Precisely Patterned Nanofibers for
High Performance Bioseparations

DATA MANAGEMENT PLAN

Contents

Data Management Plan (DMP)

Deliverable D1.5; WP1
Deliverable leader: NOVA

Document due date: 31st March 2021 (Month 6)
Dissemination level - PUBLIC

Grant Agreement Number: 899732
Project Acronym: PURE
Project title: Precisely Patterned Nanofibers for High Performance Bioseparations
Project Duration: 1st October 2020 - 30th September 2024 (48 months)
Website: https://pure-fetopen.eu

Add creator: Arménio J. M. Barbosa

Add contributor(s): Arménio J. M. Barbosa, Cecília Roque, Margarida Dias, Salima Rehemtula

Revision History

Version	Date	Modified by	Document history/approvals
01	1/3/21	NOVA	Draft version circulated to partners
02	15/3/21	NOVA	Draft version circulated to partners
03C	22/3/21	NOVA	Final complete version
03V	26/3/21	NOVA	Validation
03S	31/3/21	NOVA	Submission

Abbreviation / Acronym	
DMP	Data Management Plan
PURE	Precisely Patterned Nanofibers for High Performance Bioseparations
EU	European Union
FAIR	Findability, accessibility, interoperability, and reusability
WP	Work Package
DNA	deoxyribonucleic acid
RNA	ribonucleic acid
TB	Tera Byte
RAID	Risks, Assumptions, Issues and Dependencies
DOI	Digital Object Identifier
IP	Intellectual property
FET-Open	Future and Emerging Technologies Open
PT	Portugal
DE	Germany
PDBe	Protein Data Bank Europe
GeneBank	NIH genetic sequence database
UniProt	Database of protein sequence and functional information
Reaxys	Expert-curated chemistry database
PubChem	Database of chemical molecules
SciFinder	CAS - Division of the American Chemical Society
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting
OpenAIRE	European Open Science Infrastructure, for open scholarly and scientific communication
DataCite	International not-for-profit organization which aims to improve data citation
IUPAC	International Union of Pure and Applied Chemistry
HTTPS	Hypertext Transfer Protocol Secure
NOVA	Project Partner from the NOVA University Lisbon
BOKU	Project Partner from the University of Natural Resources and Life Sciences, Vienna
BAYREUTH	Project partner from the University of Bayreuth
IBET	Project Partner from the Instituto de Biologia Experimental e Tecnológica
DPO	Data Protection Officer
GDPR	General Data Protection Regulation – European Union

Executive Summary

The following document is the Deliverable 1.5_Data Management Plan (DMP) of the PURE Project, funded by the European Union's Horizon 2020 research and innovation programme under grant agreement Number 899732.

This document is the first version of the DMP, consisting of preliminary information regarding the type and format of data that will be collected and generated, its origin, data utility and how PURE project research data will be findable, accessible, interoperable and reusable (i.e. FAIR).

The purpose of DMP is to provide the members of the consortium with an analysis of the main elements of the data management policy regarding all the datasets generated by the project. This Data Management Plan (DMP) has been prepared by taking into account the template of the "Guidelines on Data Management in Horizon 2020".

1. Data Summary

What is the purpose of the data collection/generation and its relation to the objectives of the project?

The purpose of data generation and collection in the PURE project is to obtain quantitative and qualitative data to reach PURE´s objectives, namely:

- the pioneering of precisely functionalized purification nanofibers
- the creation of social and economic impact and market strategy for technology exploitation

The data generation and collection will comply with the EU ethics and legal requirements as well as national ethics and legal requirements.

What types and formats of data will the project generate/collect?

PURE project integrates biology, chemistry, computer science, materials science, engineering and social sciences.

The data collection in the PURE project will concern scientific literature of interest, data from scientific databases (Protein Data Bank, GenBank, and others), and data from inquiries to evaluate the socio-economic impact. Data will be generated from experimental and computational work, and from interviews and questionnaires.

In the different Work Packages of the PURE Project several types and formats of data will be used, described in table 1.

