The ultraconserved intersubunit ribosome assembly factor (InsuRAF). Data Management Plan (version 1)

Alexandre Smirnov

Alexandre SMIRNOV

The ultraconserved intersubunit ribosome assembly factor (InsuRAF)

ANR-19-CE11-0013

ANR – JCJC 2019

Data Management Plan

Version 1

July 1, 2020

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.

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. Provenance of collected and produced data is documented in the form of metadata, lab book entries, and the Central Experiment Registry (see section 2).

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)	Data retrieval, analysis & visualisation	InterPro, UniProt, PFAM, HAMAP	FAS, TXT, JPEG, PNG	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	PubMed, BiblioVie, HAL, univOAK	PDF, DOCX, XLSX, CSV, MP4, ZIP	1-10 GB
Gene properties (promoter, TSS, ORF, UTRs, SD or Kozak sequence) (text, image)	Data retrieval & analysis, prediction	RefSeq, ExPASy	DOCX, TXT, JPEG	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, pI, 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

2. Documentation & data quality

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

In addition to built-in metadata associated with instrument-specific primary files (including such information as the 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 NN are the initials of the person responsible for the experiment and ##### is the number of the experiment in the Central Experiment Registry (see below). 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 & 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).

A standard metadata file is organised in accordance with the Dublin Core format and includes 15 generic entries (Table 2), eventually followed by the concluding tag #archived#.

Table 2. Metadata file organisation

Entry	Description
1. Title	Short title of the experiment
2. Creator	Person &/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) &/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

Additional metadata are generated for the datasets deposited in the dedicated repositories (section 5), following their established conventions and requirements (ProteomeXchange Consortium Data Submission Guidelines and wwPDB Policies).

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. 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 3). These data are matched with those of the corresponding lab books, which obligatorily use the same identifiers and ensure complete cross-referencing.

Table 3. Central Experiment Registry columns

Experiment ID	Complete experiment identifier in the format NNE#####_yyyymmdd, where NN are the initials of the person responsible for the experiment, ##### is the number of the experiment, yyyymmdd is the date on which the experiment was started
Participants	Initials of the person(s) directly contributing to the experiment
Affiliations	Affiliations of the person(s)
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 4. 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 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 4. 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)
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 of in storage cryo-boxes. Each kind of resource is catalogued in a dedicated spreadsheet which ensures cross-referencing to other resource types (Table 5).

Table 5. 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 species specification (if not human), mutation in the (genotype=proteotype) format or added tags
Cloning sites	Sites used for cloning the insert
Concentration, ng/µl	Concentration of the current stock
Notes	Intended uses, provenance, construction notes & any useful observation
Oligos used for cloning	Complete list of oligonucleotides (SAO#####) used to create the plasmid
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 a 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 invariable structure, as outlined in Table 6. They necessarily cross-reference all related research resources.

Table 6. 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 hyperlinks 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 (no facultative breaks used) and maximal (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

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). The maintenance of the research infrastructure is ensured by the Lab Manager (Christelle GRUFFAZ) and supervised by the Principal Investigator (Alexandre SMIRNOV).

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 & widely used by the scientific community, their quality can be evaluated by compliance with standards accepted in the field. 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 standards established in the field, documents and interprets the experiment (in the lab book and CER), and then to the principal investigator (Alexandre SMIRNOV), who reviews the protocol, the data and the metadata and re-evaluates their compliance with the community standards and the validity of the interpretation. The Lab Manager (Christelle GRUFFAZ) ensures the identity and 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 the 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 the 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 operability; 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 & 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; 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 *in silico* into a protein that satisfies the above criteria
Genome sequence	Completely assembled bacterial, archaeal or organellar genome; at least partially assembled eukaryotic nuclear genome with at least one contig of >100 kb; for paired analyses, both criteria should be met
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 the 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 either including raw data or 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 the field conventions for each kind of analysis or justification of deviations from them,
d) compliance with the assumptions and requirements of the 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 (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 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 soon provided by the Data Centre of the University of Strasbourg.

