Sample and Data Relationship Format for Proteomics (SDRF-Proteomics)

Version 1.1.0-dev, 2026-01-17

Table of Contents

Chapter 1. Status of this document

This document provides information to the proteomics community about a proposed standard for sample metadata annotations in public repositories called Sample and Data Relationship File (SDRF)-Proteomics format. Distribution is unlimited.

Version v1.1.0 - 2025-01

Chapter 2. Abstract

The Human Proteome Organisation (HUPO) Proteomics Standards Initiative (PSI) defines community standards for data representation in proteomics to facilitate data comparison, exchange, and verification. This document presents a specification for the Sample and Data Relationship Format (SDRF-Proteomics).

Further detailed information, including any updates to this document, implementations, and examples is available at SDRF GitHub Repository. The official PSI web page for the document is the following: http://psidev.info/sdrf.

Chapter 3. Motivation

Many resources have emerged that provide raw or integrated proteomics data in the public domain. If these are valuable individually, their integration through re-analysis represents a huge asset for the community [1].

Unfortunately, proteomics experimental design and sample related information are often missing in public repositories or stored in very diverse ways and formats. For example:

- The CPTAC Consortium provides for every dataset a set of Excel files with the information on each sample including tumor size, origin, but also how every sample is related to a specific raw file (e.g. instrument configuration parameters).
- As a resource routinely re-analysing public datasets, ProteomicsDB, captures for each sample in the database a minimum number of properties to describe the sample and the related experimental protocol such as tissue, digestion method and instrument.

Such heterogeneity often prevents data interpretation, reproducibility, and integration of data from different resources. For every proteomics dataset we propose to capture at least three levels of metadata:

- (i) dataset description
- (ii) the sample metadata and data files acquisition metadata.
- (iii) The relation between the sample and the data files. The experimental design.

The general description includes minimum information to describe the study overall: title, description, date of publication, type of experiment. In ProteomeXchange partners this metadata is captured at the dataset level, in other omics resources this is captured as IDF file format (e.g. MAGE-TAB). Currently, all ProteomeXchange partners mandate this information for each dataset. However, the information regarding the sample and its relation to the data files (**Figure 1**) is mostly missing [1].

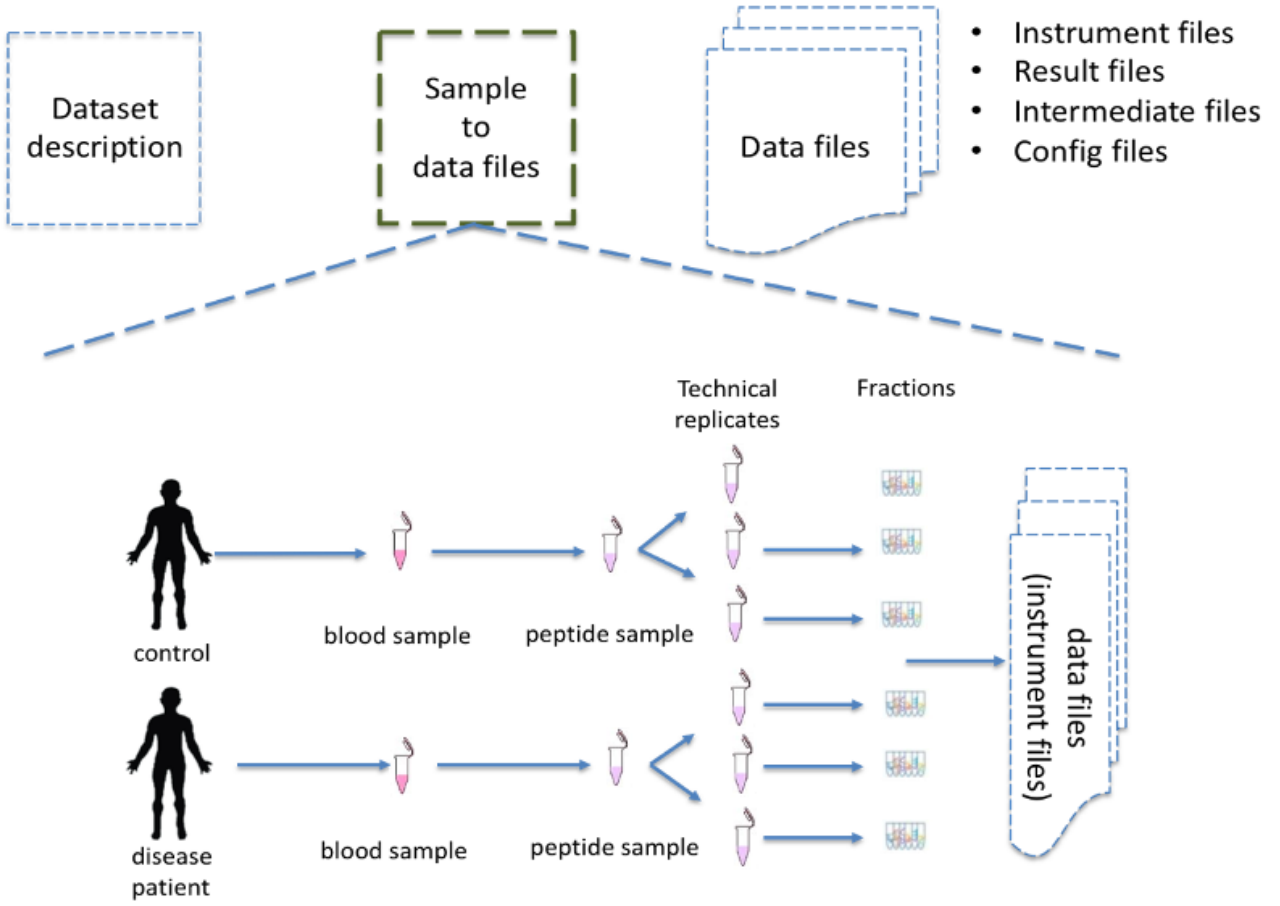

Figure 1: SDRF-Proteomics file format stores the information of the sample and its relation to the data files in the dataset. The file format includes not only information about the sample but also about how the data was acquired and processed.

Here, we introduced the Sample and Data Relationship Format (SDRF-Proteomics) to capture the sample metadata and its relation to the data files for proteomics experiments. The SDRF-Proteomics format is a tab-delimited file format that describes the sample characteristics and the relationships between samples and data files included in a dataset.

This specification, which is a community effort, aims to provide a standard for the proteomics community to annotate the sample metadata and its relation to the data files.

Chapter 4. Quick Start

If you're new to SDRF-Proteomics, here's a minimal example to get you started. An SDRF file is a tab-separated file where each row represents a sample-to-data-file relationship.

