PSI Recommendation

PSI Mass Spectrometry and Proteomics Informatics Working Groups

Status: DRAFT

Yasset Perez-Riverol, EMBL-EBI, UK

Chengxin Dai, Chongqing University Telecommunications, Chongqing, China

Timo Sachsenberg, Tübingen University, Germany

Anja Fullgrabe, EMBL-EBI, U.K

Elizaveta Solovyeva, INEPCP RAS, Moscow, Russia

Marc Vaudel, University of Bergen, Norway

Stefan Schulze, University of Pennsylvania, USA

Veit Schwämmle, University of Southern Denmark, Denmark

ProteomicsDB Team, Technical University of Munich, Germany

Juan Antonio Vizacaíno, EMBL-EBI

Johannes Griss, Medical University of Vienna, Austria

Lev Levitsky, INEPCP RAS, Moscow, Russia

Status of this document

This document provides information to the proteomics community about a proposed standard for sample metadata annotations in public repositories called SDRF. Distribution is unlimited.

Version Draft 11 - this is a draft of version 1.0

# 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 a sample metadata annotation of proteomics experiments.

Further detailed information, including any updates to this document, implementations, and examples is available at <https://github.com/bigbio/proteomics-metadata-standard>.

Contents

[Abstract 1](#_Toc52531504)

[1. Introduction 4](#_Toc52531505)

[1.1 Description of the need 4](#_Toc52531506)

[1.2 Requirements 4](#_Toc52531507)

[1.3 Issues to be addressed 5](#_Toc52531508)

[2. Notational Conventions 5](#_Toc52531509)

[3. The Peptidoform Notation Definition 5](#_Toc52531510)

[3.1 The documentation 5](#_Toc52531511)

[3.2 Relationship to other specifications 5](#_Toc52531512)

[4. The Basic Form of the Proteoform and Peptidoform Notation 6](#_Toc52531513)

[4.1 The canonical amino acid sequence 6](#_Toc52531514)

[4.2 Generic representation of protein modifications 7](#_Toc52531515)

[4.2.1 Controlled vocabulary or ontology modification names 7](#_Toc52531516)

[4.2.1.1 Definition of the Unimod modification name 8](#_Toc52531517)

[4.2.2 Controlled vocabulary or ontology protein modification accession numbers 8](#_Toc52531518)

[4.2.3 Support for cross-linkers 9](#_Toc52531519)

[4.2.3.1 Crosslink notation (within the same peptide) 9](#_Toc52531520)

[4.2.3.2 Representing inter-chain crosslinks 9](#_Toc52531521)

[4.2.3.3 Representing disulfide linkages 10](#_Toc52531522)

[4.2.4 Representation of glycans using the GNO ontology as CV 10](#_Toc52531523)

[4.2.5 Delta mass notation 11](#_Toc52531524)

[4.2.6 Specifying a gap of known mass 12](#_Toc52531525)

[4.2.7 Support for elemental formulas (e.g. for representing small molecular substructures or functional groups) 12](#_Toc52531526)

[4.2.8 Representation of glycan composition 13](#_Toc52531527)

[4.2.9 Best practises on the use of protein modifications 14](#_Toc52531528)

[4.3 Representation of special cases: N-terminal, C-terminal and labile protein modifications 14](#_Toc52531529)

[4.3.1 N-terminal and C-terminal modifications 14](#_Toc52531530)

[4.3.2 Labile modifications 15](#_Toc52531531)

[4.4 Support for the representation of ambiguity in the modification position 15](#_Toc52531532)

[4.4.1 Unknown modification position 15](#_Toc52531533)

[4.4.2 Indicating a possible set of modification positions 16](#_Toc52531534)

[4.4.3 Representing ranges of positions for the modifications 16](#_Toc52531535)

[4.4.4 Indicating modification position preference and localisation scores 17](#_Toc52531536)

[4.4.5 Representing scoring for ranges of positions for a modification 17](#_Toc52531537)

[4.5 Representation of multiple modifications in the same amino acid residue 17](#_Toc52531538)

[4.6 Representation of global modifications 18](#_Toc52531539)

[4.6.1 Use Case 1: Representation of isotopes 18](#_Toc52531540)

[4.6.2 Use Case 2: Fixed protein modifications 18](#_Toc52531541)

[4.7 The information tag 18](#_Toc52531542)

[4.8 Support for the joint representation of experimental data and its interpretation 19](#_Toc52531543)

[5. Pending Issues - Future developments 20](#_Toc52531544)

[5.1 Representation of cyclic peptides 20](#_Toc52531545)

[5.2 Representation of ambiguity when different glycans are attached to the same amino acid sequence 21](#_Toc52531546)

[5.3 Representation of rare amino acids not supported by the one letter code 22](#_Toc52531547)

[5.4 Representation of average masses 22](#_Toc52531548)

[5.5 Representation of lipids 22](#_Toc52531549)

[5.6 Distribution of isotopes in the sequence 23](#_Toc52531550)

