Version 2.3.4 (13 March 2017)

Tutorial: Submission of MS/MS datasets to ProteomeXchange *via* PRIDE

Table of Contents

Where do I start? Submission summary overview

The default PRIDE submission consists of the deposition of MS/MS proteomics datasets according to the guidelines of the ProteomeXchange (PX) consortium (1). In addition to this tutorial, more documentation is available:

A publication entitled "How to submit MS proteomics data to ProteomeXchange *via* the PRIDE database" (by T. Ternent *et al.*) (2) was published in the journal *Proteomics* (Wiley) on October 2014, explaining in detail the process using an exemplary dataset (PXD000764). The paper is open access and can be freely accessed here.

An online tutorial is available in the EBI train on-line platform, at http://www.ebi.ac.uk/training/online/course/proteomexchange-submissions-pride. Extra documentation is available in the PRIDE web pages (http://www.ebi.ac.uk/pride/help/archive). Concrete instructions to generate mzIdentML files (needed for the submissions) are available here for tools like Mascot, Scaffold and ProteinPilot. **Figure 1** shows the overall submission process (by March 2017).

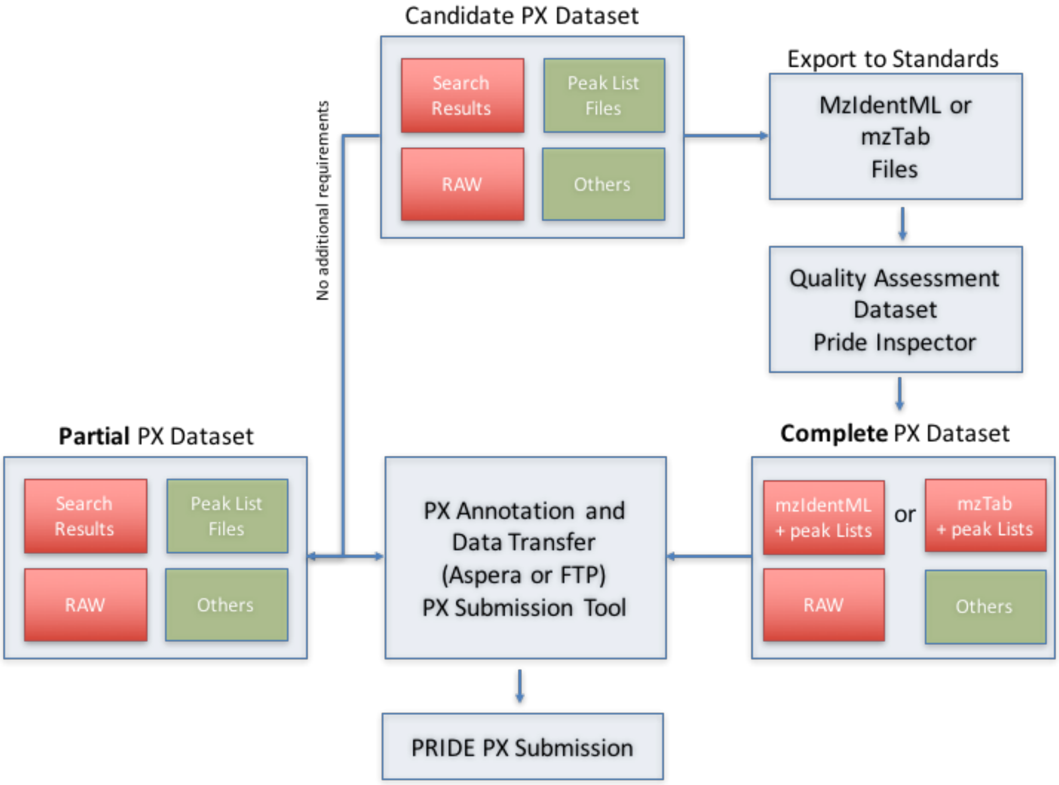


Figure 1: Overview of the data submission process to ProteomeXchange *via* PRIDE including the two default submission types: 'Complete' and 'Partial'.

What information should contain a PX Submission

Each submitted dataset to PX *via* PRIDE must contain:

- Peptide/protein identification files (called 'RESULT'),
- Mass spectrometer output files (called 'RAW'), which are either machine raw files or not heavily processed files in a XML-based format such as mzXML or mzML (3),
- Optionally other files can be included like peak list files (called 'PEAK', mandatory for 'Complete' submissions including mzIdentML files, see below), search engine output files (called 'SEARCH', mandatory for "Partial submissions", see below), quantification results ('QUANT'), gel images ('GEL'), sequence database files (FASTA), spectral libraries (SPECTRUM_LIBRARY) and any other, relevant file types ('OTHER').

In addition, a more specific procedure is now available for MS imaging datasets. For instance, some extra requirements are needed and additional file tags have been created (see Appendix VI for details). The details are also explained in this open access publication (Roempp *et al.*, *Anal Bioanal Chem*, 2015) (4), freely accessible here.

Submission types: Complete and Partial Submissions

There are two different submission workflows ('Complete' and 'Partial') depending on whether peptide/protein identification results can be submitted in a standard format that can be handled by PRIDE or not. After performing a 'complete' submission it is possible for PRIDE to connect directly the processed peptide/protein identification results with the mass spectra.

If **mzTab** 'RESULT' files or **mzIdentML** (5) plus the accompanying peak list ('PEAK') files containing the referenced spectra are provided, the 'Complete' Submission option is available. If 'RESULT' files are not available in these formats, a 'Partial' Submission can be done. In this case, the connection between the spectra and the identification results cannot be done in a straightforward way. For 'partial' submissions, the processed results are not available in a format supported by the repository. Instead, the corresponding analysis software output files ('SEARCH' files, in heterogeneous formats) are made available for download

It is important to highlight that the current version of pipeline does not support a full and standard representation of the quantification results, linked to the identification results. It is expected that data standards for quantitative proteomics (mzTab (7)) will be supported in the future. However, any quantification result output files can be submitted as accompanying 'QUANT' files.

Before a submission is started it is necessary to have a PRIDE user account (please register at http://www.ebi.ac.uk/pride/archive/register). All submissions to ProteomeXchange *via* PRIDE are private by default, and the username and password are needed to access your data. Data will be made publicly available when the submitter notify us to do it or by default when the corresponding manuscript is made available (see Section 9.3).

It is important to highlight that by default, the PX Submission Tool is using the fast Aspera upload transfer protocol (http://www.asperasoft.com/), with which terabytes of data can be potentially transferred within a day, since it can be up to 50 times faster than FTP.

As summarized above, two main submission types/workflows are available: **'Complete'** or **'Partial'** Submissions. For all types of submissions to PX *via* PRIDE, the first option for the users is to use the Java stand-alone tool "PX Submission tool" (available at http://www.proteomexchange.org/submission).

Complete Submission

This is the recommended and preferred option. 'RAW' files need to be provided together with the 'RESULT' type supported file formats mzTab or mzIdentML (version 1.1) files (5). These are the two subtypes of 'Complete' submissions.

Uploading peak list ('PEAK'), search engine output ('SEARCH'), quantification ('QUANT'), sequence database ('FASTA'), spectral library ('SPECTRUM_LIBRARY') and other post processing files ('OTHER') can also be done in order to give a near complete coverage and representation of your data and it is recommended but not enforced.

However, if the submitter chooses to submit the 'RESULT' files as mzIdentML, the corresponding peak list files ('PEAK') used in the search and referenced in the mzIdentML file/s need to be submitted as well. The reason behind is that otherwise, the mass spectra will not be submitted, since mzIdentML only contains the peptide/protein identification results.

After the submission, you will be issued with not only a ProteomeXchange accession number but also with a permanent DOI (Digital Object Identifier) to uniquely identify your dataset in the future.

Your submitted data will be fully accessible in PRIDE and allow full visualization of the data for private journal review support using the PRIDE Inspector tool (8) (it can be freely downloaded at https://github.com/PRIDE-Toolsuite/pride-inspector). Your data will be made available *via* FTP (ftp://ftp.pride.ebi.ac.uk/) to download once it has been made public.

The complete submission requires at least two sets of files in case of mzTab based submissions, and three in case of mzIdentML based submissions:

Result files fully supported by PRIDE (called 'RESULT'): Two formats are currently supported:

- **mzTab** files in Complete mode is now supported as a Result File in PRIDE. Also, mzTab files can be of: **Identification**, for Identification file the Protein and PSM section should be provided (7). If **Quantification** results are included, the PEPTIDE section should be included in addition to the Protein/PSM sections. (See mzTab format specification for more information on this format, at http://www.psidev.info)
- **mzIdentML** version 1.1 files. mzIdentML is the Proteomics Standards Initiative (PSI) standard for peptide/protein identification data (5). Many of the most popular search engine output files can be exported to mzIdentML 1.1 (see Appendix II or http://www.psidev.info/tools-implementing-mzidentml). Since the MS data is not included in mzIdentML, to have a complete submission it is also mandatory to submit the corresponding peak list files ('PEAK', see below). mzIdentML 1.0 files (the non-stable version of the standard) are not supported.

In both cases, in the PX Submission Tool both types of files should be tagged as 'RESULT' (for a comprehensive list of the formats supported by PRIDE, see Appendix III). Mass spectrometer output files (called 'RAW'): Two options are possible: MS instrument binary output files, such as BRUKER .baf files, Thermo .raw files or not heavily processed files in XML format like mzXML or mzML files (see definitions, Appendix I).

If your 'RAW' files are organized in directories instead of individual files, please compress them into one individual file (for instance to .zip) before upload. In the submission tool they should be tagged as 'RAW'.

Peak list files (called 'PEAK', only mandatory for mzIdentML 'RESULT' files): You can provide the exact version of the files that was used by the search engine to generate the experimental results, the ones that are referenced from the original mzIdentML files. In the submission tool they should be tagged as 'PEAK'. Otherwise, it would be impossible to link the identifications to the corresponding spectra.

Although not required, other types of files can be submitted optionally:

- **Search engine output files (called 'SEARCH')**: the original output files from your search engine or your analysis pipeline, such as Trans-Proteomic Pipeline (TPP) pep.xml and/or prot.xml files, or MaxQuant text output files, among many others. They should contain the peptide/protein identifications. In the submission tool they should be tagged as 'SEARCH'.
- **Quantification output files**: In the PX Submission Tool they should be tagged as 'QUANT'.
- **Gel images files**: In the PX Submission Tool they should be tagged as 'GEL'.

- **Sequence database files**: Sequence database file (usually in FASTA format) that was used to perform the mass spectral search. Sequence database files can contain both amino acid and nucleic acid sequences. In the PX Submission Tool they should be tagged as 'FASTA'
- **Spectrum libraries**: Spectral library file that was used for performing the mass spectrometry search. In the PX Submission Tool they should be tagged as 'SPECTRUM_LIBRARY'
- **Any other files**: In the PX Submission Tool they should be tagged as 'OTHER'.

It is important to highlight that if the PX Submission Tool is not used to perform the submission (for instance it is done using the command line option), an extra file is needed. The file is generated automatically and submitted by the PX submission tool, so it does not need to be created independently if the PX Submission Tool is used.

PX submission summary file: This file captures the descriptive information about a ProteomeXchange submission, such as: experimental metadata, included files, file mappings, etc. All the details about the data format can be found here.

Partial Submission

You should only choose this option if your search results cannot be converted/exported to mzIdentML v1.1 (plus the accompanying spectra). It is not the recommended option, since it will significantly reduce the reusability of your dataset.

'RAW' files need to be provided together with search engine output files ('SEARCH'). Uploading peak list ('PEAK'), and other types of files ('QUANT', 'FASTA', 'SPECTRUM_LIBRARY', 'GEL' or 'OTHER') is also possible but not enforced.

As a result, you will be issued with a ProteomeXchange accession number but not with a DOI (like it happened for 'Complete' submissions. Once it is made public, your dataset will be available to download *via* FTP but peptide/protein identification data will not be visualized in the PRIDE webpage and/or the PRIDE Inspector tool.

The partial submission requires two sets of files:

- **Search engine result files**: (called 'SEARCH'): the original output files from your search engine or your analysis pipeline, Trans-Proteomic Pipeline (TPP) pep.xml and/or prot.xml files, or MaxQuant text output files, among many others. They should contain the peptide/protein identifications. In the submission tool, they should be tagged as 'SEARCH'.

- **Mass spectrometer output files (called 'RAW')**: MS instrument binary output files, such as BRUKER .baf files, Thermo .raw files or not heavily processed mzXML or mzML files (see definitions, Appendix I). If your 'RAW' files are organized in directories instead of individual files, please compress them into one individual file (for instance to .zip) before upload. In the submission tool, they should be tagged as 'RAW'.

- Again, although not required, other types of files can be submitted optionally:

    - **Peak list files**: It is strongly recommended to provide the peak list files (e.g. mgf files) that were used for the original search since these are different from the provided mandatory raw files. In the submission tool they should be tagged as 'PEAK'.
    - **Quantification output files**: In the PX Submission Tool they should be tagged as 'QUANT'.
    - **Gel images files**: In the PX Submission Tool they should be tagged as 'GEL'.
    - **Sequence database files**: Sequence database file (usually in FASTA format) that was used to perform the mass spectral search. Sequence database files can contain both amino acid and nucleic acid sequences. In the PX Submission Tool they should be tagged as 'FASTA'.
    - **Spectrum libraries**: Spectral library file that was used for performing the mass spectrometry search. In the PX Submission Tool they should be tagged as 'SPECTRUM_LIBRARY'.
    - **Any other files**: In the PX Submission Tool they should be tagged as 'OTHER'.

The submission of MS imaging data is a special case of 'Partial' Submission with special data types and data files, and it is explained in detail in the Appendix VI. The details are also explained in this open access publication (Roempp *et al.*, *Anal Bioanal Chem*, 2015) (4), freely accessible here.

As explained earlier, if the PX Submission Tool is not used to perform the submission, an extra file is needed. The file is generated automatically and submitted by the PX submission tool, so it does not need to be created independently if the PX Submission Tool is used.

PX submission summary file: This file captures the descriptive information about a ProteomeXchange submission, such as: experimental metadata, included files, file mappings, etc. All the details about the data format can be found here.

Bulk Submissions

Independently from being complete or partial, you can make a 'Bulk Submission' if you need to submit a large set of files. This path is envisioned for labs with some bioinformatics support since some scripting work is needed. Both 'Complete' and 'Partial' Submissions can be performed through this mechanism.

The "bulk submission" requires also two sets of information:

- Experiment data files: The files you want to submit to PRIDE *via* ProteomeXchange. See section 2 for the exact files needed for each submission type (either 'Complete' or 'Partial').
- PX submission summary file: Needed if the submission is not performed using the PX submission tool. This file captures the descriptive information about a ProteomeXchange submission, such as: experimental metadata, included files, file mappings, etc. All the details about the data format can be found here.

How to make complete submissions?

As discussed earlier in Section 2.1 the two subtypes of 'Complete' submissions are either mzIdentML or mzTab based. 'Complete' submissions mixing the two types of 'RESULT' files are not allowed.

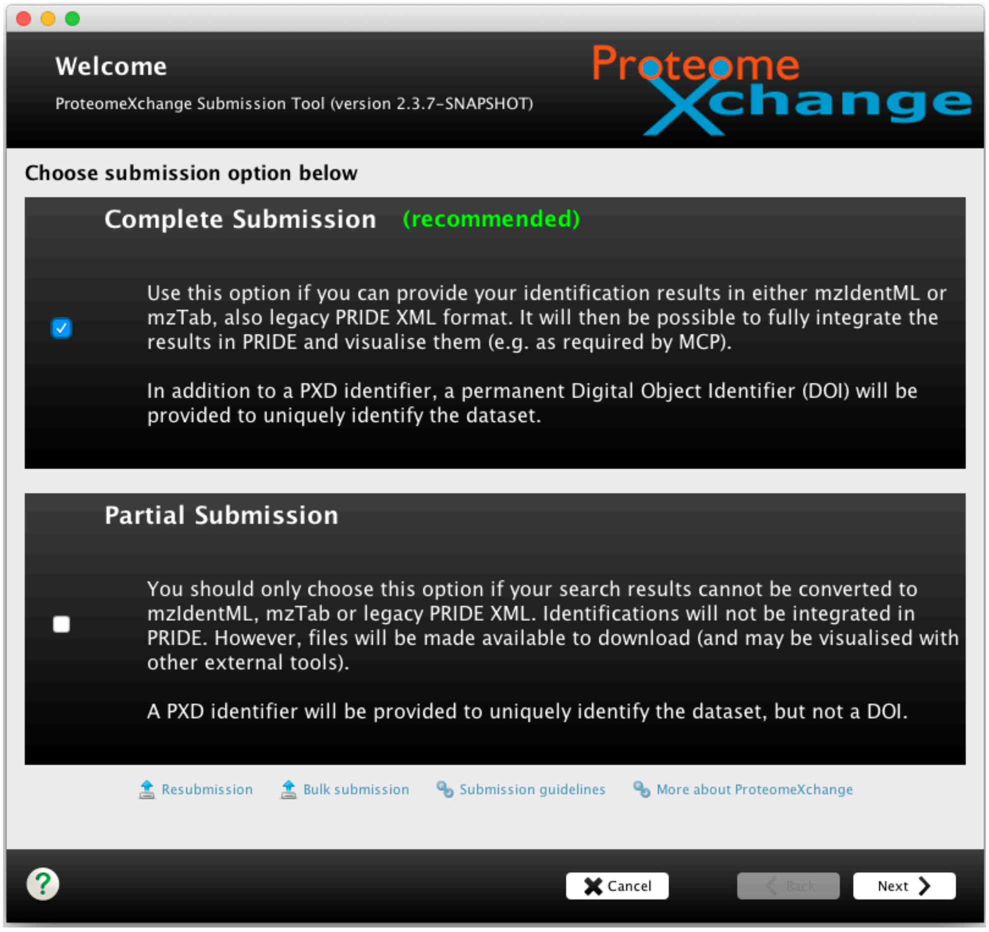
Many of the submission steps are identical for the two subtypes so these steps are going to be discussed in a uniform manner. The differences will be highlighted in case of those steps that are different. The different steps are the following: Step 5: 'Add Files and assign file types', and Step 6: 'Assign relationships between the submitted files'.

Step 1: Launch PX Submission Tool

First you need to install and launch the PX Submission Tool (available at http://www.proteomexchange.org/submission).

Step 2: Select Submission Type

You then need to select ‘Complete Submission’ in the PX Submission Tool ‘Welcome’ screen (**Figure 2**).



Welcome

ProteomeXchange Submission Tool (version 2.3.7-SNAPSHOT)

Choose submission option below

☑ Complete Submission (recommended)

Use this option if you can provide your identification results in either mzIdentML or mzTab, also legacy PRIDE XML format. It will then be possible to fully integrate the results in PRIDE and visualise them (e.g. as required by MCP).

In addition to a PXD identifier, a permanent Digital Object Identifier (DOI) will be provided to uniquely identify the dataset.

□ Partial Submission

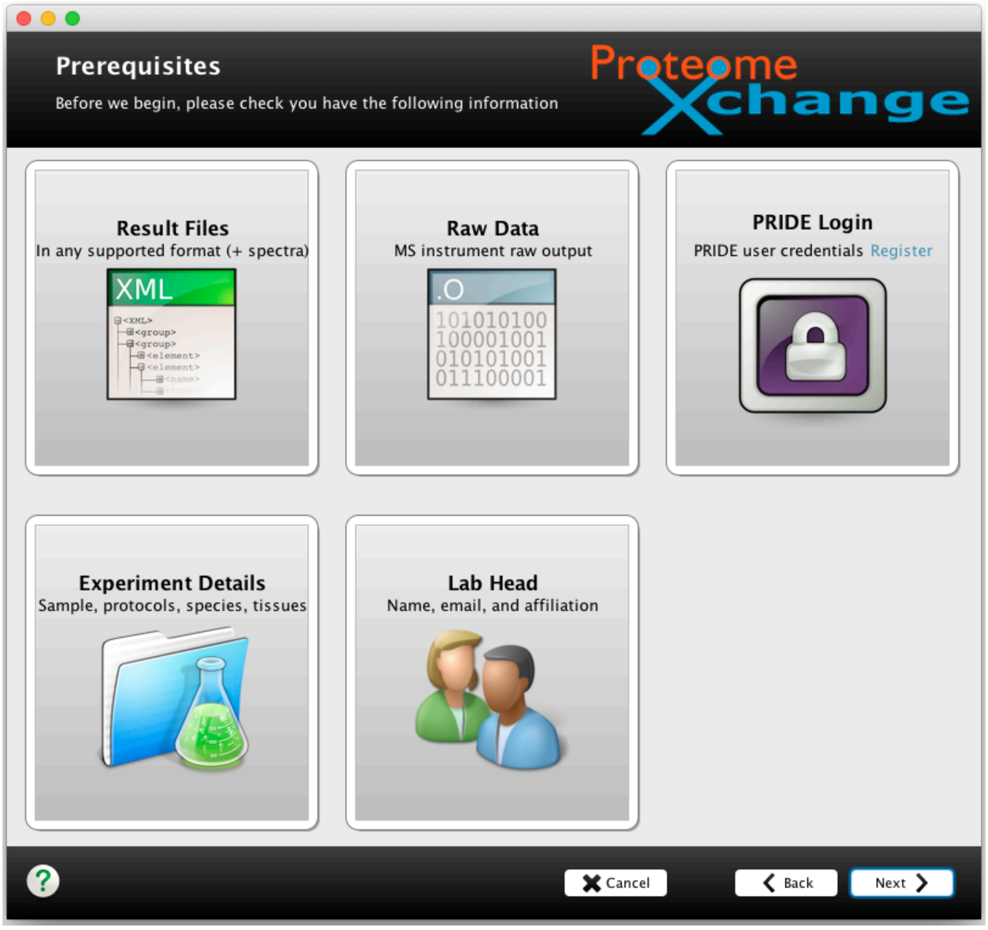
You should only choose this option if your search results cannot be converted to mzIdentML, mzTab or legacy PRIDE XML. Identifications will not be integrated in PRIDE. However, files will be made available to download (and may be visualised with other external tools).

A PXD identifier will be provided to uniquely identify the dataset, but not a DOI.

