DATA MANAGEMENT PLAN

Project Title:
"Supramolecular Agents as Radiotheranostic Drugs"

Acronym:
SMARTdrugs

Grant Agreement No:
101129886

Call:
HORIZON-EIC-2023-PATHFINDEROPEN-01

Document information

Deliverable number	D4.7 – Data Management Plan (DMP)
Deliverable date	Month 12
Due date	31/12/2024
Deliverable lead	TUM
Authors	Jordi Llop, Jason Holland, Alex Poot, Tim Witney, Angela Casini
Keywords	Data Management
WP	WP4
Nature	Report
Dissemination level	Public
Document submission date	19/12/2024
Reviewed by	Angela Casini

SMARTdrugs has received funding from the European Innovation Council (EIC) under grant agreement N°101129886.

1. Data Summary

The Data Management Plan (DMP) for the SMARTdrugs project is developed to facilitate data flow and utilization of the data between the parties, including third parties/public where appropriate, and ensure proper data preservation for future use. The DMP is developed in line with the Guidelines on FAIR Data Management in Horizon Europe projects. The purpose of the DMP is to cover the complete research data life cycle and to describe the types of data that will be generated/collected during the project, the standards that will be used, how data will be preserved and what parts will be shared for verification or use. The team is aware of the sensitivity of research data-related data protection, as well as exploitation and licensing needs, so it will keep certain data closed and stick with "as open as possible, as closed as necessary".

1.1. Purpose of the data collection/generation and its relation to the objectives of the project

From diagnostic applications in the quantification and/or characterisation of biomarker expression in cancer patients, through to molecularly targeted radionuclide therapy, radiopharmaceuticals are at the frontline of modern personalised medicine. The radical long-term vision of the project **SMARTdrugs** is to harness the untapped potential of supramolecular chemistry to create a new class of therapies - radiotheranostics - which combine both diagnostic and therapeutic radionuclides in one compound. By using the self-assembly of host-guest supramolecular coordination complexes and molecularly interlocked molecules as scaffolds for creating supramolecular radiotheranostic drugs, new methods for radiotheranostic synthesis that break away from conventional medicinal chemistry concepts will be introduced. SMARTdrugs will establish a proof-of-concept demonstrating the utility of non-covalent systems in the design of multifunctional radiotheranostic agents with tailored pharmacokinetics, and their application in challenging drug-delivery scenarios including targeted delivery to cancers of the lung and brain.

The data produced and collected will be of value for other researchers working in the fields of supramolecular chemistry, radiochemistry, molecular imaging and theranostic applications in cancer. All the detailed information on the nature and procedures of the experiments including research protocols and relevant details about suppliers and facilities will be facilitated. Dissemination of research results will be divided in four different sections to ensure maximum impact and advertisement of the research actions and other technologies derived from this project. Specifically:

1. Information for the general public.
2. Discourse within the scientific community.
3. Precompetitive information for industry.
4. Outreach and dissemination to end-users, including students and researchers, scientific and private entities (open access publications, and patents if applicable).

Per principle and following the Grant Agreement (GA) obligations, research data linked to exploitable results will not be put into the open domain if they compromise commercialisation prospects. This will be carefully scrutinised by the SMARTdrugs Management Committee, with the help of the partners support structures (e.g. Technology Transfer Offices, Research Support Offices). The rest of the research data will be deposited in an open-access repository (Zenodo). In practice, compliance with article 29.3 and the former mentioned guidelines will mean that for any of the project findings, that are highly innovative and/or have a high possibility for commercialization, different actions will be taken: 1) withhold the data for internal use; 2) embargo and restrict access; or 3) delay access to ensure protection (e.g. apply for a patent application, trademark registration, etc.) and initialize technology valorisation efforts in order to find a possible licensee. Generally, data will be restricted for no more than 12 months after initial filing. In this context, additional measures will be taken, namely the application of non-disclosure agreements whenever information is to be shared outside the consortium and/or publication delay so that patent protection can be correctly applied. Otherwise, the results will be made available Open Access by using online repository services or publishing in journals adhering to Open Access policies (green or gold). Dissemination will be made through the project's website (https://smartdrugs2024.com), and the project page on CORDIS will also be made. One notable exception is the sharing of small animal imaging data, which typically reach >5 Gb in size. Original list mode and reconstructed data will be stored locally and shared across partners. Upon publication, these data will be shared to external parties upon request.

