Data Submission Guidelines for the ProteomeXchange Consortium

Since 2012, the members of the ProteomeXchange (PX) consortium have been coordinating the submission and dissemination of mass spectrometry (MS) proteomics datasets in the public domain (http://www.proteomexchange.org). This document outlines the main policies of PX. Given the complex nature of some of the issues that can arise during submission and handling of datasets, exceptions to policy can be considered on a case-by-case basis by the leaders of the PX resources.

Table of contents

1 Components of the ProteomeXchange Consortium

The components of the PX Consortium at the moment of writing are included in Table 1, sorted in chronological order considering the date they joined PX.

Table 1. Members of the PX Consortium.

Resource Name	Institution, country	URL	Function in PX	Lead
PRIDE	European Bioinformatics Institute (EMBL-EBI), Cambridge, UK	http://www.ebi.ac.uk/pride	Archival (Universal)	Juan A. Vizcaíno
PeptideAtlas	Institute for Systems Biology, Seattle, WA, USA	http://www.peptideatlas.org/	Re-analysis	Eric W. Deutsch
PASSEL	Institute for Systems Biology, Seattle, WA, USA	http://www.peptideatlas.org/passel/	Archival (Focused)	Eric W. Deutsch
MassIVE	University of California, San Diego, CA, USA	https://massive.ucsd.edu/	Archival (Universal), Re-analysis	Nuno Bandeira
jPOST	Various Institutions, Japan	https://jpostdb.org/	Archival (Universal)	Yasushi Ishihama
iProX	Beijing Proteome Research Center, Beijing, China	http://www.iprox.org/	Archival (Universal)	Yunping Zhu
Panorama Public	University of Washington, Seattle, WA, USA	https://panoramaweb.org/public.url	Archival (Focused)	Brendan MacLean

In addition to the individual PX resources, **ProteomeCentral** (PC, available at http://proteomecentral.proteomexchange.org) is the portal for accessing all PX datasets, independently from the original resource where the datasets were stored. This queryable archive provides the users with an efficient way to identify datasets of interest. PC is also in charge of the generation of the dataset PX identifiers.

PX resources can have different functions in the framework. One resource can have more than one function:

a) ***Archival*** **resources**: Data deposition and storage related activities. There are two types of resources in this context:

   a. *Universal* resources: They can store any type of proteomics datasets, coming from any data workflow. However, they are normally focused in supporting "complete" submissions for particular data workflows, e.g. bottom-up proteomics DDA (Data Dependent Acquisition) workflows. The current examples in the Consortium are PRIDE, MassIVE, jPOST and iProX, for DDA datasets, although they can store all types of datasets.

   b. *Focused* resources: They support specifically one type of data workflow and will not store data coming from other proteomics approaches. This is exemplified by the PASSEL resource (PeptideAtlas) and Panorama Public, which are devoted to targeted proteomics approaches.

b) ***Re-analysis***: They provide re-analysis of primary datasets provided by submitters, which are made publicly available.

2 ProteomeXchange dataset entry requirements

ProteomeXchange will accept, handle and disseminate all types of proteomics datasets generated by MS approaches that meet the minimum requirements established by the consortium (see below), irrespective of their origin.

What constitutes a PX dataset?
The general rule is that a dataset should correspond to the data described in a single manuscript, if all data in the manuscript comes from the same data workflow (e.g. DDA). If a manuscript contains data coming from different proteomics workflows (e.g. DDA and Selected Reaction Monitoring, SRM), it is recommended to split the data in different datasets so this is easier to interpret for third parties. However, it should be highlighted that it is always the submitter's decision how to organise their submitted datasets, which could depend on a number of factors (e.g. organisation of future planned publications).