Figure 2: 'Welcome' screen of the PX submission Tool showing the two submission types

Step 3: Prerequisites

Please double check you have all the required information before submission as shown in **Figure 3**:



Prerequisites

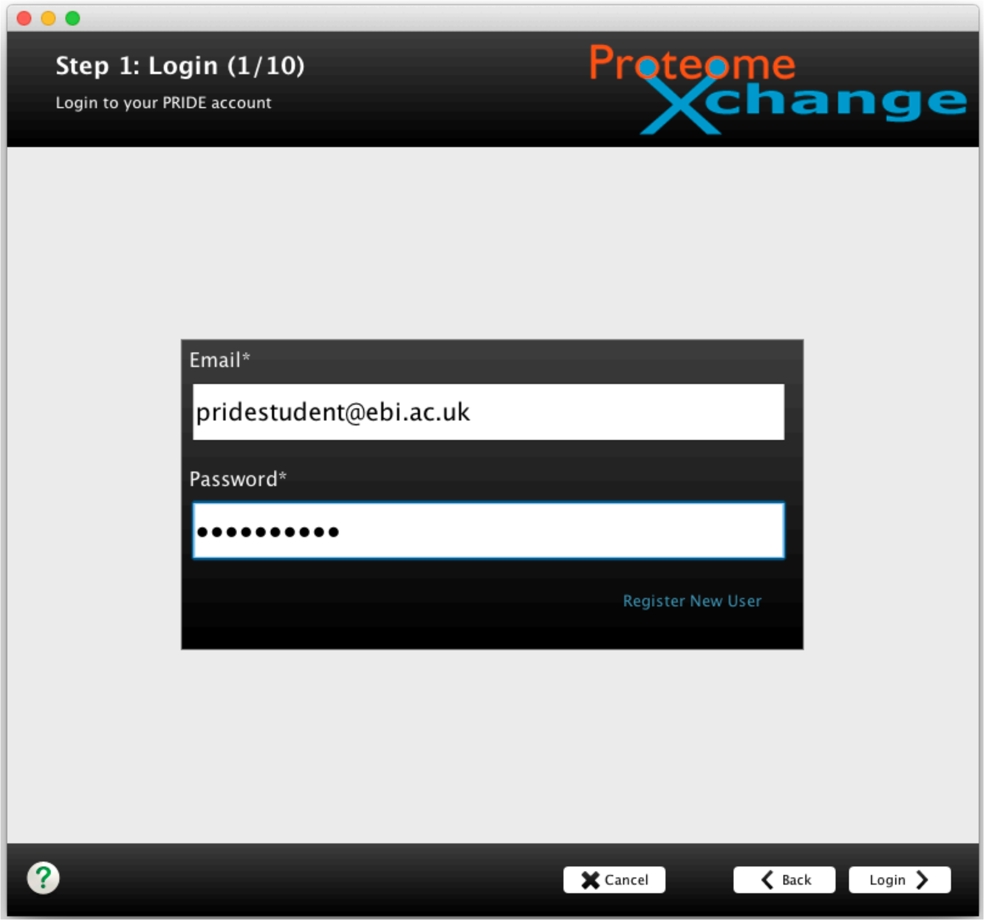
Before we begin, please check you have the following information

ProteomeXchange
Result Files
In any supported format (+ spectra)
XML
Raw Data
MS instrument raw output
PRIDE Login
PRIDE user credentials Register
Experiment Details
Sample, protocols, species, tissues
Lab Head
Name, email, and affiliation
?
Cancel
Back
Next

Figure 3 : Prerequisites screen for ‘complete’ submission in the PX submission tool

Step 4: Login

Please log in using your existing PRIDE account as shown in **Figure 4**:



Step 1: Login (1/10)

Login to your PRIDE account

Email*

pridestudent@ebi.ac.uk

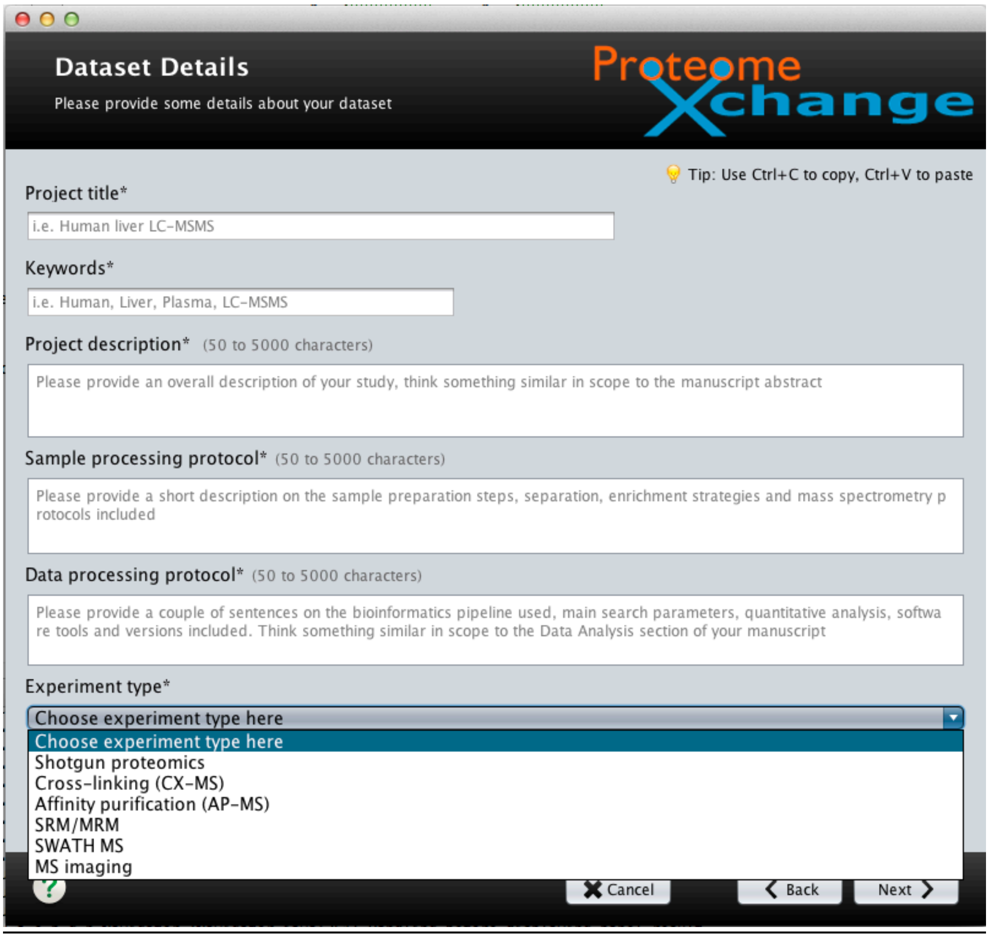
Password*

●●●●●●●●●●●

Figure 4: Login screen of the PX submission tool

Step 5: Provide submission details

The user is asked to provide some basic details about the uploaded dataset (**Figure 5**) such as the title, a list of keywords (in a comma separated format), and a brief description of the data (similar to the abstract of the corresponding publication) a sample processing and a data processing protocol. The user also picks a mass spectrometry experiment type from a drop-down menu.



Dataset Details

Please provide some details about your dataset

Tip: Use Ctrl+C to copy, Ctrl+V to paste

Project title*

i.e. Human liver LC-MSMS

Keywords*

i.e. Human, Liver, Plasma, LC-MSMS

Project description* (50 to 5000 characters)

Please provide an overall description of your study, think something similar in scope to the manuscript abstract

Sample processing protocol* (50 to 5000 characters)

Please provide a short description on the sample preparation steps, separation, enrichment strategies and mass spectrometry protocols included

Data processing protocol* (50 to 5000 characters)

Please provide a couple of sentences on the bioinformatics pipeline used, main search parameters, quantitative analysis, software tools and versions included. Think something similar in scope to the Data Analysis section of your manuscript

Experiment type*

Choose experiment type here

Figure 5: 'Dataset details' screen in the PX submission tool

Step 6: Add Files and assign file types

In this stage, you should choose the files you would like submit. As shown in **Figure 6**, you can add files by clicking on the highlighted button.

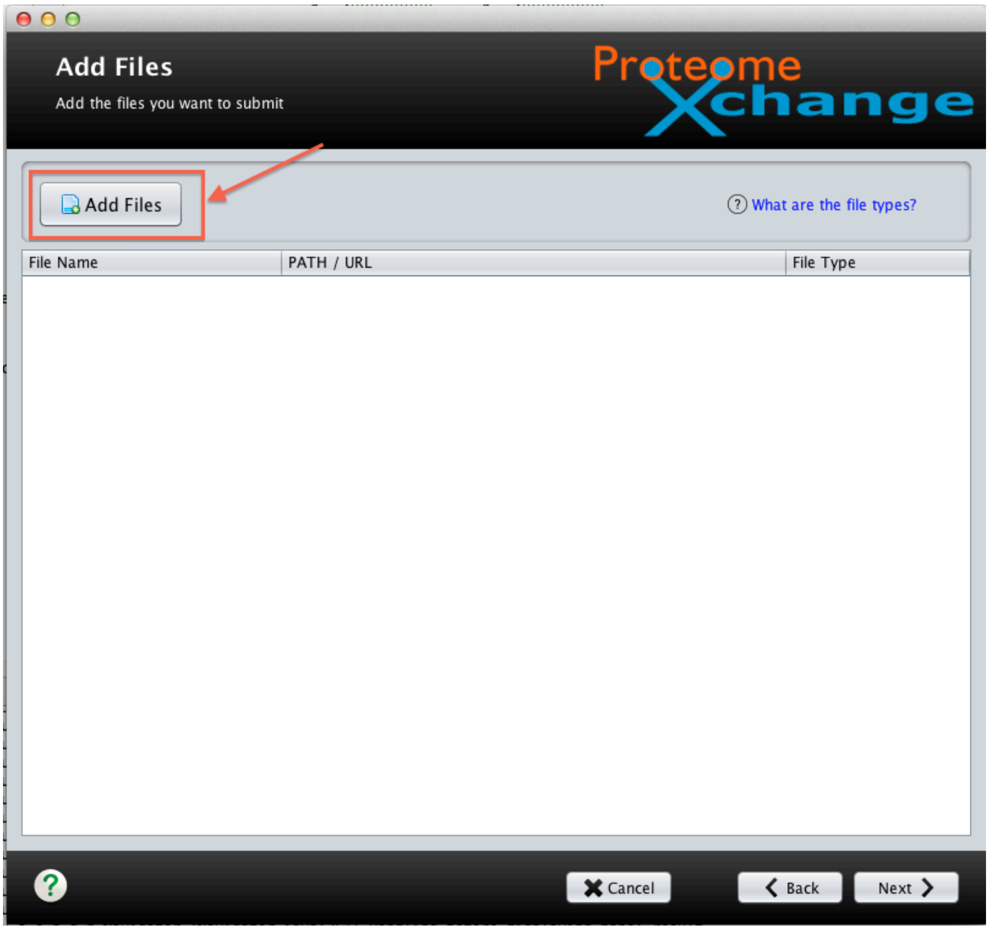
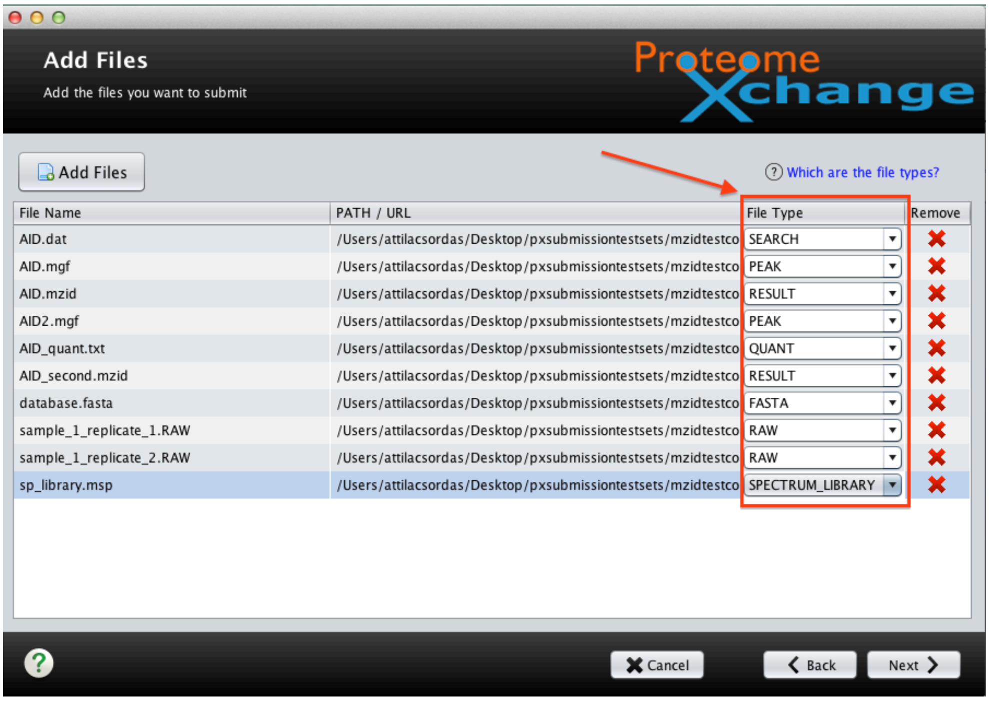


Figure 6: 'Add files' screen of the PX submission tool

There are slight differences in this step between the two subtypes of submissions so we will discuss them separately.

Step 6A: mzIdentML files

You have to make sure that at least 'RESULT' files, 'RAW files and 'PEAK' files are selected. The minimal dataset should contain at least one of the abovementioned files so 3 files in total. There could also be other file types included in the submission: 'SEARCH' (for search engine output files in case those were not mzIdentML files natively), 'QUANT', for quantification results, 'FASTA', for sequence database files, 'SPECTRUM_LIBRARY' for spectral library files, 'GEL', for gel images, or 'OTHER' (any other file e.g. protein inference, post-search files). All the files need to be selected at this stage. Once they are added, double-check that they were assigned with the correct file type, as shown in **Figure 7**.



Add Files

Add the files you want to submit

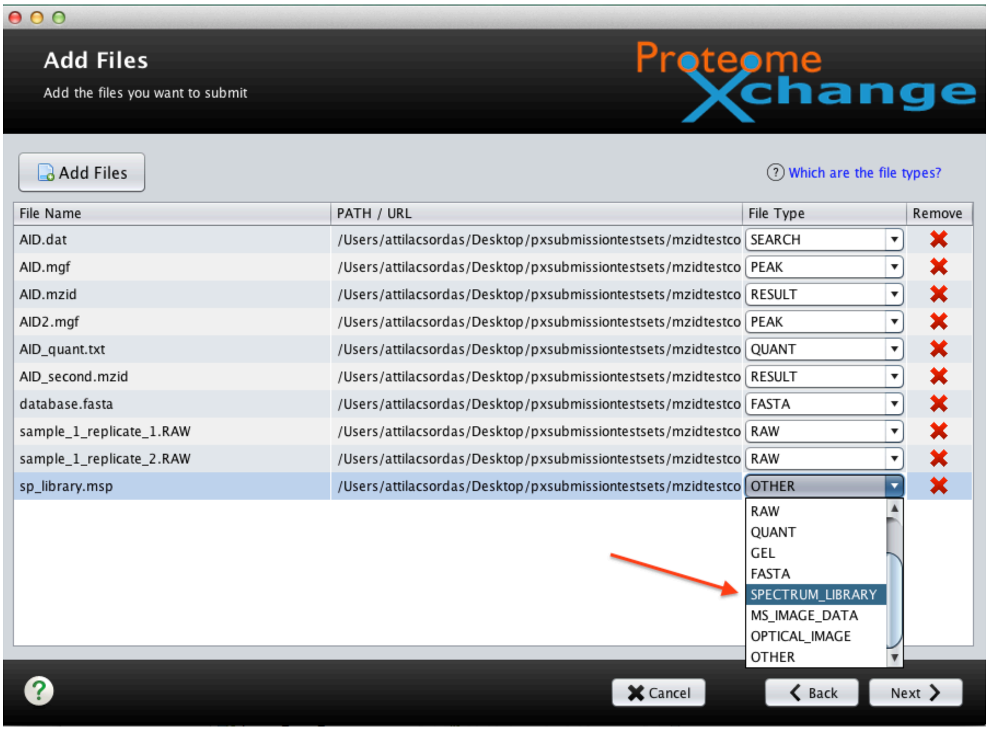
File Name	PATH / URL	File Type	Remove
AID.dat	/Users/attilacsordas/Desktop/pxsubmissiontestsets/mzidtestco	SEARCH	✖
AID.mgf	/Users/attilacsordas/Desktop/pxsubmissiontestsets/mzidtestco	PEAK	✖
AID.mzid	/Users/attilacsordas/Desktop/pxsubmissiontestsets/mzidtestco	RESULT	✖
AID2.mgf	/Users/attilacsordas/Desktop/pxsubmissiontestsets/mzidtestco	PEAK	✖
AID_quant.txt	/Users/attilacsordas/Desktop/pxsubmissiontestsets/mzidtestco	QUANT	✖
AID_second.mzid	/Users/attilacsordas/Desktop/pxsubmissiontestsets/mzidtestco	RESULT	✖
database.fasta	/Users/attilacsordas/Desktop/pxsubmissiontestsets/mzidtestco	FASTA	✖
sample_1_replicate_1.RAW	/Users/attilacsordas/Desktop/pxsubmissiontestsets/mzidtestco	RAW	✖
sample_1_replicate_2.RAW	/Users/attilacsordas/Desktop/pxsubmissiontestsets/mzidtestco	RAW	✖
sp_library.msp	/Users/attilacsordas/Desktop/pxsubmissiontestsets/mzidtestco	SPECTRUM_LIBRARY	✖

Figure 7: Adding files in case of an mzIdentML based 'Complete' submission: Assignment of the correct file types

In the case of 'PEAK' files, the tool will check and validate that all the required file(s) that were referenced in the mzIdentML file's <SpectraData> element are present. If your peak list files had an extension recognized by the tool (.mgf , .dta, .ms2, .pkl) then the tool will automatically annotate the type as 'PEAK' (see Figure 6) but in other cases you have to assign the file type yourself. For instance, .mzXML files, are recognized by the tool as them as 'PEAK' files by default but, since they can be used as 'RAW' file replacements as well, you may need to manually change that (see Figure 7 as an example of file type assignment switch). The same applies if you are using a peak list files format that is not recognized by the tool as a 'PEAK' file but as an 'OTHER' file.

In case both the referenced 'PEAK' files and the 'RAW' files are the same files (in a XML-based format) then currently you need to provide them twice, as 'RAW' and as 'PEAK'.

If you are adding a spectral library file, then please assign the file type manually (see **Figure 8**) as these files might come in many different flavors, for instance as .msp, .splib or .nist files.



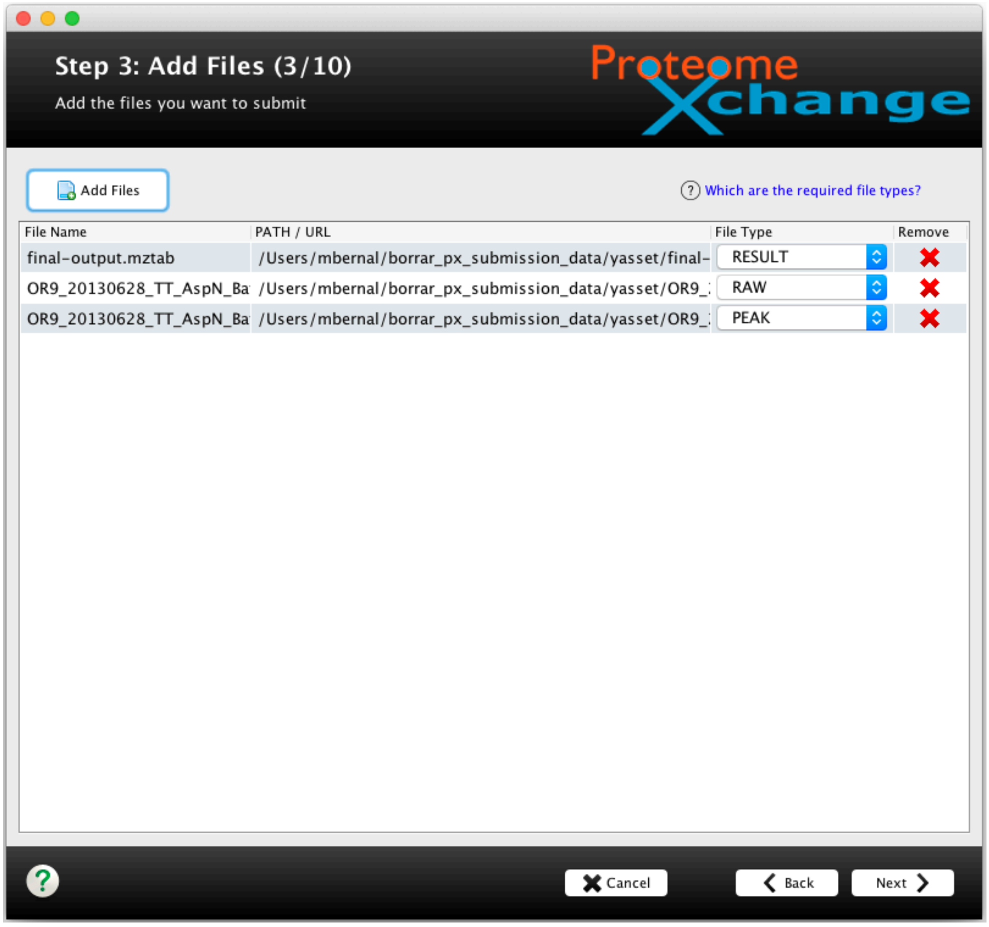
File Name	PATH / URL	File Type	Remove
AID.dat	/Users/attilacsordas/Desktop/pxsubmissiontestsets/mzidtestco	SEARCH	✖
AID.mgf	/Users/attilacsordas/Desktop/pxsubmissiontestsets/mzidtestco	PEAK	✖
AID.mzid	/Users/attilacsordas/Desktop/pxsubmissiontestsets/mzidtestco	RESULT	✖
AID2.mgf	/Users/attilacsordas/Desktop/pxsubmissiontestsets/mzidtestco	PEAK	✖
AID_quant.txt	/Users/attilacsordas/Desktop/pxsubmissiontestsets/mzidtestco	QUANT	✖
AID_second.mzid	/Users/attilacsordas/Desktop/pxsubmissiontestsets/mzidtestco	RESULT	✖
database.fasta	/Users/attilacsordas/Desktop/pxsubmissiontestsets/mzidtestco	FASTA	✖
sample_1_replicate_1.RAW	/Users/attilacsordas/Desktop/pxsubmissiontestsets/mzidtestco	RAW	✖
sample_1_replicate_2.RAW	/Users/attilacsordas/Desktop/pxsubmissiontestsets/mzidtestco	RAW	✖
sp_library.msp	/Users/attilacsordas/Desktop/pxsubmissiontestsets/mzidtestco	OTHER	✖

Figure 8: Switching the file type to the correct file type in case of an mzIdetnML based 'Complete' submission

Step 6B: mzTab Files

When adding files please make sure that at least ‘RESULT’ files and the ‘RAW files are selected. The minimal dataset should contain at least one mzTab ‘RESULT’ file and one ‘RAW’ file, so two files in total. Peak list files are not mandatory as opposed to mzIdentML based ‘Complete’ submissions. Once the files are added, double-check that they were assigned with the correct file type, as shown in **Figure 9**.

There could also be other file types included in the submission: ‘SEARCH’ (for search engine output files), ‘PEAK’ (for peak list files), ‘QUANT’ (for quantification results), ‘FASTA’ (for sequence database files), SPECTRUM_LIBRARY, for spectral library files, ‘GEL’ (for gel images) or ‘OTHER’. All these files need to be selected at this stage.



Step 3: Add Files (3/10)

Add the files you want to submit

File Name	PATH / URL	File Type	Remove
final-output.mztab	/Users/mbernal/borrar_px_submission_data/yasset/final-	RESULT	✖
OR9_20130628_TT_AspN_Ba	/Users/mbernal/borrar_px_submission_data/yasset/OR9_	RAW	✖
OR9_20130628_TT_AspN_Ba	/Users/mbernal/borrar_px_submission_data/yasset/OR9_	PEAK	✖

Figure 9: Adding files in case of a mzTab based 'Complete' submission: Assignment of the correct file types

Step 7: Assign relationships between the submitted files

This mapping step consist of assigning the relations between the 'RESULT' files and the other types of files included in the submission, for example, which 'RAW' (mandatory), 'PEAK' (mandatory for mzIdentML 1.1), 'SEARCH', 'QUANT', 'FASTA', 'SPECTRUM_LIBRARY', 'GEL' or 'OTHER' files can be linked to a given 'RESULT' file or are associated with it. This will enable others to understand how your data is connected and structured.

By default, the tool makes an attempt to generate the mapping between the 'RESULT' and the other - most importantly 'RAW' - files. For instance, if there has been only 1 'RESULT' file found during the previous 'Add Files' step (Step 5) then all the other files will be mapped to this 'RESULT' file. If there are multiple 'RESULT' files the tool maps the other files – 'RAW', 'PEAK', 'SEARCH', … - with the same file name prefix, but without the file extension, to the corresponding 'RESULT' files. This mapping is done even if the suffix part of the 'RAW' files contains different numbers (for instance indicating different replicates).

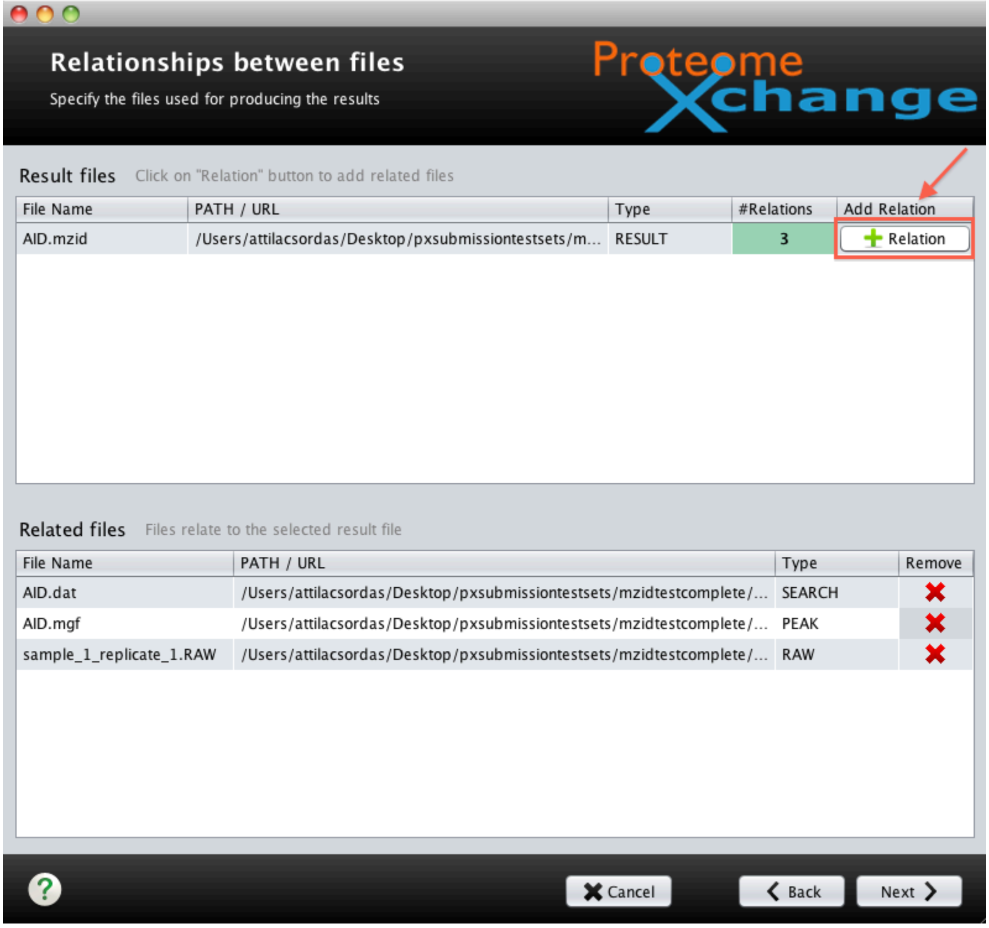
If the automatic mapping is partial only or does not apply, the submitter is asked to manually assign the relationships between the files.

