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	**Giuseppe Floris; ORCID: https://orcid.org/0000-0003-2391-5425**
Contributor name(s) (+ ORCID) & roles	
Project number [1] & title	**Using standard and next-generation pathology to advance breast cancer characterization.**
Funder(s) GrantID [2]	**1800125N**
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	Breast cancer (BC) is the most common cancer in women. It is a very heterogeneous disease characterized by a complex range of clinical evolution patterns, histological features and underlying molecular features. Despite the life expectancy of early BC patients has improved dramatically in recent years, the identification of reliable biomarkers predicting which patients are most likely to derive clinical benefit or develop treatment resistance remains an unmet clinical need. Additionally, metastatic BC prognosis remains poor, justifying the need for additional research. The overarching objective of this project is the demonstration that next-generation pathology is a key element to improve diagnostics and characterization of BC to eventually enhance patient care. This will be achieved through 3 specific objectives: 1) To advance the understanding and characterization of invasive lobular carcinoma. 2) To use spatial multi-dimensional techniques to improve BC characterization and identification of novel biomarkers. 3) To unravel BC progression in the context of an institutional post-mortem tissue donation program. My engagement for the future will be to integrate all standard and novel aspects of pathology to advance BC research in the context of a highly multidisciplinary teamwork. The generation of a breast dedicated line of next-generation pathology, will bring to the next level my translational efforts to improve BC patients treatment and survival.

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
Objective 1, WP1.1; objective 2 and objective 3	The main datatypes generated throughout the project are digital images.	☒ Generate new data ☐ Reuse existing data	☒ Digital ☐ Physical	☐ Audiovisual ☒ Images ☐ Sound ☐ Numerical ☐ Textual ☐ Model ☐ Software ☐ Other:	Spatially resolved mIHC images (czi and ome tiff format). - Spatially resolved transcriptomic images (ome tiff format). - H&E images (czi and ome tiff format). - Histopathological parameters of the patients (csv format).	☐ < 1 GB ☐ < 100 GB ☐ < 1 TB ☐ < 5 TB ☒ > 5 TB ☐ NA	
Objective 1	Retrieval of histopathological sliedes for central pathological review; retrieval of clinic-pathological	Generate new data	Digital and physical	Images, numerical and textual	Clinicopathological data csv format	<1TB	About 2000 glass slides will be retrieved from the archives of the pathology department for central pathology review.

	data for correlation between specific histotype of BC and trascriptomic analysis						

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.	

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

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: ☐ Yes, animal data; provide ECD reference number: ☐ Yes, dual use; provide approval number: ☐ No Additional information: Clinical patient information will be collected, linked to the study sampleIDs and pseudonymized before registration in a customized RedCap database, under supervision of Prof. Dr. Giuseppe Floris. Basic demographic data are recorded, including age, gender, body mass index, concomitant diseases, concomitant medication, time and stage of first diagnosis, performance status, previous therapy (type, duration and response, if applicable) and stage of disease at study entry. Clinical outcome parameters are also recorded, including disease control rate (stable disease or response) and objective response rate (partial or complete response). Finally, survival data is collected (PFS and OS). All data is processed and stored on the institutional IT infrastructure, protected by a genuine user authentication system relying on username and password. The subjects' identifiers will however be stored separately (site file) from their research data and replaced with a unique code to create a new identity for the subject. This code is stored on the UZ Leuven server which is password protected, but which also allows to consult the electronic medical chart of the patient stored on UZ Leuven Hospital servers, only if necessary. In addition, we will store all data on the central servers of the KU and UZ Leuven, which are protected against unauthorized access by firewalls. The same approach is applied to patients derived from external centers. Access to the data as well as the access level is limited on a project need and individual basis. Only researchers working on the project have access to these data. Due to the sample labelling as protective measure, the researchers are not able to decipher the identity of the donor. Data collection details and the strategy to guarantee the privacy of the study participants are specified in the research protocol approved by the Ethical Committee of UZ/KU Leuven (reference number S68408; S64410; S60100; S66045).

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: - Reference to ethical committee approval: S68408; S64410; S60100; S66045, approved by the Ethical Committee of UZ/KU Leuven. - Ethical Committee of UZ/KU Leuven approval for WP1.1 currently ongoing
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: It is not excluded that the proposed work could result in research data with potential for tech transfer and valorization. Both VIB and KU Leuven have a policy to actively monitor research data for such potential. If there is substantial potential, the invention will be thoroughly assessed, and in a number of cases the invention will be IP protected (mostly patent protection or copyright protection). As such the IP protection does not withhold the research data from being made public. In the case a decision is taken to file a patent application it will be planned so that publications need not be delayed.
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:

[4] See Glossary Flemish Standard Data Management Plan

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.	- Documentation will consist of notes in the Electronic Laboratory Notebook that refers to specific datasets. These notes will describe the biological/clinical samples used, experimental setup and protocols used, sequences generated, links to the specific computer location and the specific names of the respective datasets. Metadata sheets are maintained with the connection between lab samples, sample IDs and files in the data storage so that lab samples, data files and experimental notes remain properly linked to the corresponding samples IDs. - Research methods and practices (SOPs) are fully documented. When wet lab techniques, scripts, algorithms and software tools are finalized, they are additionally described in manuscripts and/or on GitHub.