Which dataset identifiers are provided?
All PX datasets get a unique ProteomeXchange PXD identifier (PXD + a six figure integer, for additional details see: http://www.ebi.ac.uk/miriam/main/collections/MIR:00000513), which is the one that should be cited and highlighted in the scientific literature, both by the original data producers and by other third parties. In addition, resource-specific dataset identifiers can be issued, depending on the specific resource. This is the case at present for MassIVE, PASSEL, jPOST, iProX and Panorama Public. Additionally, DOIs (Digital Object Identifiers) are provided in addition by some of the resources in the case of "Complete" submissions (see below), to facilitate the traceability of the datasets.

What happens in the case proteomics datasets do not fulfil all the PX requirements?
Datasets that do not fulfil all PX requirements (see Section 2) are not supported by the PX framework (e.g. those datasets where only raw data is submitted). In those cases, individual PX resources could support their deposition, but the dataset will always get a resource-specific identifier only, not a PXD identifier. Dissemination policies of the consortium do not apply to them either. It is however encouraged that users attempt to comply with PX requirements, so that these datasets can be included in the framework.

Does PX support non-proteomics MS datasets?
Datasets generated by MS approaches that do not include any proteomics component (e.g. MS metabolomics in the wider sense, glycomics, lipidomics, etc) are not supported by the PX framework. This situation is analogous to what was described above for proteomics datasets that do not fulfil all the requirements. There are two options for users for the deposition of these non-proteomics MS datasets:

(i) use other suitable resources outside PX that explicitly support these data types;
(ii) use resources that are part of PX which are supporting these additional data types (e.g. MassIVE/GNPS, for metabolomics data). The non-proteomics datasets will get a resource-specific identifier, not a PXD one, as explained above.

3 Data types included in a PX dataset and types of data submissions

The PX resources support two types ("complete" and "partial", see below) of dataset submissions, depending on the different proteomics data workflows and the data formats that are applicable. In all cases, it is mandatory to submit:

1- Mass Spectrometry output files: Raw data (mandatory) and (optionally) the derived peak lists (see definitions of data types in Appendix I).

2- Experimental and Technical metadata, as established in the PX XML format, which represents the minimum common denominator (see Appendix II). The different PX resources can have slightly different metadata requirements (see individual documentation for each resource), but at minimum sufficient information needs to be provided to be able to generate the PX XML format (used by the ProteomeCentral resource).

3- Processed Results: At minimum peptide/protein identification results are required. Quantification results are optional at present, although strongly encouraged. Two submission types are supported:

a) **Complete submission:** A complete (also known as "supported") submission ensures that the processed results (at least the identification data) and the corresponding mass spectra can be parsed, integrated and visualised by the PX resource, connecting the identification data to the corresponding mass spectra. To achieve that, processed identification results need to be provided (depending on the policies of each individual resource): (i) in a PSI open standard format (mzIdentML, mzTab), or optionally using an open data format that is supported by the resource (e.g. Skyline XML).

Examples include bottom-up DDA datasets where identification results were generated from any tool that can export the data standard mzIdentML or mzTab (and the corresponding peak list MS files).

b) **Partial submission**: In this case (also known as "unsupported") processed identification results are provided in other data formats than the ones indicated above for complete submissions. For the PX resource, it is then not possible to parse, integrate and visualise the identification and/or connect the processed results to the corresponding mass spectra. However, all the submitted files are made available to download. This mechanism allows data generated from any software that cannot export yet to the supported formats, or from less-mature experimental approaches to be deposited into the PX resources.

Examples include bottom-up DDA datasets where identification results were generated from any tool that cannot export the PSI data standards mzIdentML or mzTab, or other datasets coming from approaches where no open standard for the results has been implemented so far (e.g. top-down proteomics).

4- Other files: Other optional components of submitted datasets are (Appendix I):

- Output of additional analysis software used (e.g. quantification results).
- Protein Sequence database, as used in the search.
- Spectral library, if it was used during the analysis.
- Images, e.g. gel images.
- Scripts
- Additional metadata files
- Other

Table 2 includes the points of contact for each PX resource and the entry point for the external documentation for the users in the different resources.

Table 2. Contact and external documentation access for each PX resource.