Since there are differences in this step between the two subtypes we are going to discuss them separately.

Step 7A: mzIdentML files

Each mzIdentML ‘RESULT’ file must have at least two files mapped to it: a ‘RAW’ and a ‘PEAK’ file. Make sure you assign the ‘PEAK’ type to the file(s) containing spectra information and referenced in the corresponding mzIdentML files, as discussed in the previous step (5A).

As shown in **Figure 10** the file linking is done by clicking on the ‘Add Relation’ button. Many files can be assigned to the same ‘RESULT’ file.



Relationships between files

Specify the files used for producing the results

Result files Click on "Relation" button to add related files

File Name	PATH / URL	Type	#Relations	Add Relation
AID.mzid	/Users/attilacsordas/Desktop/pxsubmissiontestsets/m...	RESULT	3	+ Relation

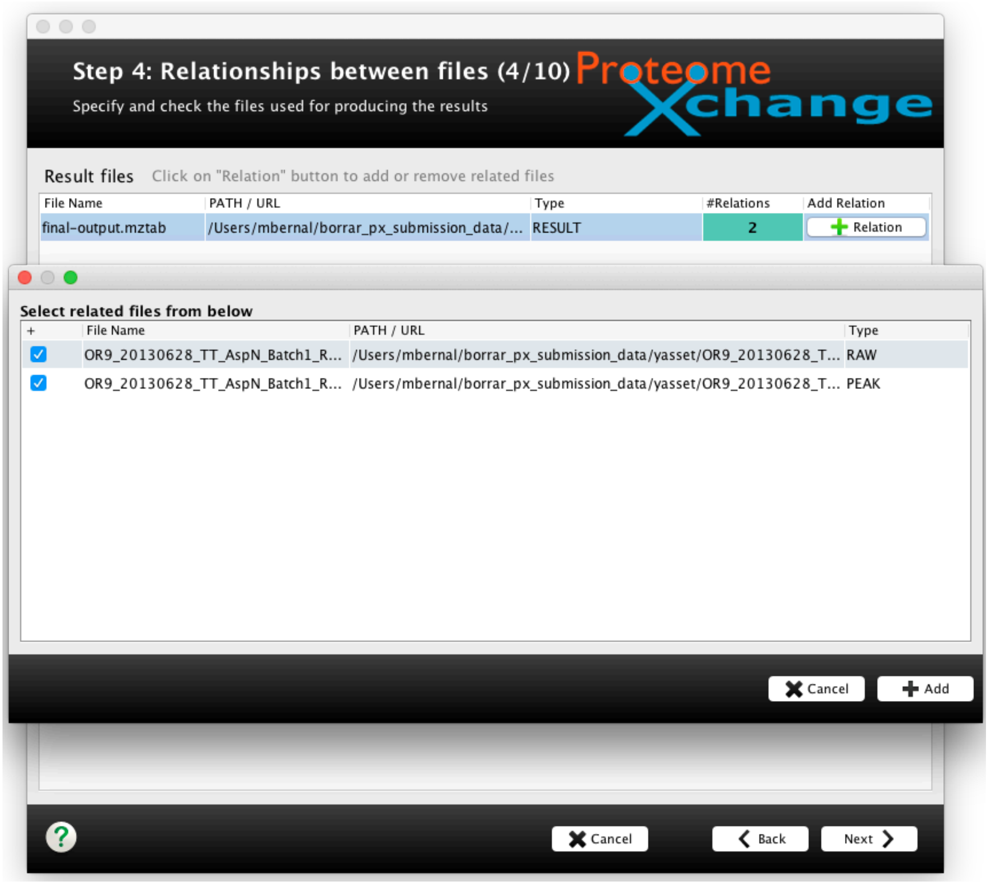
Related files Files relate to the selected result file

File Name	PATH / URL	Type	Remove
AID.dat	/Users/attilacsordas/Desktop/pxsubmissiontestsets/mzidtestcomplete/...	SEARCH	✖
AID.mgf	/Users/attilacsordas/Desktop/pxsubmissiontestsets/mzidtestcomplete/...	PEAK	✖
sample_1_replicate_1.RAW	/Users/attilacsordas/Desktop/pxsubmissiontestsets/mzidtestcomplete/...	RAW	✖

Figure 10: 'Relationships between files' screen of the PX submission tool

Step 7B: mzTab files

Each ‘RESULT’ file must have at least one ‘RAW’ file linked to it. **Figure 11** shows the dialog that allows you to map, for a selected ‘RESULT’ file, a list of other type of files.



Step 4: Relationships between files (4/10)

Specify and check the files used for producing the results

Result files Click on "Relation" button to add or remove related files

File Name	PATH / URL	Type	#Relations	Add Relation
final-output.mztab	/Users/mbernal/borrar_px_submission_data/...	RESULT	2	Relation

Select related files from below

+	File Name	PATH / URL	Type
☑	OR9_20130628_TT_AspN_Batch1_R...	/Users/mbernal/borrar_px_submission_data/yasset/OR9_20130628_T...	RAW
☑	OR9_20130628_TT_AspN_Batch1_R...	/Users/mbernal/borrar_px_submission_data/yasset/OR9_20130628_T...	PEAK

Figure 11: Assigning mappings between different and multiple file types on the 'Relationships between files'

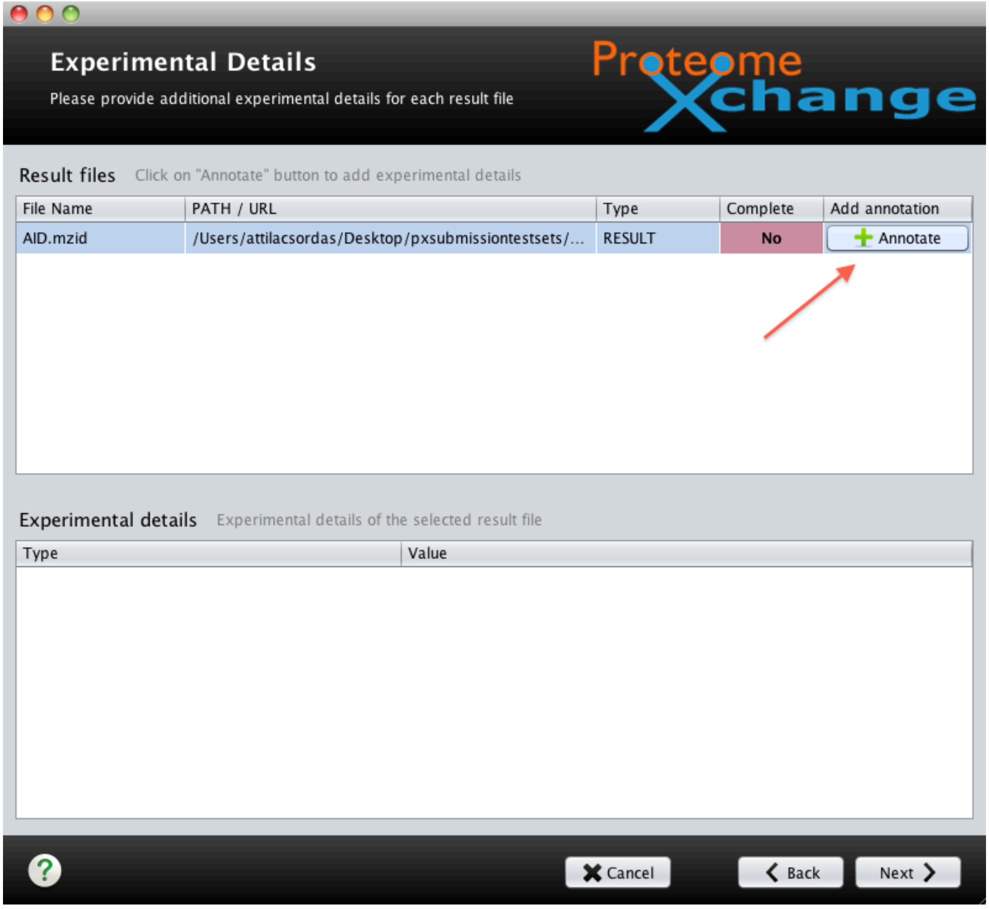
Step 8: Provide additional experimental details for each result file

Additional metadata need be provided for each ‘RESULT’ file in the case of a ‘Complete’ submission, and what is needed is the same for both subtypes of submissions (mzTab and mzIdentML).

Figure 12 shows the screen where the ‘Annotate’ button can be clicked for each ‘RESULT’ file. This information is usually imported automatically in the case of a mzTab file (if the recommended CVs/ontologies are used). For mzIdentML, the information needs to be manually annotated.

The following additional metadata are required: species, tissue, and instrument information (provided as Controlled Vocabulary (CV) terms from a drop-down menu), and experimental factor information in a free text format (**Figure 13**). Optionally, providing information about the cell type, disease and quantification method (if applicable) is recommended.

If you have more than one ‘RESULT’ file, as it is usually the case, then you can pick the ‘Apply to all’ box for species and tissue information instead of doing this many times.



Experimental Details

Please provide additional experimental details for each result file

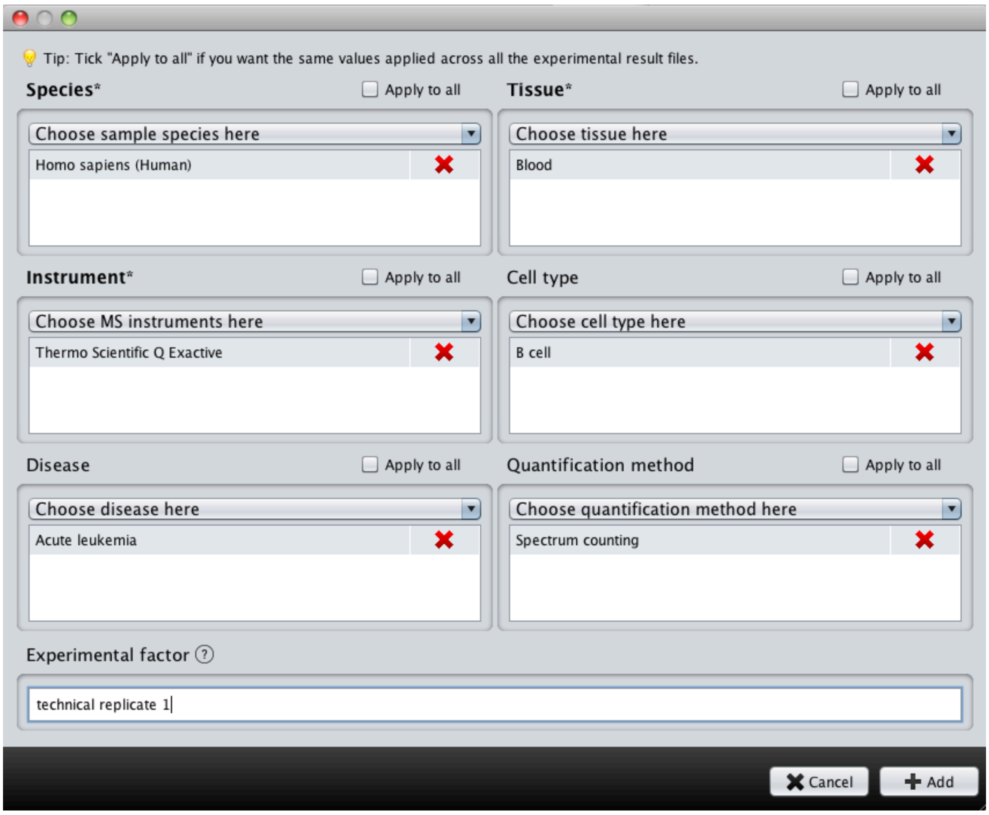
Result files Click on "Annotate" button to add experimental details

File Name	PATH / URL	Type	Complete	Add annotation
AID.mzid	/Users/attilacsordas/Desktop/pxsubmissiontestsets/...	RESULT	No	Annotate

Experimental details Experimental details of the selected result file

Type	Value

Figure 12: Please click the 'Annotate' button to add metadata to each result file



Tip: Tick "Apply to all" if you want the same values applied across all the experimental result files.

Species* ☐ Apply to all
Choose sample species here
Homo sapiens (Human) ✖

Tissue* ☐ Apply to all
Choose tissue here
Blood ✖

Instrument* ☐ Apply to all
Choose MS instruments here
Thermo Scientific Q Exactive ✖

Cell type ☐ Apply to all
Choose cell type here
B cell ✖

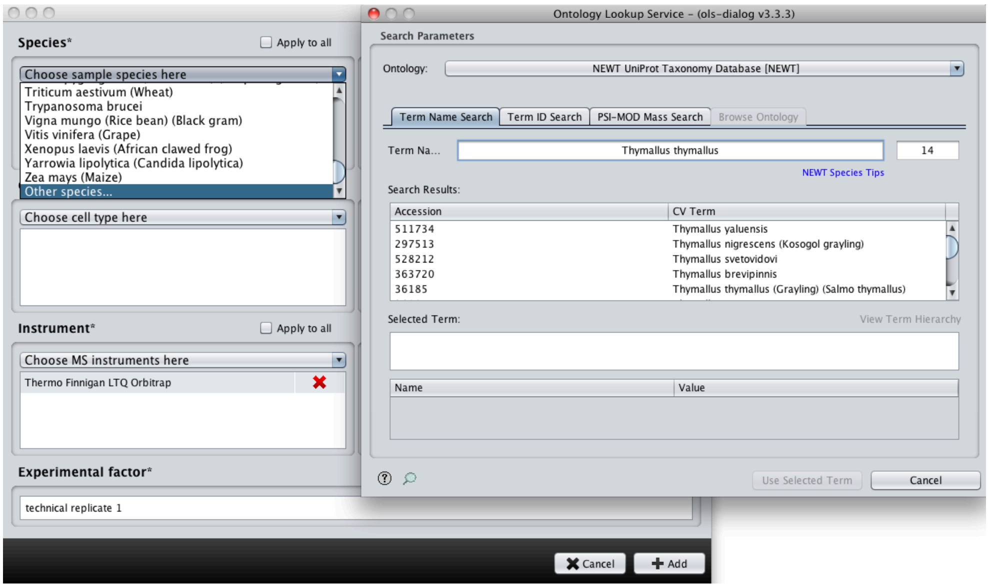
Disease ☐ Apply to all
Choose disease here
Acute leukemia ✖

Quantification method ☐ Apply to all
Choose quantification method here
Spectrum counting ✖

Experimental factor ⓘ
technical replicate 1

Figure 13: Annotating each result files with additional metadata

In the majority of the cases you will find the metadata annotation you are looking for in the drop-down menu since the elements of the drop-down menus have been selected based on frequency. But sometimes the annotation you are looking for is not going to be available from the drop-down lists. If that's the case, you have to select to the OLS (Ontology Lookup Service) panel and search for the annotation you want to provide. For the more extensive search you need to click on the "other" options on the bottom of the drop-down menu. For instance, if you have samples from e.g. the fish Grayling (*Thymallus thymallus*) the species is not available from the drop-down list menu. You have to click on "Other species" and search for *Thymallus thymallus* in the OLS panel, see **Figure 14**.



Ontology: NEWT UniProt Taxonomy Database [NEWT]

Term Na... Thymallus thymallus 14

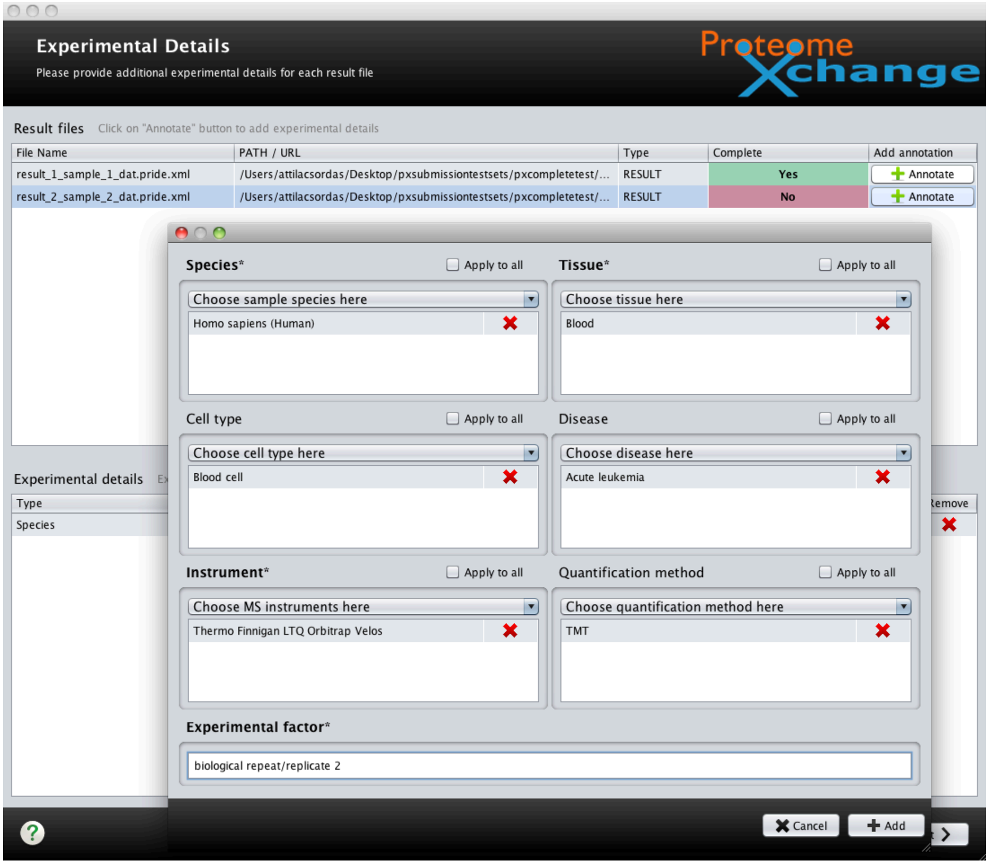
Accession	CV Term
511734	Thymallus yaluensis
297513	Thymallus nigrescens (Kosogol grayling)
528212	Thymallus svetovidovi
363720	Thymallus brevipinnis
36185	Thymallus thymallus (Grayling) (Salmo thymallus)

Instrument: Thermo Finnigan LTQ Orbitrap

Experimental factor: technical replicate 1

Figure 14: Annotating a result file with additional metadata with the help of the OLS panel

In case you have multiple 'RESULT' files you have to provide data for all of them using the same panel, see **Figure 15**.



File Name	PATH / URL	Type	Complete	Add annotation
result_1_sample_1_dat.pride.xml	/Users/attilacsordas/Desktop/pxsubmissiontestsets/pxcompletetest/...	RESULT	Yes	Annotate
result_2_sample_2_dat.pride.xml	/Users/attilacsordas/Desktop/pxsubmissiontestsets/pxcompletetest/...	RESULT	No	Annotate

Species: Homo sapiens (Human)

Tissue: Blood

Cell type: Blood cell

Disease: Acute leukemia

Instrument: Thermo Finnigan LTQ Orbitrap Velos

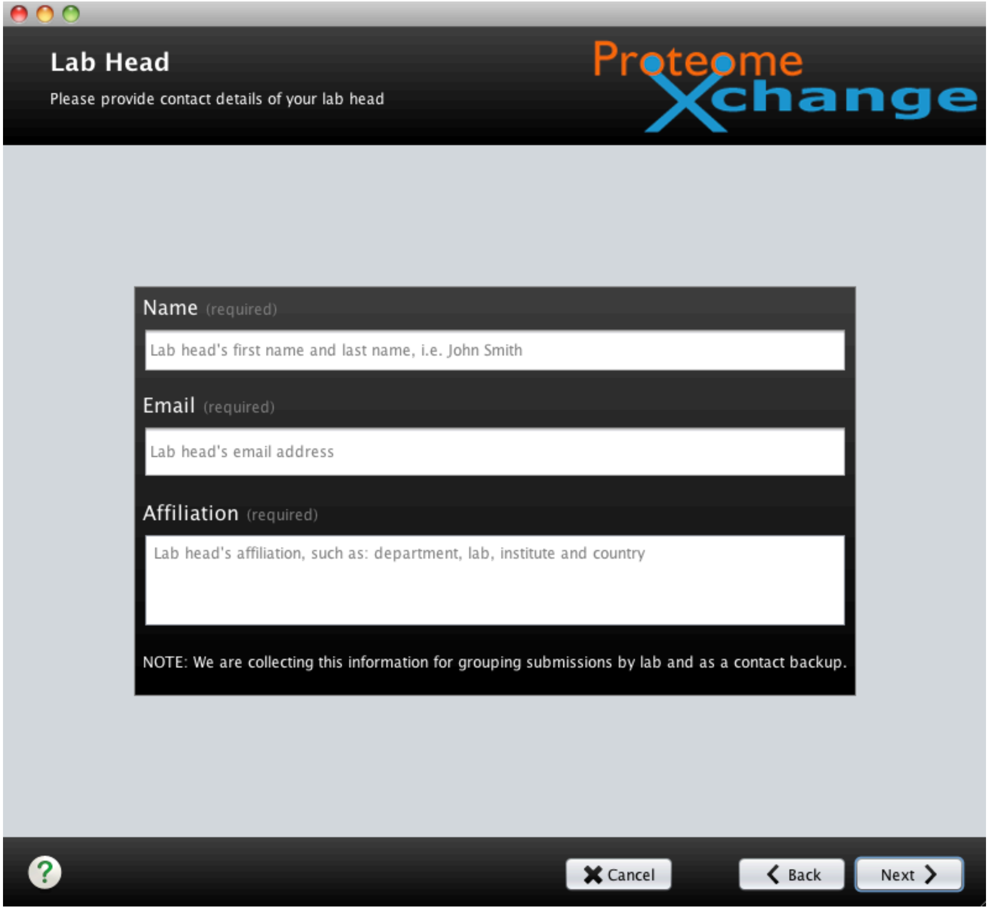
Quantification method: TMT

Experimental factor: biological repeat/replicate 2

Figure 15: Annotating multiple result files

Step 9: Add Lab Head

Please provide contact details for the Lab Head/Principal Investigator of your study. Please do it in the recommended format, see **Figure 16**.



Lab Head

Please provide contact details of your lab head

Name (required)

Lab head's first name and last name, i.e. John Smith

Email (required)

Lab head's email address

Affiliation (required)

Lab head's affiliation, such as: department, lab, institute and country

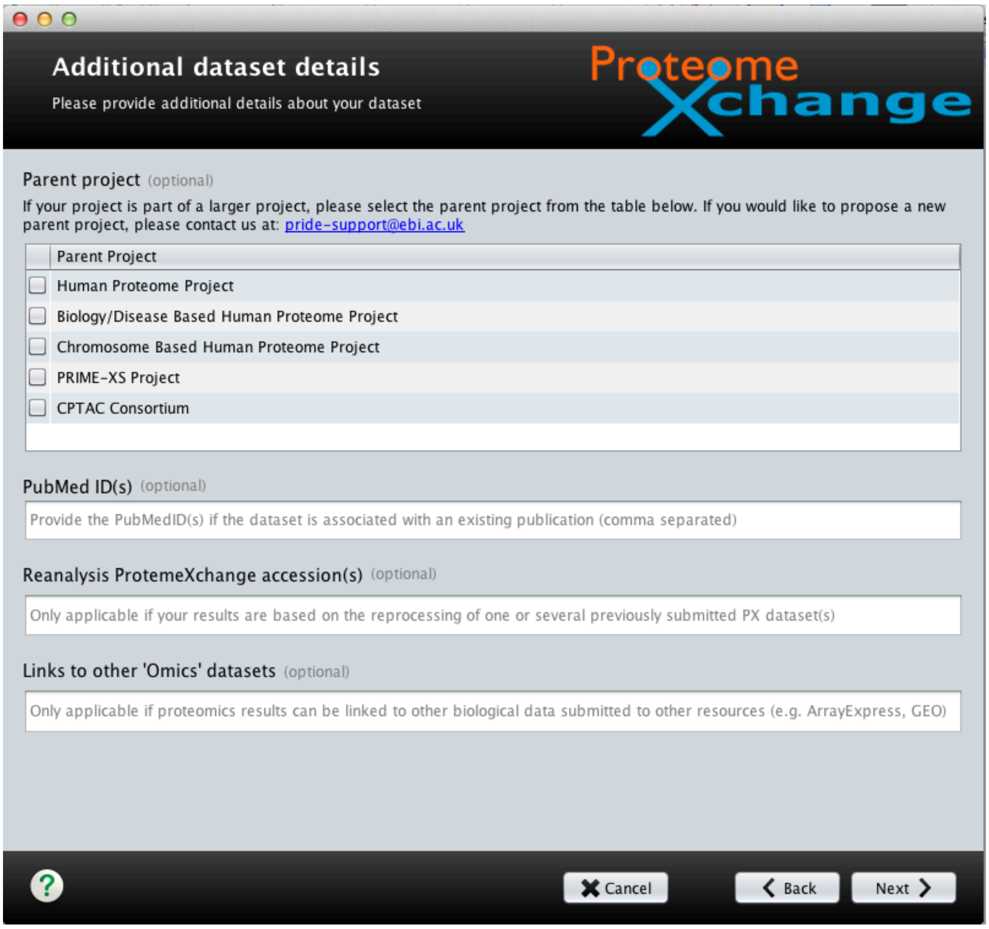
NOTE: We are collecting this information for grouping submissions by lab and as a contact backup.

