



The ultraconserved intersubunit ribosome assembly factor (InsuRAF). Data Management Plan (version 2)

Alexandre Smirnov

► To cite this version:

Alexandre Smirnov. The ultraconserved intersubunit ribosome assembly factor (InsuRAF). Data Management Plan (version 2). [Technical Report] 2, Université de Strasbourg; CNRS; Agence Nationale de la Recherche. 2021, pp.26. hal-03477332

HAL Id: hal-03477332

<https://hal.science/hal-03477332>

Submitted on 13 Dec 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

Alexandre SMIRNOV

**The ultraconserved intersubunit ribosome assembly factor
(InsuRAF)**

ANR-19-CE11-0013

ANR – JCJC 2019

Data Management Plan

Version 2

December 10, 2021

0. What changed with respect to the Version 1 of this DMP

- Throughout the text: references and links to specific documents, recommendations, norms, or legal acts are added.
- Table 1 is updated to include additional phylogenetic and sequence analysis software and formats, Zotero/Zotero RDF as an additional tool/file format to handle bibliography, CRISPR tools and formats.
- Table 4 (former Table 5): entry "Plasmids" is updated to include new information (species, links to Sanger sequencing data).
- 2a: the structure of this section has changed to facilitate progressive implementation of the DMP by the participants of the project. Namely, the description of the general organisation of the research and of the nomenclature of research resources and data now precede that of metadata, which is also enriched in comparison with the previous version. A special paragraph anticipates the transition to digital lab books and explains how it will impact the current metadata system and its interconnections with the lab book, CER, data, protocols, and research resources. Another paragraph explicitly introduces the place of the present DMP in the overall organisation of the research.
- 2a: Anna SMIRNOVA and Zhen LIAO are included in the list of the staff authorised to curate and modify research resources.
- Table 8: new inclusion guidelines for the entry "Protein sequence, conservation, phylogenetic distribution" are added, and those for the entry "Genome sequence" are made more explicit.
- 4b and 5b: the scopes of these two section are now better delimited. Section 4b focuses exclusively on the legal acts and regulations concerning intellectual property, ownership, authorship, and copyright. It also cites some of the requirements by data repositories on the re-use of third-party data. Section 5b treats more specifically on data access and sharing (including open access policies) as per repository rules, current legislation and community norms.
- 4b and 5a: relevant information on source code deposition is provided.
- 5a: a detailed plan for long-term data preservation is proposed.
- 2a, 5a, 5b and 5d: the publication modalities of the present DMP are specified.
- 5d, Tables 2, 9 and 10: ORCIDs are added as researcher identification means.
- A few grammar and typo fixes are made, and redundancies are removed.

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 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 the data will be performed in open or standard formats (TIF 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, XLSX for spreadsheets, PDB for structures), wherever possible, as recommended in [Guide Pratique de la publication en ligne et de la réutilisation des données publiques \("Open data"\)](#) elaborated by CNIL and CADA.

Table 1. 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, HAMAP, NCBI BLAST, NCBI Assembly, Python	FAS, TXT, JPEG, PNG, CSV	10-100 MB
Deposited protein structure (interactive structure)	Data retrieval, analysis & visualisation	RCSB PDB, PyMol, Jmol, NGL Viewer	PDB, mmCIF	1-10 GB
Deposited gene or genome sequence (text)	Data retrieval & analysis	RefSeq, NCBI Nucleotide, NCBI Genome	FAS, TXT	10-100 MB
Deposited MS data (text, image, spreadsheet)	Data retrieval, analysis & visualisation	neXtProt, PeptideAtlas, ProteomeXchange Consortium, Scaffold	mzML, TXT, TSV, XLSX, JPEG, PDF	10-100 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, ExPASy, CHOPCHOP, ApE, BE-Hive, CRISPOR	DOCX, TXT, JPEG, GB, CSV, XLS	10-100 MB
Predicted RNA properties (secondary structure, stability) (text, image)	Prediction	RNAfold, RNAcofold, mfold	DOCX, TXT, JPEG, PNG, PDF	10-100 MB
Predicted protein properties (MW, pl, localisation, processing & modification sites, interaction sites) (text, spreadsheet)	Prediction	ExPASy, MitoFates, TargetP, Mitoprot, RaptorX	DOCX, TXT, XLSX, JPEG	10-100 MB
Predicted protein structure (interactive structure)	Prediction	RaptorX, PyMol	PDB	10-100 MB
Sequence alignment	Analysis	NCBI BLAST	FAS, TXT	1-10 MB
Multiple sequence alignment	Analysis & visualisation	COBALT, WebLogo	FAS, TXT, DOCX, PDF, Newick, JPEG	0.1-1 GB
Nucleic acids expression, purification, modification or interaction pattern (image, spreadsheet)	Gel electrophoresis, ethidium bromide staining (G-Box), northern blotting (radioautography, Typhoon), immunoprecipitation, cleavage assay, EMSA	GeneTools, ImageQuant TL	SGD, GEL, TIFF, XLSX	10-100 GB

