ProteomeXchange provides globally co-ordinated proteomics data submission and dissemination

Juan Antonio Vizcaíno, Eric Deutsch, Rui Wang, Attila Csordas, Florian Reisinger, Daniel Ríos, José A. Dianes, Zhi Sun, Terry Farrah, Nuno Bandeira, Pierre-Alain Binz, Ioannis Xenarios, Martin Eisenacher, Gerhard Mayer, Laurent Gatto, Alex Campos, Robert J. Chalkley, Hans-Joachim Kraus, Juan Pablo Albar, Salvador Martinez-Bartolomé, Rolf Apweiler, Gilbert S. Omenn, Lennart Martens, Andrew R. Jones & Henning Hermjakob

SUPPLEMENTARY NOTES

1. Introduction

In this supplementary document, we provide documentation as it is at the moment of writing (August 2013). It is expected that formats, tools and pipelines will dynamically evolve, so we also recommend readers to visit the highlighted URLs for having access to the most updated documentation. The contents of the document can be summarized in the following sections:

- Proteome Central: Portal to access all the ProteomeXchange (PX) datasets, submitted to the different repositories. It also assigns the PX identifiers.
- Dissemination of PX datasets, including how to get subscribed to the general notification system, and a description the file format (PX XML) that it is used for generating the messages.
- A summary of the proteomics data standards implemented by the consortium.
- Tools currently available to make submissions to PX, including tools for the creation, visualization, basic analysis and submission to the corresponding repositories.
- Comprehensive list of people involved in the consortium.
- Supplementary figures.
- Sustainability.
- Membership to PX.
- Security model.
- List of relevant URLs.

2. ProteomeCentral

2.1. General concept

ProteomeCentral is the portal to access the PX datasets submitted to the different repositories. Directing a web browser to http://proteomecentral.proteomexchange.org/ provides the current listing of publicly released datasets.

ProteomeCentral (i) acts as an identifier service, generating the PX accession numbers for the different receiving repositories; (ii) stores all existing versions of the PX XML messages; and (iii) enables searches of datasets based on the available metadata present in the PX messages (see Supp. Material, Section 3).

2.2. Technical implementation

All PX submissions will be assigned an identifier from an identifier space shared among all PX partners, enabling data tracking amongst repositories. The receiving repositories may still maintain their own internal identifiers. The assignment of a guaranteed-unique identifier among the repositories is implemented as a web service. A similar mechanism has been implemented by the IMEX consortium (http://www.imexconsortium.org/) for molecular interactions data repositories, as well as by other groups that require uniform identifiers across several sites.

The second ProteomeCentral component is the dataset registration and announcement system, whereby a set of uniform metadata for each dataset is assembled into the PX XML document (see section 3) transmitted to ProteomeCentral, archived there, and then distributed to all interested parties. The announcement explicitly does not contain the experimental data or even all of the metadata, but merely some basic information along with URLs for humans and automated agents to access complete metadata as well as the full dataset, the name of the laboratory, and the submitter of the dataset. Such a scheme avoids the development of a master repository that must take into account all functionality of the member repositories, but rather provides only a basic, global indexing service into the already existing repositories. The transmission of dataset announcement to ProteomeCentral is accomplished via a web service, and the public announcement of such dataset is implemented *via* RSS.

The ProteomeCentral service generates an automated message describing essential metadata about the dataset and posts it to the PX RSS feed, which may be received by all parties who subscribe to the RSS feed (http://groups.google.com/group/proteomexchange/feed/rss_v2_0_msgs.xml, see Section 3.2).

The final component is the ProteomeCentral web site itself, which enables users to browse and search the list of available datasets as well as obtain the full set of basic information for each dataset. This can be done *via* any web browser supporting simple, sortable Javascript table widgets, or programmatically via a web service that makes the information available in XML and TSV formats. Each dataset has a unique page displaying all PX-level metadata, both in a user-friendly format and a computer-parsable XML format. Users may browse and

page through the listing, and may also search for datasets by typing any search string into the search box, and receiving back a matching list of datasets.

