Essential and universal ribosome isoaspartylation catalysed by ancient enzymes (EURICA). Data Management Plan

Alexandre Smirnov, Yannis-Nicolas François, Nicolas Leulliot, Wang-Qing Liu

Yannis FRANÇOIS, Nicolas LEULLIOT, Wanqing LIU, Alexandre SMIRNOV*

Essential and universal ribosome isoaspartylation catalysed by ancient enzymes (EURICA)

ANR-23-CE44-0039

ANR – PRC 2023

Data Management Plan

Version 1

February 10, 2025

University of Strasbourg

University Paris Cité

CNRS

INSERM

*For correspondence: alexandresmirnov@unistra.fr

1. Data description & collection or re-use of existing data

1a. What data (for example the kind, formats, & volumes), will be collected or produced?

Outlined below are the data types, formats and volumes collected or expected to be collected, produced, or re-used, as drafted in the original proposal and further updated to keep up with the progress of the research (Table 1). In the majority of cases, the type of the instrument employed to collect data dictates the proprietary software and the format in which the data are first registered and undergo primary analysis. Similarly, databases and on-line analysis tools impose built-in search engines and retrieval options. In addition to these specific formats, storage, downstream manipulations, analysis, aggregation, and visualisation of data will be performed in open or standard formats (TIFF for most images, JPEG or PNG for gels and structure snapshots, AI and PDF for assembled images, DOCX for text documents, TXT for sequences and alignments, CSV, TSV and XLSX for spreadsheets, PDB and CIF for structures), wherever possible, as recommended in Turning FAIR into reality and Guide Pratique de la publication en ligne et de la réutilisation des données publiques ("Open data") elaborated by CNIL and CADA.

Table I. Data types, methodology, software, primary formats & estimated volumes

Data type	Methodology, instrument	Software, database or repository	Primary formats	Estimated volume
Deposited protein sequence, conservation, phylogenetic distribution, domain architecture (text, image, code)	Data retrieval, analysis & visualisation	InterPro, UniProt, PFAM, NCBI BLAST, NCBI Assembly	FAS, TXT, JPEG, PNG, CSV, TSV, XLSX	100 MB – 1 GB
Deposited protein structure (interactive structure)	Data retrieval, analysis & visualisation	RCSB, PyMol, Chimera	PDB, mmCIF	100 MB – 1 GB
Deposited gene or genome sequence (text)	Data retrieval & analysis	RefSeq, NCBI Nucleotide, NCBI Genome	FAS, TXT	100 kB – 1 MB
Deposited MS data (text, image, spreadsheet)	Data retrieval, analysis & visualisation	neXtProt, PeptideAtlas, ProteomeXchange Consortium, Scaffold	mzML, TXT, TSV, XLSX, JPEG, PDF	1 – 10 MB

Data from publications (gene, RNA & protein properties, interactomics, gene expression, experimental outcomes etc) (all types)	Data retrieval, analysis, & referencing	PubMed, BiblioVie, HAL, univOAK, Zotero	PDF, DOCX, XLSX, CSV, MP4, ZIP, Zotero RDF	1 – 10 GB
Gene properties (promoter, TSS, ORF, UTRs, SD or Kozak sequence, CRISPR sites) (text, image)	Data retrieval & analysis, prediction	RefSeq, EcoCyc, ExPASy, CHOPCHOP, ApE, BE-Hive, CRISPOR	DOCX, TXT, JPEG, GB, CSV, XLS	100 kB – 1 MB
Predicted RNA properties (secondary structure, stability) (text, image)	Prediction	RNAfold, mfold	DOCX, TXT, JPEG, PNG, PDF	1 – 10 MB
Predicted protein properties (MW, pI, localisation, processing & modification sites, interaction sites) (text, spreadsheet)	Prediction	ExPASy, MitoFates, TargetP, Mitoprot, RaptorX	DOCX, TXT, XLSX, JPEG	1 – 10 MB
Predicted protein structure (interactive structure)	Prediction	AlphaFold, PyMol	PDB	10 – 100 MB
Sequence alignment	Analysis	NCBI BLAST	FA, TXT	1 – 10 MB
Multiple sequence alignment	Analysis & visualisation	COBALT, WebLogo	ALN, ASN, FA, PHYLIP, TXT, DOCX, PDF, SVG, PNG, JPEG	10 MB – 100 MB

Nucleic acid expression, purification, modification or interaction pattern (image, spreadsheet)	Gel electrophoresis, ethidium bromide staining, northern blotting (radioautography, Typhoon), immunoprecipitation, EMSA	HEROLAB, ImageQuant TL	SGD, GEL, TIFF, XLSX	1 – 10 GB
Protein expression, purification or modification pattern (image, spreadsheet)	Gel electrophoresis, Coomassie or silver staining (Epson perfection V700 Photo), western blotting (chemiluminescence, ChemiDoc Touch), immunoprecipitation, affinity chromatography, modification assay (radioautography, Typhoon)	Epson Scan, ImageLab	JPEG, TIFF, SCN, GEL, XLSX	100 – 1000 GB
DNA sequence (text, image)	Sanger sequencing	Chroma	SEQ, AB1, FAS, PDF	1 – 10 GB
Protein concentration (text, spreadsheet)	Bradford assay (Photometer), UV absorbance (Nanodrop)	Microsoft Excel, ND-1000	XLSX, NJD	1 – 10 MB
Nucleic acid concentration (text, spreadsheet)	UV absorbance (Nanodrop)	ND-1000	NJD	1 – 10 MB
Sedimentation properties (image, spreadsheet)	Velocity sedimentation, UV absorbance (Nanodrop), gel electrophoresis, ethidium bromide staining (Typhoon), northern & western blotting (Typhoon, ChemiDoc Touch)	ND-1000, ImageQuant TL, ImageLab, Microsoft Excel	NJD, GEL, SCN, XLSX	10 – 100 GB

Protein MS data (identification, quantification, modifications) (image, spreadsheet)	Immunoprecipitation, crosslinking, affinity chromatography, velocity sedimentation, LC-MS/MS (TripleTOF 5600, NanoLC-Ultra-2D-Plus, Q-Exactive Plus, EASY-nanoLC-1000)	Mascot, Proline, Scaffold, Microsoft Excel	RAW, MGF, MZID, mzML, MSF, TIFF, PDF, XLSX	100 – 1000 GB
Peptide synthesis data (text, image)	Solid-phase Fmoc chemistry, Activotec P14	ACTIVO	SYN, TEXT	10-100 MB
Peptide quality control data	HPLC analysis & visualisation, Shimadzu HPLC system	Prominence	LCD, PDF	10-100 MB
Peptide MS analytical data	ESI, Maldi-Tof	Q-Tof, Data-Explorer	PDF	10-100 MB
Peptide CE-MS analytical data, spectra, electropherograms (image, spreadsheet)	CE-MS	Analyst, Skyline, O-TOF	D, WIFF, mzXML	1 – 10 TB
Protein & ribosome structures	X-ray crystallography, cryo-EM	wwPDB, EMDB, PyMol, Chimera, Phenix, ccp4, Cryosparc, Relion	PDB, PDBx, mmCIF, mtz, map files	20-40 TB
Protein abundance & localisation (image, spreadsheet)	Immunofluorescence, confocal microscopy (LSM700, LSM780)	ImageJ, Fiji, MosaicSuit, Microsoft Excel, Microsoft Power Point	LSM, ZVI, TIFF, XLSX, TXT, PPTX	10 – 100 GB
Oxygen consumption data (image, spreadsheet)	Respirometry (Seahorse XFe96 Analyser, Seahorse XF HS Mini Analyser)	Wave Desktop, Controller 2.6, Microsoft Excel, Power Point	ASYR, ASYT, XFLR, XLSX, TIF, PPT	10 – 100 MB
Bacterial & yeast growth data (spreadsheet)	Growth curves	Tecan, SAFAS, Microsoft Excel	AMP, XSLX	10-100 MB

Statistical data (spreadsheet, image)	Statistical tests, plotters	Microsoft Excel, Microsoft Power BI, GraphPad, Physics: Tools for science, Statistics Kingdom	XLSX, PBIX, PDF, PNG	10 – 100 MB

1b. How will new data be collected or produced &/or how will existing data be re-used?

The methodologies, databases and software employed to collect, produce or re-use data are specified in Table 1. The implemented techniques rely on a maintained system of in-house protocols, research resources, instruments and software either directly available in the host laboratories or reachable via collaborations and TGIR (ESRF, Synchrotron SOLEIL). Data re-use concerns both publicly available deposited and/or published data and previously collected in-house data and samples. In the collection, production, and re-use of data, we follow general archiving principles described in La vie d'une donnée au regard des réglementations "CRPA", "RGPD" et "Patrimoine".

The provenance of re-used data is documented with the help of the associated persistent identifiers (e.g. UniProt and InterPro entries, PDB ID, PXD, RefSeq ID, article reference/PMID, HAL ID, DOI). Some of the repositories used in this project (e.g. ProteomeXchange, wwPDB) have established versioning pipelines with rich provenance metadata enabling reconstruction of detailed lifelines of the deposited data. In the case of data from continuously updated databases (e.g., phylogenetic distributions, BLAST, PDB searches), the filters and the date of the search are additionally specified. For predictions, the parameters, the filters, and the on-line software versioning information are provided. The provenance of collected and produced data is documented in the form of metadata, lab book entries, and the Central Experiment Registry (CER, see section 2).

2. Documentation & data quality

2a. What metadata & documentation (for example the methodology of data collection & way of organising data) will accompany the data?