4.1. Minimal Example

source name	characteristics [organism]	characteristics [disease]	assay name	comment [label]	comment [instrument]	comment [cleavage agent details]	comment [data file]
sample_1	homo sapiens	normal	run_1	label free sample	Q Exactive HF	NT=Trypsin;AC=MS:1001251	sample_1.raw
sample_2	homo sapiens	hepatocellular carcinoma	run_2	label free sample	Q Exactive HF	NT=Trypsin;AC=MS:1001251	sample_2.raw

Blue columns: Sample metadata (characteristics) | Green columns: Data file metadata (comments)

4.2. Key Concepts

1. **Sample metadata** uses `characteristics[…]` columns (e.g., organism, disease)
2. **Data file metadata** uses `comment[…]` columns (e.g., instrument, label)
3. **Factor values** use `factor value[…]` columns to indicate variables under study
4. Each row links one sample to one data file

4.3. Format Requirements

The SDRF-Proteomics format has the following core requirements:

- The SDRF file is a **tab-delimited format** where each row corresponds to a relationship between a Sample and a Data file.
- Each column MUST correspond to an attribute/property of the Sample or the Data file.
- Each cell value MUST be the property value for the corresponding Sample or Data file.
- The file MUST start with columns describing sample properties (e.g., organism, disease), followed by data file properties (e.g., label, fraction identifier, data file).
- Unknown values MUST be handled using `not available` (value is unknown) or `not applicable` (property doesn't apply to this sample).

4.4. Scope

The SDRF-Proteomics format aims to capture the **sample metadata** and its **relationship with data files** (e.g., raw files from mass spectrometers).

IMPORTANT	SDRF-Proteomics does **not** aim to capture downstream analysis details, including: which samples were compared to which other samples, how samples are combined into study variables, or analysis parameters such as FDR thresholds or p-value cutoffs.

4.5. Getting Started Steps

1. Choose a core template (Human, Vertebrates, Plants, etc.)
2. Fill in sample metadata (characteristics columns)
3. Fill in data file metadata (comment columns)
4. Add factor values for your experimental variables
5. Validate your file using sdrf-pipelines

For detailed guidance, continue reading the full specification below.

Chapter 5. Validating SDRF Files

The official validator for SDRF-Proteomics files is **sdrf-pipelines**, a Python tool that checks your SDRF file for errors and compliance with the specification.

Installation:

```
pip install sdrf-pipelines
```

Basic Validation:

```
# Validate an SDRF file
parse_sdrf validate-sdrf --sdrf_file your_file.sdrf.tsv

# Validate with a specific template
parse_sdrf validate-sdrf --sdrf_file your_file.sdrf.tsv --template human
```

For more information, visit: sdrf-pipelines on GitHub

Chapter 6. Specification structure

This document describes the main specification of SDRF-Proteomics, the structure of the specification (**Figure 2**), how to contribute, and extend the specification. SDRF-Proteomics uses a three-tier system for organizing metadata requirements:

- The SDRF-Proteomics core specification: This document contains the main specification, requirements and rules for the SDRF-Proteomics format. It also includes the notational conventions and the relationship to other specifications.
- Core templates: Organism-based templates (human, vertebrates, plants, etc.) that define base schemas for common proteomics experiments. See the Core Templates section.
- Specialized templates: Complete schemas for specific experiment types (cell-lines, single-cell, affinity-proteomics, crosslinking, immunopeptidomics, metaproteomics). Each template has its own directory containing:
  - A detailed README.adoc with checklists, and examples.
  - A template file ({name}-template.sdrf.tsv) with column headers.
- Annotation guidelines: Detailed documentation for specific metadata annotations (e.g., patient pre-existing condition, sample metadata, data file metadata).

[Logo] | *images/sdrf-guidelines-structure.png*

Figure 2: SDRF-Proteomics specification structure. The main specification defines the core rules and is extended by specific experiment templates and annotation guidelines.

> **NOTE**
>
> The main specification is in the `sdrf-proteomics` directory. Core templates (organism-based) are in `sdrf-proteomics/core-templates/` and specialized templates (experiment-type-specific) are in `sdrf-proteomics/templates/`. Templates are extensions of the core specification, and should follow all the rules and requirements in the main specification. If a template rule is in conflict with the specification, a note should be done in the main specification to reflect the extension or conflict.

The official website for SDRF-Proteomics project is https://github.com/bigbio/proteomics-metadata-standard. New use cases, changes to the specification and examples can be added by using Pull requests or issues in GitHub (see introduction to GitHub - https://lab.github.com/githubtraining/introduction-to-github).

A set of examples and annotated projects from ProteomeXchange can be found here: https://github.com/bigbio/proteomics-metadata-standard/tree/master/annotated-projects

Multiple tools have been implemented to validate, annotate and convert SDRF-Proteomics files. The official validator of SDRF-Proteomics is sdrf-pipelines (Python - https://github.com/bigbio/sdrf-pipelines). This tool allows to validate an SDRF-Proteomics file. In addition, it allows converting SDRF to other popular pipelines and software configure files such as MaxQuant or OpenMS.

6.1. Versioning

The SDRF-Proteomics specification is versioned using the Semantic Versioning 2.0.0 (https://semver.org/) scheme. The version number is in the format MAJOR.MINOR.PATCH, where:

- MAJOR version is incremented for incompatible changes to the specification, when major changes are done to the specification.
- MINOR version is incremented for new features that are backward compatible with the previous version. Guidelines and templates are added or modified.
- PATCH version is incremented for bug fixes and minor changes that do not affect the specification or the templates. This includes typos, formatting changes, and other minor updates.

Every change in the specification should be done in GitHub using pull requests into the dev branch. The pull request should include a description of the changes and the reason for the changes. The pull request will be reviewed by the community and merged into the main branch when approved. After the merge, the version number will be updated according to the changes made, the release will be performed, and the Zenodo record will be updated.

NOTE We added the prefix v to the version number to indicate that it is the version of the specification that was used to create the file. Examples: v1.1.0, v2.0.0, v3.0.0.

6.2. Notational Conventions

The key words “MUST”, “MUST NOT”, “REQUIRED”, “SHALL”, “SHALL NOT”, “SHOULD”, “SHOULD NOT”, “RECOMMEND/RECOMMENDED”, “MAY”, “COULD BE”, and “OPTIONAL” are to be interpreted as described in RFC 2119 (https://www.rfc-editor.org/rfc/rfc2119).

6.3. Relationship to other specifications

SDRF-Proteomics is fully compatible with the SDRF file format part of MAGE-TAB. MAGE-TAB is the file format used to store metadata and sample information for transcriptomics experiments. When the ProteomeXchange project file is converted to idf file (project description in MAGE-TAB) and is combined with the SDRF-Proteomics a valid MAGE-TAB is obtained.