1.2. Summary of data types and formats generated/collected

We expect that the data generation within SMARTdrugs will come from several sources: experimental research data, computational data, publications in scientific journals, deliverables, outreach and dissemination activities. Data types generated by the project include compound synthesis and characterization, structural data, molecular docking, density functional theory and atomistic simulations, compound screening *in vitro* and *in vivo*, including imaging data from PET/SPECT/CT/US or autoradiography studies, histology of animal tumour models, and high-end microscopy (see Annex I).

The following data formats will be generated or collected during the execution of the project:

Document description	Formats
Bibliographic citation files	.xml, .bib, .ris, .enl, .myi, .rmx
General data files	.xml, .atc
Graphs, tables, statistics	.xls, .xlsx, .xlsm, .prism, .sta, .pzfx
Images, including medical images	.tiff, .giff, .png, .jpeg, .lsm, .psd, .nd2, .dcm, .nii, .img, .hdr, .mca, .bmp, .dcm, .dc3, .dicom, .gif
Measurements/Analysis	.xlsx, .opj
High quality videos	.avi, .mov

Document description	Formats
Posters	.ppt, .pptx, .ai
Presentations	.ppt, .pptx
Standard Generalized/HyperText Markup Language	.xml, .html
Statistical analysis	.pzf, .pzfx, .prism, .sav, .spv, .voistat
Text incl. figures, graphs etc.	.pdf
Text, general results	.doc, .docx, .pdf, .txt, .rtf, .odt
Videos	.avi, .mp4, .mov, .mpg
Crystallographic data	.cif
Computational data	.mol, .mol2, .pdb, .csf, .out, .txt, .log

File formats	Associated programmes
.ai	Adobe Illustrator
.avi, .mov, .mpg	Windows Media Player, VLC media player or others
.dcm, .dc3, .dicom	DICOM online (Digital Imaging and Communications in Medicine), VivoQuant 2020
.docx	Microsoft Office Word 2010, 2007
.dwg	AutoCAD
.html, xml	WorldWideWeb
.mer	The Observer XT 11 (from Noldus)
.nd2	Nikon image reader, ImageJ
.nii, .dcm	Magnetic resonance imaging (MRI, reconstructed images NIfTI)
.opj	Origin
.pdf	Adobe Acrobat Reader/Pro
.ppt	Microsoft Office Power Point
.psd, .bmp	Adobe Photoshop
.pzf, .pzfx, .prism	GraphPad Prism
.rmx, .bib, .ris, .enl, .myi	Reference Manager, EndNote, Zotero
.sav, .spv	SPSS (Statistical Package for the Social Sciences)
.sta	Statistica
.tiff, .png, .giff	Windows Photo Viewer, Adobe Photoshop, ImageJ
.xls, .xlsx, .xlsm	Excel
.cdx	ChemDraw
.out, .log	Gaussian16 (or higher) computational software
.xlsx	Microsoft Excel 2010, 2007

2. Responsibilities

Task	Responsible
Monitoring the plan	Angela Casini (TUM) & Jordi Llop (CIC biomaGUNE)
Managing the data	**WP1** – Angela Casini & Jason Holland **WP2** – Tim Witney & Jordi Llop **WP3** – Alex Poot & Tim Witney **WP4** – Angela Casini
Preservation and conservation	Zenodo.org server will allow the long-term storage of generated data, at least as long CERN is operating. If for some reason Zenodo.org closes, the migration to another repository will be guaranteed.