Figure 16: Providing contact details for the Lab Head

Step 10: Optional metadata annotation

In this panel, it is recommended to provide additional metadata in four cases:

- Your dataset is part of a bigger project/effort (for instance the Human Proteome Project or the EU project 'PRIME-XS'). It is a way to tag your dataset to enable grouping of datasets this way.
- There is already a PubMed ID associated with it (the data has been already published).
- Your dataset represents a reanalysis of an earlier public PX dataset.
- There are other "omics" datasets (for instance transcriptomics, metabolomics data present in other repositories) that can be associated with it. In this case, please provide the accession number of the dataset in the corresponding repository.



Additional dataset details

Please provide additional details about your dataset

Parent project (optional)

If your project is part of a larger project, please select the parent project from the table below. If you would like to propose a new parent project, please contact us at: pride-support@ebi.ac.uk

	Parent Project
□	Human Proteome Project
□	Biology/Disease Based Human Proteome Project
□	Chromosome Based Human Proteome Project
□	PRIME-XS Project
□	CPTAC Consortium

PubMed ID(s) (optional)

Provide the PubMedID(s) if the dataset is associated with an existing publication (comma separated)

Reanalysis ProtemeXchange accession(s) (optional)

Only applicable if your results are based on the reprocessing of one or several previously submitted PX dataset(s)

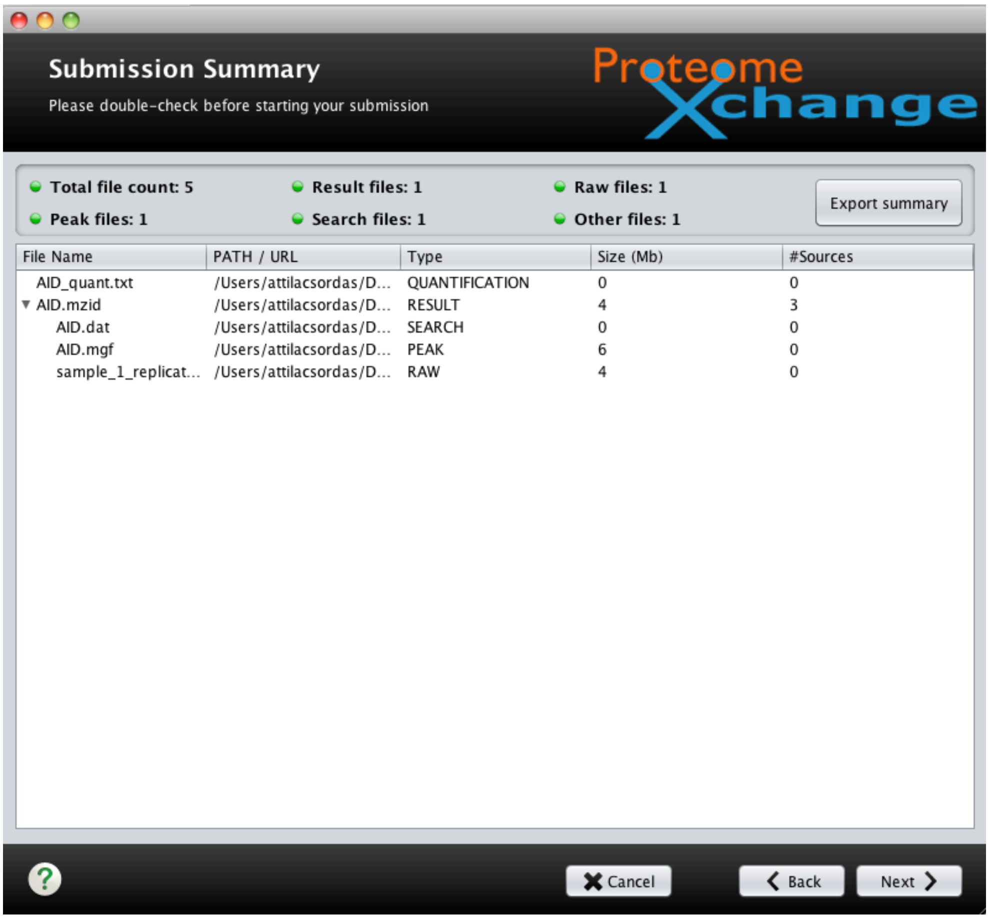
Links to other 'Omics' datasets (optional)

Only applicable if proteomics results can be linked to other biological data submitted to other resources (e.g. ArrayExpress, GEO)

Figure 17: Providing additional, applicable metadata

Step 11: Check before submission

This is the last step before the file upload actually starts. You should double-check that all the necessary files are included in the submission summary before continuing to the upload step, see an example of an mzIdentML based 'complete' submission in **Figure 18**.



Submission Summary

Please double-check before starting your submission

● Total file count: 5
● Result files: 1
● Raw files: 1
● Peak files: 1
● Search files: 1
● Other files: 1

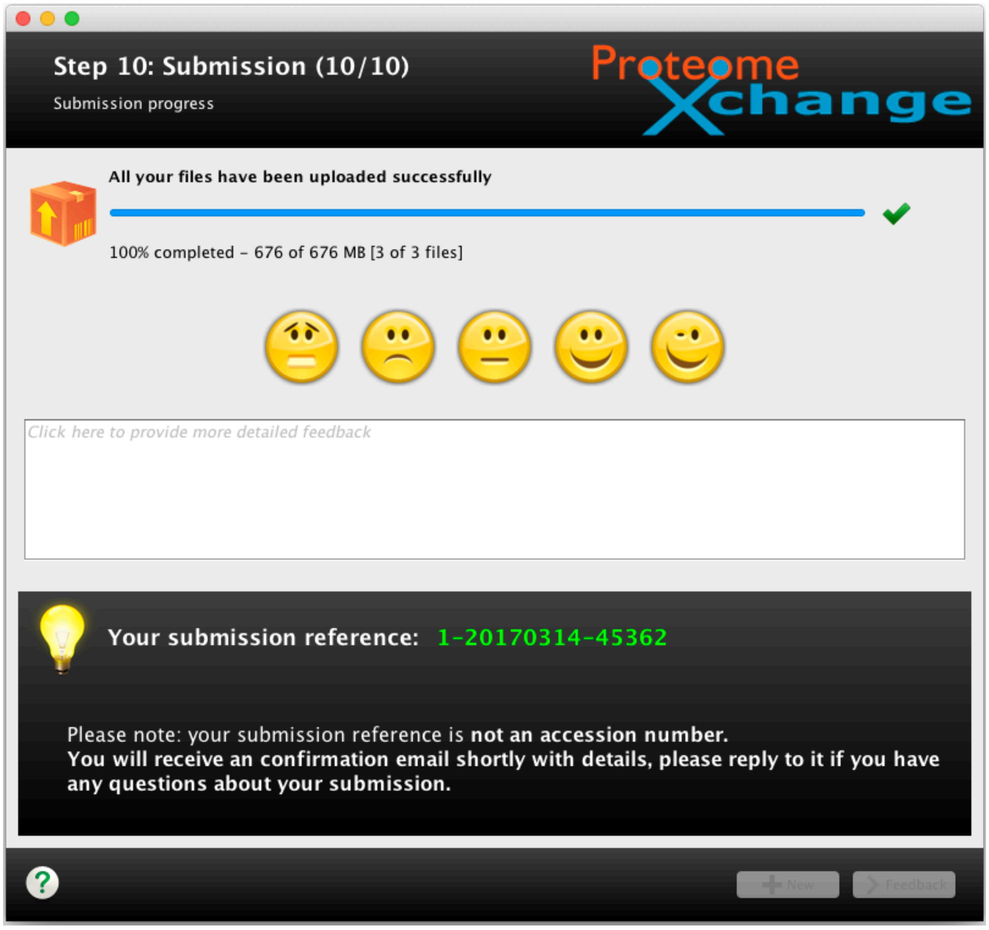
File Name	PATH / URL	Type	Size (Mb)	#Sources
AID_quant.txt	/Users/attilacsordas/D...	QUANTIFICATION	0	0
▼ AID.mzid	/Users/attilacsordas/D...	RESULT	4	3
AID.dat	/Users/attilacsordas/D...	SEARCH	0	0
AID.mgf	/Users/attilacsordas/D...	PEAK	6	0
sample_1_replicat...	/Users/attilacsordas/D...	RAW	4	0

Figure 18: Example of 'Submission Summary' screen in the PX Submission Tool with a single 'RESULT' file

Step 12: File Submission

This is the actual step when all your files are uploaded to PRIDE and ProteomeXchange (**Figure 19**). Once the upload is finished, an e-mail will be sent to you to confirm that all your files have been uploaded successfully and that are waiting to be validated. If for any reason the tool crashes at this point, the PX Submission Tool can be restarted and the file upload will restart in the same point before it crashed.

By default, the PX submission Tool (since version 2.1) is using the fast Aspera upload transfer protocol with which terabytes can be potentially transferred within a day. Aspera functionality usually provides much faster file transfer speeds than FTP (typically up to 50 times). Should there be any issues with the Aspera upload (probably due to the Internet/ data transfer local settings), submitters can always switch to the slower FTP file transfer protocol by changing the 'px.upload.protocol = aspera' line to 'px.upload.protocol = ftp' in the plain config.props text file located in the 'config' subdirectory in the PX Submission Tool's working directory. You will be also issued with a temporary submission reference, to help us to quickly identify and track your submission should you have any questions. This is not the PX accession number.



Step 10: Submission (10/10)

Submission progress

All your files have been uploaded successfully

100% completed – 676 of 676 MB [3 of 3 files]

Click here to provide more detailed feedback

Your submission reference: 1-20170314-45362

Please note: your submission reference is **not an accession number.**
You will receive an confirmation email shortly with details, please reply to it if you have any questions about your submission.

Figure 19: 'Submission' screen of the PX Submission Tool showing that a submission has been completed

Step 13: User Feedback

In PRIDE, we strive for excellence, we are always looking for ways to be at the top of most useful resources for the proteomics community, and the best way to do it, is getting as much feedback as possible from our user base.
After your data, has been uploaded, extra information will appear in the final screen:

- Smiley faces, they represent, from left to right, how satisfied the user is with the submission process in general (very dissatisfied, dissatisfied, neutral, satisfied and very satisfied).
- A free text box, where additional comments can be provided for the PRIDE Team about the submission process.

The level of satisfaction is required to be reported in order for the tool to finish or being able to start another submission, and we really encourage people to provide comments as well, so we can have as much information as possible to find out what's best for the proteomics community.

How to make Partial Submissions?

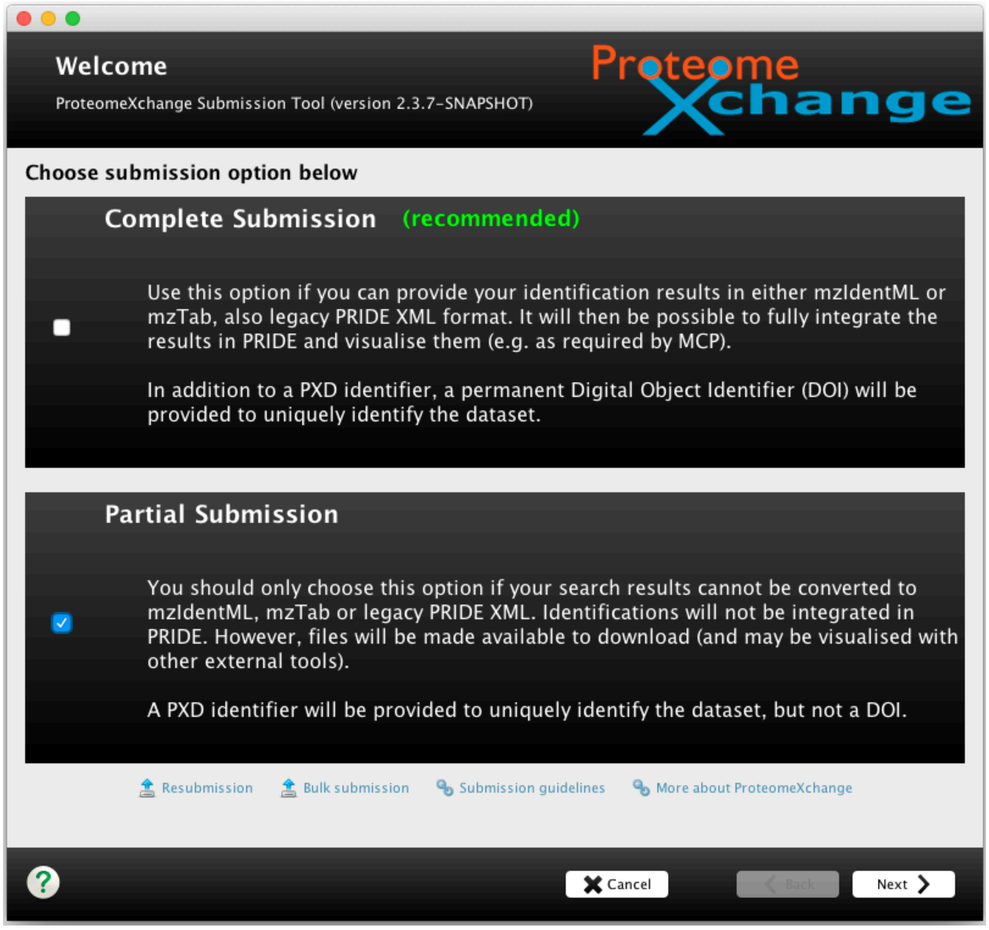
Remember that by default we recommended doing 'Complete' submissions. You should only use this option if your 'RESULT' files cannot be converted/exported to mzTab or mzIdentML 1.1. See Appendix VI for details about the special case of MS imaging datasets.

Step 1: Launch PX Submission Tool

Please install and launch the PX Submission Tool (available at http://www.proteomexchange.org/submission).

Step 2: Select Submission Type

Select ‘Partial Submission’ in the PX Submission Tool ‘Welcome’ screen (**Figure 20**).



Welcome

ProteomeXchange Submission Tool (version 2.3.7-SNAPSHOT)

Choose submission option below

Complete Submission (recommended)

□ Use this option if you can provide your identification results in either mzIdentML or mzTab, also legacy PRIDE XML format. It will then be possible to fully integrate the results in PRIDE and visualise them (e.g. as required by MCP).

In addition to a PXD identifier, a permanent Digital Object Identifier (DOI) will be provided to uniquely identify the dataset.

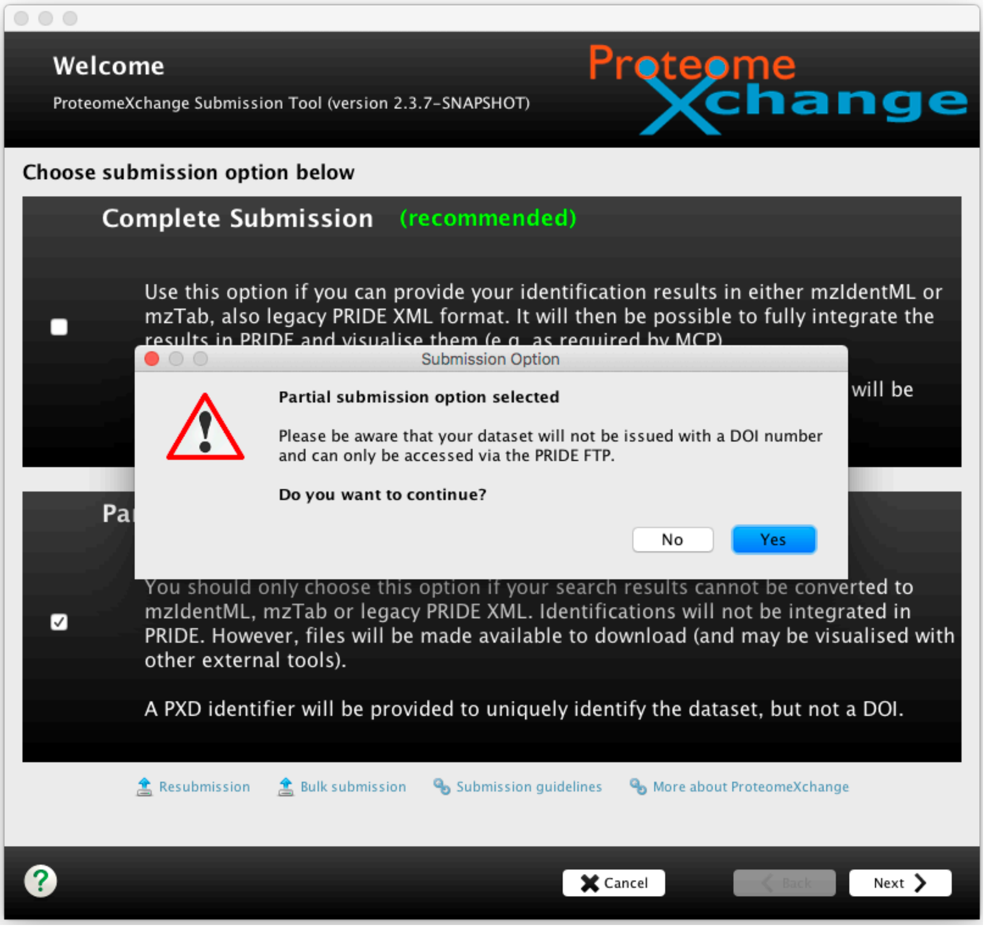
Partial Submission

☑ You should only choose this option if your search results cannot be converted to mzIdentML, mzTab or legacy PRIDE XML. Identifications will not be integrated in PRIDE. However, files will be made available to download (and may be visualised with other external tools).

A PXD identifier will be provided to uniquely identify the dataset, but not a DOI.

Figure 20: Selecting “Partial” Submission in the 'Welcome' screen of the PX Submission Tool

Upon selecting this option a warning will pop up, see **Figure 21**. Continue with clicking ‘Yes’.



Welcome

ProteomeXchange Submission Tool (version 2.3.7-SNAPSHOT)

Choose submission option below

Complete Submission (recommended)

□ Use this option if you can provide your identification results in either mzIdentML or mzTab, also legacy PRIDE XML format. It will then be possible to fully integrate the results in PRIDE and visualise them (e.g. as required by MCP)

will be

Submission Option

Partial submission option selected

Please be aware that your dataset will not be issued with a DOI number and can only be accessed via the PRIDE FTP.

Do you want to continue?

Pa

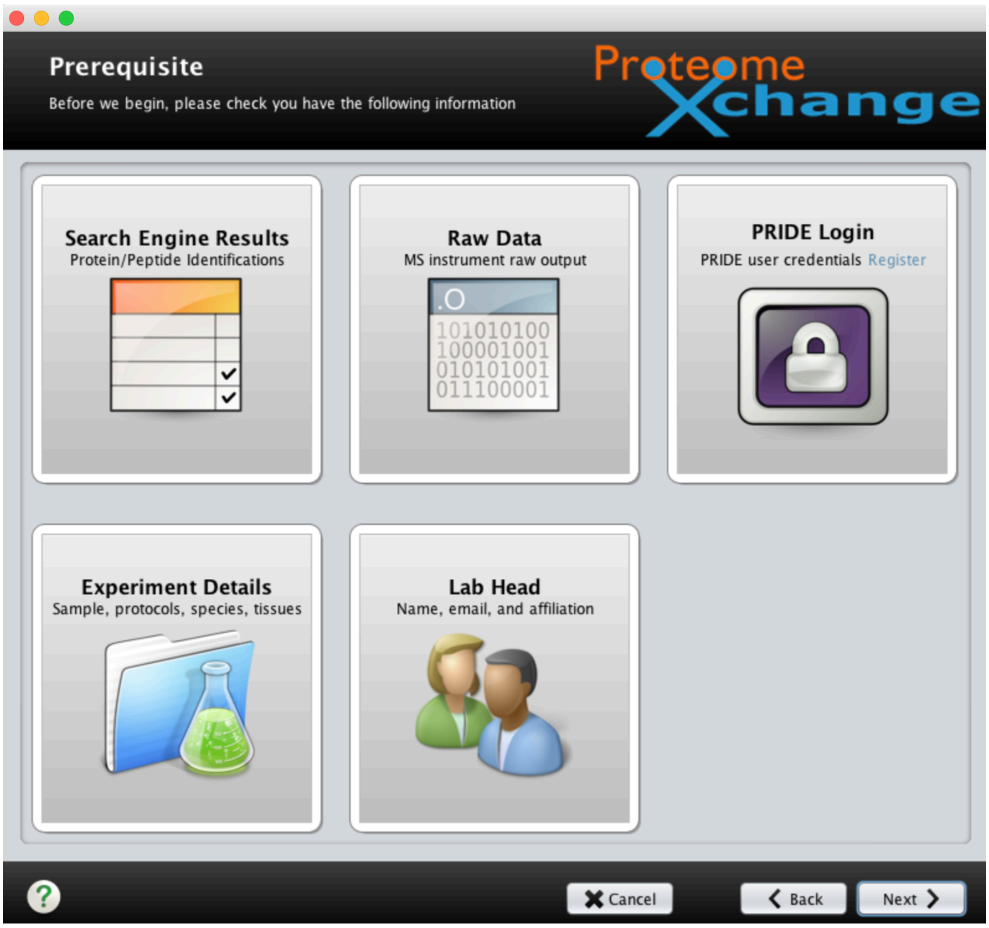
☑ You should only choose this option if your search results cannot be converted to mzIdentML, mzTab or legacy PRIDE XML. Identifications will not be integrated in PRIDE. However, files will be made available to download (and may be visualised with other external tools).

A PXD identifier will be provided to uniquely identify the dataset, but not a DOI.