SDRF-Proteomics sample information can be embedded into mzTab metadata files. The sample metadata in mzTab contains properties as the columns in the SDRF-Proteomics and values as Sample cell values.

The SDRF-Proteomics aims to capture the sample metadata and its relationship with the data files (e.g. raw files from mass spectrometers). The SDRF-Proteomics do not aim to capture the downstream analysis part of the experimental design such as what samples should be compared, how they can be combined or parameters for the downstream analysis (FDR or p-values thresholds). The HUPO-PSI community will work in the future to include this information in other file formats such as mzTab or a new type of file format.

Chapter 7. SDRF-Proteomics specification

The SDRF-Proteomics file format describes the sample characteristics and the relationships between samples and data files. The file format is a tab-delimited one where each **ROW** corresponds to a relationship between a Sample and a Data file (in an ms proteomics experiment the data file containing the mass spectra), each **COLUMN** corresponds to an attribute/property of the Sample, the Data file, or the Factor values; and the value in each **CELL** is the specific value of the property for a given Sample/Data file/Factor value (**Figure 3**).

source name	characteristics [organism]	characteristics [disease]	assay name	comment [instrument]	comment [data file]	factor value [disease]
sample_1	homo sapiens	normal	run_1	Q Exactive HF	sample_1.raw	normal
sample_2	homo sapiens	liver cancer	run_2	Q Exactive HF	sample_2.raw	liver cancer

Blue: Sample metadata | Green: Data file metadata | Orange: Factor values

Figure 3: SDRF-Proteomics in a nutshell. Each **row** links a sample to a data file. **Columns** represent sample properties (characteristics), data file properties (comments), or experimental variables (factor values).

The SDRF-Proteomics format contains three main sections:

- The first section contains the sample metadata.
- The second section contains the data file metadata.
- The third section contains the factor values properties.

7.1. Format rules

There are general scenarios/use cases that are addressed by the following rules:

- **Unknown values**: In some cases, the column is mandatory in the format, but for some samples the corresponding value is unknown. In those cases, users SHOULD use **not available**.
- **Not Applicable values**: In some cases, the column is mandatory, but for some samples the corresponding value is not applicable. In those cases, users SHOULD use **not applicable**.

Table 1. When to use "not available" vs "not applicable"

Term	Meaning	Example
not available	Value exists but is unknown or could not be determined	characteristics[age] = not available
not applicable	Value or concept does not apply to this sample	characteristics[age] = not applicable

- **Case sensitivity**: By specification the SDRF is case-insensitive for text values, but we RECOMMEND using lowercase characters throughout all the text (Column names and values).
- **Space sensitivity**: By specification the SDRF is sensitive to spaces in column names (sourcename != source name).
- **Column order**: The SDRF columns follows some structure; first the sample metadata columns in Chapter 9; then the data file metadata columns in Chapter 10; followed by the factor values columns in Chapter 13.
- **Extension**: The extension of the SDRF file SHOULD be sdrf.tsv (preferred) or .txt.

7.2. SDRF file-level metadata

Since version 1.1.0, SDRF-Proteomics supports optional file-level metadata using header comments at the beginning of the file. These header comments provide information about the SDRF file itself, such as the format version, template used, and validation status. This approach is inspired by other omics formats such as VCF (Variant Call Format) file headers and is fully compatible with pandas and other tabular data processing tools.

Header comments MUST:

- Start with `#` (single hash) followed by a key-value pair
- Appear at the very beginning of the file, before the column header row
- Use the format `#key=value`

The following header fields are supported:

Key	Description	Example	Requirement	Ontology Term
file_format	Identifier for the file format	SDRF	RECOMMENDED	PRIDE:0000831
version	SDRF-Proteomics specification version used	v1.1.0	RECOMMENDED	NCIT:C25714
template	Name of the template used	human, cell_lines, default	OPTIONAL	PRIDE:0000832
template_version	Version of the template	v1.0.0	OPTIONAL	PRIDE:0000833
source	Origin or creator of the file	PRIDE, user-generated	OPTIONAL	NCIT:C25683
validation_hash	Hash from validator certification	sha256:abc123…	OPTIONAL	PRIDE:0000834

Example of an SDRF file with header comments (simplified example showing only select columns; see Chapter 12 for complete required columns):

```
#file_format=SDRF
#version=v1.1.0
#template=human
#template_version=v1.0.0
#source=PRIDE
source name characteristics[organism]   characteristics[organism part]
characteristics[disease]    assay name  comment[data file]
sample_1    homo sapiens    liver   normal  run_1   sample_1.raw
```

NOTE

Header comments are OPTIONAL. SDRF files without header comments are still valid. When present, header comments provide valuable provenance information and enable tools to handle version-specific features appropriately. Header property names use underscores (e.g., `file_format`, `template_version`) rather than spaces to maintain consistency with the tab-delimited nature of SDRF files and avoid ambiguity when parsing.

7.3. Table Column headers

Depending on each section the column headers (property names) will be prefixed with the following prefixes:

- `characteristics`: Sample metadata (e.g. *characteristics[organism]*)
- `comment`: Data file metadata (e.g. *comment[data file]*)
- `factor value`: Factor values properties (e.g. *factor value[disease]*)

Each property name MUST be a valid ontology term or a valid controlled vocabulary term. Each section will have some specific order for column headers.

NOTE

A list of all controlled vocabularies and ontologies supported are in the Chapter 8 section. On each section we also provide a list of properties that are supported.

7.4. Table Cell values

The value for each property, (e.g. characteristics, comment, factor value) corresponding to each sample or data file can be represented in multiple ways.

- Free Text (Human readable): In the free text representation, the value is provided as text without Ontology support (e.g. colon or providing accession numbers). This is only RECOMMENDED when the text inserted in the table is the exact name of an ontology/CV term in EFO. If the term is not in EFO, other ontologies can be used.

source name	**characteristics[organism]**
sample 1	homo sapiens
sample 2	homo sapiens

- Ontology url (Computer readable): Users can provide the corresponding URI (Uniform Resource Identifier) of the ontology/CV term as a value. This is recommended for enriched files where the user does not want to use intermediate tools to map from free text to ontology/CV terms.

source name	characteristics[organism]
Sample 1	http://purl.obolibrary.org/obo/NCBITaxon_9606
Sample 2	http://purl.obolibrary.org/obo/NCBITaxon_9606

- Key=value representation (Human and Computer readable): The current representation aims to provide a mechanism to represent the complete information of the ontology/CV term including Accession, Name and other additional properties. In the key=value pair representation, the Value of the property is represented as an Object with multiple properties, where the key is one of the properties of the object and the value is the corresponding value for the particular key. An example of key value pairs is post-translational modification (see Protein Modifications):

```
NT=Glu->pyro-Glu;MT=fixed;PP=Anywhere;AC=Unimod:27;TA=E
```