The WP leaders will be responsible for the implementation of the DMP and will monitor data management activities. All partners are requested to collect and manage the data in accordance with the DMP and other common professional practices, especially the task leaders. Daily backup of research data relies on the institutional storage and backup systems (e.g. TUM uses the LRZ Cloud Storage, KCL has dedicated servers, backed up daily). These storage systems are secure and highly available. In any case, the consortium follows the standard practices in natural and life sciences: reproducibility, reliable statistical analysis, noise/error bars evaluation, consistency with other experimental results and theory. All project internal related documents like data summaries, presentations, reports etc. are shared in OneDrive and also stored in the Cloud Storage of the Leibniz Supercomputing Center (LRZ) at TUM (see section 5).

3. FAIR data

3.1. Making data findable

All research data will be named according to a specific file naming convention including a versioning system. The metadata will be documented in a .txt-File “Metadata”. The Metadata file is produced by the author of the document. By the time of publication, the data(sets) will get a DOI as unique identifier and they will be indexed by a searchable resource.

All project internal related documents like result summaries, presentations, reports etc. are stored at TUM with sufficient back-up systems in place (see below point 5. data security). These data are accessible for all members of the project through: (i) a common folder located at the DSS Storage of LRZ at TUM (DSS project for PN39HO (10TB)) accessible via Globus and DSS Web with individual named access.

All research outputs will be shared using Zenodo as a repository. A unique community folder will be generated to facilitate discoverability and accessibility. Total files size limit per record is 50GB. Higher

quotas can be requested and granted on a case-by-case basis. Metadata for both open, closed, embargoed and restricted records are available in Zenodo. Data files and data sets for restricted access records are only visible to their owners and to those the owner grants access. Restricted access allows a researcher to upload a dataset and provide the conditions under which he/she grants access to the data. Researchers wishing to request access must provide a justification for how they fulfil these conditions. The owner of the dataset gets notified for each new request and can decide to either accept or reject the request. If the request is accepted, the requester receives a secret link which usually expires within 1-12 months. The open results that are deposited in the Zenodo repository will be available at least 10 years after the conclusion of the project. According to Zenodo's general policies (http://about.zenodo.org/policies/), "*items will be retained for the lifetime of the repository. This is currently the lifetime of the host laboratory CERN, which currently has an experimental programme defined for the next 20 years at least*".

Metadata types and sources:

All metadata will be stored at Zenodo in JSON-format (JavaScript Object Notation) according to the defined JSON schema (a human- and machine-readable documentation). The Metadata will be exportable in several standard formats according to the OpenAIRE Guidelines. Metadata is licensed under CC0, except for email addresses. All metadata is exported via OAI-PMH and can be harvested. Persistent and unique identifiers such as Digital Object Identifiers (DOI) will be provided for all published research data. Data and metadata will be given in standard UK English. Each data set will be accompanied by a description (field available on Zenodo.org) to identify contents and experimental conditions.

3.2. Making data openly accessible

In line with the Open Access principle, the consortium partners will follow two main routes in SMARTdrugs:

i) Sharing research data with the public

Our common strategy is guided by the FAIR (findable, accessible, interoperable, and reusable) principles in order to foster digitalization of research and the re-usage of existing data in other contexts (e.g. screening data and structural data for medicinal chemistry programs, or imaging data for AI tool development). Accordingly, all PIs are committed to depositing data (including high-throughput and high-content data) into appropriate and professionally managed institutional or international repositories (see Annex I below) ahead of publication or latest at the point of publication and/or IP application.