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

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 researchers involved in the project in GMGM. 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 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, and by the CNRS policy in the domain of Open Science. Data collected at the synchrotron SOLEIL are subject to the SOLEIL Data Management Policy, which regulates access, re-use and publication of the data (see sections 4 and 5). Access to the data deposited in the ProteomeXchange Consortium is regulated by the Data Submission Guidelines of the repository.

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 project is regulated by the intellectual property code (section R611-12) under the French law. The data are a property of the University of Strasbourg, and access to them is controlled by the Principal Investigator (Alexandre SMIRNOV). Upon publication of an article, the associated primary and secondary data are deposited to the dedicated repositories and are open-access (see section 5). Publication of the corresponding "data papers" is also planned to increase the access of the community to primary data and analyses. 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 mandatory deposition of the accepted author versions in HAL and univOAK after an embargo period (no more than 6 months), under the French law (section 30 of LRN).

The ownership of the primary data obtained at the synchrotron SOLEIL is regulated by the SOLEIL Data Management Policy (§3). These data are open access under the CC-BY licence after an embargo period (§3.1, see section 5b). 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 agreement with §3.4 of the SOLEIL Data Management Policy, all derived products (results, data interpretation, IPR) remain a property of the University of Strasbourg.

The data deposited in the ProteomeXchange Consortium remain a property of the Principal Investigator and of the PRIDE Submitter, as specified in the Data Submission Guidelines of the repository. Unreleased data are by default private (password-protected); access to them is controlled by 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 publication (and in some journals prior to publication) of the original article, the corresponding dataset is released under the CC-BY licence without embargo.

The data deposited in PDB are open-access without copyright restrictions.

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 PDB data re-use requirements are outlined in the wwPDB Policies: the original authors of the structure should be given credit, the corresponding publications cited and the unique dataset identifier (PDB ID) provided. The ProteomeXchange Consortium data re-use requirements are outlined in the Guidelines for Handling ProteomeXchange Reprocessed Datasets: 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: 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.

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)?

The expected volume of the collected and produced data (see Table 1) permits long-term preservation of the totality of data on hard drives or in the Data Centre in house. These data are expected to be re-used by the research team and collaborators and may be shared to third parties (members of the research community), following the modalities described in section 4.2, for research purposes. Foreseeable uses of the data include but are not limited to publication, follow-up, replication or comparative studies, re-analysis, benchmarking, algorithm training. Long-term curation of such data is ensured by the Lab Manager (Christelle GRUFFAZ) and by the Principal Investigator (Alexandre SMIRNOV).

Key mass spectrometry and structural data associated with publications are deposited in trusted dedicated repositories, ProteomeXchange Consortium (via PRIDE repository) and RCSB PDB. The deposited data follow the corresponding policies and guidelines for formatting, storage and sharing, as outlined in sections 2a, 3b, 4b, 5b, 5c, 5d, 6a, and 6b. 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.

SOLEIL acts as custodian for all raw data and associated metadata collected at the synchrotron. It ensures storage and curation of the data and the associated metadata. These data and metadata follow the SOLEIL Data Management Policy for formatting, storage and sharing, as outlined in sections 2a, 3b, 4b, 5b, 5c, 5d, 6a, and 6b. 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.

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

Data discoverability and sharing are intimately connected with the processes of publication and (eventually) patenting.

Unpublished data 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 will not be released until the patent is granted or rejected.

Prior to 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 (see section 4b).

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 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.

Embargo periods imposed by journals for publication of results and certain types of associated data are overridden by the French law (Loi pour la République Numérique) and cannot exceed 6 months since publication, as described in section 4b.

All published data which could not be deposited in trustworthy repositories and were 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.

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. PDB 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.

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 9.

The Principal Investigator (Alexandre SMIRNOV) is responsible for the DMP implementation, review and revision.

This DMP is subject to mandatory updates 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 "7. DMP modifications with respect to the previous version".

Table 9. 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.