[5.7 Representation of molecular formula 23](#_Toc52531551)

[5.8 Representation of overlapping range of possible modification localizations 23](#_Toc52531552)

[5.9 Representation of ambiguous crosslinker modification positions 23](#_Toc52531553)

[6. Appendix I. Levels of Compliance 24](#_Toc52531554)

[7. Appendix II: Extensions to improve the representation of PSMs in mass spectra 26](#_Toc52531555)

[7.1 Representation of the ion charges 26](#_Toc52531556)

[7.2 Representation of multiple peptidoform assignments in chimeric spectra 26](#_Toc52531557)

[8. Appendix III. Glossary of terms used in the specification 28](#_Toc52531558)

[9. Authors Information 32](#_Toc52531559)

[10. Contributors 33](#_Toc52531560)

[11. Intellectual Property Statement 34](#_Toc52531561)

[12. Copyright Notice 34](#_Toc52531562)

[13. Glossary 35](#_Toc52531563)

[14. References 35](#_Toc52531564)

# Introduction

## **Description of the need**

Proteomics experimental design and sample related information are missing in puvlic repositories or stored in very diverse ways and formats. For example, the CPTAC consortium (<https://cptac-data-portal.georgetown.edu/>) provides for every dataset a set of excel files with the information on each sample (e.g. <https://cptac-data-portal.georgetown.edu/study-summary/S048>) 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 or instrument (e.g. <https://www.proteomicsdb.org/#projects/4267/6228>). For every proteomics dataset we should capture at least three levels of metadata: (i) dataset description, (ii) the sample to data files related information; and (iii) standard data file formats (e.g. mzIdentML, mzML, or mzTab). The general description includes a piece of minimum information to describe the study: title, description, date of publication, type of experiment (e.g. <http://proteomecentral.proteomexchange.org/cgi/GetDataset?ID=PXD016060.0-1&outputMode=XML>). The standard data files captured all the metadata associated with the dataset including search engine settings, scores, workflows, configuration files, among others. 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.

These three levels of metadata are combined in the well-established data formats ISA-TAB (<https://www.isacommons.org/>) or MAGE-TAB [1], which are used in other omics fields. In both data formats, a tab-delimited file is used to annotate the metadata about the sample and link it to the corresponding data file(s) (sample to data file format - SDRF). Both data formats encode the properties and sample attributes as columns, and each row represents a sample in the study. However, a careful review of an existing proteomics dataset annotated in ISA-TAB makes clear that not only a file format is needed, but most importantly, general guidelines about what information should be encoded in proteomics data repositories. The lack of guidelines to annotate information like disease stage, cell line code, or organism part; and the analytical information labeling channels (e.g. TMT, SILAC) or instrument configurations makes the data representation incomplete to understand the original experiment, reproduce the results or perform a re-analysis. If the information of the fraction, labeling, or enrichment method is not annotated, the reuse and reproduce of the original results will be challenging.

![A picture containing diagram

Description automatically generated]()

## **Requirements**

The SDRF (Sample and Data Relationship Format) describes the sample characteristics and the relationships between samples and data files, etc. The information in SDRF files is organised so that it follows the natural flow of a proteomics experiment. The main requirements to be fulfilled the sample to data file format are:

* The SDRF file is a tab-delimited format where each ROW corresponds to a relation between a Sample and a Data file.
* Each column MUST corresponds to an attribute/property of the Sample or the Data file.
* Each value in each cell MUST be the specific value of the property for a given Sample or Data file.
* The file MUST begin by describing the samples and finishes with the names of the data files generated from the analyses of the experimental results.
* Unknown values: In some cases, the column is mandatory but for some samples the 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 value is not applicable: In those cases, users SHOULD use **not applicable**.
* By specification the SDRF is **case insensitive**, but we RECOMMEND to use lower cases throughout all the text (column names and values).
* By specification the SDRF is insensitive to spaces (SourceName == source name). We RECOMMEND using the space representation because is more human readable (e.g. source name).

## **Issues to be addressed**

The main issues to be addressed by the SDRF are:

* It MUST be able to represent the sample metadata and the data file generated by the instruments or the analysis.
* It MUST be able to represent the experimental design including the way samples and data has been collected.

# Notational Conventions

The key words “MUST”, “MUST NOT”, “REQUIRED”, “SHALL”, “SHALL NOT”, “SHOULD”, “SHOULD NOT”, “RECOMMENDED”, “MAY”, and “OPTIONAL” are to be interpreted as described in RFC 2119 (2).