Chapter 8. Ontologies and Controlled Vocabularies

SDRF-Proteomics uses ontologies and controlled vocabularies (CVs) to standardize metadata values. The following ontologies are supported:

Category	Ontology/CV	Description	Notes
General Purpose			
General	Experimental Factor Ontology (EFO)	General experimental metadata	
General	PATO	Phenotype and Trait Ontology	
General	NCI Thesaurus (NCIT)	Biomedical terminology	
General	PRIDE Controlled Vocabulary	Proteomics-specific terms	
Organism and Taxonomy			
Taxonomy	NCBI Taxonomy (NCBITaxon)	Organism classification	
Anatomy and Cell Types			
Anatomy	UBERON	Cross-species anatomy ontology	
Cell Type	Cell Ontology (CL)	Cell type classification	
Anatomy	BRENDA Tissue Ontology (BTO)	Tissues and cell lines	
Anatomy	Plant Ontology (PO)	Plant anatomy and development	For plant samples
Anatomy	FlyBase Anatomy (FBbt)	Drosophila anatomy	For Drosophila samples
Anatomy	WormBase Anatomy (WBbt)	C. elegans anatomy	For C. elegans samples
Anatomy	Zebrafish Anatomy (ZFA)	Zebrafish anatomy and development	For zebrafish samples
Disease (see Disease Annotation Guidelines)			
Disease	Mondo Disease Ontology (MONDO)	Unified disease ontology	RECOMMENDED
Disease	Experimental Factor Ontology (EFO)	Disease terms from EFO	
Healthy samples	Phenotype And Trait Ontology (PATO)	Use `normal` (PATO:0000461) for healthy samples	
Cell Lines			
Cell Lines	Cellosaurus	Cell line knowledge resource	RECOMMENDED

Category	Ontology/CV	Description	Notes
Cell Lines	Cell Line Ontology (CLO)	Cell line ontology	Legacy support only
Mass Spectrometry and Proteomics			
MS/Proteomics	PSI Mass Spectrometry CV (PSI-MS)	Instruments, methods, parameters	
Modifications	Unimod	Protein modifications database	
Modifications	PSI-MOD CV	Protein modifications ontology	
Other			
Chemistry	ChEBI	Chemical Entities of Biological Interest	
Environment	Environment Ontology (ENVO)	Environmental sample classification	For metaproteomics
Ancestry	Human Ancestry Ontology (HANCESTRO)	Human ancestry categories	For human samples

Chapter 9. SDRF-Proteomics: Samples metadata

The Sample metadata section provides information about the samples of origin and their characteristics. Each sample contains a *source name* (unique identifier) and a set of *characteristics* columns. The first column of the file should be the *source name* and the following columns should be the characteristics of the sample. For example, for any proteomics experiment (human, vertebrate, cell line), the following characteristics should be provided:

- **source name**: Unique sample name (it can be present multiple times if the same sample is used several times in the same dataset)
- **characteristics[organism]**: The organism of the Sample of origin.
- **characteristics[organism part]**: The part of organism's anatomy or substance arising from an organism from which the biomaterial was derived, (e.g., liver)
- **characteristics[disease]**: The disease under study in the Sample. For healthy/control samples, use `normal` (see Disease Annotation Guidelines).
- **characteristics[cell type]**: A cell type is a distinct morphological or functional form of cell. Examples are epithelial, glial etc.

Example:

source name	characteristics [organism]	characteristics [organism part]	characteristics [disease]	characteristics [cell type]
sample_treat	homo sapiens	liver	liver cancer	not available
sample_control	homo sapiens	liver	liver cancer	not available

NOTE

Additional characteristics can be added depending on the type of the experiment and sample. The SDRF-Proteomics templates defines a set of templates and checklists of properties that should be provided depending on the proteomics experiment. In the core guidelines and templates, main document of SDRF-Proteomics, we explain the major sample properties for different experiments. However, SDRF-Proteomics can be extended using guidelines for specific experiments.

Some important notes:

- Each characteristic name in the column header SHOULD be a CV term from the EFO ontology. For example, the header *characteristics[organism]* corresponds to the ontology term Organism. However the values could be from EFO or other ontologies. For example, we RECOMMEND to use MONDO for diseases because it has better coverage than EFO. For healthy samples, use `normal` (PATO:0000461) - see Disease Annotation Guidelines.
- Multiple values (columns) for the same characteristics term are allowed in SDRF-Proteomics. However, it is RECOMMENDED not to use the same column in the same file. If you have multiple phenotypes, you can specify what it refers to or use another more specific term, e.g., "immunophenotype".

9.1. BioSamples database integration

BioSamples provides persistent identifiers for biological samples that enable cross-database linking [5]. Use the optional *characteristics[biosample accession number]* column to link samples to BioSamples entries (EBI format: SAMEA*, NCBI format: SAMN*).

Property	Column Name	Example Value	Description
BioSample Accession	characteristics[biosample accession number]	SAMN12345678	The unique BioSamples database accession number for the sample (NCBI format)
BioSample Accession	characteristics[biosample accession number]	SAMEA12345678	The unique BioSamples database accession number for the sample (EBI format)

Example usage:

source name	**characteristics [biosample accession number]**	**characteristics [organism]**	**characteristics [organism part]**	**characteristics [disease]**
sample_001	SAMN12345678	homo sapiens	liver	liver cancer
sample_002	SAMN12345679	homo sapiens	liver	normal
sample_003	SAMEA12345680	mus musculus	brain	normal

The BioSamples accession number enables:

- **Cross-database linking**: Connect proteomics datasets with genomics, transcriptomics, and other omics data that reference the same biological samples
- **Enhanced metadata**: Access additional sample metadata stored in the BioSamples database
- **Data provenance**: Trace the origin and history of biological samples across multiple studies
- **Improved findability**: Enable discovery of related datasets through shared sample identifiers

When a BioSample accession number is provided, the full sample metadata from the corresponding BioSamples database becomes available and can complement the information provided in the SDRF-Proteomics file. If there are discrepancies between the SDRF-Proteomics metadata and the BioSamples metadata, the SDRF-Proteomics values take precedence for the specific proteomics experiment context.

NOTE

BioSample accession numbers from NCBI follow the format `SAMNxxxxxxxxxx` and from EBI follow the format `SAMEAxxxxxxxxx`, where `x` represents digits. Either NCBI or EBI BioSample accession numbers can be used depending on where the sample is registered. The *characteristics[biosample accession number]* column is optional, but when available, providing BioSample accession numbers is RECOMMENDED to

> enhance data integration and reusability. Users must first request BioSample accession numbers from the appropriate service (NCBI or EBI) before including them in their SDRF files.