Proteins 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	10-100 GB
DNA sequence (text, image)	Sanger sequencing	Chroma	SEQ, AB1, FAS	0.1-1 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
X-ray reflection data & structure model (image, text, interactive structure)	X-ray crystallography (SOLEIL), modelling	MODELLER, PHENIX, Phaser, Coot, PyMol etc	MTZ, mmCIF, PDBx	0.1-1 TB
SAXS data & structure model (image, text, interactive structure)	SAXS (SWING, SOLEIL), modelling	FOXTROT, ATSAS, DADIMODO, SASREF	NXS, DAT, PDB	0.1-1 TB
Sedimentation properties (image, spreadsheet)	Velocity sedimentation, UV absorbance (Nanodrop), gel electrophoresis, ethidium bromide staining (Typhoon), northern & western blotting (Typhoon, ChemiDoc Touch), $s_{20,w}$ calculation	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), ETD (Orbitrap Elite)	Mascot, Proline, Scaffold, Microsoft Excel	RAW, MGF, MZID, mzML, MSF, TIFF, PDF, XLSX	1-10 GB
Protein & RNA localisation (image, spreadsheet)	Immunofluorescence, smFISH, confocal microscopy (LSM700, LSM780)	ImageJ, MosaicSuit, Microsoft Excel	ZVI, TIFF, XLSX	1-10 GB
Protein-protein FRET (image, spreadsheet)	FLIM-FRET	FLIM Nikon TE2000, Microsoft Excel	FLI, BMP, XLSX	1-10 GB
Proximity ligation assay (PLA) data (image, spreadsheet)	PLA, confocal microscopy (LSM700, LSM780)	ImageJ, MosaicSuit, Microsoft Excel	ZVI, TIFF, XLSX	1-10 GB
Nucleotide hydrolysis pattern (image)	TLC, radioautography (Typhoon)	ImageQuant TL	GEL	0.1-1 GB
Cell & mitochondrial morphology & physiology (image, spreadsheet)	Cell staining, immunofluorescence, TUNEL, confocal microscopy	ImageJ, Microsoft Excel	ZVI, TIFF, XLSX	0.1-1 GB
Oxygen consumption data (image, spreadsheet)	Respirometry (Seahorse XFe96 Analyser)	Wave Desktop, Controller 2.6, Microsoft Excel	ASYR, ASYT, XLSX	10-100 MB
RT-qPCR data (image, spreadsheet)	RT-qPCR (CFX96)	CFX Manager	PCRD, ZPCR	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 the 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 laboratory or reachable via collaborations and TGIR (Synchrotron SOLEIL). Data re-use concerns both publicly available deposited data and previously collected in-house data and samples.

The provenance of re-used data is documented with the help of the associated persistent identifiers (e.g., UniProt ID, PDB code, PXD identifier, RefSeq number, article reference/PMID etc). 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, parameters, filters, and 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 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 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, 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 comma-separated 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)

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 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).