## 

# Sample metadata

## The list of ontologies/controlled vocabularies supported

## PRIDE Controlled Vocabulary (CV)

## PSI Mass Spectrometry CV

## Experimental Factor Ontology

## Unimod protein modification database for mass spectrometry

## PSI-MOD CV

## Cell line ontology

## Drosophila anatomy ontology

## Cell ontology

## Plant ontology

## Uber-anatomy ontology

## Zebrafish anatomy and development ontology

## Zebrafish developmental stages ontology

## Plant Environment Ontology

## FlyBase Developmental Ontology

## Rat Strain Ontology

## Chemical Entitities of Biological Interest Ontology

## NCBI organismal classification

## Minimum information about Samples

## The Sample metadata has different **Categories/Headings** to organize all the attributes / column headers of a given Sample.

| Property | **M**andatory (1)**O**ptional (0) | Cardinality | Description | Example |
| --- | --- | --- | --- | --- |
| Sourcename | 1 | 1 | The Sample name. | Sample 1 |
| characteristics | 1 | 1.. \* | “characteristics” column headings should contain an ontology property term in square brackets.\* Multiple *Characteristic* columns of the same category (e.g., “characteristics[organism part]”) are allowed. Typically the usage implies whole to part from left to right. | characteristics [Organism Part] |

## Each Sample in the experiment MUST contain a *Source Name*, and a collection of *characteristics*:

| source name | characteristics[organism] | characteristics[phenotype] | characteristics[compound] |
| --- | --- | --- | --- |
| sample\_treat | homo sapiens | necrotic tissue | drug A |
| sample\_control | homo sapiens | normal | none |

## Some important notes:

## Each *characteristics* name in the column header SHOULD be a control vocabulary (CV) term from the [EFO ontology](https://www.ebi.ac.uk/ols/ontologies/efo). For example, the Header characteristics[organism] corresponds to the ontology term [Organism](http://www.ebi.ac.uk/efo/EFO_0000634).

## Multiple values (columns) for the same *characteristics* term are allowed in the SDRF. However, we RECOMMENDED not to use same column in the same SDRF. If you have multiple phenotypes, you can specify what it refers to or use another more specific term, e.g. "immunophenotype".

* Each value in the row can be free-text or a control vocabulary term. In the previous example, the characteristics[phenotype] value CAN BE the free-text *control* or the corresponding ontology identifier in [EFO](https://www.ebi.ac.uk/ols/ontologies/efo) [*http://www.ebi.ac.uk/efo/EFO\_0001461*](http://www.ebi.ac.uk/efo/EFO_0001461)

## SDRF values

## The value for each property (e.g. *characteristics*, *comment*) corresponding to each sample 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.

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

## Ontology url (Computer readable): Users can provide the corresponding URI of the ontology/CV term as a value. This is recommended for enriched files where the client 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. For example:

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

## **From Samples to Data files**

## The connection between the *Sample* to the final Data file is done by using a series of properties and attributes. All the properties needed to relate a given *Sample* to the corresponding *file* are annotated with the category **comment**. The use of *comment* is mainly aimed at differentiating sample properties from the data properties.

## 

## The following properties SHOULD be provided for each data file (MSRun) file:

## assay name: For SDRF compatibilities we cannot use MSRun but *assay name*. Examples of assay name: run 1, run\_fraction\_1\_2

## comment[fraction identifier]: The *fraction identifier* allows to record the number of a given fraction. The fraction identifier corresponds to this [ontology term](https://www.ebi.ac.uk/ols/ontologies/ms/terms?iri=http%3A%2F%2Fpurl.obolibrary.org%2Fobo%2FMS_1000858). The fraction identifier MUST start from **1** and if the experiment is not fractionated, the annotator MUST use **1** for each MSRun.

## comment[label]: The *label* describes the label applied to each Sample (if any). In case of multiplex experiments such as TMT, SILAC, and/or ITRAQ the corresponding *label* SHOULD be added. For Label-free experiments the [label free sample] term MUST be added.

## comment[data file]: The *data file* provides the name of the raw file from the instrument.

## The order of the columns are important, **assay name** SHOULD we always before the comments. We RECOMMENDED to put the last column as comment[data file].

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | assay name | comment[label] | comment[fraction identifier] | comment[data file] |
| sample 1 | run 1 | label free sample | 1 | 000261\_C05\_P0001563\_A00\_B00K\_R1.RAW |
| sample 1 | run 2 | label free sample | 2 | 000261\_C05\_P0001563\_A00\_B00K\_R2.RAW |

## All the possible *label* values can be seen in the in the PRIDE CV under the [Label](https://www.ebi.ac.uk/ols/ontologies/pride/terms?iri=http%3A%2F%2Fpurl.obolibrary.org%2Fobo%2FPRIDE_0000514&viewMode=All&siblings=false) node.

## The "comment" columns in **SDRF** are included as a basic extensibility mechanism for local implementations. The name associated with the comment is included in square brackets in the column heading, and the value(s) entered in the body of the column. *comment* columns could be used in various ways - to provide references to external files like raw files, or to include identifiers of objects in external systems.