Figure 21: Warning concerning “Partial” Submissions in the PX Submission Tool

Step 3: Prerequisites

Please double check and make sure that you have all the required information before starting the submission process as shown in **Figure 22**:



Prerequisite

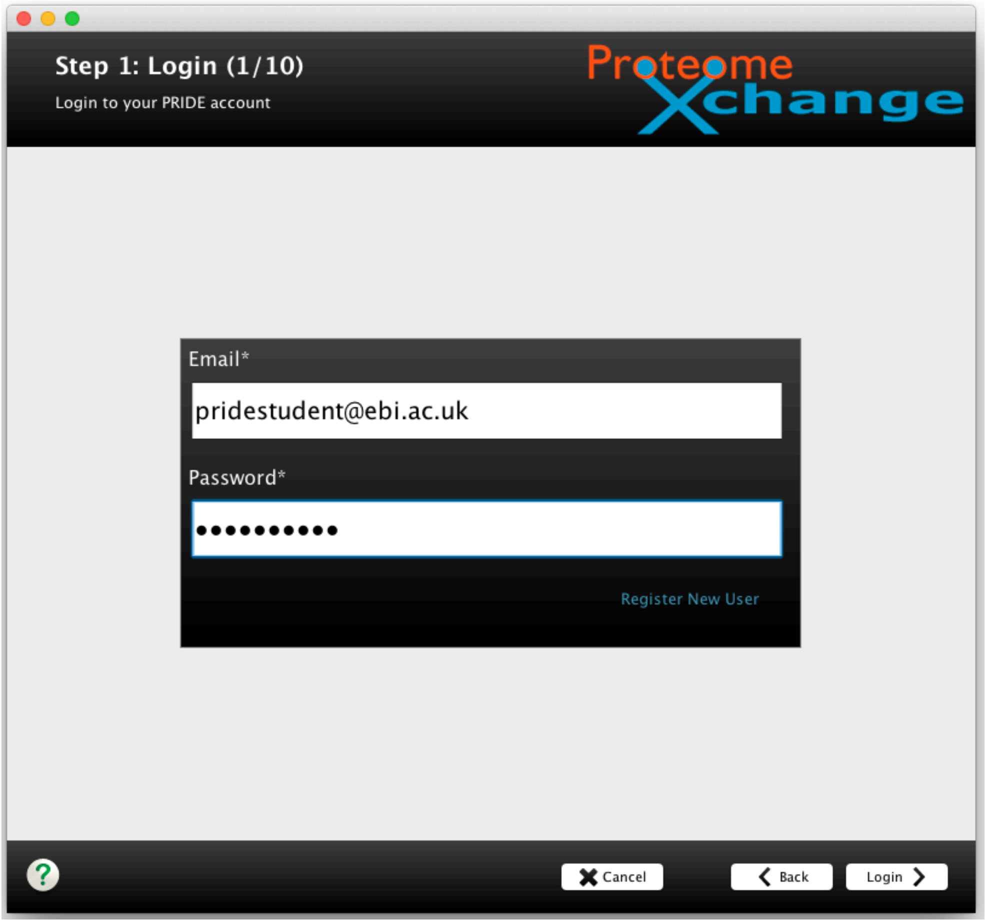
Before we begin, please check you have the following information

ProteomeXchange
Search Engine Results
Protein/Peptide Identifications
Raw Data
MS instrument raw output
PRIDE Login
PRIDE user credentials Register
Experiment Details
Sample, protocols, species, tissues
Lab Head
Name, email, and affiliation
?
Cancel
Back
Next

Figure 22: Prerequisites for “Partial” submissions

Step 4: Login

Please log in using your existing PRIDE account as shown in **Figure 23**:



Step 1: Login (1/10)

Login to your PRIDE account

Email*

pridestudent@ebi.ac.uk

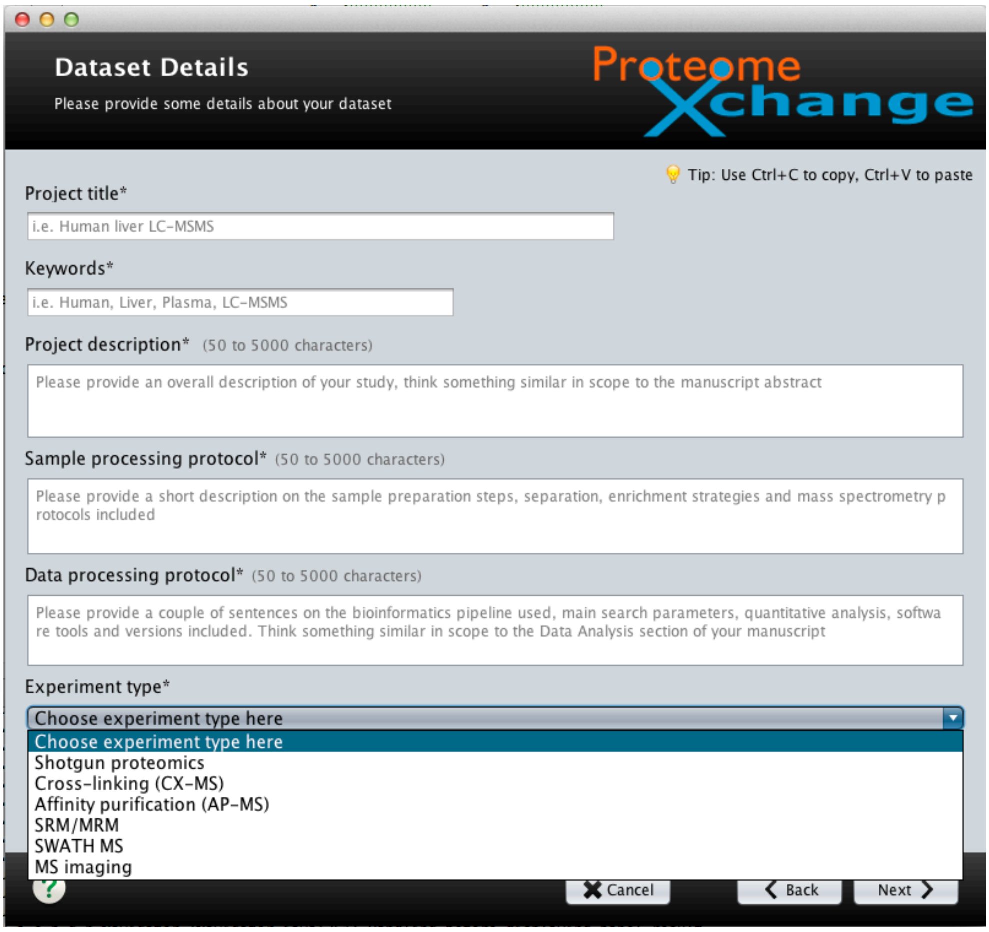
Password*

•••••••••••

Figure 23 : Login screen of the PX Submission Tool

Step 5: Provide submission details

The user is asked to provide some basic details about the uploaded dataset (**Figure 24**) such as the title, a list of keywords (in a comma separated format), and a brief description of the data (similar to the abstract of the corresponding publication) a sample processing and a data processing protocol. The user also picks a mass spectrometry experiment type from a drop-down menu.



Dataset Details

Please provide some details about your dataset

Tip: Use Ctrl+C to copy, Ctrl+V to paste

Project title*

i.e. Human liver LC-MSMS

Keywords*

i.e. Human, Liver, Plasma, LC-MSMS

Project description* (50 to 5000 characters)

Please provide an overall description of your study, think something similar in scope to the manuscript abstract

Sample processing protocol* (50 to 5000 characters)

Please provide a short description on the sample preparation steps, separation, enrichment strategies and mass spectrometry p rotocols included

Data processing protocol* (50 to 5000 characters)

Please provide a couple of sentences on the bioinformatics pipeline used, main search parameters, quantitative analysis, softwa re tools and versions included. Think something similar in scope to the Data Analysis section of your manuscript

Experiment type*

Figure 24: 'Dataset details' screen in the PX Submission Tool

Step 6: Add Files and assign file types

You should choose the files you would like submit in this step. As shown in **Figure 25**, you can add files by clicking on the highlighted button.

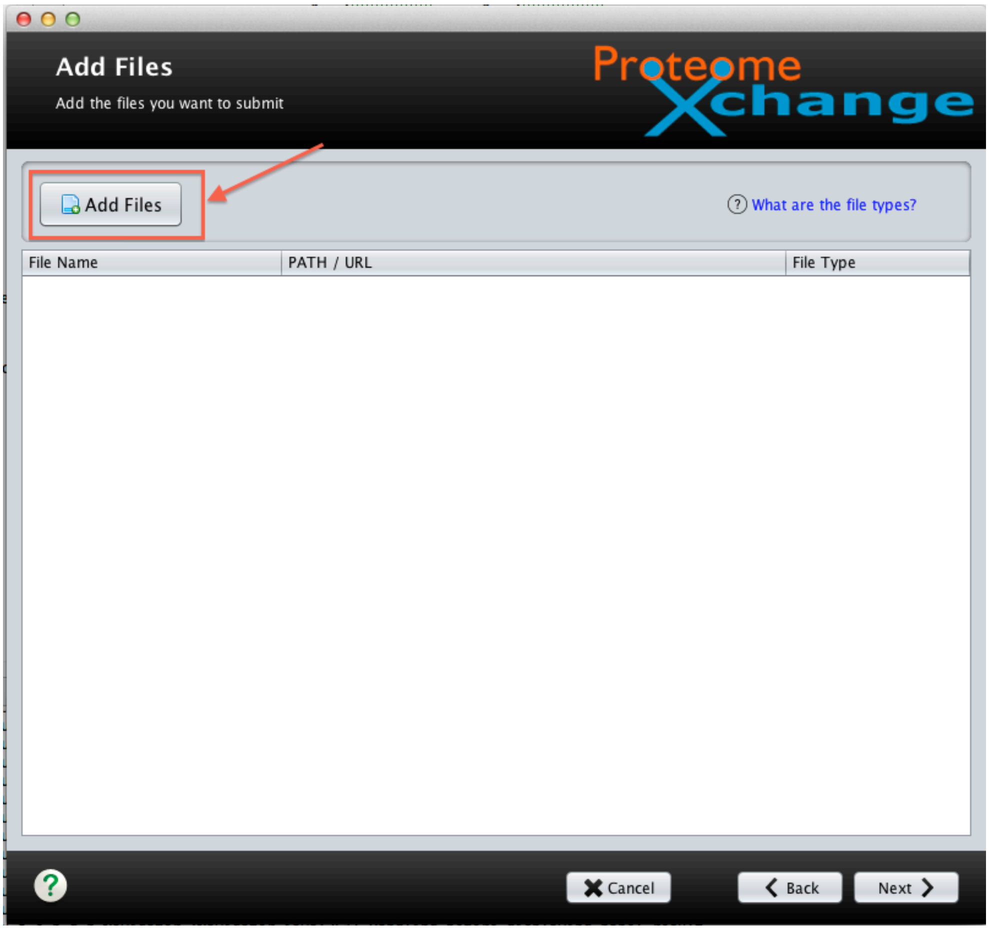
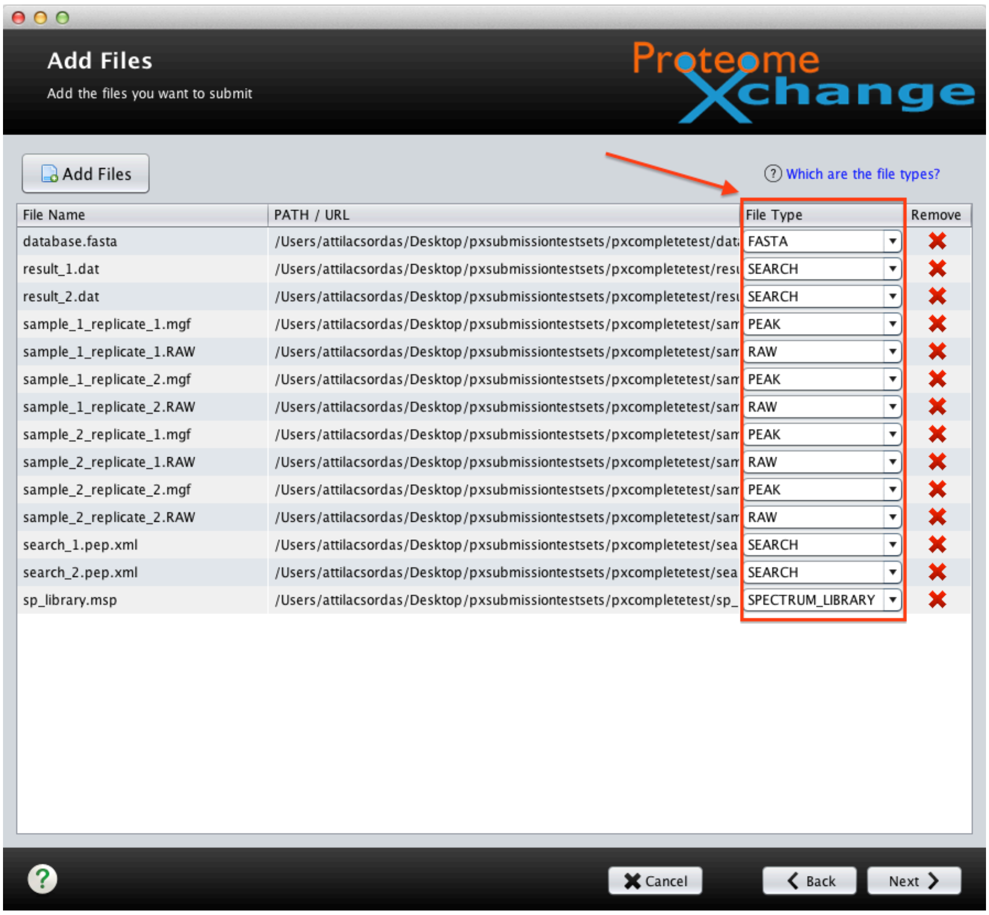


Figure 25: 'Add files' screen of the PX Submission Tool

You should make sure that both the 'SEARCH' search engine output files and the 'RAW' files are selected. The minimal dataset should contain at least one 'SEARCH' and one corresponding 'RAW' file. There could also be other files types included in the submission: 'PEAK' (for peak list files), 'QUANT', for quantification results, 'FASTA, for sequence database files, 'SPECTRUM_LIBRARY', for spectral library files, 'GEL', for gel images, or 'OTHER' (any other file). All the files need to be selected at this stage.

Once the files are added, double-check them to make sure they were assigned with the correct file types. For instance, in **Figure 26**, the pep.xml 'SEARCH' file has been recognized as 'OTHER' file and this need to be changed by selecting 'SEARCH' from the drop-down menu.



File Name	PATH / URL	File Type	Remove
database.fasta	/Users/attilacsordas/Desktop/pxsubmissiontestsets/pxcompletetest/dat	FASTA	✖
result_1.dat	/Users/attilacsordas/Desktop/pxsubmissiontestsets/pxcompletetest/resu	SEARCH	✖
result_2.dat	/Users/attilacsordas/Desktop/pxsubmissiontestsets/pxcompletetest/resu	SEARCH	✖
sample_1_replicate_1.mgf	/Users/attilacsordas/Desktop/pxsubmissiontestsets/pxcompletetest/sam	PEAK	✖
sample_1_replicate_1.RAW	/Users/attilacsordas/Desktop/pxsubmissiontestsets/pxcompletetest/sam	RAW	✖
sample_1_replicate_2.mgf	/Users/attilacsordas/Desktop/pxsubmissiontestsets/pxcompletetest/sam	PEAK	✖
sample_1_replicate_2.RAW	/Users/attilacsordas/Desktop/pxsubmissiontestsets/pxcompletetest/sam	RAW	✖
sample_2_replicate_1.mgf	/Users/attilacsordas/Desktop/pxsubmissiontestsets/pxcompletetest/sam	PEAK	✖
sample_2_replicate_1.RAW	/Users/attilacsordas/Desktop/pxsubmissiontestsets/pxcompletetest/sam	RAW	✖
sample_2_replicate_2.mgf	/Users/attilacsordas/Desktop/pxsubmissiontestsets/pxcompletetest/sam	PEAK	✖
sample_2_replicate_2.RAW	/Users/attilacsordas/Desktop/pxsubmissiontestsets/pxcompletetest/sam	RAW	✖
search_1.pep.xml	/Users/attilacsordas/Desktop/pxsubmissiontestsets/pxcompletetest/sea	SEARCH	✖
search_2.pep.xml	/Users/attilacsordas/Desktop/pxsubmissiontestsets/pxcompletetest/sea	SEARCH	✖
sp_library.msp	/Users/attilacsordas/Desktop/pxsubmissiontestsets/pxcompletetest/sp_	SPECTRUM_LIBRARY	✖

Figure 26: PX Submission Tool 'Add Files' screen: Assignment of the correct file types

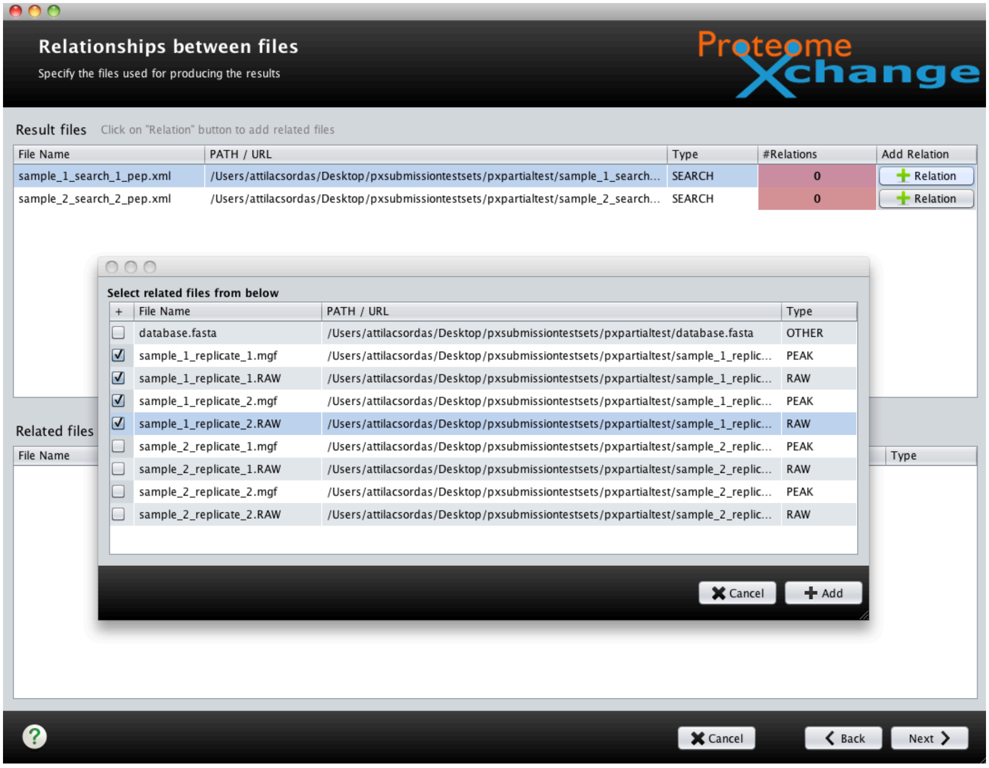
Step 7: Assign relationships between the submitted files

This mapping step consists of assigning the relations between the 'SEARCH' files and the other file types included in the submission, for example, which 'RAW' (mandatory) or 'PEAK' files have been used to produce the search engine output files ('SEARCH'). 'QUANT', 'FASTA', SPECTRUM_LIBRARY', 'GEL' or 'OTHER' files can also be added. This will enable others to understand how your files are connected.

By default, the tool makes an attempt to generate the mapping between the 'SEARCH and the other - most importantly 'RAW' - files. For instance, if there has been only 1 'SEARCH' file found during the previous 'Add Files' step (Step 6) then all the other files will be mapped to this 'SEARCH file. If there are multiple 'SEARCH' files the tool maps the other files – 'RAW', 'PEAK', … - with the same file name prefix, but without the file extension, to the corresponding 'SEARCH files. This mapping is done even if the suffix part of the ‘RAW’ files contains different numbers (for instance indicating different replicates) or the prefix contains spaces or underscores.

If the automatic mapping is partial only or does not apply, the submitter is asked to manually assign the relationships between the files.

Each ‘SEARCH’ file must have at least one file linked to it. As shown in **Figure 27**, this is done by clicking on the ‘Add Relation’ button. Many files can be assigned to the same ‘SEARCH’ file.



Relationships between files

Specify the files used for producing the results

Result files Click on "Relation" button to add related files

File Name	PATH / URL	Type	#Relations	Add Relation
sample_1_search_1_pep.xml	/Users/attilacsordas/Desktop/pxsubmissiontestsets/pxpartialtest/sample_1_search...	SEARCH	0	Relation
sample_2_search_2_pep.xml	/Users/attilacsordas/Desktop/pxsubmissiontestsets/pxpartialtest/sample_2_search...	SEARCH	0	Relation

Select related files from below

+	File Name	PATH / URL	Type
□	database.fasta	/Users/attilacsordas/Desktop/pxsubmissiontestsets/pxpartialtest/database.fasta	OTHER
☑	sample_1_replicate_1.mgf	/Users/attilacsordas/Desktop/pxsubmissiontestsets/pxpartialtest/sample_1_replic...	PEAK
☑	sample_1_replicate_1.RAW	/Users/attilacsordas/Desktop/pxsubmissiontestsets/pxpartialtest/sample_1_replic...	RAW
☑	sample_1_replicate_2.mgf	/Users/attilacsordas/Desktop/pxsubmissiontestsets/pxpartialtest/sample_1_replic...	PEAK
☑	sample_1_replicate_2.RAW	/Users/attilacsordas/Desktop/pxsubmissiontestsets/pxpartialtest/sample_1_replic...	RAW
□	sample_2_replicate_1.mgf	/Users/attilacsordas/Desktop/pxsubmissiontestsets/pxpartialtest/sample_2_replic...	PEAK
□	sample_2_replicate_1.RAW	/Users/attilacsordas/Desktop/pxsubmissiontestsets/pxpartialtest/sample_2_replic...	RAW
□	sample_2_replicate_2.mgf	/Users/attilacsordas/Desktop/pxsubmissiontestsets/pxpartialtest/sample_2_replic...	PEAK
□	sample_2_replicate_2.RAW	/Users/attilacsordas/Desktop/pxsubmissiontestsets/pxpartialtest/sample_2_replic...	RAW

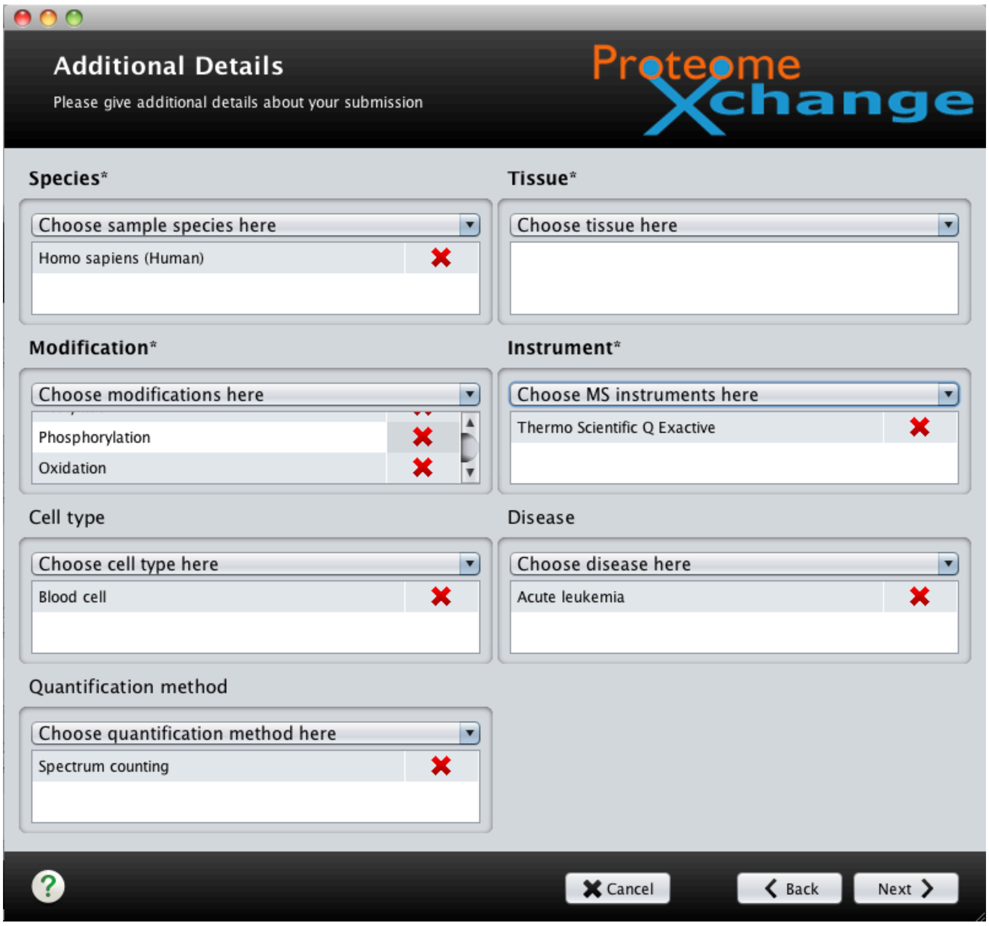
Related files

File Name	Type

Figure 27: Assigning mappings between different file types on the 'Relationships between files' screen in the PX Submission Tool

Step 8: Provide additional experimental details

In order to increase the reusability of the dataset, some additional experimental details are needed such as species, tissue, cell type, disease, MS instrument and a list of the post-translational modifications (PTMs) present in the dataset.



Additional Details

Please give additional details about your submission

Species*: Choose sample species here — Homo sapiens (Human) ✖

Tissue*: Choose tissue here

Modification*: Choose modifications here — Phosphorylation ✖, Oxidation ✖

Instrument*: Choose MS instruments here — Thermo Scientific Q Exactive ✖

Cell type: Choose cell type here — Blood cell ✖

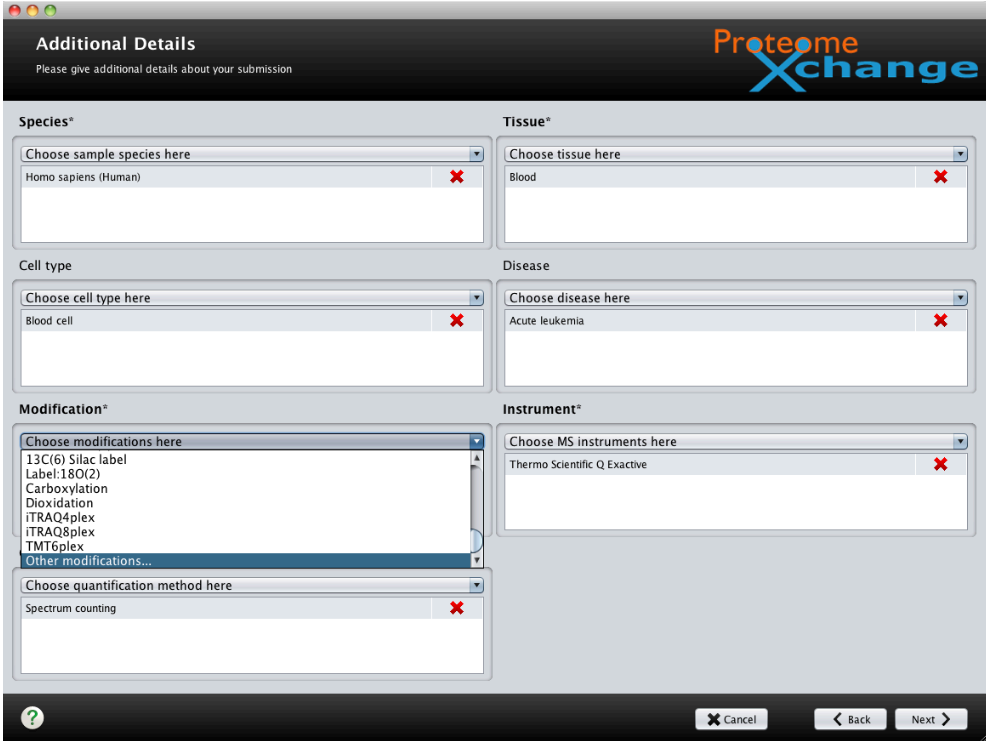
Disease: Choose disease here — Acute leukemia ✖

Quantification method: Choose quantification method here — Spectrum counting ✖

Figure 28: 'Additional details' screen in the PX Submission Tool for Partial Submissions

For each type of required experimental details, the submission tool provides a short list of commonly used values (**Figure 28**). If this list doesn't contain your experimental specific details, you should choose the 'Other' option, as shown in **Figure 29** for modifications. If that option is selected, a pop-up window will appear providing access to the 'Ontology Lookup Service' (OLS, http://www.ebi.ac.uk/ontology-lookup/).