The research data organisation in the Smirnov's lab is embodied in the form of the Central Experiment Registry (CER) and the associated digital research infrastructure. CER is organised in folders assigned to each Creator. Each such folder contains subfolders corresponding to individual experiments and named "NNE#####", where NN are the Creator's initials and ##### is the number of the experiment (continuous numbering). Each CER folder also includes a Microsoft Excel spreadsheet (or CER *sensu stricto*) with a complete list of experiments for which the Creator is responsible. Each row of this spreadsheet corresponds to one experiment. The columns contain standardised information about experiments (Table 2). These data are matched with those of the corresponding lab books and metadata files, which obligatorily use the same identifiers to ensure complete cross-referencing.

Table 2. Central Experiment Registry (CER) columns

Experiment ID	Experiment identifier in the format NNE#####
Participants	Initials of the person(s) directly contributing to the experiment
Affiliations	Affiliations of the person(s)
Started	Date of the beginning of the experiment
Completed	Date of the completion of the experiment
Description	Very short description/title of the experiment
Associated files	Complete list of all files generated in the experiment
Associated files 2, 3 etc	Additional columns (if the capacity of the previous cell happens to be insufficient to visualise all file names)
Status	The current status of the dataset. Has five possible values: "Active", "Archived", "Suppressed", "Published", or "Deposited".

File names follow pre-established conventions and are built as follows:

NNE#####_yyyymmdd_MMMMMMMM.extension

where NNE##### is the number of the experiment, yyyymmdd is the date (in the ISO 8601 format) when the file was generated, MMMMMMMM is a method-specific descriptor, as listed in Table 3. This name may be followed by a letter suffix ("a", "b", "c" etc) if it concerns a modified version of the initial file, or by a disambiguation number suffix ("1", "2", "3" etc) if the same type of data was acquired on the same day and needs to be distinguished from an already existing but unrelated file. Exceptions from this file naming system exist: files produced by some external services and collaborators retain their original names for the reasons of traceability and communication. In this case, a normal metadata file is nevertheless created to capture their provenance and experimental details.

The data obtained at synchrotron SOLEIL are subject to a specific SOLEIL Data management policy (DSI-DU-P-0056; October 2, 2018) covering standardised metadata formats and vocabulary, specific data formats, and the software required for their manipulation provided by SOLEIL, and attribution of permanent unique identifiers to experiments and datasets (see section 5d). The Main Proposer (Nicolas LEULLIOT) takes the responsibility to enter metadata correctly and as completely as possible and associate them with the experiment number, and SOLEIL provides technical means to do this.

The data obtained at ESRF are subject to a specific ESRF Data Policy 2024 (version 14/10/2023) covering data types, auxiliary data, metadata and their storage, modalities of data curation, and attribution of permanent unique identifiers to experiments and datasets (see section 5d).

Table 3. Some method-specific descriptors used in file naming

Method	Descriptors	Interpretation
Gel electrophoresis	AGE_BEt	Agarose gel stained with BEt
	PAGE_Coomassie	Polyacrylamide gel stained with Coomassie
	PAGE_Stain-free	Stain-free polyacrylamide gel
	PAGE_Ag	Polyacrylamide gel stained with silver
	PAGE_Radio	Polyacrylamide gel visualised with PhosphorImager
Blotting	NB_SAO#####	Northern blot with oligonucleotide SAO##### as probe
	WB_White	Western blot white light image
	WB_Abref_##s	Western blot with primary antibody Abref exposed for ## s; Abref = unique antibody identifier (reference number)
	WB_Radio	Western blot visualised with PhosphorImager
UV-VIS spectrum	Spectra	UV-VIS spectra report
Sanger sequencing	Seq_pSAPnnnn_PPPP	Sanger sequencing of the plasmid pSAP#### with the primer PPPP
PCR	PCR_SAO#####_SAO#####	PCR with primers SAO##### and SAO#####
RT-PCR	RT-PCR_SAO#####_SAO#####	RT-PCR with primers SAO##### and SAO#####
Quantification	Quant	Spreadsheet with quantification or measurement data
Structure	Structure	Secondary or tertiary structure of a molecule
Sequence	Sequence	Annotated sequence of a gene, a transcript or a protein
Analysis	Analysis	A free-form qualitative and quantitative analysis including images, schemes, and other kinds of data that cannot be captured under the "Quant" category

The Smirnov's lab implements the principle of separation of primary and processed datasets. Primary research data are stored in CER folders as read-only original files. By contrast, the corresponding output files – that build on the primary or re-used data,

but are not themselves primary data – are created and stored as separate digital objects (files) in dedicated "analysis" subfolders (versioned as "Analysis01", "Analysis02" etc) within each experiment folder. These "analysis" files are diverse in format (text, spreadsheet, annotated image, PyMol session etc). They cross-reference the original data and research resources they are based on using the in-house CER taxonomy (for internally produced data) or database-specific persistent identifiers (for publicly deposited data – both internally and externally produced). The "analysis" subfolders also capture, in as much detail as possible, all modifications or filtering/censoring applied to data, analysis pipelines, software versions and settings – either directly within their body (documentation) or by means of associated README files.

CER and the file naming system are rooted in the research resource taxonomy which includes unique identifiers for protocols, cell lines, bacterial strains, plasmids, DNA oligonucleotides, antibodies, and synthetic nucleic acids. Whereas the latter two categories almost invariably use manufacturer-provided references (unless the resource in question was specifically created for exclusive in-house use), the remaining in-house resources follow the conventional naming and numbering system: cell lines – SAL###, bacterial strains – SAB####, plasmids – pSAP####, DNA oligonucleotides – SAO#####, which corresponds to their physical organisation in storage cryo-boxes. Each kind of resource is catalogued in a dedicated spreadsheet which ensures cross-referencing to other resource types (Table 4).

Table 4. Key research resource catalogues & their columns

DNA oligonucleotides	
Oligo ID	Oligonucleotide identifier SAO#####
Created	Date of order
Sequence	Oligonucleotide sequence in the 5'-to-3' direction
Description	Intended uses of the oligonucleotide, its peculiarities (modification, restriction site, amplified or detected gene & organism etc)
Notes	Any useful information or observations
Plasmids	
Plasmid	Plasmid ID pSAP####
Created	Date of confirmation by Sanger sequencing
Creator	Person who created the plasmid
Experiment ID	Experiment describing the creation of the plasmid; NNE#####
Host strain	Bacterial stock strain for storage & re-isolation; SAB####
Resistance	Resistance markers
Backbone	Vector identity
Insert	Gene name with eventual mutation in the format (genotype=proteotype) or added tags

Species of the insert	Species specification for the inserted sequence
Cloning sites	Sites used for cloning the insert
Concentration, ng/µl	Concentration of the current stock
Notes	Intended uses, provenance, construction notes & any useful observations
Oligos used for cloning	Complete list of oligonucleotides (SAO#####) used to create the plasmid
Sanger file 1, 2 etc	A series of columns containing links to Sanger sequencing data of the corresponding plasmid
Strains	
Strain	Strain ID SAB####
Date created	Date of the original DMSO/glycerol stock
Creator	Person who created the strain
Experiment ID	Experiment describing the creation of the strain; NNE#####
Species & strain	Microbial species & strain
Genotype	Eventual mutations or modifications
Plasmids	Hosted plasmids pSAP#### (with an ordinary name if empty vector)
Resistance	All known resistances of the strain
Description	Intended use of the strain: "Primary stock" (genetically unmodified, usually lacking plasmids), "Storage strain" (for plasmid storage & isolation), "Expression strain" (for protein overexpression), "Constructed strain" (genetically modified in a specific way)
Notes	Any useful information & observations (e.g. growth peculiarities, unusual phenotypes & sensitivities), eventually provenance & alternative ID (if received from third parties)
Cell lines	
Cell line	Cell line ID (typically SAL###)
Line description	Free text information about the line and its *raison d'être*
Genotype	Eventual mutations or modifications or other peculiarities
Authentication	Information about the authentication of the cell line
Parental strain	The parent of the cell line from it is derived
Resistance	All known resistances of the cell line
Old name during isolation	Operational name during the creation of the cell line (before it entered the collection)

Media	Media used to grow this cell line optimally
Provenance	Provenance information of the cell line (institution)
Date of creation	Date when the cell line is added to the collection
Creator	Person who created the cell line
Usage	A short token specifying the main use of the cell line (e.g. "reference line" or "overexpression line")
Oligos for genotyping	Oligonucleotides used to authenticate the cell line
Location	Physical location of the cell line in the collection

Protocols (NNM###) are named by the Creator's initials (NN) and numbered (###). Their versions are labelled with letter suffixes ("a", "b", "c" etc). Protocols have an invariant structure, as outlined in Table 5. They necessarily cross-reference all related research resources.

Table 5. Protocol structure

Entry	**Content**
Title	Name of the protocol
Date	Date of the last edit (excluding major modifications, which require versioning)
Associated protocols	Any other protocols related to the present one
Solutions & reagents	Complete list & compositions of buffers & chemicals required for the experiment
Kits	Any kits required for the experiment (with links to manuals)
Instruments & specific consumables	Any instruments or consumables required for the experiment
Biological materials	Organism, tissue, specific strain, cell lysate, purified macromolecules, or any other material directly used in the protocol
Time estimate	Minimal (with no facultative breaks used) and maximal (with all facultative breaks included) duration of the protocol
Breaks	Steps at which experimental manipulations may of have to be interrupted for more than 12 h
Successful implementation	Examples of strains, proteins, or other biological entities for which the protocol has already been successfully used in house, with references to specific experiments
Unsuccessful implementation	Examples of strains, proteins, or other biological entities for which the protocol failed in house (with brief notes & proposed alternatives), with references to specific experiments
Protocol	Step-by-step experimental guide

In addition to built-in metadata associated with instrument-specific primary files (including such information as date, instrument, and acquisition settings), a dedicated TXT metadata file accompanies each dataset. The metadata file name is built following these conventions:

NNE#####_DublinCore_metadata.txt

where NNE##### is the number of the experiment in CER. An experiment may have several associated datasets and the corresponding number of metadata files which, in this case, assume a suffix "_1", "_2", "_3" etc. When the experiment is finished and all associated data have been collected, it is archived, and the metadata file name is extended with "_archived". In this state, it is not subject to modification any more (read-only). (At this stage, however, it is still considered as 'intermediate data' not necessarily deemed for long-storage preservation – see section 5a.)