### **Sample technical and biological replicates**

## Different measurements are often categorized as (i) *Technical* or (ii) *Biological* replicates, based on whether they are (i) matched on all variables, e.g. same sample and same protocol; or (ii) different samples matched on explanatory variable(s), e.g. different patients receiving a placebo in a placebo vs. drug trial. Technical and biological replicates have different levels of independence, which must be taken into account during data interpretation. For a given experiment, there are different levels to which samples can be matched - e.g. same sample, same sample same protocol, same sample same protocol same covariates - the definition of technical replicate can therefore vary based on the number of variables included. In addition, an experiment might be used in multiple models with different explanatory variable(s), and biological replicates in one model would not be replicates in another. Therefore, *Technical vs. Biological* considerations, while sometimes relevant to analytical and statistical interpretation, fall beyond the scope of the sdrf format. However, data providers are encouraged to provide any identifier - e.g. *Biological\_replicate\_1*, *Technical\_replicate\_2* - that would help linking the samples to their analytical and statistical analysis as comments. A good starting point for the SDRF specification is the following:

## **Technical replicate**: repeated measurements of the same sample that represent independent measures of the random noise associated with protocols or equipment.

## In MS-based proteomics a technical can be, for example, doing the full sample preparation from extraction to mass spectrometry multi times to control variability in the instrument, or sample preparation. Another valid example is run repeat, replicate only one part of the analytical method, for example, run the sample twice on the LC-MS/MS. Technical replicates indicate if your measurements are scientifically robust or noisy, and how large the measured effect must be to stand out above that noise.

## In the following example, only if the technical replicate column is provided, one can distinguish quantitative values of the same fraction but different technical replicates.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| source name | assay name | comment[label] | comment[fraction identifier] | comment[technical replicate] | comment[data file] |
| Sample 1 | run 1 | label free sample | 1 | 1 | 000261\_C05\_P0001563\_A00\_B00K\_F1\_TR1.RAW |
| Sample 1 | run 2 | label free sample | 2 | 1 | 000261\_C05\_P0001563\_A00\_B00K\_F2\_TR1.RAW |
| Sample 1 | run 3 | label free sample | 1 | 2 | 000261\_C05\_P0001563\_A00\_B00K\_F1\_TR2.RAW |
| Sample 1 | run 4 | label free sample | 2 | 2 | 000261\_C05\_P0001563\_A00\_B00K\_F2\_TR2.RAW |

## The comment[technical replicate] column is OPTIONAL.

## **Biological replicate**: parallel measurements of biologically distinct samples that capture biological variation, which may itself be a subject of study or a source of noise. Biological replicates address if and how widely the results of an experiment can be generalized. For example, repeating a particular assay with independently generated samples, individuals or samples derived from various cell types, tissue types, or organisms, to see if similar results can be observed. Context is critical, and appropriate biological replicates will indicate whether an experimental effect is sustainable under a different set of biological variables or an anomaly itself.

## In SDRF proteomics biological replicates can be annotated using characteristics[biological replicate] but is OPTIONAL.

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

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

Multiple tools like MaxQuant or MSstats recognize some of this technical and biological replicate as Groups. In MSstatsTMT for example: Technical replicate correspond to TechRepMixture, while biological replicates correspond to BioReplicate.

1. **Data properties.**

## **Type and Model of Mass Spectrometer**

## The model of the mass spectrometer SHOULD be specified as comment[instrument]. Possible values are listed under [instrument model](https://www.ebi.ac.uk/ols/ontologies/ms/terms?iri=http%3A%2F%2Fpurl.obolibrary.org%2Fobo%2FMS_1000031&viewMode=All&siblings=false) term.

* Additionally, it is strongly RECOMMENDED to include comment[MS2 analyzer type]. This is important e.g. for Orbitrap models where MS2 scans can be acquired either in the Orbitrap or in the ion trap. Setting this value allows to differentiate high-resolution MS/MS data. Possible values of comment[MS2 analyzer type] are [mass analyzer types](https://www.ebi.ac.uk/ols/ontologies/ms/terms?iri=http%3A%2F%2Fpurl.obolibrary.org%2Fobo%2FMS_1000443&viewMode=All&siblings=false).

**Label annotations**

In order to annotate quantitative projects, the SDRF file format use tags for each channel associated with the sample in comment[label]. The label values are organized under the following ontology term [Label](https://www.ebi.ac.uk/ols/ontologies/pride/terms?iri=http%3A%2F%2Fpurl.obolibrary.org%2Fobo%2FPRIDE_0000514&viewMode=All&siblings=false).

Some of the most popular labels are:

* For label-free experiments the value should be: label-free
* For TMT experiments the SDRF uses the PRIDE ontology terms under sample label. Here some examples of TMT channels:
  + TMT126, TMT127, TMT127C , TMT127N, TMT128 , TMT128C, TMT128N, TMT129, TMT129C, TMT129N, TMT130, TMT130C, TMT130N, TMT131

Please, if you need to add an additional label, create an [issue in the pride-ontology repository](https://github.com/PRIDE-Utilities/pride-ontology/issues).

In order to achieve a clear relationship between the label and the sample condition, each channel of each sample (in multiplex experiments) should be defined in a separate row: **one row per channel used (annotated with the corresponding comment[label] per file**.