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_Ag	Polyacrylamide gel stained with silver
	PAGE_Phos	Polyacrylamide gel visualised with Phosphor Imager
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)
UV-VIS spectrum	Spectra	UV-VIS spectra report
Sanger sequencing	Seq_pSAPnnnn_PPPP	Sanger sequencing of the plasmid pSAP#### with primer PPPP
PCR	PCR_SAO#####_SAO#####	PCR with primers SAO##### & SAO#####
Quantification	Quant	Spreadsheet with quantification data
Structure	Structure	Secondary or tertiary structure of a molecule
Sequence	Sequence	Annotated sequence of a gene, a transcript or a protein

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
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
Plasmid resources	[Discussion are ongoing on the best way to directly link Sanger sequencing data of the corresponding plasmid to each entry to make them easily searchable & analysable]
Bacterial strains	
Strain	Strain ID SAB####
Date created	Date of the original DMSO stock
Species	Bacterial species
Genotype	Genetic background with 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, storage, expression) or other specificities
Notes	Any useful information & observations, eventually provenance & alternative ID (if received from third parties)

Protocols (NNM###) are named by Creator's initials (NN) and numbered (###). Their versions are labelled with letter suffixes ("a", "b", "c" etc). Protocols have an invariable 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 editing (excluding major modification which requires 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, 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
Unsuccessful implementation	Examples of strains, proteins, or other biological entities for which the protocol failed in house (with brief notes & proposed alternatives)
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#.

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, main experimenter
3. Subject	Type of the experiment
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), (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 spectra)
9. Format	Name of the folder(s) in which the data are classified
10. Identifier	ID of the experiment & a complete list of the files in the dataset
11. Source	Primary or secondary (in the case of data re-use or re-analysis)
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
15. Rights	IPR Unistra

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”.

Additional metadata are generated for the datasets deposited in dedicated repositories (section 5), following their established conventions and requirements ([ProteomeXchange Consortium data submission guidelines](#), version 3.0.1; October 13, 2019; and [wwPDB Policies and processing procedures document](#), version 5; September 2021).

The possibility of introducing other, data type-specific metadata formats, among those catalogued in the [Minimum Information for Biological and Biomedical Investigations \(MIBBI\)](#) portal, is currently studied.

All the research resources 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, Cédric SCHELCHER, Anna SMIRNOVA, Zhen LIAO). The maintenance of the research infrastructure is ensured by the Lab Manager (Christelle GRUFFAZ) and supervised by the Principal Investigator (Alexandre SMIRNOV).

With the imminent introduction of digital lab books ([CNRS Research data plan](#), November 2020), linking together CERs, data, metadata, protocols, and research resource IDs and their sharing between the participants of the project will become even more straightforward. We foresee that the introduction of digital lab books 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, albeit 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) the connecting node between various data, metadata, protocols, and research resources that organises them in experiments and ensures cross-referencing;
- (ii) the contextualisation tool that explains the place of data and processes in the research;
- (iii) the 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.

The transition from the paper lab books to the digital ones will be enacted as soon as the corresponding prototypes, elaborated and/or approved by CNRS and/or the University of Strasbourg, will be clearly indicated and become available.

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 InsuRAF project. It follows the outline and requirements elaborated in the Science Europe [Practical guide to the International alignment of research data management](#) (January 2021) and is subject to updates and open-access publication on Zenodo.

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). This task 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). The experimenter ensures the identity and maintenance of collected data and metadata.

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")
Design	Presence of negative & positive controls; standards; complementation; optimal design of time or concentration series; complete combinatorial design; uniformity of measurement (e.g., all samples on the same gel/membrane or in the same run/session); 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
Documentation	The experiment is completely described in the lab book, data follow the conventional vocabulary, are accompanied by properly registered 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 etc); fidelity of digitalised images; absence of non-documented or irreversible modifications to the original files

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 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 <i>in silico</i> into a protein that satisfies the above criteria
Genome sequence	Completely assembled bacterial, archaeal or organellar genome; at least partially assembled eukaryotic nuclear genome (complete, chromosome, or scaffold 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 or at least 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),
- 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.

3. Storage & backup during the research process

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

The data and metadata are stored on the office computers and monthly backed up on hard drives which are kept in a different location. Additionally, primary data are stored on original instruments and on the servers of the services and collaborators

who generated them (e.g., SOLEIL), ensuring that several copies of the most important data are preserved in geographically distinct locations. The Seafire 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.

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](#)).

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 InsuRAF project. The access of 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 Seafire links or by e-mail. 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.

No personal or sensible data 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 in the domain of Open Science ([CNRS Research data plan](#), November 2020). 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).

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 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?