9.2. Encoding sample technical and biological replicates

SDRF-Proteomics uses two columns to track replicates [4]:

- **characteristics[biological replicate]**: Identifies independent biological samples (e.g., different patients, different cell cultures)
- **comment[technical replicate]**: Identifies repeated measurements of the same sample (e.g., multiple LC-MS/MS injections)

The following example shows 2 biological replicates, each with 2 fractions and 2 technical replicates:

source name	characteristics [biological replicate]	assay name	comment [label]	comment [fraction identifier]	comment [technical replicate]	comment [data file]
patient_001_sample	1	run_01	label free sample	1	1	P001_F1_TR1.raw
patient_001_sample	1	run_02	label free sample	2	1	P001_F2_TR1.raw
patient_001_sample	1	run_03	label free sample	1	2	P001_F1_TR2.raw
patient_001_sample	1	run_04	label free sample	2	2	P001_F2_TR2.raw
patient_002_sample	2	run_05	label free sample	1	1	P002_F1_TR1.raw
patient_002_sample	2	run_06	label free sample	2	1	P002_F2_TR1.raw
patient_002_sample	2	run_07	label free sample	1	2	P002_F1_TR2.raw
patient_002_sample	2	run_08	label free sample	2	2	P002_F2_TR2.raw

In this example:

- **Biological replicates**: `patient_001_sample` and `patient_002_sample` are different biological samples (different source names), annotated with `characteristics[biological replicate]` values 1 and 2
- **Technical replicates**: Each biological sample is measured twice (`comment[technical replicate]` = 1 and 2)

- **Fractions**: Each technical replicate has 2 fractions (`comment[fraction identifier]` = 1 and 2)

IMPORTANT	Both *characteristics[biological replicate]* and *comment[technical replicate]* columns are REQUIRED. When no replicates are performed in a study, set both columns to 1 (i.e., each sample is biological replicate 1 and technical replicate 1). For **pooled reference samples** (e.g., TMT reference channels), use `pooled` for biological replicate since these samples are mixtures of multiple replicates and assigning a specific replicate number would be misleading.

Some examples with explicit annotation of the biological replicates can be found here:

- https://github.com/bigbio/proteomics-metadata-standard/blob/c3a56b076ef381280dfcb0140d2520126ace53ff/annotated-projects/PXD006401/sdrf.tsv

9.3. Pooled samples

When multiple samples are pooled into one, the general approach is to annotate them separately, abiding by the general rule: one row stands for one sample-to-file relationship. In this case, multiple rows are created for the corresponding data file, much like in multiplexed labeling experiments (see Label Annotations).

One possible exception is made for the case when one channel (e.g., in a TMT/iTRAQ multiplexed experiment) is used for a sample pooled from all other channels, typically for normalization purposes. In this case, it is not necessary to repeat all sample annotations. Instead, the *characteristics[pooled sample]* column SHOULD be used.

9.3.1. Allowed values for characteristics[pooled sample]

The *characteristics[pooled sample]* column accepts the following values:

Value	Description	When to Use
not pooled	Sample is not pooled, represents a single biological sample	Regular individual samples
pooled	Sample is pooled but individual source samples cannot be annotated	When pooling details are unknown or samples are from external sources
SN=sample1;SN=sample2;…	Structured format listing source names of pooled samples	When individual samples are known and annotated in the same SDRF file

NOTE	The `SN` key stands for "source name" and lists the `source name` values of samples that are annotated in the same file and used in the same experiment and same MS run. Use semicolons to separate multiple entries.

9.3.2. Example with simple pooled annotation

When pooling details are unknown or samples come from external sources, use the simple `pooled` value:

source name	characteristics [pooled sample]	characteristics [organism]	assay name	comment [label]	comment [data file]
sample_1	not pooled	homo sapiens	run_1	TMT126	file01.raw
sample_2	not pooled	homo sapiens	run_1	TMT127N	file01.raw
pooled_ref	pooled	homo sapiens	run_1	TMT131C	file01.raw

9.3.3. Example with detailed pooled reference

When pooled samples are known and annotated in the same SDRF file, use the `SN=` format:

source name	characteristics [pooled sample]	characteristics [organism]	characteristics [age]	characteristics [sex]	assay name	comment [label]	comment [data file]
sample_1	not pooled	homo sapiens	45Y	male	run_1	TMT126	file01.raw
sample_2	not pooled	homo sapiens	52Y	female	run_1	TMT127N	file01.raw
sample_3	not pooled	homo sapiens	38Y	male	run_1	TMT127C	file01.raw
pooled_ref	SN=sample_1;SN=sample_2;SN=sample_3	homo sapiens	pooled	pooled	run_1	TMT131C	file01.raw

TIP

For pooled reference samples (e.g., TMT reference channels), use `pooled` for individual-specific fields including **biological replicate**, age, sex, and individual. This clearly indicates that the value represents a mixture rather than a single sample. If all pooled samples share a value (e.g., all females, or age range 40Y-50Y), that shared value MAY be used instead.

9.4. Spiked-in samples

There are multiple scenarios when a sample is spiked with additional analytes. Peptides, proteins, or mixtures can be added to the sample as controlled amounts to provide a standard or ground truth for quantification, or for retention time alignment, etc.

To include information about the spiked compounds, use *characteristics[spiked compound]*. The information is provided in key-value pairs. Here are the keys and values that SHOULD be provided:

Key	Meaning	Examples	Peptide	Protein	Mixture	Other
CT	Compound type	protein, peptide, mixture, other	Required	Required	Required	Required
QY	Quantity (molar or mass)	10 mg, 20 nmol	Required	Required	Required	Required
PS	Peptide sequence	PEPTIDESEQ	Required	-	-	-
AC	Uniprot Accession	A9WZ33	-	Required	-	-
CN	Compound name	iRT mixture, substance name	Optional	Optional	Optional	Optional
SP	Species	Escherichia coli K-12	Optional	Optional	Optional	Optional
CV	Compound vendor	in-house or vendor name	Optional	Optional	Required	Optional
CS	Compound specification URI	http://vendor.web.site/specs/kit.xlsx	Optional	Optional	Optional	Optional
CF	Compound formula	C2H2O	-	-	-	Optional

In addition to specifying the component and its quantity, the injected mass of the main sample SHOULD be specified as *characteristics[mass]*.

An example of SDRF-Proteomics for a sample spiked with a peptide would be:

characteristics [mass]	characteristics [spiked compound]
1 ug	CT=peptide;PS=PEPTIDESEQ;QY=10 fmol

For multiple spiked components, the column *characteristics[spiked compound]* may be repeated.

