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	**Isabelle Vanden Bempt 0000-0002-3433-555X**
Contributor name(s) (+ ORCID) & roles	**Not applicable**
Project number [1] & title	S66276 Optical genome mapping for comprehensive genomic analysis of sarcomas in a diagnostic setting.
Funder(s) GrantID [2]	G077723N
Affiliation(s)	X 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	Malignant soft tissue and bone tumors, also called sarcomas, are rare mesenchymal tumors that are characterized by a high heterogeneity in terms of histological presentation, clinical disease course and underlying genetic landscape. The identification of genetic aberrations significantly adds to a correct diagnosis of sarcomas, which is strongly associated with patient treatment and outcome. However, detecting these genetic aberrations in diagnostic practice remains a major challenge. While some sarcomas carry a characteristic genetic aberration, many remain poorly characterized or show multiple, complex aberrations that are difficult to detect using currently available methods. Therefore, we often need to combine different molecular and cytogenetic techniques making genetic testing of sarcomas unsatisfactory, labor-intensive and highly inefficient. We plan the implementation of Optical Genome Mapping (OGM), a next-generation cytogenetics tool that enables comprehensive, genome-wide detection of virtually all classes of structural variants including copy number alterations with high sensitivity. We assume that OGM will replace a number of standard genetic techniques by one single test which will dramatically change the current diagnostic workflow. Moreover, OGM will lead to the identification of novel genetic aberrations or events with diagnostic, prognostic or therapeutic potential in especially poorly characterized sarcomas.

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
OGM	OGM data	☒ Generate new data ☐ Reuse existing data	☒ Digital	☐ Audiovisual ☒ Images ☐ Sound ☒ Numerical ☒ Textual ☐ Model ☐ Software ☐ Other:	Data downloaded as .csv file for SVs and CNVs separately. Online bnx.gz data files.	☐ < 1 GB ☐ < 100 GB ☐ < 1 TB ☒ < 5 TB ☐ > 5 TB	-
WTS	Confirmation novel findings	☒ Generate new data	☒ Digital	☒ Numerical ☒ Textual	.vcf file	☒ < 100 GB ☐ < 1 TB	-
PCR	Confirmation novel findings	☒ Generate new data	☒ Digital	☒ Numerical ☒ Textual	.txt file converted to .xlsx	☒ < 1 GB	-
HE	Tumor content check	☒ Generate new data	☒ Physical	–	–	–	120 glass slides

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

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 correlation to results obtained from standard of care techniques and pathology, existing data will be retrieved from the local laboratory work station (LWS).
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: S66276, approved by Ethics Committee Research UZ/KU Leuven ☐ Yes, animal data; provide ECD reference number: ☐ Yes, dual use; provide approval number: ☐ No Additional information: /
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: S66276, correlation of genetic findings to histology and clinical data. More specifically, it concerns retrospective use of patient information retrieved from the LWS and Klinisch Werk Station (KWS). A GDPR questionnaire has been included into the application form for submission to the applicable ethics committee.

[4] See Glossary Flemish Standard Data Management Plan

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: There is potential tech transfer/valorization in both the novel genomic profiles we identify and the putative assays/biomarkers we measure. Whenever applicable, we will contact LRD.
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

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*.	All collected data points (i.e. histological, clinical information and genetic data, protocols, data analysis settings/filtering, etc) are stored on an encrypted M drive from UZ Leuven (secured and backed up, managed by ICT), and sample/patient data are pseudonymized. All data will be saved for 25 years. Access is strictly controlled ensuring only relevant study team members have access to the data files.
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: Throughout the entire project, metadata will be documented and listed in a standard template, with headers corresponding to fields reflecting professional guidelines or required by public repositories. Diagnostic information will be listed following the most recent WHO classification series, and versions will be listed. Technical and analytical methods used to generate the data will be documented in sufficient detail to allow for independent reproduction, including kit version numbers, analysis kit catalogue and lot numbers, bio-informatics software version, filter settings, treatment type and duration, genome build, …. When depositing data in a repository, the final dataset will be accompanied by this information under the form of a README.txt document. This file will be located in the top level directory of the dataset and will also list the contents of the other files and outline the file-naming convention used. This will allow the data to be understood by other members of the laboratory and add context to the dataset for future reuse.

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: encrypted M drive from UZ Leuven (secured and backed up, managed by ICT)
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 There is sufficient storage and back-up capacity on UZ Leuven server "M-drive", which is an easily scalable system under control of the UZLeuven IT department.
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*	The project specific map on the UZ Leuven server "M-drive" is only accessible to well defined study team members who have individually requested access, motivated and approved by the department and the study PI.

What are the expected costs for data storage and backup during the research project? How will these costs be covered?	The total estimated cost of data storage during the project has been estimated at is ~12,000 EUR. This estimation covers the following costs: the costs of digital data storage are as follows: 150€/TB/Year for the "M-drive". Raw data will be stored as Cloud storage at 25 Euro/TB/Year. We expect costs to drop slightly during the coming four years as UZLeuven ICT is working on a new data storage cloud solution that is lower in cost (estimated to be available by the end of 2023). Therefore, this calculation can be an overestimation. Laboratory budget will be used to cover the costs for the validation samples (n = 60); for the research part (n = 60) personal budget from the PI are used.

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?	The total estimated cost of data storage during the 5 years after the end of the project and extended storage is still to be determined together with IT (project ongoing to offer cheaper long term storage solutions). In any case, these costs can be covered by the laboratory (data validation samples) or the PI of the project (own 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.	Access to the data is restricted to the dedicated study team and within the context of the current study. For re-use, a novel project must be submitted describing what the data will be used for and needs approval by the ethics committee. Published results are of course open for discussion and comparison.

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: If novel biomarkers or profiles are identified; intellectual property rights. Personal data will only be published after de-identification and no patient identifiers will be published.
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) Upon publication, all (pseudonymized) patient details supporting a manuscript will be made publicly available as supplemental information. Omics datasets will be deposited in open access repositories such the NCBI Gene Expression Omnibus (GEO).
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?	We assume data management costs will be minimal 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.

7. Responsibilities	
Who will manage data documentation and metadata during the research project?	Metadata will be documented by the PhD student and PI, as well as dedicated technical staff at the time of data collection and analysis, taking careful notes in the electronic laboratory notebook (E-notebook) that refer to specific datasets.
Who will manage data storage and backup during the research project?	The PhD student, technical staff and PI will ensure data storage and back up, with support from UZ-IT staff.

Who will manage data preservation and sharing?	The PI is responsible for data preservation and sharing, with support from UZ-IT staff.
Who will update and implement this DMP?	The PI is ultimately responsible for all data management during and after data collection, including implementing and updating the DMP.