There, you can search for a specific term from a controlled vocabulary or ontology, please see **Figure 30**.



Additional Details

Please give additional details about your submission

Species*: Choose sample species here — Homo sapiens (Human)

Tissue*: Choose tissue here — Blood

Cell type: Choose cell type here — Blood cell

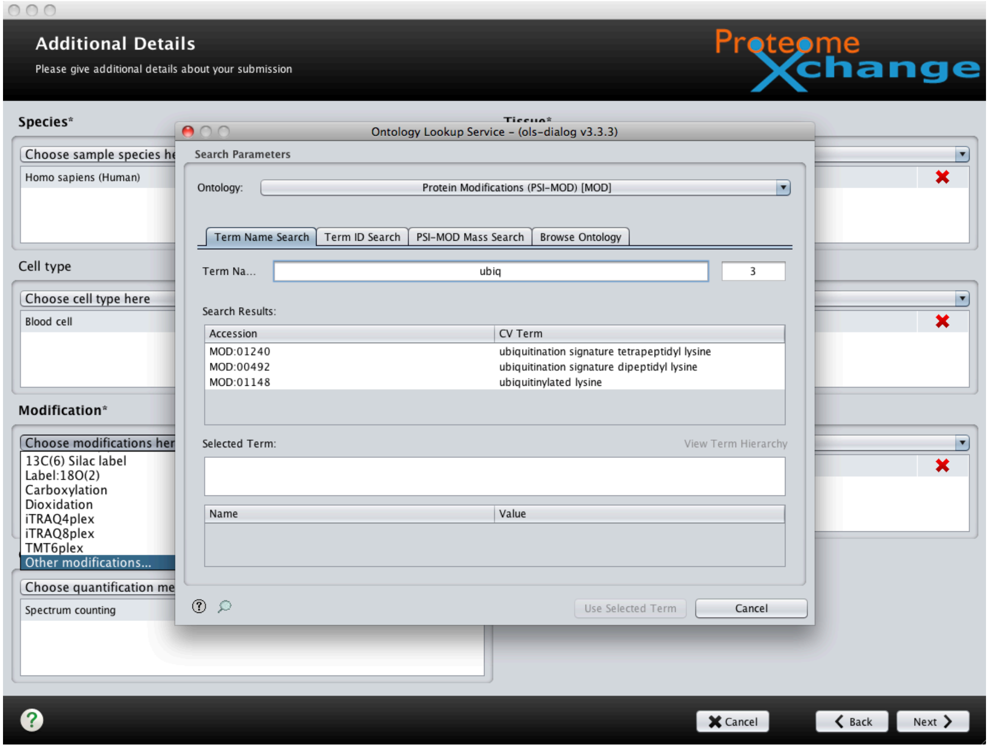
Disease: Choose disease here — Acute leukemia

Modification*: Choose modifications here — 13C(6) Silac label, Label:18O(2), Carboxylation, Dioxidation, iTRAQ4plex, iTRAQ8plex, TMT6plex, Other modifications...

Choose quantification method here — Spectrum counting

Instrument*: Choose MS instruments here — Thermo Scientific Q Exactive

Figure 29: Screenshot of the PX Submission Tool showing how to choose 'other' modifications



Ontology Lookup Service – (ols-dialog v3.3.3)

Search Parameters

Ontology: Protein Modifications (PSI-MOD) [MOD]

Term Na...: ubiq — 3

Search Results:

Accession	CV Term
MOD:01240	ubiquitination signature tetrapeptidyl lysine
MOD:00492	ubiquitination signature dipeptidyl lysine
MOD:01148	ubiquitinylated lysine

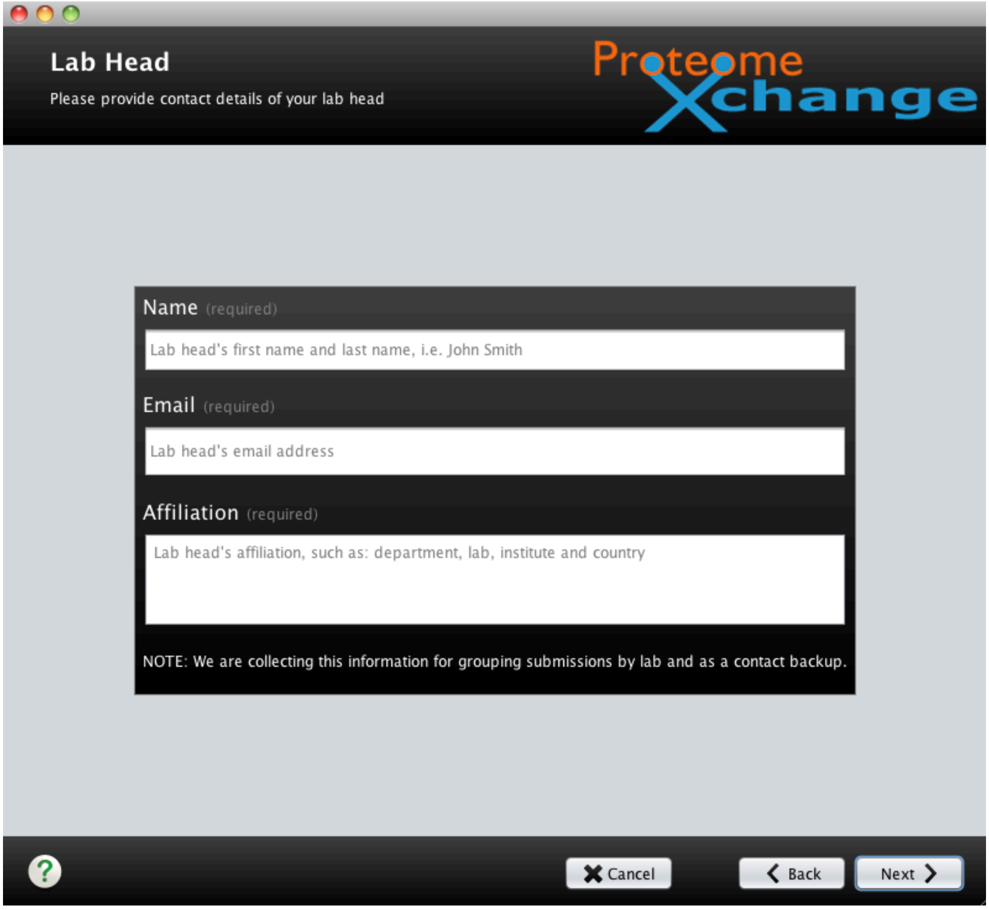
Selected Term:

Name	Value

Figure 30: Screenshot with the 'Ontology Lookup Service' (OLS) pop-up window in the PX Submission Tool

Step 9: Add Lab Head

Please provide contact details for the Lab Head/Principal Investigator of your study (**Figure 31**).



Name (required)

Lab head's first name and last name, i.e. John Smith

Email (required)

Lab head's email address

Affiliation (required)

Lab head's affiliation, such as: department, lab, institute and country

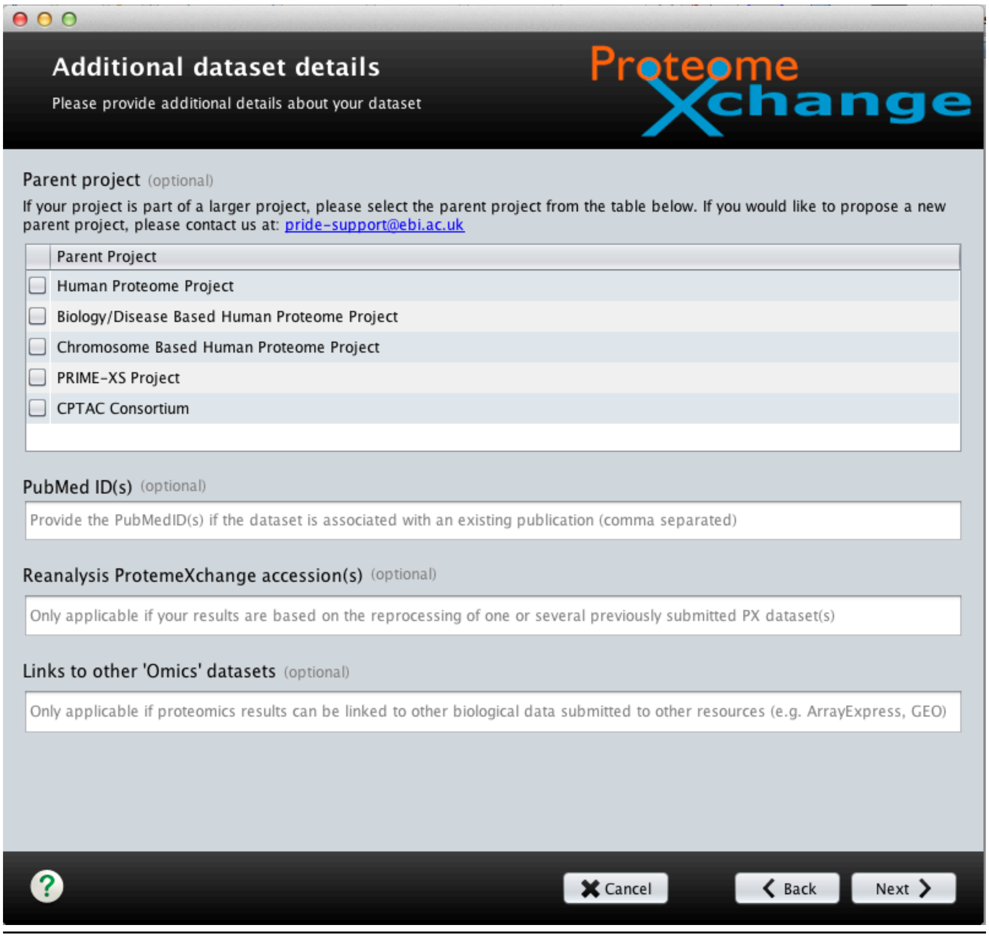
NOTE: We are collecting this information for grouping submissions by lab and as a contact backup.

Figure 31: Providing contact details for the Lab Head of your project

Step 10: Optional metadata annotation

In this panel, it is recommended to provide additional metadata in four cases:

- Your dataset is part of a bigger project/effort (for instance the Human Proteome Project or the EU project 'PRIME-XS'). It is a way to tag your dataset to enable grouping this way.
- There is already a PubMed ID associated with it (the data has been already published).
- Your dataset represents a reanalysis of an earlier public PX dataset.
- There are other "omics" datasets (for instance transcriptomics, metabolomics data present in other repositories) that can be associated with it. In this case, you need to provide the accession number of the dataset in the corresponding repository.



Additional dataset details

Please provide additional details about your dataset

Parent project (optional)

If your project is part of a larger project, please select the parent project from the table below. If you would like to propose a new parent project, please contact us at: pride-support@ebi.ac.uk

	Parent Project
□	Human Proteome Project
□	Biology/Disease Based Human Proteome Project
□	Chromosome Based Human Proteome Project
□	PRIME-XS Project
□	CPTAC Consortium

PubMed ID(s) (optional)

Provide the PubMedID(s) if the dataset is associated with an existing publication (comma separated)

Reanalysis ProtemeXchange accession(s) (optional)

Only applicable if your results are based on the reprocessing of one or several previously submitted PX dataset(s)

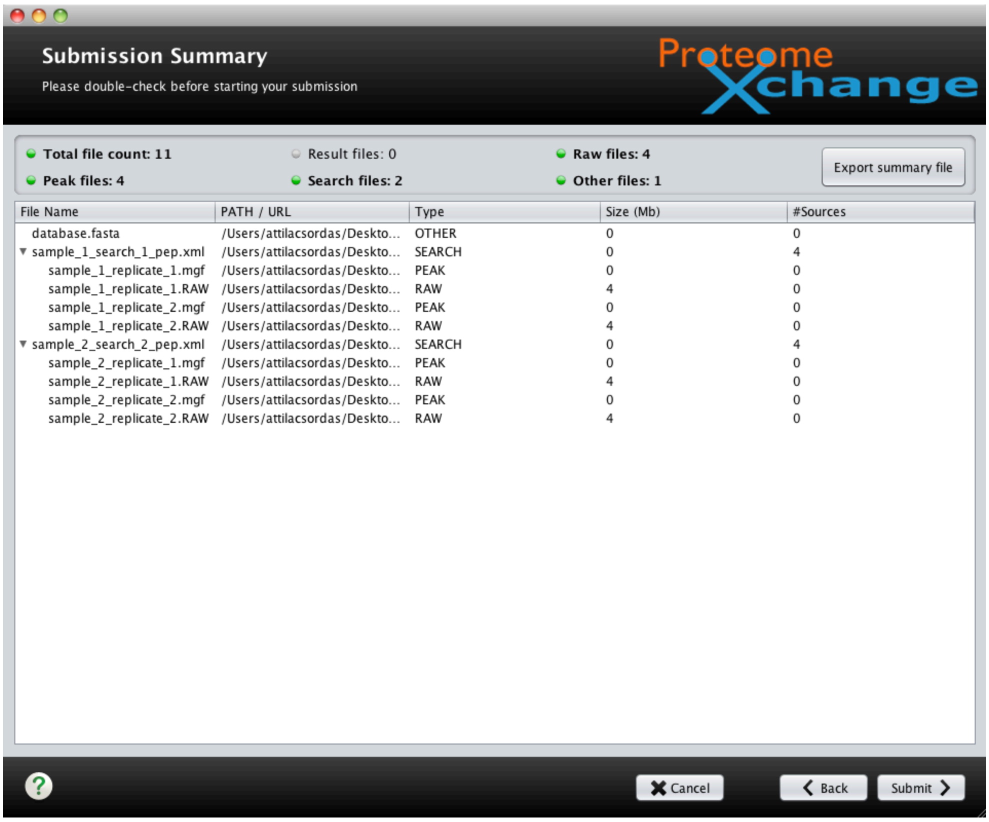
Links to other 'Omics' datasets (optional)

Only applicable if proteomics results can be linked to other biological data submitted to other resources (e.g. ArrayExpress, GEO)

Figure 32: Providing additional, applicable metadata

Step 11: Check before submission

This is the last step before the file upload actually starts. You should double-check that all the necessary files are included in the submission summary before continuing to the upload step, please see **Figure 33**.



Submission Summary

Please double-check before starting your submission

● Total file count: 11 ○ Result files: 0 ● Raw files: 4

● Peak files: 4 ● Search files: 2 ● Other files: 1

File Name	PATH / URL	Type	Size (Mb)	#Sources
database.fasta	/Users/attilacsordas/Deskto...	OTHER	0	0
▼ sample_1_search_1_pep.xml	/Users/attilacsordas/Deskto...	SEARCH	0	4
sample_1_replicate_1.mgf	/Users/attilacsordas/Deskto...	PEAK	0	0
sample_1_replicate_1.RAW	/Users/attilacsordas/Deskto...	RAW	4	0
sample_1_replicate_2.mgf	/Users/attilacsordas/Deskto...	PEAK	0	0
sample_1_replicate_2.RAW	/Users/attilacsordas/Deskto...	RAW	4	0
▼ sample_2_search_2_pep.xml	/Users/attilacsordas/Deskto...	SEARCH	0	4
sample_2_replicate_1.mgf	/Users/attilacsordas/Deskto...	PEAK	0	0
sample_2_replicate_1.RAW	/Users/attilacsordas/Deskto...	RAW	4	0
sample_2_replicate_2.mgf	/Users/attilacsordas/Deskto...	PEAK	0	0
sample_2_replicate_2.RAW	/Users/attilacsordas/Deskto...	RAW	4	0

Figure 33: 'Submission Summary' screen for a 'Partial' submission in the PX Submission Tool

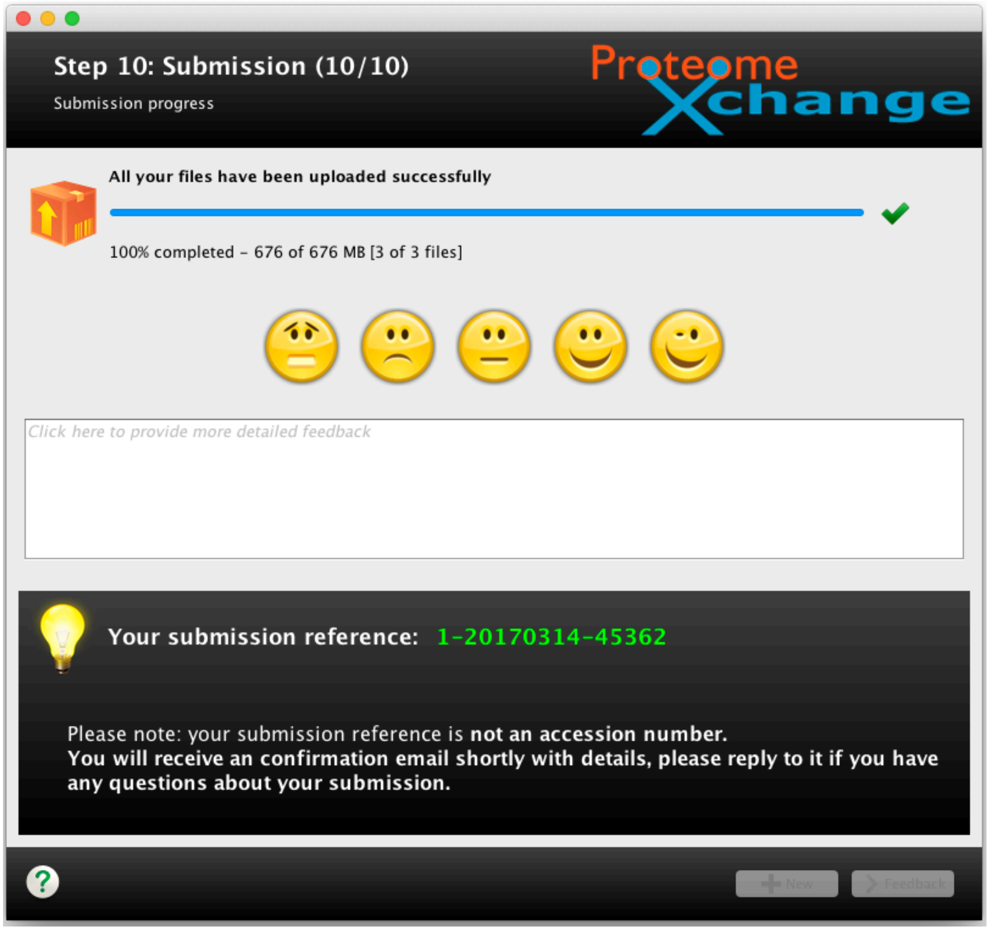
Step 12: File Submission

This is the actual step when all your files are uploaded to PRIDE and ProteomeXchange. Once the upload is finished, an email will be sent to you to confirm that all your files have been uploaded successfully and that are waiting to be validated.

If for any reason the tool crashes at this point, the PX Submission Tool can be restarted and the file upload will restart in the same point before it crashed.

Please follow the information provided in (Section 11 of Section ‘4. How to make complete submissions?’) if you need to switch from the default Aspera to the ftp upload option.

You will be also issued with a temporary submission reference, to help us to quickly identify and track your submission should you have any questions. This is neither the final PX accession number, nor a temporary one. As such it should not be used in the manuscript.



Step 10: Submission (10/10)

Submission progress

All your files have been uploaded successfully

100% completed – 676 of 676 MB [3 of 3 files]

Click here to provide more detailed feedback

Your submission reference: 1-20170314-45362

Please note: your submission reference is **not an accession number.**
You will receive an confirmation email shortly with details, please reply to it if you have any questions about your submission.

Figure 34: 'Submission' screen of the PX Submission Tool showing that a submission has been completed

For particular examples of partial submissions (e.g. software like MaxQuant or ProteinPilot), see Appendix V. As for complete submissions, there is a feedback section at the end of the process, once the data has been uploaded.

How to make bulk submissions?

Two steps are required: 'Creation of the PX submission summary file', and 'Submission using the PX submission tool'.

Creation of the PX Submission Summary File

A submission summary file (submission.px) contains two types of information needed for any PX submission:

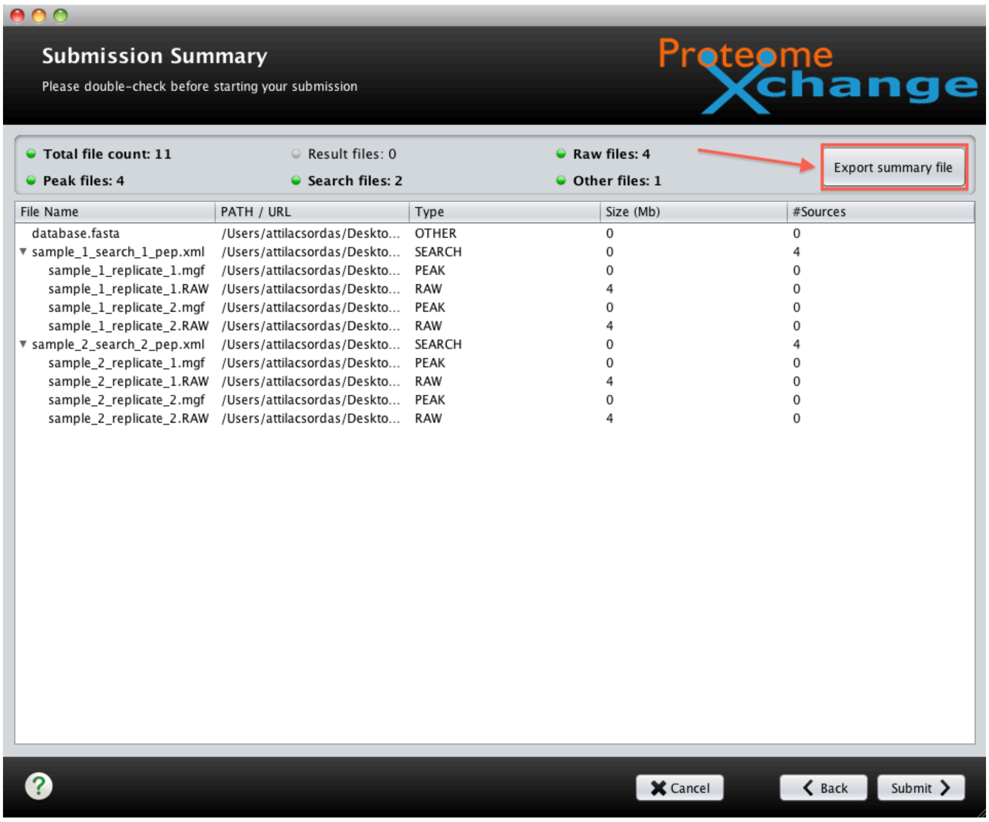
- **Metadata**: general experimental metadata like experiment description, sample taxonomy information, instruments and modifications used, experimental tags, contact information, etc.
- **Mapping between the uploaded files**: for instance, between the 'RAW' files and the corresponding 'RESULT' or search engine output files ('SEARCH').

There are two ways to create the file:

A) Generating the file independently from the PX submission tool. Some scripting work is needed. Details about the tab delimited PX submission format can be found here.

B) Using the PX Submission Tool: This is the recommended option if there are not many files, so the metadata and the file mappings can be provided with the tool but the actual data upload can be performed later.

Instead the submitters can upload their files in an alternative way (see Section 6.3) if they choose to do so. For these cases the PX Submission Tool provides an 'Export Summary' functionality. You can use this functionality when reaching the 'Submission Summary' screen, at the end of the submission process, please see **Figure 35**. The summary file can then be stored locally (usually with the extension .px).