If the spiked component is another biological sample (e.g. *E. coli* lysate spiked into human sample), then the spiked component MUST be annotated in its own row. Both components of the sample SHOULD have `characteristics[mass]` specified. Inclusion of *characteristics[spiked compound]* is optional in this case; if provided, it SHOULD be the string `spiked` for the spiked sample.

9.5. Sample Metadata Guidelines

For detailed guidance on annotating sample metadata, refer to the following conventions documents:

- Sample Metadata Guidelines - Detailed guidelines for age, sex, disease, organism part, cell type, developmental stage, and other sample characteristics
- Human Sample Metadata Guidelines - Human-specific metadata including disease staging, treatment history, demographics, and lifestyle factors

Chapter 10. SDRF-Proteomics: data files metadata

The connection between samples and data files is done using properties annotated with the **comment** prefix. All properties referring to a data file (e.g., MS run file) are annotated with the category `comment`. This differentiates data file properties from sample properties (characteristics).

The following properties MUST be provided for each data file:

Column	Requirement	Description
assay name	REQUIRED	Unique identifier for an MS run/data file
technology type	REQUIRED	Technology used to capture the data
comment[proteomics data acquisition method]	REQUIRED	DDA, DIA, PRM, SRM
comment[label]	REQUIRED	Label applied to sample (or "label free sample")
comment[instrument]	REQUIRED	Mass spectrometer model
comment[cleavage agent details]	REQUIRED	Enzyme information (use "not available" for undigested samples)
comment[fraction identifier]	REQUIRED	Fraction number (1 if not fractionated)
comment[technical replicate]	REQUIRED	Technical replicate number (1 if none)
comment[data file]	REQUIRED	Name of the raw file

Example:

source name	assay name	technology type	comment [proteomics data acquisition method]	comment [label]	comment [instrument]	comment [data file]
sample_1	sample1_run1	proteomic profiling by mass spectrometry	data-dependent acquisition	label free sample	Q Exactive HF	sample1.raw

10.1. Sample preparation properties

In order to encode sample preparation details, we strongly RECOMMEND specifying the following parameters:

- **comment[depletion]**: The removal of specific components of a complex mixture of proteins or peptides based on some specific property of those components. The values of the columns will be `no depletion` or `depletion`. In the case of depletion `depleted fraction` or `bound fraction` can be specified.
- **comment[reduction reagent]**: The chemical reagent that is used to break disulfide bonds in proteins. The values of the column are under the term reduction reagent. For example, DTT.
- **comment[alkylation reagent]**: The alkylation reagent that is used to covalently modify cysteine SH-groups after reduction, preventing them from forming unwanted novel disulfide bonds. The values of the column are under the term alkylation reagent. For example, IAA.
- **comment[fractionation method]**: The fraction method used to separate the sample. The values of this term can be read under PRIDE ontology term Fractionation method. For example, Off-gel electrophoresis.

10.2. MS/MS properties

- **comment[collision energy]**: Collision energy can be added as non-normalized (10000 eV) or normalized (1000 NCE) value.
- **comment[dissociation method]**: This property will provide information about the fragmentation method, like HCD, CID. The values of the column are under the term dissociation method.

10.3. Data acquisition

Proteomics data acquisition method can happen in multiple ways: Data Dependent Acquisition (DDA), Data Independent Acquisition (DIA), and targeted approaches. The SDRF-Proteomics file format REQUIRES capturing the method used for the data acquisition in the *comment[proteomics data acquisition method]* column. The values MUST be children of the PRIDE ontology term proteomics data acquisition method (PRIDE:0000659). The following values are commonly used:

- data-dependent acquisition
- data-independent acquisition
  - diaPASEF
  - SWATH MS
- parallel reaction monitoring
- selected reaction monitoring

IMPORTANT	The *comment[proteomics data acquisition method]* column is REQUIRED for all mass spectrometry-based SDRF files. This field must be explicitly specified and cannot be omitted or assumed.

You can find an example of a DIA experiment in the following link: DIA example

TIP	For DIA experiments, additional properties like MS1 scan range can be captured. See DIA Scan Window Limits in the Data File Metadata Guidelines.

10.4. Data File Metadata Guidelines

For detailed guidance on data file metadata, refer to the conventions document:

- Data File Metadata Guidelines - Detailed guidelines for labels, instruments, modifications, cleavage agents, mass tolerances, RAW file URIs, and other data file properties

Chapter 11. Additional SDRF Rules

11.1. Row Uniqueness Requirements

SDRF files must satisfy specific uniqueness constraints to ensure data integrity and enable proper indexing by analysis tools.

Error-level constraint (validation fails): The combination of `source name` + `assay name` + `comment[label]` MUST be unique across all rows in the SDRF file. If two rows have identical values for all three columns, validation will fail with an error. This constraint ensures that each sample-run-label combination can be uniquely identified.

Warning-level constraint (validation warns): The combination of `source name` + `assay name` SHOULD be unique across all rows. Non-unique combinations will generate a warning during validation. This constraint helps identify potential issues where the same sample appears to have multiple entries for the same MS run without distinguishing labels.

Assay name uniqueness: Each distinct MS run/data file MUST have exactly one globally unique `assay name`, and no two different data files may share an assay name. To ensure uniqueness, it is RECOMMENDED to incorporate sample-specific information in assay names, such as sample IDs or replicate numbers (e.g., "sample1_run1", "sample1_run2", "patient001_fraction01").

NOTE

For multiplexed experiments (e.g., TMT, iTRAQ), multiple SDRF rows will share the same `assay name` because multiple samples are analyzed in a single MS run. In these cases, the `comment[label]` column distinguishes between different samples within the same run, and the combination of `source name` + `assay name` + `comment[label]` remains unique.

Example of valid multiplexed experiment:

source name	…	assay name	comment [label]	…	comment [data file]
sample_A	…	TMT_batch1_run1	TMT126	…	batch1_run1.raw
sample_B	…	TMT_batch1_run1	TMT127N	…	batch1_run1.raw
sample_C	…	TMT_batch1_run1	TMT127C	…	batch1_run1.raw
sample_D	…	TMT_batch1_run1	TMT128N	…	batch1_run1.raw

In this example, all four rows share the same `assay name` and `comment[data file]` because they represent different samples multiplexed in a single MS run. The combination of `source name` + `assay name` + `comment[label]` is unique for each row.

Chapter 12. Core Templates

SDRF-Proteomics provides core templates that define the required and recommended metadata columns based on the sample organism. Each template includes both **sample metadata** (characteristics) and **data file metadata** (comments) - everything needed to create a complete SDRF file.

Choose the appropriate template based on your sample organism:

- Default template: Basic template for any proteomics experiment
- Human template: For human samples with additional clinical metadata (age, sex, ancestry)
- Vertebrates template: For non-human vertebrate species (mouse, rat, zebrafish)
- Invertebrates template: For insects (Drosophila), nematodes (C. elegans), and other invertebrates
- Plants template: For plant species (Arabidopsis, crops)

For detailed explanations of each column, see sample metadata for sample properties and data file metadata for data file properties.