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: All data and accompanying information will be stored exclusively on KU Leuven servers, Onedrive. All data will be accompanied with a README file or tab that outlines the exact data collection procedure, especially important for experimental data. Data documentation will be tailored to their ultimate deposition in public repositories, with spreadsheet headers corresponding to fields required by these public repositories. Technical and analytical methods used to generate the data will be documented in sufficient detail to allow for independent reproduction. These will include analysis package version numbers, analysis kit, disease status, treatment type and duration…. When depositing data in a repository, the final dataset will be accompanied by this information in the file format that the repository provides. This will allow the data to be understood by other members of the laboratory and add context to the dataset for future reuse. 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) ☐ Teams ☐ Sharepoint online ☐ Sharepoint on-premis ☐ Large Volume Storage ☐ ManGO ☐ Digital vault ☒ Other: Additionally, the hosting laboratory has recently purchased a network-attached storage (NAS) system with 500Tb of storage (mirrored) where an extra copy will be stored. This NAS system is attached to KU Leuven's network so the same safety measurements apply.
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) All data stored in the NAS is backed up automatically with version control and logging. Additionally, the NAS is mirrored to prevent catastrophic disaster.
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 We have 460 TB of data capacity in the NAS from which 100Tb have been allocated for this project. 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*	The data are stored on the KU Leuven servers, only accessible with double authentication.
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 will be ~12,000 EUR. All costs for data storage will be covered by own funding.

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): Additionally, the hosting laboratory has recently purchased a network-attached storage (NAS) system with 500Tb of storage (mirrored) where an extra copy will be stored. This NAS system is attached to KU Leuven's network so the same safety measurements apply.
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 will be ~ €10000. -The storage after the project is much smaller because during the project a large working space is needed, and post-publication data will be made accessible via open access platforms. All costs for data preservation will be covered by our own funding.

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: Whenever possible, datasets and appropriate metadata will be made publicly available through repositories that support FAIR data sharing. Personal data will be double coded and no reference to subject name will be made. KU Leuven has recently joined the iRODS consortium (https://irods.org/), an open-source data management software that will be fully adopted as DMP. All datasets will be made available upon publication of the research results, pending potential embargo periods due to IPR-related confidentiality clauses (no more than 2 years). Research results will be further discussed with the Leuven Research and Development (LRD) department before being made available. All research results will be uploaded in csv format in Zenodo and GitHub as an open-access dataset under a CC-BY license. Therefore, it will be available to anyone for any purpose, provided that they give appropriate credit to the creators. Sharing policies for specific research outputs are detailed below: - Double-coded patient data: Upon publication, all double-coded patient details supporting a manuscript will be made publicly available as supplemental information. - Research documentation: All protocols used to generate published data will be described in the corresponding manuscript(s), and the related documentation will be included as supplementary information. These data and all other documents (raw data) deposited in the E-Notebook are accessible to the PI and the research staff, and will be made available upon request. - Manuscripts: All scientific publications will be shared openly. Manuscripts submitted for publication will be deposited in a pre-print server such as bioRxiv. (Pre-print) publications will be automatically added to our institutional repository, Lirias 2.0, based on the authors name and ORCID ID.

	- Algorithms, scripts and software: All the relevant algorithms, scripts and software toosls driving the project will be described in manuscripts and/or on GitHub (https://github.com). - Data that do not support publication will be either deposited in an open access repository or made available upon request by email after having signed a DTA between the different parties according to institutional KUL and UZL regulations. Data will be reused by transfer via Belnet Filesender or secure copy.
If access is restricted, please specify who will be able to access the data and under what conditions.	
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) ☐ 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)
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?	It is the intention to minimize data management costs by implementing standard operating procedures (SOPs) 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. All data management costs will be covered by own funding.

7. Responsibilities

Who will manage data documentation and metadata during the research project?	(Meta)data will be documented by the research and technical staff at the time of data collection and analysis, by taking careful notes in the E-notebook that refer to specific datasets. The research and technical staff includes the PhD student(s) (Ulysse Henriot, Tim Van Assche, Gitte Zels, Anirudh Pabba), technical assistants (Kathleen Van Den Eynde, Nikolina , Sarha Cumps, Dubroja) and bio-informaticians (Asier Antoranz) directly involved with this research project.
Who will manage data storage and backup during the research project?	The research and technical staff will ensure data storage and back up, with support from ICTS, gbiomed-IT staff, and UZ-IT staff. Final responsibility for data storage & back-up lies with promotor of this project, supported by ICTS, HPC, gbiomed-IT staff and UZ-IT staff.
Who will manage data preservation and sharing?	The research and technical staff will ensure data preservation and sharing, with support from ICTS, gbiomed-IT staff, and UZ-IT staff. Final responsibility for ensuring data preservation and sharing lies with the promotor of this project, supported by ICTS, HPC, gbiomed-IT staff, and UZ-IT staff.
Who will update and implement this DMP?	The bursary of this FWO FKM carries the end responsibility for updating and implement this DMP (Giuseppe Floris).