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	Promotor: Frederik De Smet (KU Leuven) - https://orcid.org/0000-0002-6669-3335
Contributor name(s) (+ ORCID) & roles	Co-promotors • Jeroen Lammertyn (KU Leuven) - http://orcid.org/0000-0001-8143-6794 • Thierry Voet (KU Leuven) - http://orcid.org/0000-0003-1204-9963 • Steven De Vleeschouwer (KU Leuven) - http://orcid.org/0000-0002-9576-9917 • Alejandro Sifrim (KU Leuven) - http://orcid.org/0000-0001-8247-4020 • Paul Clement (KU Leuven) - http://orcid.org/0000-0001-7600-0806 • Robert Raedt (UGhent) - https://orcid.org/0000-0002-8939-0169
Project number [1] & title	3M240531 - GlioSELECT: A multi-omic single-cell, functional precision oncology platform to guide personalized medicine in high-grade brain tumors
Funder(s) GrantID [2]	S000825N
Affiliation(s)	X KU Leuven ☐ Universiteit Antwerpen X 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	Glioblastoma (GBM) remains among the most difficult-to-treat cancers with 5-year survival rates of <5% despite intensive standard-of-care therapy. The grim reality is that virtually all clinical trials failed over the past 20 years, which can be attributed to the large differences among GBM patients and the heterogeneous and plastic nature of each individual tumor. The identification of small groups of exceptional responding patients in many trials highlights that better selection procedures could significantly improve clinical outcomes. While the classical approach of matching patients and therapies by bulk genomic profiling did not lead to improvements in GBM outcome, functional precision oncology (FPO) assays offer an attractive alternative to speed up the gathering of actionable insights. In this project, we aim to leverage next-generation FPO assays using single-cell drug response profiling in GBM, for which workflows and instrumentation will be developed. Using this FPO workflow, we will collect single-cell molecular signatures related to a multitude of therapeutic perturbations in patient-derived GBM cultures and fresh surgical samples and link these to clinical outcome by performing an observational clinical trial. The resulting algorithms will lay the foundation for a new diagnostic framework to better match patients to the most active drug/combination. Finally, by incrementally collecting thousands to millions of single-cell drug response profiles in the 'Single-cell drUg ResPonse AnalySiS' (SURPASS) database, we will create a state-of-the-art data framework against which newly designed drug/combinations can be mapped/screened, while enabling us to identify novel drug targets using advanced artificial intelligence screening.

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
SURPASS datalake	Database containing single-cell RNA-seq drug response profiles (> 4M profiles), cytotoxicity response and clinical data	☒ Generate new data ☐ Reuse existing data	☒ Digital ☐ Physical	☐ Audiovisual ☐ Images ☐ Sound ☒ Numerical ☐ Textual ☐ Model ☐ Software ☐ Other:	.fastq, .bam, .h5ad, .RDS, .txt, .xls	☐ < 1 GB ☐ < 100 GB ☐ < 1 TB ☐ < 5 TB ☒ > 5 TB ☐ NA	
AI/ML models	AI/ML models trained in the SURPASS datalake to enable drug response prediction and identification of new biomarkers	☒ Generate new data ☐ Reuse existing data	☒ Digital ☐ Physical	☐ Audiovisual ☐ Images ☐ Sound ☐ Numerical ☐ Textual ☒ Model ☐ Software ☐ Other:	.h5ad	☐ < 1 GB ☐ < 100 GB ☒ < 1 TB ☐ < 5 TB ☐ > 5 TB ☐ NA	
DepMap/	Bulk L1000 drug	☐ Generate new	☒ Digital	☐ Audiovisual	.h5ad, .RDS, .txt	☐ < 1 GB	

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

LINCS database	response transcriptomic profiles and cytotoxicity data from DepMap and LINCS for model pretraining	data ☒ Reuse existing data	☐ Physical	☐ Images ☐ Sound ☒ Numerical ☐ Textual ☐ Model ☐ Software ☐ Other:		☐ < 100 GB ☒ < 1 TB ☐ < 5 TB ☐ > 5 TB ☐ NA	
Microfluidics device	A microfluidics platform to measure drug responses effectively	☒ Generate new data ☐ Reuse existing data	☐ Digital ☒ Physical				Proprietary microfluidics platform (tabletop format)
PDX models	PDX models from glioblastoma PDCLs	☒ Generate new data ☐ Reuse existing data	☐ Digital ☒ Physical				Mouse models
patient samples	Cryopreserved, viable dissociated tumor samples	☒ Generate new material ☐ Reuse existing data	☐ Digital ☒ Physical				Cryopreserved tubes in the KULeuven biobank (~esstimate of 500-1000 tubes)

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.	DepMap database: https://depmap.org/portal/ LINCS database: https://clue.io/
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: ECD25-05 (UGent) ☐ Yes, dual use; provide approval number: ☐ No Additional information: S59804 and S61081 (human data) and ECD25-05 (animal data; UGent)
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: G-2021-3789 and G-2021-3895
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: We expect to form a spin-off at the end of this mandate (SBO project). The following items have potential commercial valorization: 1) microfluidics device, 2) SURPASS datalake and trained AI models, and 3) novel biomarkers and drug targets.
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: An agreement has been established with UGhent for the PDX models experiments (which use cell lines generated at KU Leuven). For reusing LINCS/DepMap data for commercial purposes a license can be obtained, but we will first assess whether pretraining with this data improves the model.