Work Package	Data Type	Data Format	Generated	Collected
WP1	Text data, presentations, tabular, image	Pdf, docx, pptx, xls, csv, JPEG (.jpeg, .jpg, .jp2), GIF (.gif), TIFF (.tif, .tiff), RAW image format (.raw), Photoshop files (.psd), BMP (.bmp), PNG (.png), Adobe Portable Document Format (PDF/A, PDF) (.pdf)	Yes	Yes
WP2	Text data, presentations, tabular, image	Pdf, docx, pptx, xls, csv, pdb, sdf, fasta, genbank, JPEG (.jpeg, .jpg, .jp2), GIF (.gif), TIFF (.tif, .tiff), RAW image format (.raw), Photoshop files (.psd), BMP (.bmp), PNG (.png), Adobe Portable Document Format (PDF/A, PDF) (.pdf)	Yes.	Yes
WP3	Text data, presentations, tabular, image	Pdf, docx, pptx, xls, csv, pdb, sdf, fasta, genbank, JPEG (.jpeg, .jpg, .jp2), GIF (.gif), TIFF (.tif, .tiff), RAW image format (.raw), Photoshop files (.psd), BMP (.bmp), PNG (.png), Adobe Portable Document Format (PDF/A, PDF) (.pdf)	Yes	No
WP4	Text data, presentations, tabular, image	Pdf, docx, pptx, xls, csv, pdb, sdf, fasta, genbank, JPEG (.jpeg, .jpg, .jp2), GIF (.gif), TIFF (.tif, .tiff), RAW image format (.raw), Photoshop files (.psd), BMP (.bmp), PNG (.png), Adobe Portable Document Format (PDF/A, PDF) (.pdf)	Yes	No
WP5	Text data, presentations, tabular, image, audio data, video data	Pdf, docx, pptx, xls, csv, JPEG (.jpeg, .jpg, .jp2), GIF (.gif), TIFF (.tif, .tiff), RAW image format (.raw), Photoshop files (.psd), BMP (.bmp), PNG (.png), Adobe Portable Document Format (PDF/A, PDF) (.pdf), Free Lossless Audio Codec (FLAC) (.flac), MPEG-1 Audio Layer 3 (.mp3), Audio Interchange File Format (.aif), Waveform Audio Format (.wav), MPEG-4 (.mp4), OGG video (.ogv, .ogg), motion JPEG 2000 (.mj2)	Yes.	Yes

Table 1 – Types and formats of data generated and collected during the PURE project.

Will you re-use any existing data and how?

During the project existing data from several scientific databases will be used. In WP2, WP3 and WP4, data from Genebank, Uniprot and Protein Data Bank, will be collected to analyse sequence and structural data of DNA, RNA and/or proteins of interest. Chemical data of interest for WP2, WP3 and WP4, may also be re-used from chemical databases like Reaxys, SciFinder, Pubchem, and other similar sources. All re-used data will be referenced according to the owner's terms and conditions. These data are essential for comparative analysis of results, preparing successful pipelines, and studies involving modelling.

What is the origin of the data?

There are two main origins of the data generated and collected during the PURE project:

1- Open source data available from scientific databases, namely gene and protein databases
2- New data generated during project implementation
3- Data obtained through interviews and questionnaires

What is the expected size of the data?

The total expected size of data in the PURE Project will be around 100TB.

To whom might it be useful ('data utility')?

The produced data will be useful to the PURE consortium partners, and to dissemination, communication and exploitation activities.

2. FAIR data

2.1. Making data findable, including provisions for metadata

Are the data produced and/or used in the project discoverable with metadata, identifiable and locatable by means of a standard identification mechanism (e.g. persistent and unique identifiers such as Digital Object Identifiers)?

PURE data will be shared between partners. Data will be stored in files that have a common template for all partners in the project – Reports, RAID logs and presentations. To facilitate search using a search engine, documents must follow a specific naming and contain:

- a list of abbreviations used in the specific Work Package
- a list of keywords
- a list of the references used in the work package

Previous files will not be deleted just saved in a folder identified as "Older Versions".

PURE open data will be deposited in the European Commission open access repository – Zenodo (https://zenodo.org/). This repository allows for Digital Object Identifier (DOI) assignment to datasets and other types of research outputs, and also discoverability by human and machine-readable metadata.

What naming conventions do you follow?