Examples:

## [Label free experiment](https://github.com/bigbio/proteomics-metadata-standard/blob/c69665600d5e0ddaf6099b4660cc70764ef6cddf/annotated-projects/PXD000612/sdrf.tsv)

## [TMT experiment](https://github.com/bigbio/proteomics-metadata-standard/blob/c69665600d5e0ddaf6099b4660cc70764ef6cddf/annotated-projects/PXD011799/sdrf.tsv)

## [SILAC experiment](https://github.com/bigbio/proteomics-metadata-standard/blob/a141d6bc225e3df8d35e36f0035307f0c7fadf1d/annotated-projects/PXD017710/sdrf-silac.tsv)

## **3.4. Additional RAW file properties**

## We RECOMMEND to include the public URI of the file if available. For example, for ProteomeXchange datasets the URI from the FTP can be provided:

|  | **…​** | **comment[file uri]** |
| --- | --- | --- |
| sample 1 | …​ | ftp://ftp.pride.ebi.ac.uk/pride/data/archive/2017/09/PXD005946/000261\_C05\_P0001563\_A00\_B00K\_R1.RAW |

## **3.5. MSRun additional properties**

## 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](https://www.ebi.ac.uk/ols/ontologies/pride/terms?iri=http%3A%2F%2Fpurl.obolibrary.org%2Fobo%2FPRIDE_0000550). Example, Off-gel electrophoresis.

## comment[depletion]: The removal of specific components of a complex mixture of proteins or peptides on the basis of some specific property of those components. The values of the columns will be no depletion or depletion.

## 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](https://www.ebi.ac.uk/ols/ontologies/ms/terms?iri=http%3A%2F%2Fpurl.obolibrary.org%2Fobo%2FMS_1000044&viewMode=All&siblings=false).

# MSRun technical details properties

## We RECOMMEND to encode some of the technical parameters of the MS experiment as \_comment\_s ([Check what is a comment in SDRF](https://www.ebi.ac.uk/arrayexpress/help/creating_a_sdrf.html)) including the following parameters:

## Protein Modifications [Section 4.1](https://github.com/bigbio/proteomics-metadata-standard/tree/master/sample-metadata#encoding-protein-modifications)

## Precursor and Fragment mass tolerances [Section 4.3](https://github.com/bigbio/proteomics-metadata-standard/tree/master/sample-metadata#encoding-tolerances)

## Digestion Enzyme [Section 4.2](https://github.com/bigbio/proteomics-metadata-standard/tree/master/sample-metadata#encoding-enzymes)

## **4.1. Protein Modifications**

## Sample modifications (including both chemical modifications and post translational modifications, PTMs) are originated from multiple sources: **artifacts modifications**, **isotope labeling**, adducts that are encoded as PTMs (e.g. sodium) or the most **biologically relevant** PTMs. The most common and widely studied PTMs include phosphorylation and glycosylation, among many others. Many of these PTMs are critical to a given protein’s function.

The current specification RECOMMENDS providing Sample modifications including the aminoacid affected, if is Variable or Fixed (also Custom and Annotated modifications are supported) and other properties such as mass shift/delta mass and the position (e.g. anywhere in the sequence).

The RECOMMENDED name of the column for sample modification parameters is: comment[modification parameters]

The modification parameters is the name of the ontology term [MS:1001055](https://www.ebi.ac.uk/ols/ontologies/ms/terms?iri=http%3A%2F%2Fpurl.obolibrary.org%2Fobo%2FMS_1001055)

For each modification, we will capture different properties in a key=value pair structure including name, position, etc. All the possible features available for modification parameters:

| **Property** | **Key** | **Example** | **Mandatory(1)****Optional(0️)** | **comment** |
| --- | --- | --- | --- | --- |
| Name of the Modification | NT | NT=Acetylation | 1 | \* Name of the Term in this particular case Modification, for custom modifications can be a name defined by the user. |
| Modification Accession | AC | AC=UNIMOD:1 | 0️ | Accession in an external database UNIMOD or PSI-MOD supported. |
| Chemical Formula | CF | CF=H(2)C(2)O | 0️ | This is the chemical formula of the added or removed atoms. For the formula composition please follow the guidelines from [UNIMOD](http://www.unimod.org/names.html) |
| Modification Type | MT | MT=Fixed | 0️ | This specifies which modification group the modification should be included with. Choose from the following options: [Fixed, Variable, Annotated]. *Annotated* is used to search for all the occurrences of the modification into an annotated protein database file like UNIPROT XML or PEFF. |
| Position of the modification in the Polypeptide | PP | PP=Any N-term | 0️ | Choose from the following options: [Anywhere, Protein N-term, Protein C-term, Any N-term, Any C-term]. Default is **Anywhere**. |
| Target Amino acid | TA | TA=S,T,Y | 1 | The target amino acid letter. If the modification targets multiple sites, it can be separated by ,. |
| Monoisotopic Mass | MM | MM=42.010565 | 0️ | The exact atomic mass shift produced by the modification. Please use at least 5 decimal places of accuracy. This should only be used if the chemical formula of the modification is not known. If the chemical formula is specified, the monoisotopic mass will be overwritten by the claculated monoisotopic mass. |
| Target Site | TS | TS=N[^P][ST] | 0️ | For some software, it is important to capture complex rules for modification sites as regular expressions. These use cases should be specified as regular expressions. |