A standard metadata file is organised in accordance with the Dublin Core format and includes 15 generic entries (Table 6), eventually followed by the concluding tag #archived#.

As an alternative metadata format, an MS PowerPoint file (PPT), containing the experimental image with detailed annotation and relevant information on the experimental design, is used by some collaborators. This file is named in the same way as the original file it refers to (see below), followed by "_annoté".

An additional metadata file is generated for enzymatic assays. It uses the EnzymeML macros template (Lauterbach & al., "EnzymeML: seamless data flow and modelling of enzymatic data", *Nat Methods* 2023) based on Systems Biology Markup Language (SBML). The metadata file name is built following these conventions:

NNE#####_EnzymeML_metadata.xlsm

where NNE##### is the number of the experiment in CER. Once the template is filled out, the file is saved as a regular .xlsx file, as recommended by the developers. The EnzymeML metadata and the associated infrastructure are designed to support the FAIR principles. The metadata are formatted in a way that directly enables downstream kinetic modelling in COPASI and data deposition in the EnzymeML Dataverse, SABIO-RK, or STRENDA DB repositories.

In a similar manner, additional metadata are generated for other datasets deposited in dedicated repositories (section 5), following their established conventions and requirements (e.g. ProteomeXchange Consortium data submission guidelines, version 3.0.1; October 13, 2019; or Image Data Resource submission guidelines, version 1.0; March 2017; BioImage Archive submission – Lab Guide). In particular, REMBI metadata will be created to accompany published microscopy datasets selected for deposition to BioImage Archive. The possibility of introducing other data type-specific FAIR-compatible metadata formats, approved by the corresponding communities, continues to be studied.

Table 6. Metadata file organisation (for research data)

Entry	Description
1. Title	Short title of the experiment
2. Creator	Person (name, ORCID) &/or institution who created the metadata file, usually the main experimenter
3. Subject	Type of the experiment (method)
4. Description	Concise description of the experiment, including (i) relevant IDs of in-house research resources (cell lines, strains, plasmids, oligonucleotides, antibodies, protocols with eventual deviations) & related experiments, (ii) all generated files in the dataset & their descriptions (type of data, order & amount of samples, technical observations), (iii) any additional textual information which does not make part of a separate file (e.g., OD measurements, concentrations, qualitative remarks)
5. Publisher	Université de Strasbourg
6. Contributor	Person(s) (names, ORCIDs) &/or institution(s) who directly contributed to the experiment
7. Date	Date of the last modification
8. Type	Nature of the data (e.g. gel scan, 3D model, UV absorbance spectrum)
9. Format	Name of the folder & eventually subfolder(s) in which the data are classified
10. Identifier	ID of the experiment & a complete list of the files & subfolders in the dataset
11. Source	Primary (for original data) or secondary (for re-used or re-analysed data)
12. Language	English
13. Relation	IDs of all other experiments &/or data or metadata to which the present dataset is related in terms of deliverables, continuation, replication, modification, re-use, or re-analysis
14. Coverage	Time period between the beginning of the dataset & its completion; geographical coverage (city, state)
15. Rights	IPR Unistra

All the research resources in the Smirnov's lab are shared between the participants of the project via the protected Seafile cloud space hosted by the University of Strasbourg. They are retrievable by all participants of the project. They are curated and modified by authorised participants (Alexandre SMIRNOV, Christelle GRUFFAZ, Anna SMIRNOVA, Zhen LIAO, Théo MARKEZIC, Tia ALHAJ ABDO). The maintenance of the research infrastructure is ensured by the Lab Manager (Christelle GRUFFAZ) and supervised by the Principal Investigator (Alexandre SMIRNOV).

In the François's lab, data are organised in folders according to the date of analysis, which ensures their easy and unequivocal handling. The files are named by the date in the ISO 8601 format (yyyymmdd). The folders also contain a metadata Excel file containing key information about the experiment, such as the sample ID (e.g. "S07R01), CE-MS/MS data file name and description, the date and time of the analysis, and instrument calibration data.

In the Liu's lab, data on synthesised peptides are organized in folders named "XY#####", where XY are the Creator's initials and ##### is the number of the experiment (continuous numbering). The Creator keeps a laboratory notebook and writes in a free form where and how the peptide synthesis and the analysis results are registered. The original data are conserved on the computers associated with the instruments used to collect them; the spectra are analysed by corresponding software (Table 1) and converted and saved in the PDF format.

CNRS has recently begun to deploy, on voluntary bases and at the scale of entire research units, digital lab books (eLabFTW). This tool would permit linking together CERs, data, metadata, protocols, and research resource IDs, and their sharing between the participants of the project would become even more straightforward. The Leulliot's and Liu's labs have already switched to the eLabFTW lab books. In 2024, the Laboratory Council of UMR7156-GMGM approved the introduction of the eLabFTW lab books in the Smirnov's team. We foresee that this change may require reshaping of the metadata system to avoid redundancy between them (since the digital lab book is expected to contain most of the information currently stored within metadata files, even though in a less formal way). However, the essential external function of metadata to make data FAIR (Turning FAIR into reality. Final report and action plan from the European Commission Expert Group on FAIR data, November 2018) and the necessity to be able to share intermediate data selectively, without disclosing other, unrelated and/or sensitive intermediate data impose the metadata as the primary entities which necessarily have to be complete and self-sufficient. In this perspective, the lab book retains the roles of:

(i) a connecting node between various data, metadata, protocols, and research resources that organises them in experiments and ensures cross-referencing;
(ii) a contextualisation tool that explains the place of data and processes in the research;
(iii) a space for data post-analysis (e.g., plotting, display of adjusted images to visualise specific features, association of non-data objects and links etc) as well as for free-style interpretation and hypotheses.

Finally, the present Data Management Plan (DMP) can be considered as another piece of metadata describing the entire organisation of the data, metadata, protocols, and research resources produced and/or enacted in the frame of the EURICA project. It follows the outline and requirements elaborated in the Science Europe Practical guide to the International alignment of research data management (January 2021), further detailed in Additional guidance to the Science Europe DMP assessment rubric, and is subject to publication on Zenodo under the CC BY Attribution 4.0 International license.

2b. What data quality control measures will be used?

Three cases should be distinguished: collected, re-used, and produced data.

Since all the methods employed to collect data in this project have been extensively characterised and widely used by the scientific community, their quality can be evaluated by compliance with field standards. Those include technical, design, reproducibility, documentation, and digital measures and criteria (Table 7). In addition to these general criteria, further quality checks (implemented in specific metadata and data presentation modes) are applied to select types of data with community-endorsed "minimal information" requirements (see Minimum Information for Biological and Biomedical Investigations) and "best-practice" guidelines, such as microscopy (Community-developed checklists for publishing images and image analyses), qPCR (MIQE), molecular interaction (MIMIx) or enzymatic activity (STRENDA) data.

Table 7. Data quality measures & criteria

Technical	Completeness (all samples are present, all parameters & settings are known & kept constant); functionality of the instrument & reagents (calibration, absence of obvious artefacts, maintenance); authentication &/or certification of internal & external research materials (cell lines, strains, plasmids, antibodies etc); compliance with a registered protocol; compliance of the protocol with established standards; sample integrity (absence of degradation, aggregation, or another undesirable modification); signal-to-noise ratio; dynamic range; saturation; number of events (e.g., in MS); CI of measurement; FDR; resolution; coverage; compliance of the behaviour of standards & controls with previously reported observations ("typical picture" or "ground truth"); for synthetic peptides, quality is evaluated with HPLC & HRMS (purity >95%, molecular mass error <5 ppm, & unambiguous retention time are required)
Design	Presence of negative & positive controls; standards; complementation; optimal design of time or concentration series (within the dynamic range); complete combinatorial design; uniformity of measurement (e.g., all samples on the same gel/membrane or in the same run/session); statistical power & sample size estimation; correspondence between the experimental design & the statistical analysis approach (independence, pairedness, distribution, homoscedasticity, & other assumptions; continuous, count, or categorical data)
Reproducibility	Technical & biological replication; single blinding (where possible); cross-experimenter replication (where possible); cross-batch replication; statistical convergence (absence of overdispersion)
Documentation	The experiment is completely described in the lab book, data follow the conventional vocabulary, are accompanied by properly registered & complete metadata & referenced in CER
Digital	Data can be opened & modified with original software; data can be extracted & converted into a widely accepted format without losing information, quality, or interoperability; unambiguous association between data, metadata & the corresponding research resources (protocol, cell lines, strains, plasmids, oligonucleotides, antibodies, external datasets); fidelity of digitalised images; absence of non-documented or irreversible modifications to original files

In the Smirnov's lab, the main task of data quality control falls first to the experimenter, who ensures the compliance of the protocol and the experimental outcome with the established standards, documents and interprets the experiment (in the lab book and CER), and then to the principal investigator (Alexandre SMIRNOV), who reviews the protocol, data and metadata and re-evaluates their compliance with the community standards and the validity of the interpretation. The Lab Manager (Christelle GRUFFAZ) ensures the identity, digital and physical maintenance of the associated research resources (reagents, protocols, cell lines, strains, plasmids, oligonucleotides, synthetic RNAs, antibodies) and equipment (pipettes, instruments). The experimenter ensures the identity and maintenance of collected data and metadata.