The SMARTdrugs management committee will decide specifically the access procedures, the embargo periods, the necessary software, and other tools for enabling re-use of the data produced during this

project. If the dataset cannot be shared, the reasons for this will be mentioned. Patentable and sensitive (IP at risk) data protection will be responsibility of the IP owners.

ii) Publication of scientific findings

In the frame of Horizon Europe program, Open Access, as a step towards Open Science, is mandatory. Concurrently, in line with the Horizon Europe rules, open access (green and gold) Journals and media will be chosen for the scientific publications that result from the project. Copies of non-open access articles (if any) related to this project will be kept, when allowed by the publisher, in the Zenodo repository to enable availability to the research community and the general public. Moreover, the SMARTdrugs website will provide the links to the journal articles, public deliverable reports and produced dissemination/communication materials.

The partners have also identified the following acceptable choices for self-archiving:

- *Subject-based repositories:* arXiv.org, chemRxiv.org, bioRxiv.org.

-*Centralized repositories:* SMARTdrugs-related publications will be made available in the institutional repositories (e.g. mediaTUM, PURE) and on the project website (www.smartdrugs2024.com). To determine what repository to choose, the Access Infrastructure for Research in Europe (OpenAIRE) will be used (www.openaire.eu), where the National Open Access Desks provides further assistance. Besides the published article, the overlaying data necessary to validate the results will also be deposited in the Zenodo data repository.

3.3. Making data interoperable

In accordance with the EU regulations, the project partners will make sure that the research data produced in the project will be interoperable. Therefore, data repositories will be based on the DataCite-standard. Common interfaces like OAI-PMH will be contained, for allowing the exchange and harvesting of metadata into other data collections and into the repository of the EU “OpenAIRE”. The data created in this project will be made available in standardized, wide-spread formats. Standard vocabulary will be used for all data types to allow inter-disciplinary interoperability. All datasets generated are expected to map to commonly used ontologies.

3.4. Increase data re-use

All published data will be licenced by using standard licences (e.g. CC-BY 4.0), defining the terms and conditions of re-use while protecting the intellectual property rights (IPR). The SMARTdrugs management committee will also decide on which data an embargo period will apply, as well as for the duration of such embargo, especially after the project has finished (up to 5 years).

Data quality will be assured by following appropriate quality control and curation methods e.g., rigorous control of any incoming data by well-managed data profiling (formats, value distributions and data consistency and completeness will be assessed for any incoming data); logically defined data pipeline with centralized data management preventing duplicate data entering the system; capturing and documenting data conditions and scenarios with their dependencies and conditions; maintaining data integrity with checksums and triggers, if necessary; enhancing data and metadata lineage traceability for the pipeline, thus, enabling more effective data governance. Research teams will regularly check the quality of not just the data, but also related software, algorithms, and workflows when and if changes are made in them.

4. Coverage of the ethics review procedure context

All the experiments undertaken in this project will be carried out to conform with the EU legislation. All regulations comply with the European ethic guidelines for use of animals and human tissue/cells for research purposes.

5. Data security and storage

All the raw data will be collected, stored, and analysed in the respective internal institutional storage platforms with named access and with sufficient back-up systems in place. For example, at TUM the Research Data Hub offers a virtual research environment (currently TUM Workbench), which is based on the Leibniz Supercomputing Center's (LRZ) storage and backup system. Research data can be uploaded, structured, versioned, provided with metadata, shared and archived in a web-based platform. Transparent rights management allows access and editing rights to be granted to project partners. The LRZ Munich provides various services for data storage and exchange for internal as well as external users, and thus is the ideal partner to facilitate the planning, collection, processing, analysis, preservation, sharing, and reuse of data. Key services are Sync + Share, a platform for data sharing and collaborative editing of text documents, Cloud Storage (with temporary cloud storage options for data exchange), and the Data Science Storage (DSS), a service to store, manage, and protect data oceans in collaborative research. In addition, the LRZ provides the necessary tools and infrastructure for the backup and long-term archiving of research data, e.g. the simple shift from DSS to DSA (Data Science Archive) which also enables sharing of data via download links.

All project internal related documents like data summaries, presentations, reports etc. are shared in OneDrive and also stored via the LRZ intranet system. Here, files deleted by mistake can be recovered up to twelve months afterwards.