Files will be identified with the following elements:

- PURE_Labbook_number_authorname_institution_date_version
- PURE_Data_Authorname_institution_date_version
- PURE_ RAIDSLog_Institution_date_version
- PURE_Report_Number_Authorname_institution_date_version
- PURE_Presentation_Authorname_institution_date_version

The date format used follows ISO 8601 (YYYYMMDD).

Will search keywords be provided that optimize possibilities for re-use?

Yes, a list of keywords will be provided by PURE Project regarding data and publications.

Do you provide clear version numbers?

Version numbers are provided in the file name. Draft versions will not be deleted and stored in "Older Versions" folder.

What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

Metadata will include a collection of information to describe the document, namely the title, author, description, file size and format, type, publication date, keywords, access rights, license, digital object identifier, related identifiers and grant reference.

2.2. Making data openly accessible

Which data produced and/or used in the project will be made openly available as the default? If certain datasets cannot be shared (or need to be shared under restrictions), explain why, clearly separating legal and contractual reasons from voluntary restrictions.

The data produced and/or used in the PURE project will be revised and approved by all consortium partners before making it openly available.

All data published in scientific journals, presented orally or as poster in scientific conferences, used to produce communication material, used to elaborate questionnaires and interview questions, will be in agreement with the FET-OPEN guidelines and made available in open access through Zenodo repository. The consortium may decide that some data should not be made openly available before IP rights protection or access is clarified.

How will the data be made accessible (e.g. by deposition in a repository)? What methods or software tools are needed to access the data? Is documentation about the software needed to access the data included? Is it possible to include the relevant software (e.g. in open source code)?

Data shared among the PURE consortium partners is made accessible through a dedicated cloud service, PURE-Cloud, available only for project members and allocated at the "PTServidor" web host service, using the Nextcloud platform. Documentation on how to access, deposit and search for data in the project's cloud is available to all project members. The Nextcloud platform is opensource, both the web host (PT) and cloud platform (DE) are EU based.

Open data will be shared trough Zenodo repository in an open format.

Where will the data and associated metadata, documentation and code be deposited? Preference should be given to certified repositories that support open access where possible.

All processed data produced during the project will be available to the consortium through the dedicated cloud service. Other data related with the project that is agreed to be openly available, will be deposited in the current openly available repositories/databases: Zenodo, protein structure – PDB, genetic sequences – Genebank, protein sequences – Uniprot, etc.

Scientific publications resulting from project scientific achievements will be published as open access by default and deposited in Zenodo.

Have you explored appropriate arrangements with the identified repository? If there are restrictions on use, how will access be provided?

Restrictions on data collected and/or generated are mainly related to the questionary in WP5. This data will be collected and analysed in the project's dedicated cloud with a limited user access, i.e., only the members collecting and analysing this data will have access to it during the project period. This data will have higher security protection as, with the possibility of being sensible data, its origin must be untraceable.

Only open data will be made available in Zenodo, so there will be no restrictions to re-use.

Is there a need for a data access committee?

The Impact Committee and Data management responsible (Arménio Barbosa) will assure that the Data Management Plan is correctly implemented.

Are there well described conditions for access (i.e. a machine readable license)?

Since Zenodo repository will be used for deposition of open data, a machine-readable license (Creative Commons License) will be associated with each dataset.

How will the identity of the person accessing the data be ascertained?

The PURE cloud has a log of data assessment and several levels of security and user permissions, so that any sensible data may be secure in the platform. Open data made available on Zenodo, will follow Zenodo's guidelines for ascertaining of people accessing the data.

2.3. Making data interoperable

Are the data produced in the project interoperable, that is allowing data exchange and re-use between researchers, institutions, organisations, countries, etc. (i.e. adhering to standards for formats, as much as possible compliant with available (open) software applications, and in particular facilitating re-combinations with different datasets from different origins)?

By default, data produced in the PURE project will be interoperable by the project members. This will be achieved by always using open or globally used data formats for data exchange. During the project, the use of open software is recommended, although for certain tasks licensed software may be used when needed. The final data will always be available in open format.