[4] See Glossary Flemish Standard Data Management Plan

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: Intellectual property developed at UGhent will be owned by UGhent (PDX models), while intellectual property developed at KU Leuven will be owned by KU Leuven (everything else). We do not expect issues in commercialization, as the PDX models will not be taken to the spin-off.

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*.	Data will be stored primarily at HPC. We have applied for a Tier-1 Data Project Grant for this project, which will give us 200Tb of storage. Tier-1 Data, as ManGo, supports file metadata. We will create specific templates for each data type, that will be filled by the wet-lab and dry-lab staff to keep track of experimental and bioinformatics details. Metadata can be downloaded in JSON format. We will store relevant code in a private repository in Github. All collected clinical data points are and will be stored in a coded manner in LabCollector (an online lab management system with integrated Electronic Lab Notebook (ELN)) which runs on a secured and backed up server of KULeuven (managed by ICT of the Biomedical Sciences Group). All patient information (age, sex, disease) will be registered in a pseudonymized way in the file containing all collected samples. Every patient will receive an identification number which can only be decoded by the responsible data manager. This system also provides a logging system so no data can ever be erased, making that everything will be traceable and stored long-term (way beyond the common 5-year requirement). Only coded information will be extracted and used for the downstream research analyses. Patient material, patient-derived cell lines and PDX mouse tissues are and will be stored in -80°C freezers or liquid nitrogen within the KULeuven Cryotheek, where the exact location is defined in LabCollector and the Biobank software system.

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 is searchable and includes various levels of metadata. The LabCollector platform is structured according to projects and topics where all relevant information is directly linked to each experiment. The integration of a lab inventory to the ELN system in LabCollector, makes that information can be retrieved both from the experiment point of view or from the sample point of view (e.g., which experiments have been performed with cell line A, or which cell lines were used in experiment B). The used medical terminology (e.g., diagnosis) is according to the standards in the neuro-oncology and neuropathology fields. Cell biology-based terms are explained in the database. For computational data (e.g. SURPASS datalake, AI models) we will create metadata schemas to fill for each data type and experiment, including experimental and data processing details, which can be queried. 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: Tier-1 Data

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): In Tier-1 Data, the data itself is synchronized on two separate storage systems, each with 27 PB of usable capacity, located at ICTS KU Leuven data centers in Heverlee and Leuven. The data is protected against calamities at either site by synchronizing it in real-time at hardware level. One system does not function as a backup for the other, so this is no protection against accidental instructions (i.e. user mistakes) to delete data. Snapshots are made at regular intervals (hourly, daily and monthly) in case data needs to be recovered.
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*	In Tier-1 Data, files can be shared with users and user groups via a system of permissions. Permissions can be managed on file and folder level, allowing for fine-tuned access control. Only authorized users will have access to Tier-1 Data project data. LabCollector is protected by the KU Leuven firewall and requires user two-factor authentification.
What are the expected costs for data storage and backup during the research project? How will these costs be covered?	The Tier-1 Data Project is free and lasts for 4 years (as the length of the project). Files in OneDrive (i.e. manuscripts, figures, presentations) will be stored within the space provided by KU Leuven.

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): A cold storage solution in Tier-1 Data/ManGO is under development. At the end of the project we will move inactive data to this location.
What are the expected costs for data preservation during the expected retention period? How will these costs be covered?	Costs directly depend on the pricing of storage drives of HPC/KU Leuven for large volume and cold storage. These costs will be covered through the general lab budget (non-related grant applications). We expect a price of <20€/Tb, based on current ManGO prices. LabCollector has been purchased by the coordinating laboratory (only updates are required).

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.	Data in publications will be made available through EGA, and could be shared for research purposes upon a data transfer agreement. Pretrained published models will be shared in Zenodo. Unpublished data and models trained with unpublished data will not be shared, but will become part of the spin-off assests.
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: Since we will generate human data, proper agreements must be set up prior to sharing data sets. Regarding the AI models and the full database, these components will be a key IP asset of the spin-off company. Appropriate discussions with CTC and LRD will define the conditions of data transfers.
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): EGA, Zenodo ☐ 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): For (public) AI models, we will have a license for commercial use only (free for academic use).
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?	None.

7. Responsibilities	
Who will manage data documentation and metadata during the research project?	Frederik De Smet, Yanti de Visser & Carmen Bravo González-Blas

Who will manage data storage and backup during the research project?	Frederik De Smet, Yanti de Visser & Carmen Bravo González-Blas
Who will manage data preservation and sharing?	Frederik De Smet, Yanti de Visser & Carmen Bravo González-Blas
Who will update and implement this DMP?	Frederik De Smet, Yanti de Visser & Carmen Bravo González-Blas