Further, OneDrive will serve as a sufficient back-up system due to following reasons:

- Access controlled by the data owner
- Unlimited storage on request
- Data encrypted in transit and at rest

6. Allocation of Resources

Making research data quality-controlled, FAIR-compatible and as open as possible should be considered by the consortium members while allocating resources to the project. Costs related to making data and other research outputs FAIR may include direct and indirect costs. Direct costs are included in the budget estimates, considering the eligibility rules and the usual accounting practises of each project participant.

Data storage and backup is provided by the respective institutions – cost is included in standard indirect costs or overheads. If additional storage will be needed – cost server/ disk space, as well as the cost of setting up and maintenance, we will consider the following estimates - University drive €0.80 per GB/y Cloud: €0.30 per GB/y2 x Harddrive: €0.14 per GB (single purchase) as for https://www.openaire.eu/estimating-costs-rdm-tool.

We consider that the data storage and backup of imaging data will require the highest storage capacity: considering ca. 200 PET+CT images produced within the duration of the network, with average size of 100 and 400 MB, respectively, can generate ca. 20 GB for CT and 80 GB for PET studies. The total of 100 GB could correspond to ca. 500 Euro cost (see estimate at https://zenodo.org/records/4548344).

7. Other Issues

To benefit from and utilize standards and ontologies developed by the respective communities, we will also interact and collaborate with various relevant NFDIs, particularly the NFDI4Chem, GHGA, NFDI4Bioimage, and NFDI4Health. Further, we are supported by the TUM Research Data Hub, which offers training, consultation, network and infrastructure.

8. Annex I

Types of data and storage

Compound Synthesis and Characterization (WP1):

including the synthetic procedures and the analytics data (e.g. NMR, ESI-MS data) are stored using the Open Enventory server or an electronically archived lab journal, which is also printed and witnessed in a timely manner to ensure suitability of the documentation for generating and supporting IP. Analytical data are stored on university servers, backed up regularly and connected to synthesized batches via unambiguous IDs.

Crystallographic data (WP1):

All X-ray structures of the crystallized compounds related to SMARTdrugs will be deposited in the Cambridge Crystallographic Data Centre (CCDC). CCDC is the world's repository for small molecule organic and metal-organic crystal structures. Structures deposited with CCDC are made publicly

available for download at the point of publication or at consent from the depositor. Each crystal structure undergoes extensive validation and cross-checking by expert chemists and crystallographers. Every entry is enriched with bibliographic, chemical and physical properties information. On receipt, all depositions are stored in the secure CCDC Supplementary Data Archive. Each deposited crystal structure will be assigned with a seven-digit CCDC number. This number will be the identification number to unequivocally refer to each crystal structure (e.g. in any subsequent publication).

Computational data (WP1):

Visualisation and molecular docking studies will be obtained by dedicated software solutions, such as PyMol, Glide or BioSolveIT. Computational output from DFT calculations using commercial software including Gaussian16 (or higher) will be in the form of human readable text files (.txt or standard output extensions including .out and .log). Computational output data will be stored locally but where space becomes limited, the output files will be deleted but the input files retained so that the calculation can be readily reproduced.

Imaging data (WP2 & WP3):

- PET/SPECT/CT imaging data will be obtained after reconstruction of raw data using the vendors' software in standard format (.dicom;.dcm; .dc3). Original list mode data will be stored on local servers for the duration of the project. Metadata for each study will be generated as text files (.txt).

- Ultrasound (US) imaging data will be stored as .dicom files with metadata recorded as human readable text files (.txt or similar).

- Data of tissue slides will be routinely digitized, stored and managed as virtual slides on designated servers at the respective institutions. At both locations, high throughput slide scanners (TUM: Leica AT2 & CS Scanners) are available to create high-quality digitized slides. These will be uploaded automatically to dedicated servers (e.g. TUM: Leica Aperio eSlide Manager) and hosted at the local data centers of the respective universities.