Re-used external data are normally expected to follow the same standards. However, data-specific checks need to be performed and eventual deviations from the norm flagged in order to establish whether the external data in question require censoring. Inclusion guidelines for such data are provided in Table 8. Filtering parameters and the decision to censor select data are systematically justified and included in the metadata. For all types of external data, permanent unique identifiers and duly documented provenance are required for inclusion.

Table 8. Inclusion guidelines for external data

Protein sequence, conservation, phylogenetic distribution	The sequence is complete; experimental design, as reported in the original publication, makes unlikely data contamination; independent sequencing of other isolates of the same strain, other strains of the same species, or other species of the same genus strongly support the sequencing variant; specifics of the genetic code are correctly captured (verified with ExpasyTranslate); NCBI BLAST confirms the identity of the entry & finds highly similar sequences in sister clades; COBALT confirms their clustering against selected outer groups; hallmarks of the protein family are clearly recognisable; the choice of the canonical splice isoform is based on well-described homologues from model species; significantly deviating sequences (COBALT) are inspected individually for misannotation
Protein structure	Quality of the model is evaluated based on the experimental details in the original publication & with the help of NGL Viewer (built in RCSB PDB) for the quality of fit & chain geometry
Gene sequence	The sequence is complete & is translated *in silico* into a protein that satisfies the above criteria
Genome sequence	The genome must be assembled to at least the chromosome level
MS data	Accompanied by proper metadata reporting on provenance, experimental protocol, & original publication
Data from publications	High technical standard, as commonly required in the field; independent corroboration (wherever applicable)

Produced data (i.e., those resulting from analysis of collected or re-used data) should be themselves based on high-quality data, as outlined above, and additionally satisfy a set of specific criteria. Those include:

a) unequivocal association with the corresponding primary data (by including raw data and referring to their unique identifiers or databases),
b) metadata with clear information about the date and the analysis protocol (including definitions, prediction parameters, information about binning, thresholding, normalisation, transformation, statistical analyses, software etc) captured as within-file documentation or in a separate README file,
c) respecting field conventions for each kind of analysis or justification of deviations from them,
d) compliance with the assumptions and requirements of adopted statistical analyses and visualisation modes,
e) sample inclusion and exclusion criteria,
f) portable and convertible format permitting facile re-run or alternative visualisation with desktop or online tools,
g) proper organisation in an "analysis" subfolder with a transparent structure that captures the processing workflow.

3. Storage & backup during the research process

3a. How will data & metadata be stored & backed up during the research?

In the Smirnov's lab, data and metadata are stored on the office computers and monthly backed up on two 1-TB hard drives which are kept in a different location. Additionally, primary data are stored on original instruments and on the servers of external services who generated them (e.g., ESRF, SOLEIL, Plateforme Protéomique Esplanade), ensuring that several copies of most important data are preserved in geographically distinct locations (see below). The Seafile cloud storage of the University of Strasbourg is used to store and share between collaborators the research resources of the laboratory and some data. A more reliable, permanent, and spacious solution for heavy data storage is expected to be provided by the Data Centre of the University of Strasbourg.

In the François's lab, data and metadata are stored and backed up on the office computers by each staff member involved in the project. The strategy put in place consists in making two weekly back-ups of the data and metadata on a Network Attached Storage (NAS) dedicated to the project and on an external hard disk of at least 1 TB provided to each project staff member. This hard disk is stored in a location other than the laboratory for security reasons. The metadata are also stored on the computers associated with the instruments used to acquire them. Similarly, two weekly back-ups of these data are made on the NAS and on an external hard drive dedicated to each machine. To avoid keeping these back-ups in the same place, two separate locations are used for the NAS and the various hard drives.

In the Leulliot's lab, data and metadata are stored on the CNRS cloud sDrive. The Smirnov's lab is also considering this storage alternative for some data types.

In the Liu's lab, data and metadata are stored on the computers associated with the instruments used to collect them, in folders assigned to individual Creators, and additionally on the Creators' office computers.

All primary data are read-only to avoid unintentional modification.

The storage of intermediate data and metadata is regulated by the French law acts (Circulaire du 2 novembre 2001 relative à la gestion des archives dans les services et établissements publics de l'Etat).

The ESRF Data Policy 2024 (version 14/10/2023) makes ESRF the custodian of the data stored at the ESRF. The ESRF guarantees archiving of raw, processed, and auxiliary data for at least 5 years (and up to 10 years, if possible) and will store all metadata indefinitely, but it reserves the right to restrict the storage period or data sets in consultation with the beamline responsible for the high data rate instruments. For high data volume experiments, lossy compression may be used to store data, if required because of raw data storage issues (on approval by the Principal Investigator, Nicolas LEULLIOT). The ESRF declines responsibility for unavailability or loss of data or analysis software.

The SOLEIL Data management policy (DSI-DU-P-0056, October 2, 2018) makes SOLEIL the custodian of the raw data and the associated metadata. SOLEIL will make its best to ensure storage, curation, and access to experimental data and metadata during the duration of the proposal but cannot fully guarantee them from technical or human mistakes. Raw data and metadata are migrated or copied to long-term storage facilities, that could be sub-contracted, for long-term storage. Reduced data and their metadata and processed data from interim analysis steps are not curated by SOLEIL for long-term. Access to data can be temporarily limited or precluded due to maintenance or overhaul of services or failure of third-party providers. SOLEIL declines responsibility in case of loss, damage, or unavailability of data, metadata, or results. If SOLEIL decides to no longer act as a custodian of the data, it will inform the Main Proposer and Principal Investigators (Nicolas LEULLIOT) and provide them with effective means to make a copy of the raw data, metadata, and results.

3b. How will data security & protection of sensitive data be taken care of during the research?

The recoverability of the data is ensured by the multiplicity of stored copies resulting from monthly backups of all data. In the event of an incident, a complete copy of all the data will be reconstituted from backup on a reliable machine.

The access to the research infrastructure is open to all GMGM researchers involved in the EURICA project. The access of the GMGM team members to the data is controlled by the Principal Investigator (Alexandre SMIRNOV) or directly by the Creators of the data. External collaborators receive relevant data and metadata upon request handled by the Principal Investigator (Alexandre SMIRNOV) via protected Seafile links. This policy is reciprocal. Non-collaborators in general do not have access to the data during the research, unless those are already deposited in open repositories and/or published (see section 5). Exceptions are handled individually by the Principal Investigator (Alexandre SMIRNOV) upon consultation with collaborators. Lab books are stored in a securely locked team's office; only in exceptional cases, the Principal Investigator (Alexandre SMIRNOV) may authorise their removal from the laboratory by Creators or other authorised participants of the EURICA project.

The access to the electronic lab book (Leulliot's and Liu's labs) is secured on a by-user basis for each experiment and accessible through the use of a personal security

key and a password. The hard disks used for data storage are encrypted to prevent unauthorised access to the data in case of theft.

No sensitive data were treated or are expected to be treated in the frame of this project.

General guidelines for data handling, protection, and sharing are governed by a specific policy of the University of Strasbourg, as a member of the SupDPO network (La Vie d'une Donnée au regard des réglementations "CRPA", "RGPD" et "Patrimoine", v1.0, June 2019), and by the CNRS Policy of Security of Information Systems (PSSI).

Data collected at the synchrotron ESRF are subject to the ESRF Data Policy 2024 (version 14/10/2023), which specifies, among other provisions, that all data and metadata curated at ERSF are subject to the data protection legislation applicable in France. While the ESRF staff may access data to improve the performance and processes of the facility, it guarantees their confidentiality throughout the entire embargo period.

Data collected at the synchrotron SOLEIL are subject to the SOLEIL Data management policy (DSI-DU-P-0056, October 2, 2018), which regulates the access to, re-use, and publication of the data (see sections 4 and 5). SOLEIL makes the Main Proposer (Nicolas LEULLIOT) responsible for the compliance of metadata with the French and European data protection regulations regarding personal data. Authorised SOLEIL staff has unlimited and unrestricted right to use data and metadata to the extent necessary to curate them and make them available under their data management policy; it guarantees the confidentiality of data with restricted access.

4. Legal & ethical requirements, code of conduct

4a. If personal data are processed, how will compliance with legislation on personal data & on security be ensured?

No personal data were treated or are expected to be treated in the frame of this project.

4b. How will legal issues, such as intellectual property rights & ownership, be managed? What legislation is applicable?

Ownership of the data collected and produced in this research is regulated by the Code de la Propriété Intellectuelle, Code de la Recherche and, more specifically, the Loi pour une République Numérique (LOI n° 2016-1321 du 7 octobre 2016) under the French law. The research data obtained primarily in the Smirnov's or the François's labs are the property of the University of Strasbourg while those obtained primarily in the Leulliot's or the Liu's labs are the property of the University Paris Cité (Code de la Recherche; September 19, 2024; Partie réglementaire, livre V, titre III, chapitre III). The access to the data is controlled by the corresponding Principal Investigators (Alexandre SMIRNOV, Yannis FRANÇOIS, Nicolas LEULLIOT, Wangqing LIU). The intellectual property rights for any inventions stemming from activities in the frame of the EURICA project are regulated by the French law (Code de la propriété intellectuelle; October 3, 1996; section R611-12).

Authorship and intellectual property rights on all published documents belong to their authors (physical persons), as stipulated by the French law (loi n° 2006-961 du 1 août 2006 relative au droit d'auteur et aux droits voisins dans la société de l'information; see also the comprehensive Guide pratique de la publication en ligne et de la réutilisation des données publiques ("Open data ») elaborated by CNIL and CADA) and internationally promoted through the Rights Retention Strategy (RRS) by cOAlition S (ANR is a member of cOAlition S and aligns its Open Science policy accordingly).