We RECOMMEND using for the modification name the UNIMOD interim name or the PSI-MOD name. For custom modifications, we RECOMMEND using an intuitive name. If the PTM is unknown (custom), the *Chemical Formula* or *Monoisotopic Mass* MUST be annotated.

## An example of a **SDRF** file with sample modifications annotated:

|  | **comment[modification parameters]** | **comment[modification parameters]** |
| --- | --- | --- |
| sample 1 | NT=Glu→pyro-Glu; MT=fixed; PP=Anywhere; AC=Unimod:27; TA=E | NT=Oxidation; MT=Variable; TA=M |

## **4.2. Enzyme**

## The REQUIRED comment [cleavage agent details] property is used to capture the Enzyme information. Similar to protein modification [Section 4.1](https://github.com/bigbio/proteomics-metadata-standard/tree/master/sample-metadata#encoding-protein-modifications) we will use a key=value pair representation to encode the following properties for each enzyme:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Property** | **Key** | **Example** | **Mandatory(1)****Optional(0️)** | **comment** |
| Name of the Enzyme | NT | NT=Trypsin | 1 | \* Name of the Term in this particular case Name of the Enzyme. |
| Enzyme Accession | AC | AC=MS:1001251 | 0️ | Accession in an external PSI-MS Ontology definition under the following category [Cleavage agent name](https://www.ebi.ac.uk/ols/ontologies/ms/terms?iri=http%3A%2F%2Fpurl.obolibrary.org%2Fobo%2FMS_1001045). |
| Cleavage site regular expression | CS | CS=(?⇐[KR])(?!P) | 0️ | The cleavage site defined as a regular expression. |

## An example of a **SDRF** with sample enzyme annotated:

|  | **comment[cleavage agent details]** |
| --- | --- |
| sample 1 | NT=Trypsin; AC=MS:1001251; CS=(?⇐[KR])(?!P) |

## **4.3. Precursor and Fragment mass tolerances**

## Encoding precursor and fragment tolerances, for proteomics experiments is important to encode different tolerances (Precursor and fragment).

|  | **comment[fragment mass tolerance]** | **comment[precursor mass tolerance]** |
| --- | --- | --- |
| sample 1 | 0.6 Da | pm |

## **Specific use cases and conventions**

## **How to encode age**

## One of the characteristics about the sample is the age of an individual. It is RECOMMENDED to provide the age in the following format: {X}Y{X}M{X}D. Some valid examples are:

## 40Y (forty years)

## 40Y5M (forty years and 5 months)

## 40Y5M2D (forty years, 5 months and 2 days)

## When needed, weeks can also be used:

## 8W (eight weeks)

## Age interval:

## Some times the sample do not have an exact time but biologically it is more important to define a range of age. In order to annotate an age range the following standard MUST be followed:

## 40Y-85Y

## This means that the subject (sample) can be group between 40 and 85 years old.

## Other temporal information can be encoded in a similar way.

## **Phosphoproteomics and PTMs experiments**

## In phopshoproteomics experiments the sample is enriched to detect phosphorylation sites. In those experiments the characteristics[enrichment process] should be provided.

## The different values already included in EFO are:

## enrichment of phosphorylated Protein

## enrichment of glycosylated Protein

## This characteristic can be used as factor value[enrichment process] to differentiate the expression between proteins in the phosphoproteomics sample compare with control.

## **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.

## 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, a special characteristic can be used:

| **source name** | **characteristics[pooled sample]** | **assay name** | **comment[label]** | **comment[data file]** |
| --- | --- | --- | --- | --- |
| sample 1 | not pooled | run 1 | TMT131 | file01.raw |
| sample 2 | not pooled | run 1 | TMT131C | file01.raw |
| sample 10 | SN=sample 1,sample 2, …​ sample 9 | run 1 | TMT128 | file01.raw |

## SN stands for source names and lists source name fields of samples that are annotated in the same file and **used in the same experiment and same MS run**.

## Another possible value for characteristics[pooled sample] is a string pooled for cases when it is known that a sample is pooled but the individual samples cannot be annotated.

## **5.4. Derived samples (such as PDX)**

## In cancer research, patient-derived xenografts (PDX) are commonly used where the patient’s tumour is transplanted into a, for example, mouse. In these cases, the metadata, such as age and sex MUST refer to the original patient and not the animal.

## PDX samples SHOULD be annotated by using the column characteristics[xenograft]. The value should then describe the growth condition, such as pancreatic cancer cells grown in nude mouse.