Submission Summary

Please double-check before starting your submission

Total file count: 11 | Result files: 0 | Raw files: 4

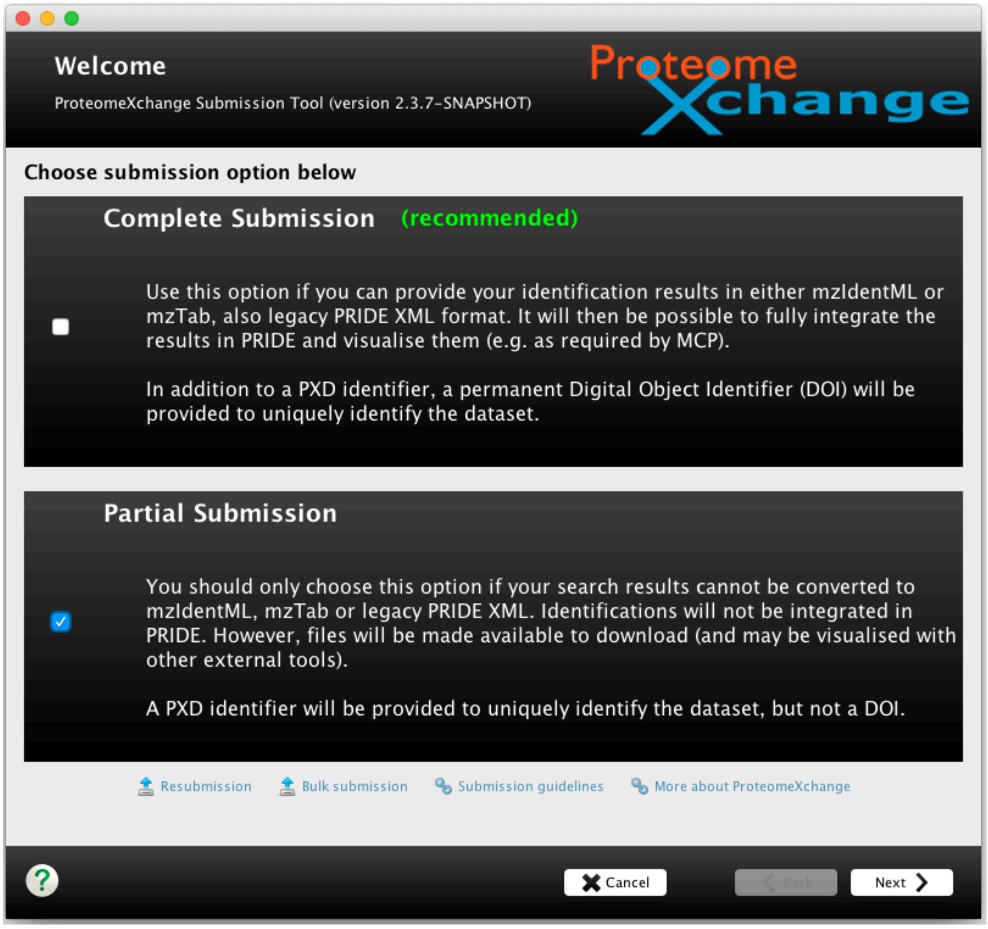
Peak files: 4 | Search files: 2 | Other files: 1

File Name	PATH / URL	Type	Size (Mb)	#Sources
database.fasta	/Users/attilacsordas/Deskto...	OTHER	0	0
▼ sample_1_search_1_pep.xml	/Users/attilacsordas/Deskto...	SEARCH	0	4
sample_1_replicate_1.mgf	/Users/attilacsordas/Deskto...	PEAK	0	0
sample_1_replicate_1.RAW	/Users/attilacsordas/Deskto...	RAW	4	0
sample_1_replicate_2.mgf	/Users/attilacsordas/Deskto...	PEAK	0	0
sample_1_replicate_2.RAW	/Users/attilacsordas/Deskto...	RAW	4	0
▼ sample_2_search_2_pep.xml	/Users/attilacsordas/Deskto...	SEARCH	0	4
sample_2_replicate_1.mgf	/Users/attilacsordas/Deskto...	PEAK	0	0
sample_2_replicate_1.RAW	/Users/attilacsordas/Deskto...	RAW	4	0
sample_2_replicate_2.mgf	/Users/attilacsordas/Deskto...	PEAK	0	0
sample_2_replicate_2.RAW	/Users/attilacsordas/Deskto...	RAW	4	0

Figure 35: 'Submission Summary' screen in the PX Submission Tool, highlighting how to export and store locally the PX summary file

Submission using the PX Submission tool

You have already created a PX submission summary file for your dataset by scripting. In this case, you can use the PX Submission Tool to perform the submission. In the 'Welcome' screen of the PX submission tool, please select the option 'Bulk submission' highlighted in **Figure 36**, and proceed as indicated by the tool. You will need to load the created PX summary file.



Welcome

ProteomeXchange Submission Tool (version 2.3.7-SNAPSHOT)

Choose submission option below

□ **Complete Submission** (recommended)

Use this option if you can provide your identification results in either mzIdentML or mzTab, also legacy PRIDE XML format. It will then be possible to fully integrate the results in PRIDE and visualise them (e.g. as required by MCP).

In addition to a PXD identifier, a permanent Digital Object Identifier (DOI) will be provided to uniquely identify the dataset.

☑ **Partial Submission**

You should only choose this option if your search results cannot be converted to mzIdentML, mzTab or legacy PRIDE XML. Identifications will not be integrated in PRIDE. However, files will be made available to download (and may be visualised with other external tools).

A PXD identifier will be provided to uniquely identify the dataset, but not a DOI.

Figure 36: 'Welcome screen' of the PX Submission Tool highlighting the 'Bulk submission' mode

Command line Aspera upload option

As mentioned earlier the PX Submission Tool is using by default the fast Aspera upload transfer protocol with which terabytes can potentially be transferred within a day. Nevertheless, it is also possible to use the Aspera protocol *via* a command line upload option. This option is available for submitters with bioinformatics support who prefer not to use the PX Submission Tool, due to the manual work involved (e.g. if the submission contains a large number of files). Some command line skills are needed in order to use this option. Please follow the steps below.

Requirements: Please download the Aspera Connect Web Browser Plug-in. Although you download a Browser Plug-in you will be using the 'ascp' command line transfer program distributed with it.

- Operating System: Windows XP / 2003 / Vista / 2008 / 7 / 8, Mac OS Intel 10.5 / 10.6 / 10.7 / 10.8 (You don't have to register in order to download the Browser Plug-in and the download is free of charge.)

- Check the command line transfer usage for more configuration details. This is the location of the 'ascp' program in the file system:
- Mac: on the desktop go cd /Applications/Aspera\ Connect.app/Contents/Resources/ (there you'll see the command line utilities where you're going to use 'ascp')
- Windows: the downloaded files are a bit hidden. For instance, in Windows 7 the ascp.exe is located in the users home directory at: AppData\Local\Programs\Aspera\Aspera Connect\bin\ascp.exe

How to upload a directory of files

Step 1. Ask PRIDE support (at pride-support@ebi.ac.uk) for a target directory and a password.

The PRIDE curators will specify a target directory for you, see <name-of-target-dir-specified-by-PRIDE> in the following commands, and you will be provided with this information.

Step 2. The upload command and process.
When preparing your dataset please be sure to unambiguously assign a unique file name to all of your files. Please also upload the submission summary file into the same folder.

- Mac: ./ascp -QT -l500m --file-manifest=text -k 2 -o Overwrite=diff <path-to-folder-to-be-uploaded> pride-drop-006@ah01.ebi.ac.uk:<name-of-target-dir-specified-by-PRIDE>

- Windows: ascp.exe -QT -l500m --file-manifest=text -k 2 -o Overwrite=diff <path-to-folder-to-be-uploaded> pride-drop-006@ah01.ebi.ac.uk:<name-of-target-dir-specified-by-PRIDE>

The <path-to-folder-to-be-uploaded> should not have any blank spaces in it.

Please set the '--file-manifest=text -k 2' flags as well.

This will generate an Aspera progress file on your side that will contain a report on the files that have been uploaded, also you can interrupt the process and then it will only upload the ones that were not there so no more overwriting files. It will also skip the ones that are already in the target directory.

If -l500m ~ 500 Mb/s is unstable and leads to timeouts then we suggest to go back to -l250m as the maximum transfer rate, even that is fast enough to transfer theoretically 2 TBs within a day.

Once upload has been finished you will be prompted to enter the password provided earlier.

Step 3. Notify the PRIDE Team

E-mail pride-support@ebi.ac.uk in case your upload has been successfully finished.

What happens after the submitter has uploaded all the data?

Once your dataset has been uploaded into the EBI, the PRIDE internal submission pipeline will validate your files. The results of the validation will be checked by a curator and, if no problems are found, the dataset will be submitted to PRIDE and the relevant information will be stored. The process varies for 'complete' and 'partial' submissions. As a result, you will be issued with a ProteomeXchange accession number.

In addition, a DOI will also be assigned if a 'complete' submission was performed. PRIDE assay accession numbers will also be provided for mzTab and mzIdentML result files in case of 'complete' submissions. A confirmation e-mail will be sent to you with all the relevant details once your submission is complete, including a username and password for potential journal reviewers and editors to be able to access your data privately. Please note all submissions are private by default.

Accessing Private Data

Submitted datasets are private by default, which means you need to be logged-in to view your data. We will however also create a PX reviewer account and a password for your dataset, which you should include in your manuscript. Again, the PX reviewer account will give you access to all of the files belonging to your submission. For that you can use the new PRIDE Archive web site or the PRIDE Inspector tool.

PRIDE Archive web page

PRIDE Archive web site is available at http://www.ebi.ac.uk/pride/archive. Registered submitters can use their personal accounts or the reviewer accounts to access and download the individual PX datasets. For every submission there is a separate reviewer account generated.

Please navigate first to the login page available at http://www.ebi.ac.uk/pride/archive/login (see **Figure 37**):

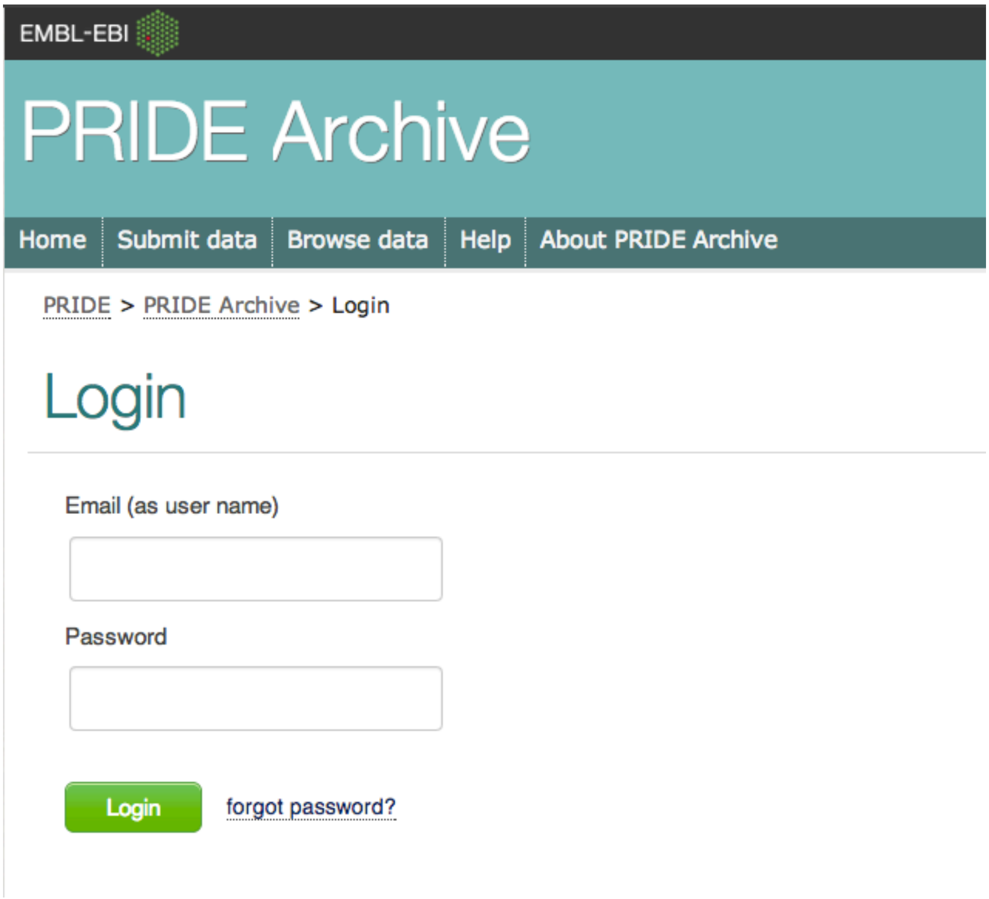


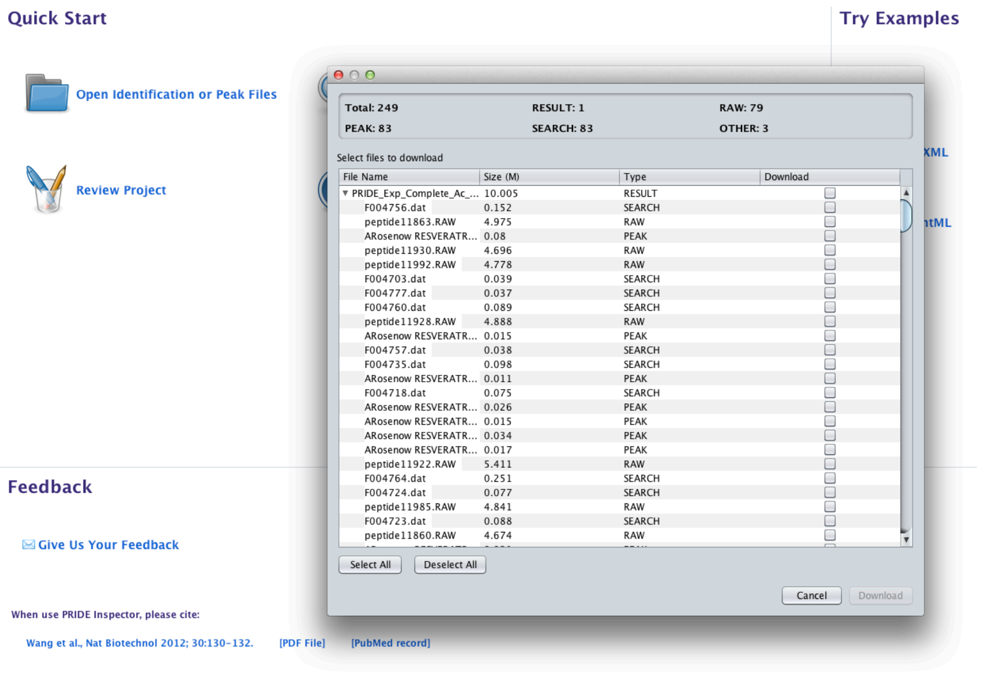
Figure 37: PRIDE Archive 'Login' page

Once logged in with your registered User (the e-mail account you used to register in PRIDE) or an issued Reviewer Account you are going to see the private dataset/s listed.

PRIDE Inspector

PRIDE Inspector is a stand-alone tool developed by the PRIDE team. It can be downloaded here:
https://github.com/PRIDE-Toolsuite/pride-inspector/releases
for further information please see Appendix 2.

In order to access private datasets, first open PRIDE Inspector by clicking on the pride-inspector-<version-number>.jar file in the tool's working directory and go to Review Project-> Reviewer account details. You can open the mzIdentML (plus spectra files) or mzTab result files with PRIDE Inspector or just download all the files that you wish to investigate.



Total: 249 RESULT: 1 RAW: 79
PEAK: 83 SEARCH: 83 OTHER: 3

Select files to download

File Name	Size (M)	Type	Download
▼ PRIDE_Exp_Complete_Ac_...	10.005	RESULT	□
F004756.dat	0.152	SEARCH	□
peptide11863.RAW	4.975	RAW	□
ARosenow RESVERATR...	0.08	PEAK	□
peptide11930.RAW	4.696	RAW	□
peptide11992.RAW	4.778	RAW	□
F004703.dat	0.039	SEARCH	□
F004777.dat	0.037	SEARCH	□
F004760.dat	0.089	SEARCH	□
peptide11928.RAW	4.888	RAW	□
ARosenow RESVERATR...	0.015	PEAK	□
F004757.dat	0.038	SEARCH	□
F004735.dat	0.098	SEARCH	□
ARosenow RESVERATR...	0.011	PEAK	□
F004718.dat	0.075	SEARCH	□
ARosenow RESVERATR...	0.026	PEAK	□
ARosenow RESVERATR...	0.015	PEAK	□
ARosenow RESVERATR...	0.034	PEAK	□
ARosenow RESVERATR...	0.017	PEAK	□
peptide11922.RAW	5.411	RAW	□
F004764.dat	0.251	SEARCH	□
F004724.dat	0.077	SEARCH	□
peptide11985.RAW	4.841	RAW	□
F004723.dat	0.088	SEARCH	□
peptide11860.RAW	4.674	RAW	□

When use PRIDE Inspector, please cite:

Wang et al., Nat Biotechnol 2012; 30:130-132. [PDF File] [PubMed record]

Figure 38: Downloading data with the reviewer account using PRIDE Inspector private download option

Post-submission steps

How to do a resubmission of a dataset?

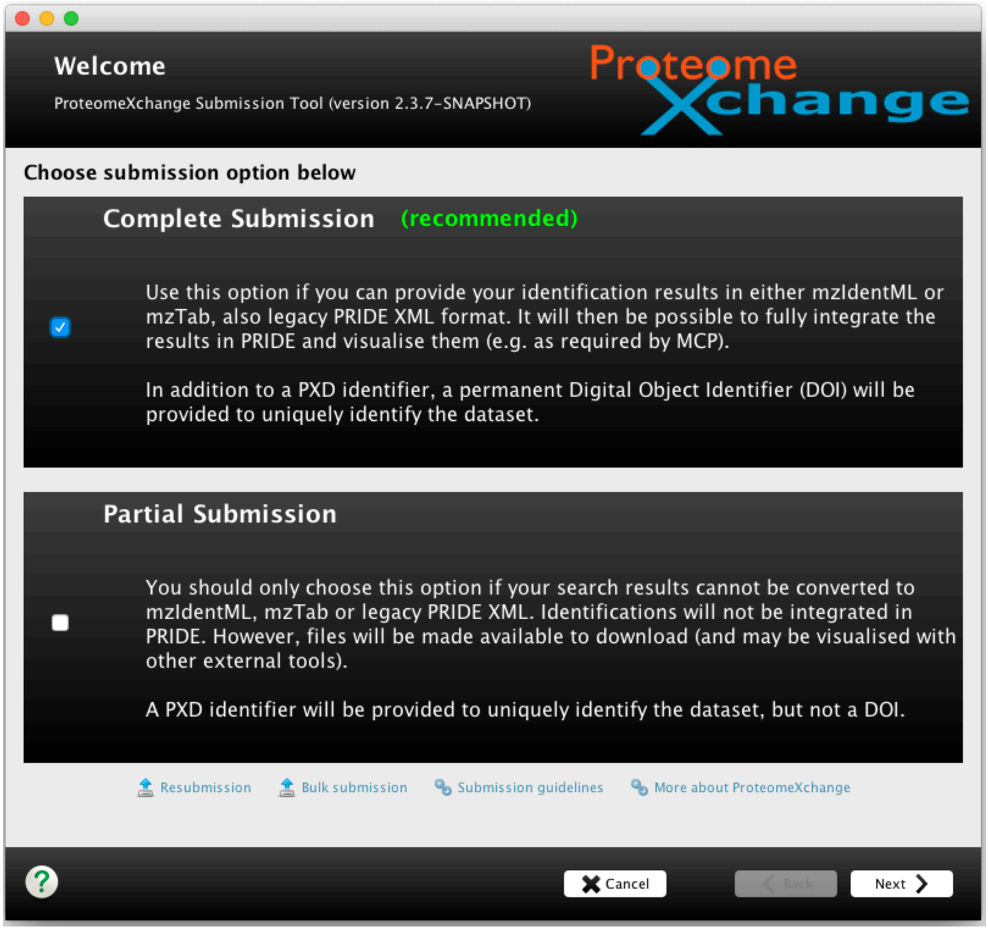
While the data is still private (during the manuscript review process) it is possible to resubmit the whole dataset by keeping the previously issued PX identifier. Data resubmissions consisting in a subset of the previous submission are not currently supported.

Resubmission with the PX Submission Tool

Install and launch the PX Submission Tool as explained before (available at http://www.proteomexchange.org/submission).

Step 1: Click resubmission on the 'Welcome' page

The option is highlighted in **Figure 38**.



Welcome

ProteomeXchange Submission Tool (version 2.3.7-SNAPSHOT)

Choose submission option below

☑ Complete Submission (recommended)

Use this option if you can provide your identification results in either mzIdentML or mzTab, also legacy PRIDE XML format. It will then be possible to fully integrate the results in PRIDE and visualise them (e.g. as required by MCP).

In addition to a PXD identifier, a permanent Digital Object Identifier (DOI) will be provided to uniquely identify the dataset.

☐ Partial Submission

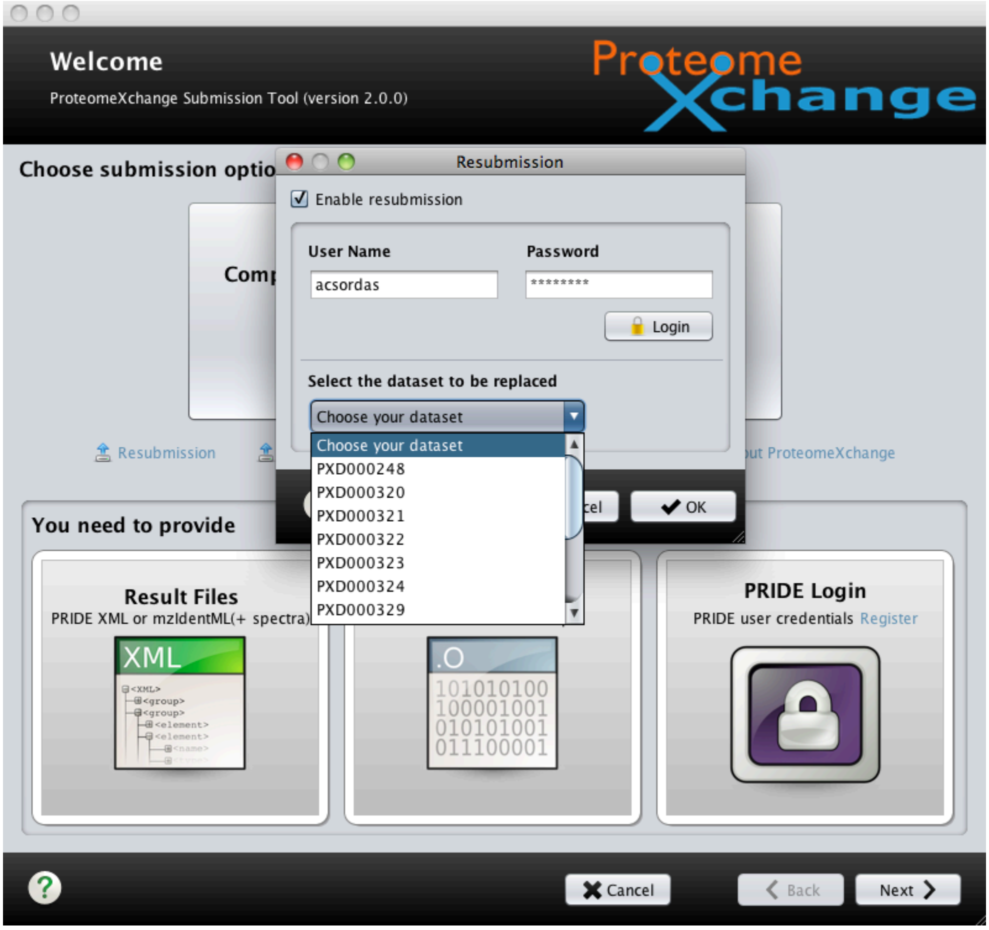
You should only choose this option if your search results cannot be converted to mzIdentML, mzTab or legacy PRIDE XML. Identifications will not be integrated in PRIDE. However, files will be made available to download (and may be visualised with other external tools).

A PXD identifier will be provided to uniquely identify the dataset, but not a DOI.

Figure 39: 'Welcome screen' of the PX Submission Tool highlighting the resubmission mode

Step 2: Enable resubmission and provide resubmission details

In the pop-up dialog box please provide your PRIDE login details and select the PX identifier of the dataset you want to resubmit, please see **Figure 40**.



Resubmission

☑ Enable resubmission

User Name: acsordas

Password: ********

Select the dataset to be replaced: Choose your dataset

- Choose your dataset
- PXD000248
- PXD000320
- PXD000321
- PXD000322
- PXD000323
- PXD000324
- PXD000329

Figure 40: Screenshot showing how to select the dataset that needs to be resubmitted

After these two steps the resubmission follows the same steps described for a regular submission.

Resubmission *via* Aspera command line option

If you have done a bulk submission using the command line Aspera fast transfer option resubmission of the whole dataset is possible *via* the command line again. You will upload the whole modified dataset with the submission summary file into the same target directory again. You can use the PX Submission Tool to export the summary file as explained before but in that case, you need to use the “Resubmission” option of the tool and specify the PX Identifier that will be used for resubmission, please see the 9.1.1 section above. This way the summary file will contain the required resubmission information.