12.1. Default Template

The default template is the most basic template that can be used for any proteomics experiment. Use this when no other template is applicable.

Checklist:

Column	Category	Requirement	Example
Sample Metadata			
source name	Sample	Required	sample_1
characteristics[organism]	Sample	Required	homo sapiens
characteristics[organism part]	Sample	Required	liver
characteristics[disease]	Sample	Required	normal
characteristics[biological replicate]	Sample	Required	1
Data File Metadata			
assay name	Data	Required	run_1
technology type	Data	Required	proteomic profiling by mass spectrometry
comment[proteomics data acquisition method]	Data	Required	data-dependent acquisition

Column	Category	Requirement	Example
comment[label]	Data	Required	label free sample
comment[instrument]	Data	Required	Q Exactive HF
comment[cleavage agent details]	Data	Required	NT=Trypsin;AC=MS:1001251
comment[fraction identifier]	Data	Required	1
comment[technical replicate]	Data	Required	1
comment[data file]	Data	Required	sample_1.raw

Template file: sdrf-default.sdrf.tsv

NOTE

The characteristics[cell type] column is RECOMMENDED across all templates when the cell type is known or can be determined. Use `not available` if the cell type cannot be determined (e.g., whole tissue samples, mixed cell populations). For cell line experiments, use the cell-lines template which provides more specific guidance.

12.2. Human Template

The human template extends the default template with clinical and demographic metadata required for human samples.

Checklist:

Column	Category	Requirement	Example
Sample Metadata			
source name	Sample	Required	patient_001
characteristics[organism]	Sample	Required	homo sapiens
characteristics[organism part]	Sample	Required	liver
characteristics[disease]	Sample	Required	hepatocellular carcinoma
characteristics[cell type]	Sample	Recommended	hepatocyte
characteristics[biological replicate]	Sample	Required	1
characteristics[age]	Sample	Required	45Y
characteristics[sex]	Sample	Required	male

Column	Category	Requirement	Example
characteristics[ancestry category]	Sample	Recommended	European
characteristics[individual]	Sample	Recommended	P001
Data File Metadata			
assay name	Data	Required	patient_001_run1
technology type	Data	Required	proteomic profiling by mass spectrometry
comment[proteomics data acquisition method]	Data	Required	data-dependent acquisition
comment[label]	Data	Required	label free sample
comment[instrument]	Data	Required	Orbitrap Exploris 480
comment[cleavage agent details]	Data	Required	NT=Trypsin;AC=MS:1001251
comment[fraction identifier]	Data	Required	1
comment[technical replicate]	Data	Required	1
comment[data file]	Data	Required	patient_001.raw

Template file: sdrf-human.sdrf.tsv

> **NOTE** The characteristics[individual] column is optional and is only used when a code for the individual is available; if not, use **not available**. For age encoding format, see Sample Metadata Guidelines.

12.3. Vertebrates Template

The vertebrates template is used for non-human vertebrate species such as mouse, rat, zebrafish, and other model organisms.

Checklist:

Column	Category	Requirement	Example
Sample Metadata			
source name	Sample	Required	mouse_001
characteristics[organism]	Sample	Required	mus musculus

Column	Category	Requirement	Example
characteristics[organism part]	Sample	Required	brain
characteristics[disease]	Sample	Required	normal
characteristics[cell type]	Sample	Recommended	neuron
characteristics[biological replicate]	Sample	Required	1
characteristics[developmental stage]	Sample	Recommended	adult
Data File Metadata			
assay name	Data	Required	mouse_001_run1
technology type	Data	Required	proteomic profiling by mass spectrometry
comment[proteomics data acquisition method]	Data	Required	data-dependent acquisition
comment[label]	Data	Required	label free sample
comment[instrument]	Data	Required	timsTOF Pro
comment[cleavage agent details]	Data	Required	NT=Trypsin;AC=MS:1001251
comment[fraction identifier]	Data	Required	1
comment[technical replicate]	Data	Required	1
comment[data file]	Data	Required	mouse_001.raw

Template file: sdrf-vertebrates.sdrf.tsv

12.4. Invertebrates Template

The invertebrates template is used for non-vertebrate animal species such as insects (Drosophila), nematodes (C. elegans), and other invertebrate model organisms.

Checklist:

Column	Category	Requirement	Example
Sample Metadata			
source name	Sample	Required	fly_sample_1

Column	Category	Requirement	Example
characteristics[organism]	Sample	Required	drosophila melanogaster
characteristics[organism part]	Sample	Required	head
characteristics[disease]	Sample	Required	normal
characteristics[cell type]	Sample	Recommended	neuron
characteristics[biological replicate]	Sample	Required	1
Data File Metadata			
assay name	Data	Required	fly_sample_1_run1
technology type	Data	Required	proteomic profiling by mass spectrometry
comment[proteomics data acquisition method]	Data	Required	data-dependent acquisition
comment[label]	Data	Required	label free sample
comment[instrument]	Data	Required	Q Exactive Plus
comment[cleavage agent details]	Data	Required	NT=Trypsin;AC=MS:1001251
comment[fraction identifier]	Data	Required	1
comment[technical replicate]	Data	Required	1
comment[data file]	Data	Required	fly_sample_1.raw

Template file: sdrf-invertebrates.sdrf.tsv

NOTE For Drosophila samples, use FBbt (FlyBase anatomy ontology) for organism part. For C. elegans, use WBbt (WormBase anatomy ontology).

12.5. Plants Template

The plants template is used for plant species including model organisms like Arabidopsis thaliana and crop species.

Checklist:

Column	Category	Requirement	Example
Sample Metadata			

Column	Category	Requirement	Example
source name	Sample	Required	arabidopsis_col0_1
characteristics[organism]	Sample	Required	arabidopsis thaliana
characteristics[organism part]	Sample	Required	leaf
characteristics[disease]	Sample	Required	normal
characteristics[cell type]	Sample	Recommended	guard cell
characteristics[biological replicate]	Sample	Required	1
Data File Metadata			
assay name	Data	Required	arabidopsis_col0_run1
technology type	Data	Required	proteomic profiling by mass spectrometry
comment[proteomics data acquisition method]	Data	Required	data-dependent acquisition
comment[label]	Data	Required	label free sample
comment[instrument]	Data	Required	Orbitrap Fusion Lumos
comment[cleavage agent details]	Data	Required	NT=Trypsin;AC=MS:1001251
comment[fraction identifier]	Data	Required	1
comment[technical replicate]	Data	Required	1
comment[data file]	Data	Required	arabidopsis_col0.raw

Template file: sdrf-plants.sdrf.tsv

> **NOTE** For plant samples, use PO (Plant Ontology) for organism part and cell type annotations.