The ownership of the data collected and produced in this research is regulated by the Intellectual property code under the French law. The data are a property of the University of Strasbourg ([Code de la propriété intellectuelle; October 3, 1996; section](#)

[R611-12](#)), and the access to them is controlled by the Principal Investigator (Alexandre SMIRNOV).

The authorship rights on all published documents belong to their authors (physical persons), as stipulated by the French law (see the comprehensive [Guide pratique de la publication en ligne et de la réutilisation des données publiques](#) ("Open data ») elaborated by CNIL and CADA). By contrast, the copyright on publications (e.g., journal articles, book chapters) is regulated by specific publishing policies and the corresponding signed agreements between the authors and the publisher; therefore, it may belong to either party.

The ownership of the 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 InsuRAF project, remain a property of the University of Strasbourg ([Code de la propriété intellectuelle; October 3, 1996; section R611-12](#)).

The data deposited in the ProteomeXchange Consortium remain a 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).

The data deposited in the wwPDB are open-access without copyright restrictions. The [wwPDB policies and processing procedures document](#) (version 5.0; September 2021) reserves the authorship rights to Entry Authors, Contact Authors and Citation Authors which necessarily include the principal investigator (Alexandre SMIRNOV).

No restrictions on the re-use of third-party data are expected. Specifically, no third-party intellectual property rights are expected to be affected. However, the re-use of deposited third-party data is subjected 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 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 InsuRAF 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.

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 has already been obtained and partially published, 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 will be 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 InsuRAF project or for other ongoing research projects;
- 2) high-quality published or ready-to-publish data;
- 3) other high-quality data which are not expected to be used in the foreseeable future;
- 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 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 ProteomeXchange Consortium (via PRIDE repository). Structural data are deposited in wwPDB (CoreTrustSeal certified, 2020-2022). Source codes are deposited on GitHub. The deposited data follow the corresponding policies and guidelines for formatting, storage, and sharing, as outlined in sections 4b, 5b, 5c, and 5d. They are free of charge. 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.

SOLEIL acts as 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.

[wwPDB policies and processing procedures document](#) (version 5.0; September 2021) defines the rules of structural data deposition in the wwPDB. The policies permit withdrawing unreleased datasets by the authors' decision. 'Problem structures', presenting anomalies, which have not been fixed or published can be withdrawn by wwPDB upon expiration of a one-year hold period. The withdrawn entries will still be

listed as unreleased entries in the PDB archive. By contrast, released datasets cannot be withdrawn any more. In particular, publication of an article citing a PDB entry (preprints included) automatically triggers its release. The concerned journal may request the release of a published PDB entry directly from wwPDB.

[Data submission guidelines for the ProteomeXchange Consortium](#) (version 3.0.1; October 13, 2019) regulates the deposition of mass spectrometry data. 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 the corresponding paper is retracted). Preprints are not considered publications under the ProteomeXchange Consortium policies.

Other trusted repositories (e.g., for primary image data associated with publications) 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 a Creative Commons Attribution 4.0 International licence and communicated to ANR as a specific project deliverable.

The [Deuxième Plan national pour la science ouverte](#) (2021-2024) announces the creation of a research data national platform, Recherche Data Gouv (expected in March 2022), which should permit the deposition of data types for which no trustworthy repositories currently exist as well as the cataloguing/linking to data deposited in other trusted national and international repositories.

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).

Similarly, 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. Finally, 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 ([wwPDB policies and processing procedures document](#), version 5.0; September 2021).

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

The **Group 3 data** undergo archiving. Some of them may be later re-used by the research team or collaborators and may be shared to third parties (members of the research community), following the modalities described in section 3b, for research purposes. 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 are expected to be collected in this research, there is no legally imposed limit on the duration of data preservation.

The archiving procedure includes two components: (i) copying and transferring the dataset in question to the archive hard drive and (ii) 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 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) &/or institution 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; 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 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 & 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 the localisation of the archived data and the reconstitution of their record management.

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 is organised 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)
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 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 & 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 yearly basis and will only concern the datasets (CER experiments) that have not been in use for at least one year.