Resource	E-mail contact	URL external documentation
PRIDE	pride-support@ebi.ac.uk	https://www.ebi.ac.uk/pride/help/archive
PeptideAtlas	http://www.peptideatlas.org/feedback.php	http://www.peptideatlas.org/overview.php
PASSEL	http://www.peptideatlas.org/feedback.php	http://www.peptideatlas.org/passel/
MassIVE	ccms-web@cs.ucsd.edu	http://proteomics.ucsd.edu/service/massive/documentation/
jPOST	jpostdb@gmail.com https://repository.jpostdb.org/contact	https://repository.jpostdb.org/help
iProX	iprox@iprox.org https://www.iprox.org/page/helpInformation.html	https://www.iprox.org/page/helpEn.html
Panorama Public	panorama@proteinms.net	https://panoramaweb.org/public.url

4 Data workflow for original datasets

The overall ProteomeXchange data workflow is summarized in Figure 1.

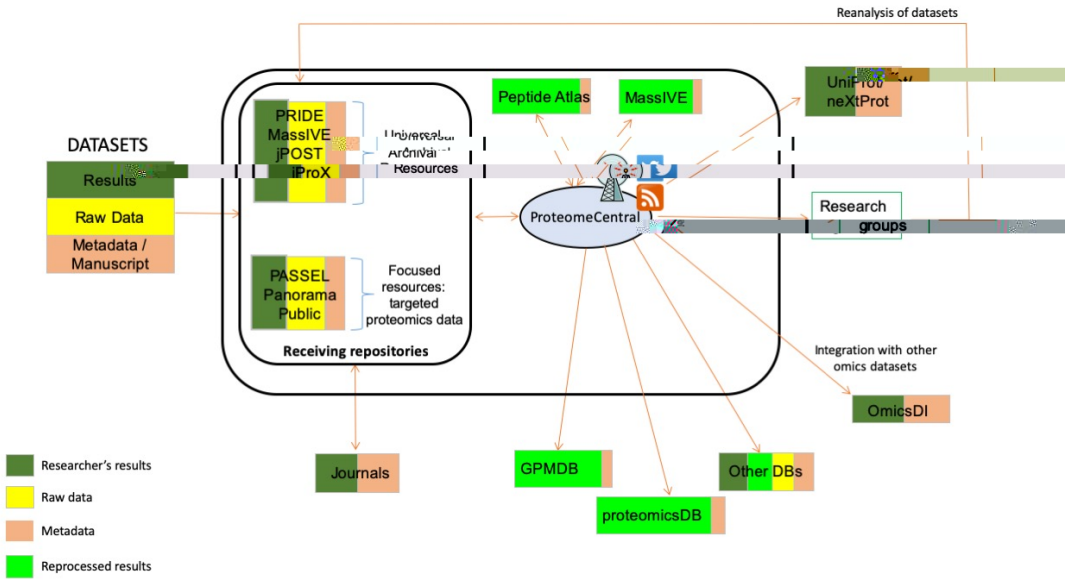

Figure 1: Overview of the ProteomeXchange data flow.

Original datasets coming from any proteomics data workflow can be submitted to any of the *universal* Archival Resources (PRIDE, MassIVE, jPOST, and iProX). However, it is highly encouraged that datasets

from data workflows explicitly supported by existing *focused* archival resources, other than shot-gun DDA proteomics (the most widely used approach), are submitted to that resource, and not to any of the available *universal archival* resources. A summary of the data types and submission types available in each resource are provided in Table 3.

Users can then choose freely the *universal Archival* resource for the submission of their datasets. User preferences can be based for instance on geographical proximity, availability of "complete" submissions for particular workflows, or technical specifications (e.g. speed for data uploads and downloads), among other considerations.

Table 3. Summary of submission guidelines for each PX resource, depending on the data workflow and submission type involved.