## For experiments where both, the PDX and the original tumour was measured, the PDX entry SHOULD reference the respective tumour sample’s source name in the characteristics[original source name] column. Non-PDX samples SHOULD contain the not applicable value in the characteristics[xenograft] and the characteristics[original source name] column. Both tumour and PDX samples SHOULD reference the patient using the characteristics[individual] column. This column should contain some sort of patient identifier.

## **Spike-in samples**

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

## To include information on 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 | 1 | 1 | 1 | ✅ |
| QY | Quantity (molar or mass) | 10 mg, 20 nmol | 1 | 1 | 1 | ✅ |
| PS | Peptide sequence | PEPTIDESEQ | 1 |  |  |  |
| AC | Uniprot Accession | A9WZ33 |  | 1 |  |  |
| CN | Compound name | iRT mixture, substance name |  | 0 | 0 | 0️ |
| CV | Compound vendor | in-house or vendor name | 0️ | 0️ | 1 | 0️ |
| CS | Compound specification URI | <http://vendor.web.site/specs/coomercial-kit.xlsx> | 0️ | 0️ | 0 | 0️ |
| CF | Compound formula | C2H2O |  |  |  | 0️ |

## 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 for a sample spiked with a peptide would be:

| **characteristics[mass]** | **charateristics[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.

## **Synthetic peptide libraries**

## Proteomics and mass spectrometry use synthetic peptide libraries for multiple use cases including:

## Benchmark of analytical and bioinformatics methods and algorithms.

## Improve of peptide identification/quantification using spectral libraries.

## When describing synthetic peptide libraries most of the sample metadata can be declare as not applicable. However, some authors can annotate the organism for example because they know the library has been design from specific peptide species, see example [Synthetic Peptide experiment](https://github.com/bigbio/proteomics-metadata-standard/blob/master/annotated-projects/PXD000759/sdrf.tsv).

## It is important to annotate that the sample is a synthetic peptide library, this can be done by adding the characteristics[synthetic peptide]the possible values are: synthetic or not synthetic.

## **Normal, healthy samples**

## Samples from healthy patients or individuals normally appear in manuscripts and annotations as healthy or normal. We RECOMMENDED to use the word normal mapped to term PATO\_0000461 that is in EFO: [normal PATO term](https://www.ebi.ac.uk/ols/ontologies/efo/terms?iri=http%3A%2F%2Fpurl.obolibrary.org%2Fobo%2FPATO_0000461). Example:

| **source name** | **characteristics[organism]** | **characteristics[organism part]** | **characteristics[phenotype]** | **characteristics[compound]** | **factor value[phenotype]** |
| --- | --- | --- | --- | --- | --- |
| sample\_treat | homo sapiens | Whole Organism | necrotic tissue | drug A | necrotic tissue |
| sample\_control | homo sapiens | Whole Organism | normal | none | normal |

## **Sample/MSRun templates**

## The **sample metadata templates** are a set of guidelines to annotate different type of proteomics experiments to ensure that a Minimum Metadata and characteristics are provided to understand the dataset. These templates respond to the distribution and frequency of experiment types in public databases like [PRIDE](http://www.ebi.ac.uk/pride/archive) and [ProteomeXchange](http://www.proteomexchange.org/):

## Default: Minimum information for any proteomics experiment [Template](https://github.com/bigbio/proteomics-metadata-standard/blob/master/templates/sdrf-default.tsv)

## Human: All tissue-based experiments that use Human samples [Template](https://github.com/bigbio/proteomics-metadata-standard/blob/master/templates/sdrf-human.tsv)

## Vertebrates: Vertebrate experiment. [Template](https://github.com/bigbio/proteomics-metadata-standard/blob/master/templates/sdrf-vertebrates.tsv)

## Non-vertebrates: Non-vertebrate experiment. [Template](https://github.com/bigbio/proteomics-metadata-standard/blob/master/templates/sdrf-nonvertebrates.tsv)

## Plants: Plant experiment. [Template](https://github.com/bigbio/proteomics-metadata-standard/blob/master/templates/sdrf-plants.tsv)

## Cell lines: Experiments using cell-lines. [Template](https://github.com/bigbio/proteomics-metadata-standard/blob/master/templates/sdrf-cell-line.tsv)

