPF) first indicated that genetic factors underlie the risk of IPF. Rare pathogenic variants in telomere-related genes or surfactant genes are detected in 25-30% of familial IPF kindreds. Furthermore, genome-wide association studies have identified multiple association signals for IPF. However, much about the genetics of IPF remains unknown. The large proportion of unexplained familial IPF kindreds and the occurrence of early IPF in 20-25% of non-mutation carrier relatives highlight these knowledge gaps. Hence, these missing pieces make proper screening and follow-up of relatives very challenging. The purpose of this project is to create possibilities for accurate risk prediction, counseling and personalized follow-up for family members of IPF patients through genetic risk stratification. To achieve this, the project combines identification of novel disease causing variants and quantification of polygenic risk.

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
Clinical screening data	Patient data on demographics, availability of samples in the biobank, ILD diagnosis, comborbidities, family history, genetic testing, exposures, pulmonary function tests, concomitant therapy, adverse events, patient status. These data are stored in a REDCap database.	☒ Generate new data ☐ Reuse existing data	☒ Digital ☐ Physical	☐ Audiovisual ☐ Images ☐ Sound ☒ Numerical ☒ Textual ☐ Model ☐ Software ☐ Other:	.csv	☒ < 1 GB ☐ < 100 GB ☐ < 1 TB ☐ < 5 TB ☐ > 5 TB ☐ NA	
WES data	Whole exome sequencing will produce	☒ Generate new data	☒ Digital ☐ Physical	☐ Audiovisual ☐ Images ☐ Sound	.xlsx	☐ < 1 GB ☒ < 100 GB ☐ < 1 TB	

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

	multiannotated excel files with numerical data	☐ Reuse existing data		☒ Numerical ☐ Textual ☐ Model ☐ Software ☐ Other:		☐ < 5 TB ☐ > 5 TB ☐ NA	
SNP array data	SNP array will produce	☒ Generate new data ☐ Reuse existing data	☒ Digital ☐ Physical	☐ Audiovisual ☐ Images ☐ Sound ☒ Numerical ☐ Textual ☐ Model ☐ Software ☐ Other:	.idat .bin .map .phenotype .script .bpm .xlsx .bsc .egt	☐ < 1 GB ☒ < 100 GB ☐ < 1 TB ☐ < 5 TB ☐ > 5 TB ☐ NA	
WGS data	Whole genome sequencing will produce multiannotated excel files with numerical data	☒ Generate new data ☐ Reuse existing data	☒ Digital ☐ Physical	☐ Audiovisual ☐ Images ☐ Sound ☒ Numerical ☐ Textual ☐ Model ☐ Software ☐ Other:	.xlsx	☐ < 1 GB ☒ < 100 GB ☐ < 1 TB ☐ < 5 TB ☐ > 5 TB ☐ NA	
Biological patient samples	1 (or 2) EDTA blood sample(s) per study participant, 1 (or 2) extracted DNA samples per study	☒ Generate new data ☐ Reuse existing data	☐ Digital ☒ Physical	☐ Audiovisual ☐ Images ☐ Sound ☐ Numerical ☐ Textual ☐ Model ☐ Software ☐ Other:			DNA samples are stored in freezers in BREATHE lab: 1 (or 2) extracted DNA samples per study participant Blood samples are stored in freezers

	participant (full DNA and/or aliquot)						in the Biobank: 1 (or 2) 10 ml EDTA tubes per study participant. At this time +- 335 study participants are included.

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.	For the interpretation of genetic analyses the project will use databanks such as gnomAD (https://gnomad.broadinstitute.org/), HGMD (https://apps.ingenuity.com/ingsso/login?service=https%3A%2F%2Fmy.qiagendigitalinsights.com%2Fbbp%2Flogin%2Fcas), ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/) For further steps in the project, we might also apply for access to the UK Biobank data.
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: S63694 ☐ Yes, animal data; provide ECD reference number: ☐ Yes, dual use; provide approval number: ☐ No Additional information: This project involves further analysis of personal (genetic) data of human participants, as well as ongoing collection of human biological samples. The Ethics Committee Research UZ Leuven/KU Leuven approved the use of these samples (S63694).

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: S63694 This project involves analysis and ongoing collection of clinical screening data (e.g. demographics, ILD diagnosis, comborbidities, family history, exposures, pulmonary function tests, results of CT scans, etc.). These sensitive personal data are pseudonymized and stored in a REDCap database.
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*.	- Experimental protocols exists for specific lab tasks (e.g. DNA extraction, sample preparation for Sanger sequencing,…), both physically in laboratory books, and in Word files on the shared drive, describing in detail materials and methods. - A sample inventory of locations of blood sample and extracted DNA samples is stored on the shared drive in an Excel sheet. - Data folders containing raw and processed data will be organized based on date of the data generation and source of the data. Each main data folder will contain a ReadMe.txt file containing all the necessary information to keep the data understandable and usable. - All of these files will be stored in the KU Leuven shared storage space (J-drive) and will be backed up regularly to the KU Leuven large storage space (L-drive) and an external hard drive.
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: If no, please specify (where appropriate per dataset or data type) which metadata will be created: See above.

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: REDCap database, the supercomputer server for calculations
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) : the data will also be backed-up on an external hard drive. ☐ 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 RedCap is hosted on central ICTS webservices and provides unlimited capacity. Digital data are stored at the KUL University's secure environment, of which daily backup is made by the ICT to secure the data. The storage and back-up capacity is available on the KU Leuven shares. In case additional storage is required, the capacity of the KU Leuven shares can be increased. 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*	- Storage facilities such as RedCap and the L:drive are incorporated within secured KU / UZ Leuven environments, are password-protected (including smartphone-based multi-factor identification) and are only accessible by registered researchers. - Sensitive personal data are pseudonymized by means of subject ID codes.
What are the expected costs for data storage and backup during the research project? How will these costs be covered?	The costs for data storage will be covered by previous funding obtained by the laboratory (e.g. the PI's TBM budget available for this project) and/or by the FWO fellowship bench fee.

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) Clinical data from patients will be kept for at least 25 years. All other research data will be kept for at least 10 years.

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?	The data generated during this project and accompanying documentation can eventually be stored on the K:drive for long-term data archiving (managed by KU Leuven ICTS with automatic back-up procedures). Given the expected data volume, we foresee that this cost will be covered by the project budget.

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:
If access is restricted, please specify who will be able to access the data and under what conditions.	All generated data will be used in published articles and in the PhD thesis manuscript. Scripts can be shared upon request.

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: - Access may be restricted for some unpublished data. - Sensitive personal data will be pseudonymized.
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) ☐ Other (specify)
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) The use of specific data usage licenses is not yet known.
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?	The expected cost for data sharing is not yet known.

7. Responsibilities	
Who will manage data documentation and metadata during the research project?	The PhD researcher will be responsible for the documentation and metadata.
Who will manage data storage and backup during the research project?	The PhD researcher will be responsible for storage of the data. The The PI and lab manager will manage the data storage facilities.
Who will manage data preservation and sharing?	The PI and lab manager will be responsible for data preservation and sharing.
Who will update and implement this DMP?	Both the PhD researcher and PI will update and implement this DMP.