Data types/ submission types	PRIDE	PASSEL	MassIVE	jPOST	iProX	Panorama Public
DDA MS/MS						
Partial	Yes	No	Yes	Yes	Yes	No
Complete: mzIdentML	Yes	No	Yes	Yes	Yes	No
Complete: mzTab	Yes	No	Yes	Yes	Yes	No
Complete: TSV	No	No	Yes	No	No	No
Other data workflows						
Targeted SRM/MRM/PRM	No explicit support (Partial only)	Partial and complete	No explicit support (Partial only)	No explicit support (Partial only)	No explicit support (Partial only)	Complete
Targeted DIA/DDA	No explicit support (Partial only)		No explicit support (Partial only)	No explicit support (Partial only)	No explicit support (Partial only)	Complete
DIA MS/MS	Partial only	No	Partial and complete	Partial only	Partial only	No
Top-down	Partial only	No	Partial only	Partial only	Partial only	No
Mass spectrometry imaging	Partial only	No	Partial only	Partial only	Partial only	No

5 Data workflow for reprocessed datasets

Current guidelines for reprocessed datasets (analysis performed of existing public PXD datasets) are available at http://www.proteomexchange.org/docs/reprocessed_guidelines_px.pdf.

6 Ownership, privacy and release of datasets to the public

All PX resources **do not assume editorial control or ownership over the submitted datasets**. Instead the original submitter and the corresponding Principal Investigator (PI) are the owners of these data. It is then required that a submitter and PI are explicitly identified for each dataset. Upon public

availability of the data, the original data ownership is maintained, although we aim that dissemination and reuse of the released data are no longer restricted at that point.

All submitted datasets are private (password protected) by default

All datasets submitted to PX resources remain private, password protected by default. Username and password must be included in the corresponding submitted manuscript/s so that reviewers and editors can access the data at review time during the journal review process. Some PX resources allow at present datasets to be kept private for any duration of time (e.g. PRIDE), although in others (e.g. PASSEL and jPOST), data become automatically available on the date that the submitter specifies. The data submitter has the option of adjusting this date in case of review delays, etc.

Changes in submitted datasets

Authors may change the content of datasets while they are unreleased (non-

that this 6-month extension does not consider the requirements of the scientific journal where the article has been published, which may mandate that the data is released immediately anyway.

7 Data licenses of PX resources

PX resources aim that dissemination and reuse of the released data are no longer restricted once datasets are released to the public. At present, the different resources have different terms of use and/or data licenses in place, which reflect the practises of the institutions and of the different countries where they are located. See Table 4 for the details.

However, all PX resources have decided to move towards a default Creative Commons CC0 license as a minimum level, making available globally datasets without any restrictions. It is important to

The field of proteomics has mostly been focused on protein expression analysis and only recently have there been larger efforts to also analyse human genetic variation on the protein level, driven mainly by proteogenomics and immuno-peptidomics projects. The presence of genomic variants on the proteome level implies that clinical proteomics data has the potential to be patient-identifiable and thus specially protected (e.g., under EU GDPR guidelines) in the same way as genomics data.

In the view of the Consortium, the proteomics community needs to develop rules and best-practice guidelines for dealing with this type of datasets and, moreover, to evaluate the alignment of these efforts with the genomics community. Relevant genomics and transcriptomics datasets are deposited in specialised, access-controlled resources such as the EGA (European Genotype Archive, https://www.ebi.ac.uk/ega/), dbGaP (http://www.ncbi.nlm.nih.gov/gap/) and JGA (https://www.ddbj.nig.ac.jp/jga/index-e.html). Access to these data is only granted after applications are reviewed by an ethics committee.

At present there is not a unified policy for the data management of sensitive proteomics data internationally. Policy and guidelines are starting to be discussed and they will take some time to develop formally. Data management practises will therefore evolve accordingly. PX resources will then adhere to the views of the scientific community in this context.

Therefore, at present, ProteomeXchange resources are committed to completely open data. Authors that have been advised to follow different data management practises for potentially sensitive proteomics datasets are advised to contact resources such as EGA, dbGAP and JGA.