Ownership of primary data obtained at the synchrotron SOLEIL is regulated by the SOLEIL Data management policy (DSI-DU-P-0056; October 2, 2018) and should be interpreted in the sense that all derived products (results, data interpretation, IPR), being performed by participants of the EURICA project, remain the property of the University Paris Cité (Code de la propriété intellectuelle; October 3, 1996; section R611-12).

The ESRF Data Policy 2024 does not define data ownership explicitly but otherwise complies with the French legislation.

The data deposited in the ProteomeXchange Consortium remain the property of the Principal Investigator and of the PRIDE Submitter even after their public release, as specified in the Data submission guidelines for the ProteomeXchange Consortium (version 3.0.1; October 13, 2019).

IDR encourages authors to submit imaging datasets under the CC BY license (with licensing information included in the accompanying metadata) to facilitate their widest re-use (Submission Guidelines v1.0; March 2017). Authors retain copyright, while IDR receives a license to publish the data on behalf of the authors.

BioImage Archive policies (version 1.1) and EMPIAR policies (version 1.0.8; October 23, 2023) make deposited data available under the CC0 license by default. This is potentially problematic since, unlike CC BY, CC0 does not protect author rights and leaves citation of the data at re-users' discretion. However, BioImage Archive policies (version 1.1) acknowledge that some deposited datasets are available under different licenses (indicated as dataset attributes). EURICA will strive to deposit microscopy data under the CC BY license. BIA has a removal procedure if the ownership or provenance of datasets are called into question (BioImage Archive Policies, version 1.1; see also section 5a).

Zenodo General Policies (version 1.0) recognise the ownership for the parties prior to submission and do not imply any transfer of property rights.

Recherche Data Gouv uses the open-access etalab2.0 license by default, but permits the use of alternative/additional licenses, including CC BY. etalab2.0 is designed to be compatible with CC BY.

There are no restrictions on the re-use of third-party data deposited to public databases within the perimeter and at the scale foreseen in this study (see section 1), as guaranteed by French law (article L342-3 du Code de la Propriété Intellectuelle) that implements the Directive 96/9/EC of the European Parliament and of the Council. Specifically, no third-party intellectual property or *sui generis* rights were or are expected to be affected.

The data deposited in the wwPDB are open-access without copyright restrictions. However, the wwPDB Deposition Policies and wwPDB Biocuration Procedures (version 5.3; April 2024) reserve the authorship rights to Entry Authors, Contact Authors and Citation Authors. In the frame of the EURICA project, these rights are reserved to Nicolas LEULLIOT.

The EMDB policies recognise the entry authorship for the Principal Investigator(s) and any persons assigned by the Principal Investigator(s) and listed by the depositor at submission. It also distinguishes Contact Authors and Citation Authors. In the frame of the EURICA project, these roles are fulfilled by Nicolas LEULLIOT. The Principal Investigator ensures the consent of all the authors of the deposited entries.

The EMPIAR policies (version 1.0.8; October 23, 2023) reserve the scientific ownership to the Principal Investigator(s) who are scientifically responsible for the study that generated the data. It must be the same person as the owner of the corresponding EMDB entry, i.e., in the frame of the EURICA project, Nicolas LEULLIOT. The policies also recognise authorship for any person, designated by the owner, who has in any way contributed to the data in an EMPIAR entry. The corresponding author of the article describing the corresponding EMPIAR entry must be one of the authors of this entry. The Principal Investigator is responsible for ensuring the consent of all the authors of the deposited entries. The names and the ORCIDs of the authors are made public upon release.

Furthermore, the re-use of deposited third-party data is subject to specific rules, as defined by the corresponding repositories.

The ProteomeXchange Consortium data re-use requirements are outlined in the Guidelines for handling ProteomeXchange reprocessed datasets (version 1.0.2; September 2019): the unique dataset identifier (PXD) should be cited, reprocessed data linked to the source PX dataset, reprocessed PX XML files should respect specific formatting.

The wwPDB Deposition Policies and wwPDB Biocuration Procedures (version 5.3; April 2024) define the authorship rights for re-refined entries and require acknowledgement and citation of the original PDB entry (and the associated publication, if available) in the re-refined PDB entry.

The SOLEIL data re-use requirements are outlined in the SOLEIL Data management policy (DSI-DU-P-0056; October 2, 2018): wherever appropriate, the Main Proposer or the Principal Investigator of the dataset should be contacted in order to inform them, propose collaboration or co-authorship; the source of the data, citing the unique persistent identifier and any publications linked to the same raw data, should be acknowledged.

Section 5b more specifically deals with the open-access publication of datasets and documents associated with the EURICA project and cites the corresponding legal norms and community standards.

4c. What ethical issues & codes of conduct are there, & how will they be taken into account?

Not applicable (no human or animal subjects, communities, their cultural and/or genetic resources, personal or other ethically sensitive data were used in the study).

5. Data sharing & long-term preservation

5a. How will data for preservation be selected, & where will data be preserved long-term (for example a data repository or archive)?

At this stage of the project, when a significant body of intermediate data have already been obtained and prepared for publication, and the quality and value of each piece of data can be evaluated with more confidence and in a broader context, the selection of data for ultimate archiving can be initiated. To this end, the intermediate data are parsed in four groups:

1) actively used data, independently of their quality, associated with or potentially important for ongoing or future research tasks within the EURICA project or for other ongoing research projects;
2) high-quality ready-to-publish data;
3) other high-quality data which are not expected to be used in the foreseeable future but are required to fully document the research process and, if needed, support published data or priority claims;
4) low-quality data without significant scientific value which are not expected to be used in the foreseeable future.

The destiny of each category of data is different.

The **Group 1 data** are not subject to archiving and continue to be managed, as described in the section 3.

The **Group 2 data** are formatted and deposited to appropriate trusted data repositories, according to Science Europe recommendations (Practical guide to the International alignment of research data management, January 2021) and established standards in the field. Key mass spectrometry data are deposited in the ProteomeXchange Consortium (via the PRIDE repository). Microscopy data are deposited to BioImage Archive (BIA) or Image Data Resource (IDR). X-ray crystallography and cryo-EM data are deposited to wwPDB. Cryo-EM data will be deposited in EMPIAR and EMDB. Source codes are deposited in GitHub. The deposited data follow the corresponding policies and guidelines for formatting, storage, and sharing, as outlined in sections 4b, 5b, 5c, and 5d. The deposition to these selected databases is free of charge and, at the scale of the EURICA project, without effective size constrains. All the data are released in open access. Foreseeable uses of the data include, but are not limited to, follow-up, replication or comparative studies, re-analysis, benchmarking, algorithm training, and database creation. Below are outlined some specific rules governing data deposition, archiving, and withdrawal in the repositories concerned by the present DMP.

The ESRF acts as a custodian of the data stored at the ESRF. ESRF Data Policy 2024 (version 14/10/2023) foresees archiving of all raw, processed, and auxiliary data for at least 5 years (and up to 10 years, if possible) and storing the associated metadata indefinitely, but it may restrict the storage period or data sets in consultation with the beamline responsible for the high data rate instruments. Foreseeable uses of the data include, in addition to the ones specified above, optimisation of processes and performance of the synchrotron.

SOLEIL acts as a custodian for all raw data and associated metadata collected at the synchrotron. It ensures long-term storage and curation of the data and the

associated metadata. These data and metadata follow the SOLEIL Data management policy (DSI-DU-P-0056; October 2, 2018) for formatting, storage and sharing. SOLEIL does not curate reduced or processed data and the associated metadata. SOLEIL stores but does not curate results issued from analyses performed on raw data and metadata using SOLEIL means. Foreseeable uses of the data include but are not limited to re-analysis, benchmarking, and algorithm training.

Data submission guidelines for the ProteomeXchange Consortium (version 3.0.1; October 13, 2019) regulate the deposition of mass spectrometry data. The ProteomeXchange Consortium similarly permits withdrawing unreleased datasets by the authors' decision. The withdrawn datasets will still be present in the list of unreleased entries, and their internal copy will be kept by ProteomeXchange Consortium but will remain inaccessible to external users. Released datasets associated with published research cannot be withdrawn (unless there are exceptional circumstances, such as major errors, retraction of the corresponding article, or violation of property rights). Preprints are not considered publications under the ProteomeXchange Consortium policies, and the repository does not encourage releasing datasets prior the acceptance of the final paper for publication. PRIDE (as part of the ProteomeXchange Consortium) in its General guidelines for submitted datasets (version 1.0, May 2022) aligns with these policies and engages to maintain all published datasets publicly accessible as part of scientific record. The PRIDE team also offers the creation of additional metadata (e.g. Sample and Data Relationship Files, mzTab and mzIdentML files) to facilitate the re-use of deposited datasets.

wwPDB Deposition Policies and wwPDB Biocuration Procedures (April 2024, version 5.3) describe the data deposition procedure and the principles of data preservation and curation by the repository. They establish the quality standards, formats, and the minimal set of metadata accompanying each submission. wwPDB requires that, in cases where the submitter seeks to deposit a 3D model for a cryo-EM structure, the map volumes are first deposited in EMDB (see below). Only the authors can access unreleased datasets, but their title, authorship, status, PDB ID, experimental data status and sequence availability are visible to anyone in the PDB archive. wwPDB *de facto* distinguishes 'revisions' to existing data and 'new data', depending on whether the corresponding new experimental data have been produced before or after deposition. A 'revision' may include updates of coordinates, structure factors, and related header information introduced prior to release. By contrast, if the authors wish to replace their original dataset with other experimental data, they must withdraw the former and obtain a new PDB ID. Released datasets may be subject to minor or major revisions and in this case are versioned under the same PDB accession code. If new experimental data are used, a new PDB entry is created that supersedes the previous one. Obsolete entries remain available to the public through the ftp archive. The authors can withdraw their unreleased entries if they have not been cited in a publication. In this case, the latest version of the processed files will be made available to the authors for eventual re-deposition. Withdrawn entries will remain in the list of unreleased entries in the PDB archive. Additionally, entries that have been on hold for more than 1 year should be either released or withdrawn. Furthermore, problem structures, as identified by wwPDB biocuration staff, can be withdrawn by the end of the 1-your on-hold period, unless the issue is resolved or the structure is published (in the latter case it is released with a database_PDB_caveat record). If a publication appears that cites a recently withdrawn entry, its withdrawn status may be reversed. wwPDB reviews the entire archive on regular bases and

remediates PDB data as required without contacting the authors. Remediation results in versioning of the concerned datasets.