In case you are generating the summary file using scripting (see section 6.1) the following line need be added to the Metadata section of the submission.px file to indicate that the dataset is a resubmission of an earlier submitted one:

```
MTD	resubmission_px PXD000444
```

Referencing the dataset in the paper

By default, we recommend to add the following formula to your manuscript (typically in "Material and Methods” or just before/in the “Acknowledgements”):

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository [1] with the dataset identifier <PXD000xxx>."

[1] and also for general PRIDE reference, please use: Vizcaino JA, Cote RG, Csordas A, Dianes JA, Fabregat A, Foster JM, Griss J, Alpi E, Birim M, Contell J, O'Kelly G, Schoenegger A, Ovelleiro D, Perez-Riverol Y, Reisinger F, Rios D, Wang R, Hermjakob H. The Proteomics Identifications (PRIDE) database and associated tools: status in 2013. Nucleic Acids Res. 2013 Jan 1;41(D1):D1063-9. doi: 10.1093/nar/gks1262. Epub 2012 Nov 29. PubMed PMID:23203882.

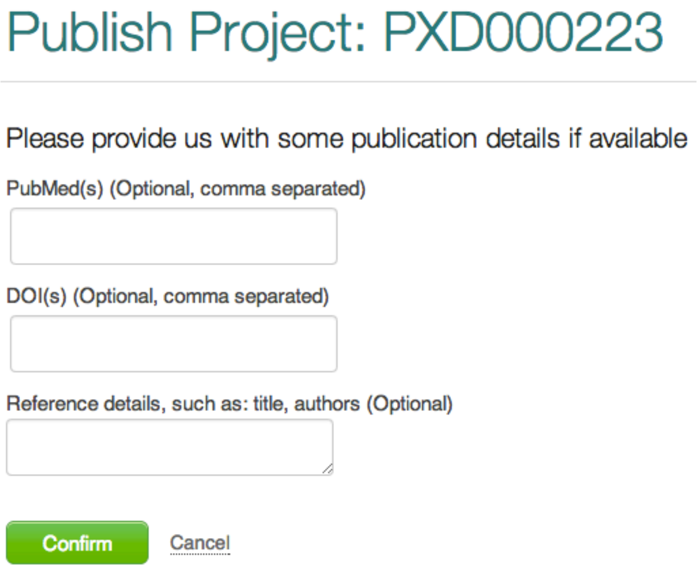
Additionally, and if it is feasible we'd like to ask our submitters to reference the dataset in a much-abridged form in the abstract itself, like this: "The data have been deposited to the ProteomeXchange with identifier <PXD000xxx>."

See for example this Chromosome-Centric Human Proteome Project dataset and paper: http://www.ncbi.nlm.nih.gov/pubmed/?term=23312004, and other examples on PubMed. In our experience, a PX Identifier in the abstract makes the dataset much more visible and accessible.

Public release of the dataset

By default, your data will be made publicly available after your manuscript has been accepted, or when we have your instructions to do so. While we may also receive acceptance notifications from some journals, we would like to ask all submitters to kindly notify us separately. Otherwise, it can happen that we don’t know that your manuscript is already published. You can notify us two ways:

A) Via the new PRIDE Archive web site (http://www.ebi.ac.uk/pride/archive). Once you have logged in with your user account at http://www.ebi.ac.uk/pride/archive/login you can click the green “Publish” buttons located next to your unpublished datasets. Here you can provide details for your dataset and submit a web form, please see **Figure 41**.



Publish Project: PXD000223

Please provide us with some publication details if available

PubMed(s) (Optional, comma separated)

DOI(s) (Optional, comma separated)

Reference details, such as: title, authors (Optional)

Figure 41: Providing publication details using the PRIDE Archive web

B) Contacting pride-support@ebi.ac.uk.

Upon making the project public, a project page will be released over at ProteomeCentral (http://proteomecentral.proteomexchange.org) and from a particular dataset page an FTP location will be available.

Appendix I: Definitions

Proteomics data come in a variety of forms, which are defined here:

Mass spectrometer output files: the data and metadata generated by mass spectrometers, usually one file per run (although some instruments put multiple runs per file). The data may be the original profile mode scans or may already have had some basic processing like centroiding applied. They may be:

- raw data (as described below).
- peak list spectra in a standardized format such as mzML, mzXML or mzData (see below), but they cannot be 'processed peak lists' (see below).

However, it is important that all of the scans that were generated are included with applicable metadata.

Raw data: the binary, vendor-specific output files directly created by the instrument software. These files are typically large (several gigabytes) and require specialized software in order to be read.

Standardized MS data formats: There are currently three widely known mass spectrometry data formats in Proteomics: mzXML (developed at the Institute of Systems Biology (ISB), Seattle, USA), mzData (now made obsolete, originally developed by the HUPO Proteomics Standards Initiative (PSI)), and the successor to both of the above: mzML (currently v1.1, jointly developed by the ISB and PSI, http://www.psidev.info/mzml). These data formats can be used to represent processed peak lists, as well as raw data. In addition to the mass spectra, they contain detailed metadata that provide context to the measurements.

Processed peak lists: Heavily processed form of mass spectrometry data, usually derived from the raw data files through various (semi-)automatic steps, e.g.: centroiding, deisotoping, and charge deconvolution. These files are formatted in plain text, with typical formats like dta, pkl, ms2 or mgf. They usually contain only a subset of only the MS2 scans (MS1 scans are excluded), and are missing significant amounts of metadata that were present in the source format.

Protein/peptide identifications: Proteomics mass spectra can be matched to peptides or proteins, resulting in identifications for those spectra. Typically a spectrum is considered identified if the score attributed to a peptide or protein match qualifies against an *a priori* or *a posteriori* defined threshold. In the case of fragmentation spectra, the initial identification will consist of a peptide sequence; subsequent steps will derive a list of proteins from the identified peptides. The protein assembly step can be a discernible process with its own input and output files, or it can be implicit in the overall identification software. This information can be represented by a variety of data formats called search engine output files (see below).

Protein/peptide quantification: Protein/peptide expression values can also be obtained from a MS-based proteomics experiment. There is a high diversity of approaches that result in the existence of very heterogeneous software and data analysis pipelines. Some search engines are able to perform both identification and quantification, and produce 'search engine output files' containing both types of data. However, there is software that only performs the quantification part of the analysis and the generated data is represented in quantification software output files (see below).

Search engine output files: They contain the data and metadata generated by the software (usually called search engines) used for performing the identification and quantification of peptides and proteins. Each search engine has its own specific output file. The formats are typically formatted in either plain text or XML, with typical formats like mascot .dat, OMSSA xml, etc.

In addition to each specific format, a data standard format called mzIdentML (currently v1.1, https://github.com/HUPO-PSI/mzIdentML) has been developed by the PSI to represent this kind of information. Some search engine output files can represent as well quantification results, but this is not the case of mzIdentML. A second standard data format called mzTab (https://github.com/HUPO-PSI/mzTab), currently under development, can represent both identification and basic quantification results.

Supported identification results: This definition includes all protein/peptide identification processed data that can be fully represented by the receiving repository. For PRIDE database, as the PX submission point for tandem MS/MS datasets, the supported data formats are mzTab and mzIdentML version 1.1. PRIDE database can represent both mass spectra data and protein/peptide identifications. Search engine output files need to be converted/exported to mzTab or mzIdentML 1.1 to allow a full representation of the processed results in PRIDE database and the PX consortium.

Quantification software output files: the data and metadata generated by the software used for performing exclusively the quantification analysis of peptides and proteins. As mentioned before, a second data format called mzTab (https://github.com/HUPO-PSI/mzTab) can represent basic quantification results.

Gel image files: In case two-dimensional gel electrophoresis has been used as a separation method the gel image files generated.

Metadata: Whereas mass spectra present the core output of any mass spectrometer, a simple collection of spectra does not provide sufficient information for confident interpretation. Something similar happens for the peptide and protein identifications and their expression values. This lack of context can be solved by providing relevant metadata along with the spectra and/or the identifications and quantification data. Mass spectrometer, search engine, and quantification software output files (see above) typically accommodate this information.

Appendix II: Available tools to help you with the submission

Creation of mzIdentML files

mzIdentML is the HUPO-PSI standard for protein/peptide identifications coming from MS-based proteomics approaches. The stable version is 1.1, which is supported by PRIDE. It does not contain the mass spectra, which must be provided in external files referenced from the mzIdentML files (XML based files like mzML, mzXML or mzData, or peak lists like mgf, dta, ms2, or pkl).

At the time of writing this document, these are the tools that can export mzIdentML v1.1 (see an updated list at http://www.psidev.info/tools-implementing-mzidentml). Up-to-date information is also available at http://www.ebi.ac.uk/pride/help/archive/submission/mzidentml.

- Mascot (Matrix Science, http://www.matrixscience.com/). From version 2.4. See detailed instructions here.
- Scaffold (Proteome Software). Detailed instructions are available here.
- MS-GF+ (http://proteomics.ucsd.edu/Software/MSGFPlus.html#pubs).
- ProteinPilot (ABSciex). From version 5.0. Detailed instructions are available here.
- PeptideShaker (http://compomics.github.io/projects/peptide-shaker.html) (10). The output of additional open source search engines are fully supported via the PeptideShaker mzIdentML export functionality: X!Tandem, MS Amanda, OMSSA, Tide and Comet.
- ProCon: Converter for Sequest .out, ProteomeDiscoverer (Thermo) v1.2/1.3/1.4 .msf files and ProteinScape 2.1 (Bruker) database content (http://www.medizinisches-proteom-center.de/procon).
- TPP (pep.xml and prot.xml files): The idConvert tool from can be downloaded from ProteoWizard, or is bundled with the TPP directly starting with version 4.6.3.
- OpenMS.
- MIAPE MSI Extractor (http://proteored.org/miape/, ProteoRed, Madrid).
- Tools from D. Tabb's lab: Myrimatch, Pepitome (spectral library search), TagRecon and IDPicker.
- PEAKS
- Tools developed by the PRIDE team

Checking the files before submission (initial quality assessment)

Tool developed by the PRIDE team
PRIDE Inspector(https://github.com/PRIDE-Toolsuite/pride-inspector). This is an open source rich client application for inspecting MS-based proteomics data. Experiments can be examined based on different views emphasizing either metadata, identified proteins or peptides, mass spectra, or quantification results.

Apart from its powerful visualization features, the major strength of PRIDE Inspector is the possibility to perform a first assessment of data quality using e.g. the ‘Summary charts’, which are generated based on different aspects of the data. Currently, PRIDE Inspector supports fast data retrieval on standard file formats: mzML, mzIdentML (plus the corresponding peak list files) and mzTab. In addition, it also gives the user direct access to a PRIDE public database instance. As a key point, it provides journal reviewers/editors access to (privately available) experiments during the review process.

External tool developed by collaborators

- mzML validator (link to Java Web Start to be done if necessary): A Java-based tool to validate semantics and MIAPE compliance of mzML files.
- mzIdentML validator (https://github.com/HUPO-PSI/mzIdentML): A Java-based tool to validate semantics and MIAPE compliance of mzIdentML files.

File submission to PRIDE

As described before in this tutorial, the PX Submission Tool can be used (http://www.proteomexchange.org/submission). It creates the relations between the different types that can be part of a dataset and uploads the data into PRIDE *via* FTP.

Appendix III: Summary of formats supported by PRIDE for PX MS/MS submissions

a) As raw data, supported formats:

- mzML, mzXML, mzData. These files must not be heavily processed to be considered 'raw'.
- Thermo .RAW, ABSCIEX .wiff, .wiff.scan, Agilent .d/, Waters .raw/ imzML, Shimadzu .run/, Bruker .yep, Bruker .baf

All peak lists formats (mgf, dta, ms2, pkl) can be supported but they will not be considered raw data. They will be considered as 'peak list processed files' or simply 'peak'.

b) as processed identification results:

Two formats are supported: mzTab and mzIdentML.

- b.1) mzTab: support for exporting results in mzTab format will be incorporated into popular tools and equipement soon.
- b.2) mzIdentML (version 1.1): There are a number of tools that can export mzIdentML 1.1 (see Appendix 1). Formats supported this way:
- Tandem XML (using mzidLibrary, https://github.com/PGB-LIV/mzidlib)
- OMSSA .csv (using mzidLibrary, https://github.com/PGB-LIV/mzidlib).
- Mascot .dat (direct export functionality available from Mascot 2.4).
- Sequest .out files (using the ProCon tool, http://www.medizinisches-proteom-center.de/procon).
- ProteomeDiscoverer .msf files (using the ProCon tool, http://www.medizinisches-proteom-center.de/procon).
- ProteinScape 2.1 (Bruker) database content (using the ProCon tool, http://www.medizinisches-proteom-center.de/procon).
- MS-GF+ (direct export functionality available).
- Phenyx (direct export functionality available).
- Trans-Proteomic Pipeline (pep.xml files). The idConvert tool from can be downloaded from ProteoWizard, or is bundled with the TPP directly starting with version 4.6.3.
- Scaffold (direct export functionality available). From version 4.0.
- OpenMS output.
- Output files from Myrimatch, Pepitome (spectral library search), TagRecon and IDPicker.

All accompanying peak lists formats.

c) as search engine output files:

Only those data formats that cannot be converted/exported to mzTab/mzidetnML are considered to be 'unsupported formats' and can use this alternative approach (datasets type B, Datasets containing raw data and search engine output files). At present, there are no reliable converters to mzTab/mzIdentML for the following formats amongst others:

MaxQuant output files,

ProteinPilot .group files

d) as quantification results

The current version of the submission pipeline can accept quantification results expressed in mzTab format, but no further processing is being carried on them. Plus, any other quantification result file format can be submitted as accompanying ‘QUANT’ files.

e) as gel images

Gel images (in any format) tagged as ‘GEL’ can be included in the submission.

f) as sequence database files

Sequence database file (usually in FASTA format) that was used to perform the mass spectral search. Sequence database files can contain both amino acid and nucleic acid sequences. In the PX Submission Tool they should be tagged as ‘FASTA’

g) as others

Any other type of files is optional and can be supported as part of a PX submission together with the other files.

Appendix IV: Metadata requirements for MS/MS submissions

Proteomics data are substantially enriched when sufficient metadata are provided. Metadata will be as extensive as possible and will aim to comply with the MIAPE (Minimum Information About a Proteomics Experiment) guidelines. However, the presence of the metadata required in this Appendix will be enforced for any PX submission (they are mandatory in the PX Summary File format). They can be provided using the PX Submission tool.

The user will need to provide:

- Contact name and e-mail for the submission. The contact details of the data submitters need to be provided, allowing interested users to contact the original authors if desired.
- Lab Head or Principal Investigator.
- Name of the PX dataset.
- Project description: it could be considered as the abstract information of the dataset (provided as free text).
- Summary of the Sample Protocol (provided as free text).
- Summary of the Data Analysis Protocol (provided as free text).
- Experiment type. Chosen from a drop-down menu.
- Keywords: A list of keywords that describe the content and type of the experiment being submitted. Multiple entries should be comma separated.
- Sample annotation: species. At least one NEWT Controlled Vocabulary (CV) term is mandatory per dataset.
- Sample annotation: tissue. Using the BRENDA Tissue ontology (BTO), accessible at http://obo.cvs.sourceforge.net/obo/obo/ontology/anatomy/BrendaTissue.obo)
- Instrument details. Using the PSI-MS CV. It is accessible at http://psidev.cvs.sourceforge.net/viewvc/psidev/psi/psi-ms/mzML/controlledVocabulary/psi-ms.obo.
- Quantification method (if applicable).
- Protein post-transcriptional modifications (PTMs). They are reported using the PSI-MOD ontology (accessible at http://psidev.cvs.sourceforge.net/psidev/psi/mod/data/PSI-MOD.obo).

Optional information:

- Sample annotation: cell type. Use the “Cell Type” ontology.
- Sample annotation: Disease. Use the “Human Disease” ontology (DOID).

Dataset optional details:

- your dataset is part of a bigger project/effort (for instance the Human Proteome Project or ‘PRIME-XS’). It is a way to tag your dataset to enable grouping this way.

- there is already a PubMed ID associated with it (the data has been already published).
- your dataset represents a reanalysis of an earlier public PX dataset
- there are other “omics” datasets (for instance transcriptomics, metabolomics data present in other repositories) that can be associated with it. In this case, please provide the accession number of the dataset in the corresponding repository.

Appendix V: Recommended Partial Submission search engine identification results for particular software tools

There are software tools and workflows whose results can't be exported in any of the supported formats. In this case, search/peptide/protein identification results can be provided as partial submissions.

Here we describe the workflow for two popular tools: MaxQuant (PubMed ID: 19029910) and ProteinPilot^{TM} (AB SCIEX).

MaxQuant

If you are using the latest version of MaxQuant (1.3.0.5) there is a txt folder generated and by default you can just zip this text folder and upload as a 'SEARCH' file.

If this is complicated, we would recommend uploading the following particular text output files:

parameters.txt
peptides.txt
modifiedPeptides.txt
proteinGroups.txt
[Modification]Sites.txt
and your 'Experimental Design Template file' saved as a tab delimited file.

[Modification] will be replaced by the names of the protein modifications under study selected in MaxQuant (e.g. like in the "Phospho (STY)Sites.txt" files).

The 'Experimental Design Template file' can be generated by hitting the "Write template" button in the "Raw files" tab of MaxQuant once everything is ready for launching the analyses. By doing that, a pop-up window will appear stating "Done writing a template file to user_specific_data_folder\combined\experimentalDesignTemplate.txt".

ProteinPilot

From version 5.0, it is possible to export mzIdentML files from ProteinPilot (see instructions here). From previous versions, see the explanations below:

For ProteinPilot as peptide/protein identification files we strongly recommend providing human readable files instead of the binary '.group' file. Please export the group files into XML files using:

http://www.absciex.com/products/software/proteinpilot-software
"Command Line Control and Open Results. To support users and third-party software vendors that want to integrate ProteinPilot Software, it is possible to script searches *via* command line and decrypt the '.group' file results into clear XML for full access to all the data it contains."

Here is a 'how to 'on the conversion process from one of our submitters:

1. Create a txt file in Notepad entitled say "group2XML_Example.bat.txt" and save it in the ProteinPilot folder (where the group2xml.exe is located).

2. Rename "group2XML_Example.bat.txt" to "group2XML_Example.bat", giving it a Windows batch file extension.

3. Open this batch file in ‘Notepad’ and type in the following command line instructions:
group2XML.exe XML <full path to the .group file to be converted> <full path to the .xml file the .group file will be converted into>

for instance

```
group2XML.exe XML "C:\AB SCIEX\ProteinPilot Data\Results\Example.group" "C:\AB SCIEX\ProteinPilot Data\Results\Example.xml"
```

The command has the following argument structure: group2XML.exe <Type> <Result.group> <Output.file>
where:
- <Type> specifies the type of output.
- <Result.group> is a .group file created by ProteinPilot Software.
- <Output.file> is the name of the file to be created.

4. Save and close the file.

5. Double-click on the file to run the conversion.

Appendix VI: Partial Submission mechanism for Mass Spectrometry imaging datasets

The default PX submission protocol has been changed for MS Imaging datasets. Only ‘partial’ submissions are supported.

These are the main specific points to consider for this type of submissions:

> (i) Additional file tags have been created: metadata information about the images (labeled as ‘MS_IMAGE_DATA’) and an optical image (labeled as ‘OPTICAL’).
> (ii) It is mandatory to provide the MS raw data (called ‘RAW’).

It is recommended to submit MS imaging data in imzML format as it offers the most flexible options for viewing, but proprietary data formats are also accepted.

There is the possibility to submit two different mass spectral related files for one dataset, as required for several MS imaging data formats (e.g. imzML and Analyze). The mass spectral data file (*.ibd for imzML or *.img file in Analyze format) must be labeled as ‘RAW’. The file that contains metadata (such as pixel dimensions and additional information) must be labeled as ‘MS_IMAGE_DATA’ (e.g. *.imzml file for imzML or *.hdr file for Analyze). If an ‘ibd file (imzML format) is submitted as ‘RAW’ an ‘MS_IMAGE_DATA’ (*.imzml) is mandatory.

However, in the case of ‘RAW’ proprietary formats that only consist of one file, a ‘MS-IMAGE_DATA’ file is not required.

> (iii) In addition, PRIDE requires a mandatory ‘SEARCH’ file for ‘partial’ submissions, which corresponds to the processed results. There is currently no strict definition for the format of this mandatory file, but it should contain a list of *m/z* values, names of (tentatively) identified compounds and additional information that were used to the generate MS images in the published work.
> (iv) It is also supported the inclusion of an optical image (‘OPTICAL’) of the measured sample, which can allow validation and/or interpretation. The ‘OPTICAL’ file could contain a photograph of the imaged sample or an adjacent section that shows comparable spatial features. Native samples, classical histological techniques (H&E, toluidine) or immunohistochemistry staining (antibody staining) can be provided for this purpose.

References

1. Vizcaino, J.A., Deutsch, E.W., Wang, R., Csordas, A., Reisinger, F., Rios, D., Dianes, J.A., Sun, Z., Farrah, T., Bandeira, N. *et al.* (2014) ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat Biotechnol*, 32, 223-226.

2. Ternent, T., Csordas, A., Qi, D., Gomez-Baena, G., Beynon, R.J., Jones, A.R., Hermjakob, H. and Vizcaino, J.A. (2014) How to submit MS proteomics data to ProteomeXchange via the PRIDE database. *Proteomics*, 14, 2233-2241.

3. Martens, L., Chambers, M., Sturm, M., Kessner, D., Levander, F., Shofstahl, J., Tang, W.H., Rompp, A., Neumann, S., Pizarro, A.D. *et al.* (2011) mzML--a community standard for mass spectrometry data. *Mol Cell Proteomics*, 10, R110 000133.

4. Rompp, A., Wang, R., Albar, J.P., Urbani, A., Hermjakob, H., Spengler, B. and Vizcaino, J.A. (2015) A public repository for mass spectrometry imaging data. *Anal Bioanal Chem*, 407, 2027-2033.

5. Jones, A.R., Eisenacher, M., Mayer, G., Kohlbacher, O., Siepen, J., Hubbard, S.J., Selley, J.N., Searle, B.C., Shofstahl, J., Seymour, S.L. *et al.* (2012) The mzIdentML data standard for mass spectrometry-based proteomics results. *Mol Cell Proteomics*, 11, M111 014381.

6. Walzer, M., Qi, D., Mayer, G., Uszkoreit, J., Eisenacher, M., Sachsenberg, T., Gonzalez-Galarza, F.F., Fan, J., Bessant, C., Deutsch, E.W. *et al.* (2013) The mzQuantML data standard for mass spectrometry-based quantitative studies in proteomics. *Mol Cell Proteomics*, 12, 2332-2340.

7. Griss, J., Jones, A.R., Sachsenberg, T., Walzer, M., Gatto, L., Hartler, J., Thallinger, G.G., Salek, R.M., Steinbeck, C., Neuhauser, N. *et al.* (2014) The mzTab data exchange format: communicating mass-spectrometry-based proteomics and metabolomics experimental results to a wider audience. *Mol Cell Proteomics*, 13, 2765-2775.

8. Wang, R., Fabregat, A., Rios, D., Ovelleiro, D., Foster, J.M., Cote, R.G., Griss, J., Csordas, A., Perez-Riverol, Y., Reisinger, F. *et al.* (2012) PRIDE Inspector: a tool to visualize and validate MS proteomics data. *Nat Biotechnol*, 30, 135-137.

9. Cote, R.G., Griss, J., Dianes, J.A., Wang, R., Wright, J.C., van den Toorn, H.W., van Breukelen, B., Heck, A.J., Hulstaert, N., Martens, L. *et al.* (2012) The PRoteomics IDEntification (PRIDE) Converter 2 framework: an improved suite of tools to facilitate data submission to the PRIDE database and the ProteomeXchange consortium. *Mol Cell Proteomics*, 11, 1682-1689.

10. Vaudel, M., Burkhart, J.M., Zahedi, R.P., Oveland, E., Berven, F.S., Sickmann, A., Martens, L. and Barsnes, H. (2015) PeptideShaker enables reanalysis of MS-derived proteomics data sets. *Nat Biotechnol*, 33, 22-24.

11. Ghali, F., Krishna, R., Lukasse, P., Martinez-Bartolome, S., Reisinger, F., Hermjakob, H., Vizcaino, J.A. and Jones, A.R. (2013) Tools (Viewer, Library and Validator) that facilitate use of the peptide and protein identification standard format, termed mzIdentML. *Mol Cell Proteomics*, 12, 3026-3035.

12. Dasari, S., Chambers, M.C., Martinez, M.A., Carpenter, K.L., Ham, A.J., Vega-Montoto, L.J. and Tabb, D.L. (2012) Pepitome: evaluating

improved spectral library search for identification complementarity and quality assessment. *Journal of proteome research*, 11, 1686-1695.