## **Sample attributes**: Minimum sample attributes for primary cells from different species and cell lines

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Default | Human | Vertebrates | Non-vertebrates | Plants | Cell lines |
| Source Name | 1 | 1 | 1 | 1 | 1 | 1 |
| characteristics[organism] | 1 | 1 | 1 | 1 | 1 | 1 |
| characteristics[strain/breed] |  |  |  | 0️ |  | 0️ |
| characteristics[ecotype/cultivar] |  |  |  |  | 0️ |  |
| characteristics[ancestry category] |  | 1 |  |  |  |  |
| characteristics[age] |  | 1 | 0️ |  | 0️ |  |
| characteristics[developmental stage] |  | 0️ | 0️ |  | 0️ |  |
| characteristics[sex] |  | 1 | 0️ |  |  |  |
| characteristics[disease] | 1 | 1 | 1 | 1 |  | 1 |
| characteristics[organism part] | 1 | 1 | 1 | 1 | 1 | 1 |
| characteristics[cell type] | 1 | 1 | 1 | 1 | 1 | 1 |
| characteristics[individual] |  | 0️ | 0️ | 0️ | 0️ | 0️ |
| characteristics[cultured cell] |  |  |  |  |  | 1 |
|  |  |  |  |  |  |  |
| comment[data file] | 1 | 1 | 1 | 1 | 1 | 1 |
| comment[fraction identifier] | 1 | 1 | 1 | 1 | 1 | 1 |
| comment[label] | 1 | 1 | 1 | 1 | 1 | 1 |
| comment[cleavage agent details] | 1 | 1 | 1 | 1 | 1 | 1 |
| comment[instrument] | 1 | 1 | 1 | 1 | 1 | 1 |

## 1 : Required Attributes for each sample Type (e.g. Human, Vertebrates).

## 0️: Optional Attribute

## **SDRF study variable**

## The variable/property under study should be highlighted using the **factor value** category. For example, the **factor value[disease]** is used when the user wants to compare expression across different diseases.

|  |  |  |  |
| --- | --- | --- | --- |
| factor value | 0..\* | “factor value” columns should indicate which experimental factor / variable are use to perform the quantitative data analysis. The “factor value” columns should occur after all characteristics and the attributes of the samples. | Factor Value [phenotype] |

We RECOMMEND using for the modification name the UNIMOD interim name or the PSI-MOD name. For custom modifications, we RECOMMEND to use an intuitive name. If the PTM is unknown (custom), the *Chemical Formula* or *Monoisotopic Mass* MUST be annotated.

## **Multiple projects into one annotation file**

## Curators can decide to annotate multiple ProteomXchange Projects into one big sdrf for reanalysis purpose. If that is the case, we RECOMMENDED to use the *comment[proteomexchange accession number]* to differentiate between projects.

## **Examples of annotated datasets**

|  |  |  |
| --- | --- | --- |
| Dataset Type | ProteomeXchange / Pubmed Accession | SDRF URL |
| Label-free | PXD008934 | <https://github.com/bigbio/proteomics-metadata-standard/blob/master/annotated-projects/PXD008934/sdrf.tsv> |
| TMT | CPTAC PMID27251275 | <https://raw.githubusercontent.com/bigbio/proteomics-metadata-standard/master/annotated-projects/PMID27251275/sdrf.tsv> |

# Authors Information

Yasset Perez-Riverol

European Bioinformatics Institute (EMBL-EBI), Hinxton, Cambridge, United Kingdom

[yperez@ebi.ac.uk](mailto:yperez@ebi.ac.uk)

Juan Antonio Vizcaíno

European Bioinformatics Institute (EMBL-EBI), Hinxton, Cambridge, United Kingdom

[juan@ebi.ac.uk](mailto:juan@ebi.ac.uk)

# Contributors

A full list of contributors can be found here: <https://github.com/bigbio/proteomics-metadata-standard#core-contributors-and-collaborators>

# 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).

# Copyright Notice

Copyright (C) Proteomics Standards Initiative (2020). All Rights Reserved.

This document and translations of it may be copied and furnished to others, and derivative works that comment on or otherwise explain it or assist in its implementation may be prepared, copied, published and distributed, in whole or in part, without restriction of any kind, provided that the above copyright notice and this paragraph are included on all such copies and derivative works. However, this document itself may not be modified in any way, such as by removing the copyright notice or references to the PSI or other organizations, except as needed for the purpose of developing Proteomics Recommendations in which case the procedures for copyrights defined in the PSI Document process must be followed, or as required to translate it into languages other than English.

The limited permissions granted above are perpetual and will not be revoked by the PSI or its successors or assigns.

This document and the information contained herein is provided on an "AS IS" basis and THE PROTEOMICS STANDARDS INITIATIVE DISCLAIMS ALL WARRANTIES, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO ANY WARRANTY THAT THE USE OF THE INFORMATION HEREIN WILL NOT INFRINGE ANY RIGHTS OR ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE."

# Glossary

All non-standard terms are already defined in detail in section 3.

# References

[1] T.F. Rayner, P. Rocca-Serra, P.T. Spellman, H.C. Causton, A. Farne, E. Holloway, R.A. Irizarry, J. Liu, D.S. Maier, M. Miller, K. Petersen, J. Quackenbush, G. Sherlock, C.J. Stoeckert, Jr., J. White, P.L. Whetzel, F. Wymore, H. Parkinson, U. Sarkans, C.A. Ball, A. Brazma, A simple spreadsheet-based, MIAME-supportive format for microarray data: MAGE-TAB, BMC Bioinformatics 7 (2006) 489.