12.6. Column Cardinality

Some columns can appear multiple times for the same sample. The cardinality rules are:

- **Single (1)**: Column appears exactly once per sample (e.g., biological replicate)
- **Multiple (*)**: Column can appear multiple times (e.g., organism part can specify both "heart" and "heart left ventricle")

Example of multiple organism part columns:

source name	…	characteristics [organism part]	characteristics [organism part]	…
sample-1	…	heart	heart left ventricle	…

The template files can be downloaded from the core-templates folder.

Chapter 13. Factor Values (Study Variables)

Factor values identify the experimental variables being studied - the conditions you want to compare in your analysis. They highlight which sample characteristics are the focus of your experiment.

13.1. Column Format

`factor value[{variable name}]`

13.2. When to Use Factor Values

Use factor values to indicate:

- The primary variable(s) under investigation
- Conditions being compared (e.g., disease vs. normal, treated vs. untreated)
- Variables that define experimental groups

> **NOTE** Use `normal` (not "control") in the disease field for healthy samples. "Control" is an experimental design concept, not a disease state. See Disease Annotation Guidelines for details.

13.3. Rules

- Factor value columns SHOULD appear after all characteristics and comment columns
- Multiple factor values can be used when studying multiple variables
- The value in a factor value column typically mirrors a characteristics column value

13.4. Example

In an experiment comparing tumor vs. normal tissue across different cancer stages:

source name	...	characteristics [disease]	characteristics [disease stage]	...	factor value [disease]	factor value [disease stage]
tumor_sample_1	…	breast carcinoma	stage II	…	breast carcinoma	stage II
normal_sample_1	…	normal	not applicable	…	normal	not applicable
tumor_sample_2	…	breast carcinoma	stage III	…	breast carcinoma	stage III

In this example, both `disease` and `disease stage` are factor values because the experiment aims to compare expression differences between disease states and across cancer stages.

Chapter 14. Experiment-specific Templates

For specialized proteomics experiments that require additional metadata beyond the core templates, experiment-specific templates provide detailed guidelines and checklists. These templates define experiment-specific metadata fields while maintaining full compatibility with the core SDRF-Proteomics format.

Choose the appropriate template based on your experiment type:

- **Affinity Proteomics**: Olink, SomaScan, Luminex, and other affinity-based methods
- **Cell Lines**: Standardized cell line annotation using Cellosaurus
- **Crosslinking**: XL-MS structural proteomics experiments
- **Immunopeptidomics**: MHC-bound peptide identification
- **Metaproteomics**: Microbial community proteomics
- **Single Cell**: Single cell proteomics experiments

For detailed specifications and examples, visit the Experiment-specific Templates documentation.

Chapter 15. Examples of Annotated Datasets

The following table provides links to example SDRF files for different experiment types. These can serve as references when creating your own SDRF files.

Experiment Type	Dataset	Description	SDRF URL
Label-free	PXD008934	Human proteome label-free quantification	View SDRF
TMT	PXD017710	TMT-labeled quantitative proteomics	View SDRF
SILAC	PXD000612	SILAC-based quantification	View SDRF
DIA	PXD018830	Data-independent acquisition	View SDRF
Phosphoproteomics	PXD000759	PTM enrichment study	View SDRF
Cell lines	PXD001819	Cell line proteomics	View SDRF

A comprehensive collection of annotated projects is available at: Annotated Projects Repository

Chapter 16. Ongoing template discussions

We have created a file in GitHub Ongoing template discussions where we aggregate all the ongoing discussions about the format and new templates.

Chapter 17. Intellectual Property Statement

The PSI takes no position regarding the validity or scope of any intellectual property or other rights that might be claimed to pertain to the implementation or use of the technology described in this document or the extent to which any license under such rights might or might not be available; neither does it represent that it has made any effort to identify any such rights. Copies of claims of rights made available for publication and any assurances of licenses to be made available or the result of an attempt made to obtain a general license or permission for the use of such proprietary rights by implementers or users of this specification can be obtained from the PSI Chair.

The PSI invites any interested party to bring to its attention any copyrights, patents or patent applications, or other proprietary rights which may cover technology that may be required to practice this recommendation. Please address the information to the PSI Chair (see contacts information at PSI website).

Chapter 18. Copyright Notice

Chapter 19. How to cite

Please cite this document as:

Dai C, Füllgrabe A, Pfeuffer J, Solovyeva EM, Deng J, Moreno P, Kamatchinathan S, Kundu DJ, George N, Fexova S, Grüning B, Föll MC, Griss J, Vaudel M, Audain E, Locard-Paulet M, Turewicz M, Eisenacher M, Uszkoreit J, Van Den Bossche T, Schwämmle V, Webel H, Schulze S, Bouyssié D, Jayaram S, Duggineni VK, Samaras P, Wilhelm M, Choi M, Wang M, Kohlbacher O, Brazma A, Papatheodorou I, Bandeira N, Deutsch EW, Vizcaíno JA, Bai M, Sachsenberg T, Levitsky LI, Perez-Riverol Y. A proteomics sample metadata representation for multiomics integration and big data analysis. Nat Commun. 2021 Oct 6;12(1):5854. doi: 10.1038/s41467-021-26111-3. PMID: 34615866; PMCID: PMC8494749. [Manuscript - https://www.nature.com/articles/s41467-021-26111-3]

References

- [1] Y. Perez-Riverol, S. European Bioinformatics Community for Mass, Toward a Sample Metadata Standard in Public Proteomics Repositories, J Proteome Res 19(10) (2020) 3906-3909. doi:10.1021/acs.jproteome.0c00376
- [2] A. Gonzalez-Beltran, E. Maguire, S.A. Sansone, P. Rocca-Serra, linkedISA: semantic representation of ISA-Tab experimental metadata, BMC Bioinformatics 15 Suppl 14 (2014) S4. doi:10.1186/1471-2105-15-S14-S4
- [3] T.F. Rayner, P. Rocca-Serra, P.T. Spellman, et al., A simple spreadsheet-based, MIAME-supportive format for microarray data: MAGE-TAB, BMC Bioinformatics 7 (2006) 489. doi:10.1186/1471-2105-7-489
- [4] P. Blainey, M. Krzywinski, N. Altman, Points of significance: replication, Nat Methods 11(9) (2014) 879-80. doi:10.1038/nmeth.3091
- [5] D. Gupta, I. Liyanage, Y. Perez-Riverol, et al., BioSamples database: the global hub for sample metadata and multi-omics integration, Nucleic Acids Res (2025). doi:10.1093/nar/gkaf1133