The concept interoperable demands that both data and metadata must be machine-readable and that a consistent terminology is used. For open data, the deposition in Zenodo will assure this interoperability. The metadata model used in Zenodo is based on Datacite`s metadata schema that is compatible with the Dublin Core (https://dublincore.org/) metadata standard and thus can be interpreted by OAI-PMH (https://www.openarchives.org) harvesters like those used by OpenAIRE (https://www.openaire.eu/).

What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable? Will you be using standard vocabularies for all data types present in your data set, to allow inter-disciplinary interoperability? In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies?

Open data deposition in Zenodo, a domain-agnostic repository, will allow for inter-disciplinary interoperability in terms of metadata and vocabularies. For chemical compounds terminology the IUPAC (https://iupac.org/) regulations will be followed.

3. Allocation of resources

What are the costs for making data FAIR in your project?

How will these be covered? Note that costs related to open access to research data are eligible as part of the Horizon 2020 grant (if compliant with the Grant Agreement conditions).

Who will be responsible for data management in your project? Who?

Are the resources for long term preservation discussed (costs and potential value, who decides and how what data will be kept and for how long)?

The project partners have allocated resources to cover costs associated with open access publications. Partners will also use Open Access Publishing Platforms namely the Open Research Europe platform.

The person responsible for data management is Arménio J. M. Barbosa.

The costs for data storage during and after project finishes is the responsibility of the Coordinator partner (NOVA) for internal data, and each consortium partner for open data. Open data will be deposited at Zenodo assuring its accessibility as it is archived permanently.

4. Data security

What provisions are in place for data security (including data recovery as well as secure storage and transfer of sensitive data)?

The project data will follow the three-copies-rule to guarantee data security if a data backup is needed. The main platform for data in the project is the dedicated PURE Cloud, which is access is strictly reserved to the project members. The host and data transfer are secured via HPPTS protocol. A copy of this data will be executed regularly in a dedicated NAS located at NOVA (PT). A third copy will be deposited in a server in one of the PURE partners outside of Portugal – BOKU or BAYREUTH.

Sensitive data will be separated as early as possible to create an anonymized dataset. Access to sensitive data is granted only for project members dedicated to that task in WP5, with secure passwords and encoding of the folders and files within cloud storage. Data transfer is secured via HTTPS protocol.

Is the data safely stored in certified repositories for long term preservation and curation?

PURE internal data security, recovery and storage will be maintained for 2 years after the project finishes. Open data will be stored in Zenodo for long term preservation.

5. Ethical aspects

Are there any ethical or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).

Is informed consent for data sharing and long term preservation included in questionnaires dealing with personal data?

In WP5 interviews and questionnaires will be conducted for socio-economic evaluation.

The host institution has appointed Rodrigo Adão da Fonseca as its Data Protection Officer (DPO) and plans to make his contact details available to all data subjects at the moment when personal data are collected from each individual (interviews/questionnaires).

At the moment of data collection, a consent form and all other relevant information about the data processing carried out by the PURE Project are going to be provided to each individual that is interviewed or that fills a questionnaire. The consent forms and informative sheets in question are being prepared at the moment and will be ready prior to the beginning of the data processing activities.

The PURE project does not entail the processing of sensitive personal data. It focuses mainly on the processing of personal opinions, marketing preferences and prior knowledge of people interviewed and of PURE participants (i.e. members of the PURE consortium). Basic information and contact details will also be processed but pseudonomysed. This data will contribute to opportunities identification at early stage of the project, which could impact the direction of the experimental tasks.

As it is easily understandable, the information gathered by the researchers at PURE project is, as required by article 5, paragraph 1, (c) of the GDPR, adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed (the collection of these opinions, prior knowledge and preferences in order to further disseminate the results of the project and to enable future applications of said results).

Pursuant to article 32 of the GDPR, access to all the information gathered in the context of PURE project will be restricted and controlled. Researchers will only access personal data stored in a need-to-know basis. In addition, personal data will be pseudonymized in order to significantly restrict the possibility of identifying the data subjects whose information was and is being processed.

The PURE project will adopt data pseudonymization techniques that are in accordance with the best practices implemented in this kind of investigative projects. We expect to use the "mapping table" technique in which the identifiers regarding the data subjects are substituted by a number/sequence of numbers and letters/pseudonyms to avoid the re-identification of the data subject. This "mapping table" will be kept safe and with limited access as explained above.