EMDB is a wwPDB core archive and follows the same general policies. In addition, it has its own comprehensive policy regarding the deposition and curation of three-dimensional cryo-EM data where the accepted data types, formats, quality requirements, and metadata are described. The Principal Investigator (Nicolas LEULLIOT) is responsible for the EMDB entries he deposited; he can request changes to the entries and will notify wwPDB if a dataset is obsoleted or withdrawn. Validation reports can be generated by wwPDB to provide the quality assessment of EM maps and 3D models. Although after deposition entries are normally locked for further modifications, they can be changed prior to release upon request to the wwPDB staff. After release, changes are possible under special circumstances, upon request to the wwPDB staff, and may include modifications in metadata, addition of new data, or deletion of accidentally uploaded files. If the primary map of a dataset is updated, the corresponding released EMDB entry is obsoleted and a new one is created to supersede it. Obsoletion of an EM entails the obsoletion of the derived atomic model. The obsoleted entry is available in a separate section of the public EMDB archive. By contrast, for withdrawn EMDB entries, no data or metadata are available in the public EMDB archive. A list of withdrawn entries is publicly available. A structure cited in a publication cannot be withdrawn and, if required, will undergo obsoletion. Preprints are considered publications under this policy.

2D images for the same cryo-EM experiments can be deposited to EMPIAR which has its own policies (version 1.0.8; October 23, 2023) regarding quality requirements and formats for deposited data, auxiliary data, and metadata and the curation procedure. EMPIAR entries are normally tied to the corresponding EMDB entries and/or publications. Depositions are locked after submission but can be modified by the depositor upon request to the EMPIAR staff. Unlocked entries are automatically resubmitted in 1 month, locked, and, at discretion of the EMPIAR staff, either removed or released. A dataset can be released when all requirements have been met, upon approval by the depositor. Alternatively, the release of an EMPIAR dataset can be triggered by the release of the corresponding EMDB entry, or by the publication of the corresponding article (preprints are considered publications under these policies). If the corresponding EMDB entry has not been released 1 year post deposition, the corresponding EMPIAR entry is deleted (a one-time 6-month extension can be granted upon the owner's request). The maximal hold period for an unreleased EMPIAR entry is 5 months; after 12 months, if the situation does not change, the entry is withdrawn. After release, changes in entries are possible under special circumstances and when strictly necessary upon request to the EMPIAR staff and may include modifications in metadata (e.g. to cite a publication), addition of new data, or deletion of accidentally uploaded files. A version history is maintained and distributed in the public archive. The obsoletion of released entries and the withdrawal of unreleased entries follow the same principles as in EMDB. A released EMPIAR entry can be removed after publication under exceptional circumstances (retraction, decision of an official body investigating scientific misconduct).

BioImage Archive (BIA) functions as a permanent public repository of imaging data associated with peer-reviewed publications or of value beyond a single experiment. It also provides archiving services for other bioimaging databases, e.g. IDR. It supports FAIR Sharing and implements REMBI guidelines for FAIR data. BioImage

Archive Policies (version 1.1) provide general guidelines about deposition of datasets. BIA allows for up to 3 months of transient storage for uploaded data, after which the dataset is either formally submitted for permanent deposition or deleted. The data can be released either right away or upon article publication (and in any case no later than 2 years after submission). All deposited data are permanently accessible for scientific record. They can be amended or updated, or even removed from visibility by search (in case of an erroneous submission). BIA reserves the right to obsolete datasets due to integrity, correctness, ownership, or provenance issues (e.g. article retraction or other official decision of integrity bodies). In either case, the data in question will nevertheless remain accessible via their PID. The quality and accuracy of the deposited data are the responsibility of the authors.

Image Data Resource (IDR) is a more specialised repository focused on reference data from published (or only submitted) studies with particularly rich metadata to enable their re-use by third parties (Williams et al., *Nat Methods*, 2017). IDR Submission Guidelines (v1.0, March 2017) require the datasets to be complete and integrated with other databases and biomolecular resources via PIDs. IDR supports a wide variety of proprietary and open file formats (with the preference for the latter), including those implemented in the EURICA project (section 1a).

BIA scope and submission to related resources explains the difference in scope between BioImage Archive and other, more specialised repositories, such as IDR, EMPIAR and EMDB, thus guiding the choice of repository for different kinds of imaging data. BIA and IDR actually overlap on this point and work together to coordinate and link submissions between the two repositories; more specialised IDR deposits are imported into BIA. BIA plans to provide a possibility of parallel submission to both the databases.

Zenodo permits the deposition of any kinds of research data in any format under any access conditions. Zenodo does not perform quality checks, does not guarantee usability or understandability of deposited data. This repository stores metadata in the JSON format, with the possibility of export to other widely used formats, e.g. Dublin Core, via OAI-PMH. Submitters themselves specify the desired license in the process of submission, which determines use and re-use possibilities. Metadata are licensed under the CC0 license. Zenodo General Policies (version 1.0) reserve the right to remove datasets and revoke their DOI in cases where the datasets in question do not fall into the scope of the repository or violate its terms of use. Under exceptional circumstances, the submitter can request withdrawal of one of his datasets, which requires full justification. If the withdrawal is granted, the dataset page is replaced by a "tombstone" page, but the original DOI and URL are retained.

Finally, the general-use French national platform for research data, Recherche Data Gouv, permits deposition of any types of data for which no trustworthy repositories currently exist as well as cataloguing/linking to data deposited in other trusted national or international repositories. The deposited data must be complete and accompanied by structured metadata. Recherche Data Gouv does not store intermediate data and does not perform scientific quality checks; however, it does provide curation (including post-deposition, long-term curation) and formal validation of completed datasets submitted to the repository. Recherche Data Gouv policies enable deposition either in open or reserved access (the latter is possible for datasets that constitute a legal exception to the open-access principle, e.g.

professional or commercial secret, personal data). As in the other repositories mentioned above, data deposited to Recherche Data Gouv are findable and citable. Curators reserve the right to delete provisional submissions, modify submissions, or withdraw published datasets or their versions from circulation, upon justification.

Other trusted repositories will be considered in agreement with Science Europe guidelines and journal recommendations.

Some Group 2 data (e.g., images, tabulated or graphed data) are directly published in the corresponding articles. The present DMP is deposited and updated, as described in section 6a, on Zenodo under Creative Commons Attribution 4.0 International (CC BY) licence and communicated to ANR as a specific project deliverable.

As a general rule, the default license for deposited EURICA data is, wherever possible, Creative Commons Attribution 4.0 International (CC BY).

The versioning of deposited data generally follows RDA recommendations (Klump & al. Principles and best practices in data versioning for all datasets big and small, version 1.1; April 6, 2020). In particular, ProteomeXchange Consortium announces new versions of deposited PX datasets and enables the preservation and tracking of all versions of a dataset in ProteomeCentral (Data submission guidelines for the ProteomeXchange Consortium, version 3.0.1; October 13, 2019). Moreover, ProteomeXchange Consortium distinguishes the terms 'revision' (previously 'version') and 'reanalysis' of PX datasets. The former refers to technical updates on data or metadata without changing the original intent of the submission; new revisions render the previous ones obsolete. By contrast, 'reanalysis' is reprocessing of original raw data to achieve new results; it does not imply that other reanalyses are obsolete and does not supersede them. The ProteomeXchange Consortium has a specific policy regarding PIDs associated with reprocessed datasets (Guidelines for handling ProteomeXchange reprocessed datasets, version 1.0.2; September 2019).

Zenodo versions all deposited documents and assigns each new revision a unique DOI.

Recherche Data Gouv has its own versioning policy that permits adding, removing or deleting files, adding or modifying metadata, and changing license terms. Any change results in republication of the dataset as a version under the same DOI (DOIs are only linked to the latest version of the dataset). The repository distinguishes minor (labelled 1.1, 1.2 etc) and major revisions (labelled as 2.0, 3.0 etc). The latter option is used to accommodate any changes in files or citations.

The **Group 3 data** undergo archiving. They may be later re-used by the research team or collaborators and shared with third parties (members of the research community or governing bodies), following the modalities described in section 3b, for research purposes, or to ensure the integrity of the research record, or to support priority claims. Foreseeable uses of such data include, but are not limited to, publication, follow-up, replication or comparative studies, re-analysis, benchmarking, algorithm training. The expected volume of the collected and produced data (see Table 1) permits their long-term preservation on hard drives or in the Data Centre in house. Since no personal data were collected or are expected to be collected in this research, there is no legally imposed limit on the duration of data preservation.

The archiving procedure in the Smirnov's lab includes three components: (i) evaluation and, if needed, fixing of the quality and completeness of the dataset (see section 2b), (ii) copying and transferring the dataset in question to the archive hard drive and (iii) creation of the corresponding records management metadata. All archived data and the associated metadata are by default read-only. The archive hard drive is preserved in a separate location. The archived copy of the dataset is accompanied by a records management metadata file in a TXT format, based on the Dublin Core model and named according to the following conventions:

NNE#####_RM_metadata.txt

where NNE##### is the number of the archived experiment in CER (see section 2a). This file is organised as described in Table 9.