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 and in [La Vie d'une Donnée au regard des réglementations "CRPA", "RGPD" et "Patrimoine"](#) (version 1.0; Juin 2019) compiled by SupDPO. They are also regulated by international norms (see Maday & Taillefer. [Les métadonnées du records management du point de vue des normes ISO](#). In: *La Gazette des archives*, n°228, 2012-4). Namely, [ISO 23081-1](#) requires the documentation of any changes introduced in the content, the context, or the structure of the record; [ISO 15489-2](#) guarantees the accessibility of archived data and metadata; [ISO 15489-1](#) requires the approval of a hierarchical authority for any permanent destruction of records.

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

Access to the data and documents associated with the InsuRAF 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 "Patrimoine"](#), version 1.0; June 2019; [Guide pratique de la publication en ligne et de la réutilisation des données publiques](#) ("Open Data"); CNIL, CADA; [CNRS Research data plan](#); November 2020). This said, the data discoverability and sharing are intimately connected with the processes of publication and eventually patenting. 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 Investigator (Alexandre SMIRNOV) upon consultation with collaborators and eventually SATT Conectus Alsace. 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. Upon publication (and in some journals prior to publication), they are rendered open-access 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 the Principal Investigator (Alexandre SMIRNOV) and by the Main Proposer (Benoît MASQUIDA). Additionally, §4.18 of the [SOLEIL Data Management Policy](#) grants SOLEIL access to all collected data and metadata for curation and sharing purposes.

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 at least a CC0 licence without embargo.

The wwPDB similarly distinguishes unreleased and released data in its policies ([wwPDB policies and processing procedures document](#), version 5.0; September 2021). Unreleased data are distributed only to the authors of the entry who are responsible for their communication to journal editors and/or reviewers along with wwPDB validation reports (see also section 5a for further details). Released data are open-access.

The source codes and other documents associated with the InsuRAF project (e.g., the present DMP) are deposited on GitHub and Zenodo in open-access without embargo.

All articles and the data contained within them are either directly open-access, due to the journal policy, or *de facto* open-access due to the mandatory deposition of the accepted author versions in HAL and univOAK after an embargo period (no more than 6 months), 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](#)).

All published data which could not be deposited in trustworthy repositories and are therefore preserved in house are made freely available by the Principal Investigator (Alexandre SMIRNOV) upon request under CC-BY licence.

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 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.

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).

According to the [SOLEIL Data management policy](#), SOLEIL engages to make available for registered users means to read, reduce and/or process raw data.

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

The system of unique and persistent identifiers for data, the related metadata and research resources adopted in the laboratory is described in section 2a. wwPDB and the ProteomeXchange Consortium associate all deposited datasets with their own unique and persistent identifiers (PDB ID and PXD, respectively). SOLEIL is set to assign unique persistent identifiers to both its experiments and datasets. Zenodo assigns DOI to all deposited documents. See also section 5a for the special case of data versioning. ORCIDs identifying 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 Investigator (Alexandre SMIRNOV) is responsible for the DMP implementation, review, and revision.

This DMP is subject to mandatory updates due on 01.01.2022 and on 31.12.2023, as required by ANR. It is also updated every time as new details or circumstances arise or corrections need to be introduced. All subsequent versions (referred to as "InsuRAF_DMP_v###") must feature the additional section "0. What changed with respect to the Version # of this DMP".

Table 11. Data stewardship responsibilities

Responsibility	Responsible
Data capture	All Creators, SOLEIL, PDB, ProteomeXchange Consortium
Metadata production	All Creators, SOLEIL, SOLEIL Main Proposer (Benoît MASQUIDA), PRIDE Submitter
Maintenance of associated digital research resources	Lab Manager (Christelle GRUFFAZ)
Data quality	All Creators, SOLEIL, Principal Investigator (Alexandre SMIRNOV)
Storage & backup	All Creators, IT Manager (Bruno PARTOUCHE), SOLEIL, PDB, ProteomeXchange Consortium
Data archiving	Principal Investigator (Alexandre SMIRNOV)
Data sharing	Principal Investigator (Alexandre SMIRNOV), SOLEIL Main Proposer (Benoît MASQUIDA), SOLEIL, PDB, PRIDE Submitter, ProteomeXchange Consortium

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 (estimated ~600 €) and eventual access to the Data Centre (in process of negotiation). Data storage in PDB and ProteomeXchange Consortium repositories and in SOLEIL is free of charge.