9 Appendix I: Data types 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:
  i) raw data (as described below).
  ii) peak list spectra in a standardized open format such as mzML (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 and require specialized software in order to be read.

- **Standardized MS data formats:** There are currently two widely known mass spectrometry data formats in proteomics: mzXML (developed at the Institute of Systems Biology (ISB), Seattle, USA, but now obsolete) and its successor, the PSI data standard mzML (currently v1.1, 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, if there is software that only performs the quantification part of the analysis, 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 often the 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, X!Tandem xml, etc.
  In addition to each specific format, a data standard format called mzIdentML (currently v1.1 or v1.2, http://www.psidev.info/mzidentml) has been developed by the PSI to represent this kind of information. Some search engine output files can also encode quantification results, but this is not the case of mzIdentML. A second standard data format called mzTab (tab delimited format, http://www.psidev.info/mztab) can represent both identification and quantification results. mzIdentML files are widely supported by the relevant PX resources and can be exported by a number of tools (see updated list at http://www.psidev.info/tools-implementing-mzidentml).

- **Quantification software output files**: the data and metadata generated by the software used for performing exclusively the quantification analysis of peptides and proteins. mzTab (http://www.psidev.info/mztab) can also be used to represent quantification results.

- **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 identification and quantification data. See details for the common metadata required currently by all PX resources at Appendix 2.

10 Appendix II: Minimum metadata required in data submission and the correspondence with the PX XML message

An XML XSD (XML Schema Definition) file has been drafted for use in the generation of the XML messages, which are used by ProteomeCentral. The PX XML schema contains the agreed common metadata by all the PX members. The philosophy behind the design of the proposed schema was to keep it as flexible as possible with an overall structure based on the heavy use of controlled vocabulary (CV) terms.

Most elements in the schema are mandatory apart from a few (, , and). The corresponding .xsd file is available at https://raw.githubusercontent.com/proteomexchange/proteomecentral/master/lib/schemas/proteomeXchange-1.4.0.xsd and a more human readable version is available at http://proteomecentral.proteomexchange.org/schemas/proteomeXchange-1.4.0.html

This is the list of elements in the schema:

- : This is the root element with mandatory attributes.
- The *formatVersion* attribute is just used for specifying the version of the format.
- : This element lists all CVs/Ontologies that were used to populate the file. This ensures that used CV terms can be traced to their origin and definition. This information is generated after the data submission and is provided by the individual PX resource.
- : This element contains some basic information about the submission, like 'title', 'announcement date' or 'project description'. Moreover, some additional information about the type of submission (fully supported ('complete') or not ('partial') by the receiving repository), and whether a related manuscript has already been published is also included in this element. This information has to be provided by the Submitter at submission time.
- : This element includes the identifiers that will unambiguously characterize the dataset: for instance, the PX accession number and the Digital Object Identifier (DOI), if relevant. This information is generated after the data submission and is provided by the individual PX resource.
- : The aim of this element is to know if the dataset constitutes a new submission, or the submission describes the reprocessing of a previously submitted dataset. Every reanalysis performed on a particular dataset gets a different PX accession number. This information can be provided by the Submitter at submission time, or *a posteriori* by the corresponding PX resource.
- : Contains information about the species included in the dataset. This information has to be provided by the Submitter at submission time.
- : Element holding the overall information about the instrumentation used in the generation of the data. This information has to be provided by the Submitter at submission time.
- : All protein modifications (natural and artificial) are listed in this record (specified as CV terms). If a dataset does not contain any modifications, it is also explicitly announced here with a specific CV term. This information has to be provided by the Submitter at submission time.
- : Information about the researchers involved in the generation and submission of the dataset. This information has to be provided by the Submitter at submission time.
- : The list of publications that the dataset has generated. This information can be provided by the Submitter at submission time, or *a posteriori* by the corresponding PX resource.
- : One or more CV terms that define a list of keywords that may be attributed to the dataset. This information has to be provided by the Submitter at submission time.
- : List of links that will allow access to the data. Different links may be used