Table 9. Records management metadata file organisation (for archiving)

Entry	Description
1. Title	"Records management metadata for the experiment NNE#####"
2. Creator	Person (name, ORCID) who created the metadata file
3. Subject	"Records management metadata"
4. Description	Concise & dated description of significant events accompanying the archiving procedure (e.g., revision of metadata; format conversion, movement or suppression of select data & the reasons for this) & the subsequent history of the archived dataset (e.g., copying, re-use in another experiment in house or by authorised external users, publication etc)
5. Publisher	"Université de Strasbourg"
6. Contributor	Person(s) (names, ORCIDs) &/or institution(s) who perform archiving and manage the archive
7. Date	Date of the last modification
8. Type	"CER experiment dataset"
9. Format	Name of the folder in which the archived data are classified ("NNE#####")
10. Identifier	ID of the experiment ("NNE#####")
11. Source	Primary or secondary (in the case of data re-use or re-analysis)
12. Language	"English"
13. Relation	IDs of other experiments &/or data or metadata to which the present dataset is related in terms of deliverables, continuation, replication, modification, re-use or re-analysis
14. Coverage	Time period between the beginning of the dataset & the last modification of its records management metadata; physical location of the archive
15. Rights	"IPR Unistra"

Another copy of the archived dataset continues to be preserved on the original hard drive but is moved to a dedicated folder "Archive"; it is not subject to backup any more. Its place in CER is taken by a read-only records management metadata file identical to the one in the archive, to enable localisation of the archived data and reconstitution of their record management. In case the dataset needs to be extracted from the archive, its records management metadata file will be updated accordingly to capture this event, its circumstances, intervening persons, and reasons.

The **Group 4** data are suppressed. The suppression procedure includes three components: (i) deleting the data in question from the corresponding CER folder while preserving all the metadata of the suppressed dataset, (ii) transferring the folder with the metadata to the archive, and (iii) creation of the corresponding records management metadata. The latter follows the Dublin Core model, as described in Table 10.

Table 10. Records management metadata file organisation (for suppression)

Entry	Description
1. Title	"Records management metadata for the experiment NNE#####"
2. Creator	Person (name, ORCID) &/or institution who created the metadata file
3. Subject	"Records management metadata"
4. Description	Concise & dated description of the reasons & the circumstances of the dataset suppression & the subsequent history of the remaining metadata (e.g., copying, damage, or loss); person (name, ORCID) who authorised the suppression (usually, Principal Investigator, Alexandre SMIRNOV)
5. Publisher	"Université de Strasbourg"
6. Contributor	Person(s) (names, ORCIDs) &/or institution(s) who performed the dataset suppression
7. Date	Date of the last modification
8. Type	"CER experiment dataset"
9. Format	Name of the folder in which the archived data are classified ("NNE#####")
10. Identifier	ID of the experiment ("NNE#####") & a complete list of all suppressed files
11. Source	Primary or secondary (in the case of data re-use or re-analysis)
12. Language	"English"
13. Relation	IDs of other experiments &/or data or metadata to which the present dataset is related in terms of deliverables, continuation, replication, modification, re-use or re-analysis
14. Coverage	Time period between the beginning of the dataset & the suppression; physical location of the archive
15. Rights	"IPR Unistra"

As in the case of archived datasets, the place of the suppressed folder in CER is taken by a read-only records management metadata file identical to the one in the archive, to enable the localisation of the remaining metadata of the suppressed dataset and the reconstitution of their record management.

The sorting of intermediate data and the decision to eventually archive/suppress them will be taken on a monthly basis and will only concern the datasets (CER experiments) that have not been in use for at least one year. The decision to suppress a dataset must be systematically approved by a hierarchical authority (in this case, the Principal Investigator, Alexandre SMIRNOV).

Note a difference in the granularity of the suppression procedure between the Group 3 and the Group 4 data. Group 4 datasets are suppressed entirely, whereas individual files from Group 3 datasets may be suppressed prior archiving, following a similar suppression procedure, if they are damaged or devoid of scientific value, provided that the suppression does not affect the completeness of the archived dataset. In either case, suppression is systematically motivated and authorised in the accompanying records management metadata.

Long-term curation of the archive is ensured by the Lab Manager (Christelle GRUFFAZ) and by the Principal Investigator (Alexandre SMIRNOV). The archived datasets are not subject to suppression; their preservation is assured by the University of Strasbourg. In the case where the activity of the latter is ceased, the archived datasets should move to an appropriate state archive, according to the French law regulations (Code du patrimoine, article L212-5).

The legal requirements for the re-use of published data, data archiving, and destruction are fixed by the French law and can be found in the Guide pratique de la publication en ligne et de la réutilisation des données publiques ("Open data") elaborated by CNIL and CADA, in La Vie d'une Donnée au regard des réglementations "CRPA", "RGPD" et "Patrimoine" (version 1.0; Juin 2019) compiled by SupDPO, and in Ouverture des données de recherche – Guide d'analyse du cadre juridique en France (version 2, December 2017). Specifically, creation, conservation, and protection of archives is regulated by the Régime général des archives – Chapitre 2 ("Collecte, conservation et protection" - Articles L212-1 to L212-29) of the Code du Patrimoine and general guidelines are given in the Circulaire du 2 novembre 2001 relative à la gestion des archives dans les services et établissements publics de l'Etat. They are also regulated by international norms (see Maday & Taillefer. Les métadonnées du records management du point de vue des norms ISO. In: *La Gazette des archives*, n°228, 2012-4). Namely, ISO 23081-1 requires documentation of any changes introduced in the content, the context, or the structure of the record; ISO 15489-1 requires the approval of a hierarchical authority for any permanent destruction of records. Importantly, while the French law requires all data produced by public institutions to be publicly accessible (with a few exceptions that are irrelevant in the context of the EURICA project), this only concerns completed ("*achevés*") datasets, i.e. Group 1 data under this DMP are not subject to communication (see an overview of these norms in Ouverture des données de recherche – Guide d'analyse du cadre juridique en France, version 2, December 2017).

5b. How & when will data be shared? Are there possible restrictions to data sharing or embargo reasons?

Access to completed datasets and documents associated with the EURICA project is by default open and free, in the spirit on European, French, supDPO and CNRS recommendations and legal acts (Turning FAIR into reality; November 2018; Deuxième Plan national pour la science ouverte; July 2021; Vie d'une Donnée au regard des réglementations "CRPA", "RGPD et "Patromoine", version 1.0; June 2019; Guide pratique de la publication en ligne et de la reutilisation des données publiques ("Open Data"), CNIL, CADA; Ouverture des données de recherche – Guide d'analyse du cadre juridique en France, version 2, December 2017; CNRS Research data plan; November 2020). This said, the data discoverability and sharing are intimately connected with the processes of publication and eventually patenting. Access to intermediate data is restricted to the participants of the project, but can be granted to collaborators or prospective collaborators outside the consortium by the Principal Investigators (Yannis FRANÇOIS, Wangqing LIU, Nicolas LEULLIOT, Alexandre SMIRNOV). Unpublished data and metadata as well as lab books and data arrays susceptible to become object of IPR can be shared with third parties upon request directly handled by the Principal Investigators upon consultation with collaborators and eventually SATT Conectus Alsace or the University Paris Cité PI office. Pursuing a patent constitutes a reason for which the relevant data may not be released until the patent is granted or rejected.

Prior to article publication, relevant datasets are deposited in dedicated repositories (see section 5a) and become discoverable by their unique identifiers (see section 5d) and key words through search engines. Upon publication (or, in some journals, prior to publication), they are rendered open-access (under the CC BY license, wherever possible) and follow the policies of the corresponding repositories.

The data, metadata and results obtained at the synchrotron SOLEIL are discoverable via a searchable on-line catalogue (restricted to registered users). The embargo period is defined in the SOLEIL Data management policy (DSI-DU-P-0056; October 2, 2018) as 3 years after the end of the beamtime (default), and cannot exceed 5 years. It can be shortened upon request from the Principal Investigator (e.g., if the corresponding results are published). Past this time, all the data and metadata are automatically rendered open-access under the CC BY licence. Access to the data during the embargo period is controlled by Nicolas LEULLIOT. Additionally, §4.18 of the SOLEIL Data Management Policy grants SOLEIL access to all collected data and metadata for curation and sharing purposes.

The ESRF data policy 2024 (version 14/10/2023) makes all non-proprietary data obtained as a result of peer-reviewed access to the ESRF open access after a 3-year embargo period since the end of the experiment (during which only the Principal Investigator can access the data). The experiment report submitted by the user after each beam time session becomes part of the auxiliary data and is subject to the same embargo rules. The Principal Investigator can extend the embargo period by submitting a justified written request to ESRF; the ESRF Directors of Research make a decision as to whether to grant such an extension or not. Data can be released before the 3-year embargo period, and EURICA will make every effort to follow this path. Data are released under an open license. Access to data, results, and

metadata is via the ESRF data portal. Raw and processed data are accessed on the same basis.

wwPDB policies and processing procedures document (version 5.3, April 2024) distinguish three data release situations: REL, HPUB, and HOLD. REL entries are released as soon as the processed files have been approved by the authors (and no later than three weeks after the validation report is made available to the authors, provided there are no outstanding issues with the entry); citation information is not required for the release. HPUB and HOLD entries are placed on hold until publication (HPUB) and/or up to 1 year after deposition; they can be released upon request by the authors or by the journal. If past this date the publication does not follow, the authors are invited either to publish or to withdraw the entry. Once the wwPDB is aware of a publication in any form, it will automatically release the entry. Any revision of released entries is managed under the PDB archival versioning system. Unpublished and unreleased entries can be withdrawn by the authors, who receive the latest version of the processed files for eventual redeposition. Problem structures should be fixed by the authors (in consultation with the wwPDB). Otherwise, unpublished problem structures are withdrawn by the end of the 1-year hold period (the withdrawn status can be reversed if a publication appears), while published problem structures are released with a database_PDB_caveat record. In all cases, experimental data and coordinates can only be released simultaneously. Files scheduled for release undergo a final data integrity check. Data are released weekly. EMDB implements very similar policies.

