# Upload of proteomics data to repositories

The availability of raw data is crucial for the transparence of scientific research. It also allows for data reanalysis for QC, data mining and comparisons. Sharing of data through public repositories is now requested by most journals as condition for manuscript submission.

For mass spec-based proteomics data we recommend the Proteomexchange.org (PX) repository, run by the EBI and linked to the PRoteomics IDEntifications (**PRIDE**) database (<https://www.ebi.ac.uk/pride/>).

Note: submissions to Proteomexchange/PRIDE **must** include :

1. Full raw MS data
2. description of methods for sample preparation and data processing
3. protein identification and quantitation results (outputs from MASCOT, MaxQuant, Spectronaut,…) including all the parameters used for these tools
4. comprehensive sample and experiment metadata, experimental design, relationships between raw data files and result files

At the PAF ([www.unil.ch/PAF](http://www.unil.ch/PAF)) , **we will do our best to do the upload to the repository for you.** However we need to collect all the information listed above. Especially for point 4) we need user’s input !

**IMPORTANT : TO DO list for PAF users** (to make our life just a little bit easier ) :

1. Please contact us with **sufficient advance (2 weeks or more)** before starting submission of a manuscript to request PX data upload.
2. Specify as accurately and completely as possible **which experiment/dataset/samples/results** are mentioned (or not !) in your manuscript and for which the data must be prepared and uploaded.
3. If a search for PTMs (phosphorylation, ubiquitination, …) was performed, tell us **which (if any) PTM is presented in your manuscript** and which one is (are) not
4. please **try to** **fill in the maximum of all points in the checklist below**. We will complete the rest.

# INFO needed for Proteomexchange.org (PX) upload

1. **Group leader**

Name : …

E-mail address : …

Affiliation (institutional address) : …

1. **Project Title  ( 50-5000 characters):**

**…**

1. **Keywords  (4-5 ) :**

…

1. **Project description ( 50-5000 characters) :**

*Please provide an overall description of your study, like the manuscript abstract (a short version can be sufficient)*

…

1. **Sample processing protocol ( 50-5000 characters) :**

*Description of sample preparation, separation and enrichment strategies as well as detailed LC-MS analysis steps. Not general manuscript methods, focus on the proteomics sample prep and analysis. This has normally been prepared before, to be included in the manuscript methods section.*

…

1. **Data processing protocol ( 50-5000 characters) :**

*Should include : MS data processing, database search, filtering (FDR), quantitation parameters. Also downstream processing, stats etc. up to a certain point. More interpretation-related steps do not need to be described. This has normally been prepared before, to be included in the manuscript methods section.*

…

1. **Species**

*When multiple species are present try to specify all of them.*

…

1. **Tissue \*\*\***

1. **Cell type \*\*\***

1. **MS instrument \*\*\***

 *(usually done by PAF )*

1. **PTMs used in search**

 *(usually done by PAF )*

1. **Quantification strategy**

*(usually done by PAF )*

1. **Sample Table**

*A table (.xlsx is OK) is highly recommended, listing the sample, conditions etc. Ideally raw data filenames should be present to clarify who is who. This can be loaded as «other».*

*Hint for us : to make list of files in dir use “dir /a-d /b >filelist.txt” in cmd window*

**IMPORTANT : What happens after submission of the data is completed ?**

* The data remains private until the article is published.
* The PRIDE administrators give us a username and password that should be included either in the text of the manuscript or the cover letter. This access will be given to reviewers in case they want to check the raw data.
* Once the article is published, the owner of the dataset has to request release in the public domain. To do this, it is usually necessary to have the DOI of the published article so that the data is automatically linked to the publication.
* The person that has done the data upload (the PAF) is per default the owner of the dataset. In principle we will transfer the ownership to the PI of the project, so that she/he can manage all these further steps. But first the PI will have to create her/himself a PRIDE account (<https://www.ebi.ac.uk/pride/> )

**Acknowledgements :**

Please add in the acknowledgement section a text about our work and support. For example :

“*Mass spectrometry-based proteomics work was performed by the Protein Analysis Facility of the Faculty of Biology and Medicine, University of Lausanne, Lausanne, Switzerland.*”

**APPENDIX**

**Files and upload : other important points** (for the person doing the submission, usually @ PAF)

* Minimum for a partial submission : one search and one raw file type must be present
* Bruker .d files need to be zipped (use winRaR for automated zipping of each dir to a separate file)
* *Checksum.txt* file must be in the list (if you startover by emptying the list it is eliminated and must be put back manually. It can be found in the PX tool folder)
* Avoid spaces and brackets in filenames ! Filenames with spaces block submission before checksum.
* When loading files in this window, their size is normally shown immediately in the list. If not, there is a problem with the files.
* For DIA with library : one should also include the .kit library file
* It is also a good practice to include the .fasta files used
* A search function is available in the PX submission tool for ***\*\*\**** these points
* Last page before starting upload : use the “verify” function is advisable