The EMPIAR policies (version 1.0.8; October 23, 2023) distinguishes five data release situations: REL, EMDBPUB, HPUB, HPRE, and HOLD. REL datasets are released as soon as the annotation is complete and approved by the depositor. EMDBPUB means release triggered by the release of the corresponding EMDB entry. HPUB and HPRE correspond to release upon the appearance of a primary publication or a preprint, respectively. For EMDBPUB, HPUB, and HPRE cases associated with EMDB entries, failure to release the corresponding EMDB entries entails the deletion of the EMPIAR datasets 1 year after the EMDB deposition (with a possible extension of up to 6 months, if justified by the owners). HOLD applies only to datasets for which there are no corresponding EMDB entries; this option delays the release of an EMPIAR dataset up to 1 year (with a possible extension of up to 6 months, if requested by the owners on reasonable grounds).

In the case of the ProteomeXchange Consortium (Data submission guidelines for the ProteomeXchange Consortium, version 3.0.1; October 13, 2019), unreleased data are by default private (password-protected); access to them is controlled by the Principal Investigator (Alexandre SMIRNOV) and the Submitter. Access to unreleased data can be granted to third-parties upon consultation with collaborators. It is systematically granted to journals for the convenience of referees. Upon the publication of the original article, the corresponding datasets are immediately released under the CC0 licence without embargo (note, however, that ProteomeXchange Consortium envisages moving to the CC BY license as default). PRIDE (as part of the ProteomeXchange Consortium) in its General guidelines for submitted datasets (version 1.0, May 2022) aligns with these policies. Of note, regardless the choice of the CC0 license, PRIDE insists that users re-using a deposited dataset give appropriate credit to the original authors/submitters by citing the dataset or the associated paper.

IDR Submission Guidelines (version 1.0, March 2017) permit datasets to remain private over a defined period of time (typically until article publication) but require them to be released thereafter. Following IDR recommendation, all datasets will be submitted under the CC BY license to facilitate their widest re-use. Once deposited data have been validated, their publication takes at least 3 weeks and DOI minting another week. There is no opportunity to generate private access tokens for referees.

BioImage Archive Policies (version 1.1) do not allow for embargo to be placed on deposited datasets. Both the BioImage Archive policies (version 1.1) and EMPIAR policies (version 1.0.8; October 23, 2023) make deposited data available under the CC0 license by default. EURICA will strive to deposit microscopy data under the CC BY license.

Datasets deposited to Recherche Data Gouv can be embargoed for up to 18 months. There is a possibility to create a private URL to enable access to unpublished datasets by referees. Recherche Data Gouv implements the open-access etalab2.0 license by default, but permits the use of alternative/additional licenses, including CC BY.

Zenodo permits to reserve access to deposited data for indefinite time.

EURICA will not recur to embargoing or other access limitation forms for its deposited datasets beyond what might be imposed by their publication and patenting pipelines.

All articles and the data contained within them are either directly open-access, due to the journal policy, or *de facto* open-access without delay due to the mandatory deposition of the accepted author versions in HAL simultaneously with publication, under the French law (Code de la Recherche, article L. 533-4; Loi n° 2016-1321 du 7 octobre 2016 pour une République numérique, section 30) and according to recommendations of Plan S Rights Retention Strategy (RRS) by cOAlition S (ANR is a member of cOAlition S).

All published data which could not be deposited in trustworthy repositories and are therefore preserved in house are made freely available by the Principal Investigators (Yannis FRANÇOIS, Wangqing LIU, Nicolas LEULLIOT, Alexandre SMIRNOV) upon request under the CC BY license.

5c. What methods or software tools are needed to access & use data?

Wherever possible, the data are stored in both the original and alternative widely accepted open and portable formats. Therefore, potential users can choose to use either (i) the original (often proprietary) software (as specified in Table 1) or (ii) more common and sustainable alternatives working with general formats specified in section 1a and Table 1. The latter include the Microsoft Office suite, PyMol, ImageJ, Adobe Suite etc.

Analytical MS data in the Francois's lab are collected in .D or .WIFF formats used by the primary instrument. These file formats allow raw data to be stored safely without reprocessing. It is not technically possible to subsequently modify the raw data acquired on the instrument, and the data can be reliably traced back. For automated data processing, it is possible to convert these files to the mzXML format, which is an open alternative for MS/MS data. The 'O-TOF control' and 'Analyst'

software comply with the Good Laboratory Practice directives of the European Medicines Agency (EMA).

Specific data types and the software required to access and re-use the data deposited in the ProteomeXchange Consortium are described in the Data submission guidelines for the ProteomeXchange Consortium (version 3.0.1; October 13, 2019) and in PRIDE recommendations.

According to the SOLEIL Data management policy, access to raw data, metadata and results stored by SOLEIL is possible for registered users via a searchable on-line catalog. SOLEIL makes a commitment to provide means to read, reduce and/or process raw data. Download of the data may be subject to volume restrictions.

In their ESRF data policy 2024 (version 14/10/2023), ESRF makes a commitment to provide a means for the Principal Investigator to access and use data, including the necessary software. However, ESRF declines any responsibility in case of data loss or unavailability; the same applies to analysis software.

The wwPDB provides detailed information about the file types implemented in this repository and extensive software-related documentation (see also PDBx/mmCIF Dictionary Resources).

EMDB provides a detailed overview of the data models and file types used by the repository and code to facilitate their reading or download and enables access to EMDB and EMPIAR data with the EMDB API.

EMPIAR Policies and Processing Procedures (version 1.0.8; October 23, 2023) provide a structured description of data and file types implemented by EMPIAR and links to software to open each file type.

BIA, while accepting a number of proprietary formats, encourages submission in widely supported formats, e.g. OME-TIFF or OME-NGFF. All users can access and download deposited data.

Image Tools Resource (ITR) of IDR provides and maintains links to a wide variety of software tools to work with imaging data. IDR provides a detailed manual on how to download deposited data.

Recherche Data Gouv (Displaying and exploring data, August 9, 2024) offers multiple options to preview, explore, and download deposited data in function of their format and access conditions. Any user can explore and download published datasets; only authorised users can access unpublished datasets.

Zenodo General Policies (version 1.0) guarantee access to deposited data (in their original format) to unregistered users.

5d. How will the application of a unique & persistent identifier (such as a Digital Object Identifier, DOI) to each dataset be ensured?

All selected trusted repositories assign persistent identifiers (PIDs) to deposited datasets with modalities following A Persistent Identifier (PID) policy for the European Open Science Cloud (October 2020).

The ProteomeXchange Consortium associates all deposited datasets with its own unique and persistent identifiers (PXD and RPXD assigned by ProteomeCentral).

Similarly, BIA assigns its own accession codes to each directly submitted dataset; by contrast, BIA keeps the PIDs of datasets imported from partner repositories.

IDR Submission Guidelines (version 1.0, March 2017) introduced *idr*-codes as informal accession numbers unique to each dataset; however, they encourage using DOI (automatically minted for each dataset) as a formal way of dataset citation.

SOLEIL is set to assign unique persistent identifiers to both its experiments and datasets.

Experimental datasets of the ESRF are assigned a DOI, which must be quoted in case of the data publication.

wwPDB implements a four-character PDB code (PDB ID) as a unique and persistent identifier, which is poised to evolve into a 12-character code before 2029.

EMDB issues a unique and unchangeable EMDB accession code (EMD-xxxx) for each submitted dataset.

EMPIAR issues a citable EMPIAR accession code for each submitted deposition.

Recherche Data Gouv and Zenodo assign DOI to all deposited documents. See also section 5a for the special case of data versioning.

ORCIDs to identify authors and contributors are added to metadata wherever possible or required by repositories.

6. Data management responsibilities & resources

6a. Who (for example role, position, & institution) will be responsible for data management (i.e. the data steward)?

Data stewardship responsibilities and their assignments are outlined in Table 11.

The Principal Investigators (Yannis FRANÇOIS, Wangqing LIU, Nicolas LEULLIOT, Alexandre SMIRNOV) are responsible for the DMP implementation, review, and revision.

This DMP version is subject to revision in 2 years after the beginning of the EURICA project (i.e. in December 2025).

Table 11. Data stewardship responsibilities

Responsibility	Responsible
Data capture	All Creators, SOLEIL, ESRF, Plateforme Protéomique Esplanade
Metadata production	All Creators, SOLEIL, SOLEIL Main Proposer (Nicolas LEULLIOT), ESRF, PRIDE Submitter
Maintenance of associated digital research resources	Lab Managers (Christelle GRUFFAZ) and all Creators on the eLabFTW book
Data quality	All Creators, Principal Investigators, SOLEIL, ESRF, wwPDB, EMDB, EMPIAR
Storage & backup	All Creators, data stewards (Valéry LARUE for the Leulliot's & the Liu's labs), IT Managers (Bruno PARTOUCHE for the Smirnov's lab), SOLEIL, ESRF, ProteomeXchange Consortium
Data archiving	Principal Investigators
Data sharing	Principal Investigators, PRIDE Submitter

6b. What resources (for example financial & time) will be dedicated to data management & ensuring that data will be FAIR (Findable, Accessible, Interoperable, Re-usable)?

The time effort from each data steward is estimated at ~1 person·month/year. It makes part of contract or statutory obligations of the stewards. Financial costs include expenses for hard drives (~1500 €) and eventual access to the Data Centre (in process of negotiation). Data storage in all selected trusted repositories is free of charge.