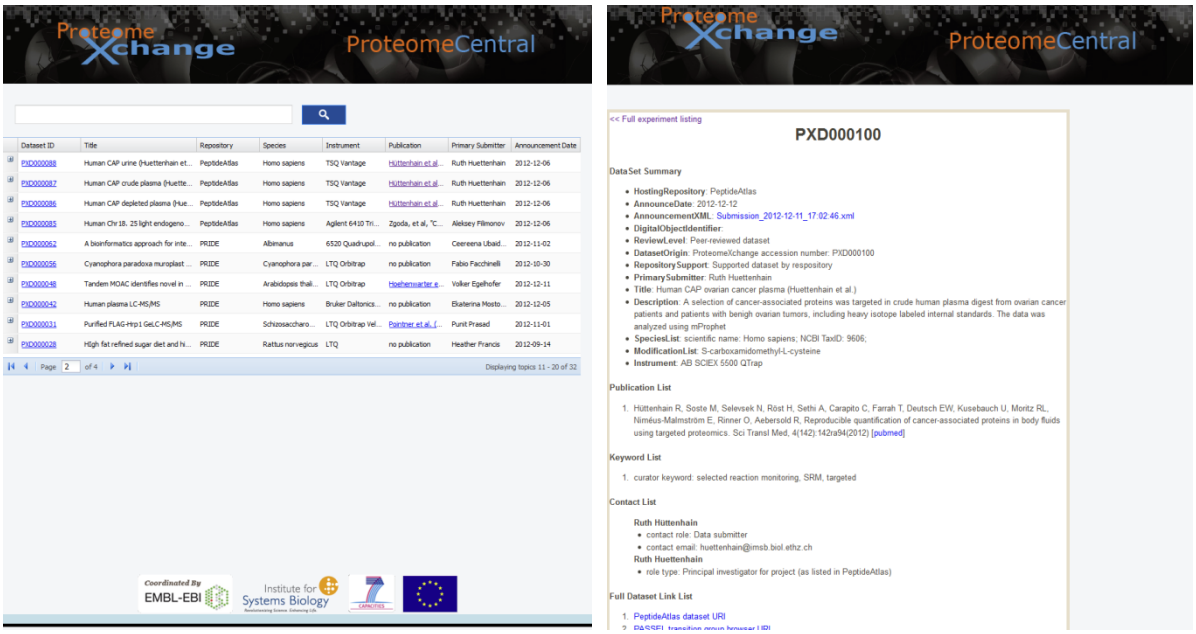


Figure left panel: Listing of 10 recently released datasets on the main ProteomeCentral browsing page. **Right panel**: Details for the 100th experiment to receive a PX accession number. PXD000100 is already public, although many experiments with a lower accession number are not yet public.

3. Dissemination of PX datasets

3.1. How to get notified about new PX datasets

Each PX dataset becomes publicly available on acceptance or publication of the manuscript supported by the dataset. When a submission becomes publicly available, a short summary is released though a public announcement system, *via* a RSS feed containing a link to a file with a defined XML schema (PX XML file). The PX XML file contains key experimental metadata such as: dataset identifiers, sample details (e.g. species and protein modifications are mandatory), mass spectrometer, publication, list of keywords, etc (see details below).

In addition, this file contains links to all the data, and allows PeptideAtlas, UniProt, and/or other resources to evaluate, reprocess and integrate the data. In fact, any member of the community can subscribe to this service. There are two ways to do it:

1) One can receive these updates by e-mail. If you would like to do that, you need to join the PX Google Group:

- Login to Google with your preferred e-mail.

- Go to https://groups.google.com/group/proteomexchange/

- Click on "Join the Group" button (the exact location depends on your preferences for how the groups are displayed in your web browser).

- Choose your preferred option for receiving the e-mails with the new datasets.

2) One can subscribe to the following RSS feed:

http://groups.google.com/group/proteomexchange/feed/rss_v2_0_msgs.xml

3.2. The PX XML message

An XML XSD (XML Schema Definition) file has been drafted for use in the generation of the XML message broadcasted. 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.

All elements in the schema are mandatory apart from the last ones (`ChangeLog`, `DatasetFileList`, `RepositoryRecordList` and `AdditionalInformation`). The corresponding .xsd file is available at http://code.google.com/p/proteomexchange/source/browse/schema/proteomeXchange-1.1.0.xsd.

This is the list of elements in the schema:

- `ProteomeXchangeDataset`: This is the root element with mandatory attributes. The

formatVersion attribute could be used if an announcement has to be repeated with some (minor) changes, e.g. the addition of a publication reference.

- `CvList`: 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.

- `DatasetSummary`: 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.

- `DatasetIdentifierList`: 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.

- `DatasetOriginList`: 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.

- `SpeciesList`: Contains information about the species included in the dataset.

- `InstrumentList`: Element holding the overall information about the instrumentation used in the generation of the data.

- `ModificationList`: 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.

- `ContactList`: Information about the researchers involved in the generation and submission of the dataset.

- `PublicationList`: The list of publications that the dataset has generated.

- `KeywordList`: One or more CV terms that define a list of keywords that may be attributed to the dataset.

- `FullDatasetLinkList`: List of links that will allow access to the data. Different links may be used for different ways of accessing the data (for example FTP download or repository web link) or for different repositories hosting the same data.

- `DatasetFileList`: Optional element to provide individual links to all the submitted files (mass spectrometer output files, search engine output files, etc) belonging to the dataset.

- `RepositoryRecordList`: This optional element allows a repository to report information with more granularity if available. For example links and information could be provided for each part/result file of a larger dataset.

- `AdditionalInformation`: Optional element that includes any other CV terms that can be used to describe the dataset.

- `ChangeLog`: An element that records comments for all changes made to the file since its first release. This element is optional for the first release of the PX XML only, all successive releases must provide a change log entry.

Different versions of the PX XML announcement for the same PX datasets can be made available to ProteomeCentral. This happens if some information included there is updated (for instance, the final version of the reference of a publication). All the versions are tracked and kept in ProteomeCentral.

After reprocessing of a dataset, if the resulting new results are submitted to PX, a new PX identifier will be generated but also the original PX accession number will be retained, to allow coordinated search for different views of data from one submission. This ensures that a simple one-time submission from a contributor is automatically distributed to all PX repositories with sufficient information.

4. Data standards and file formats implemented in PX

A challenging aspect of proteomics research for many years has been a lack of standardisation of file formats used for different stages of a typical analysis pipeline. Each instrument vendor stores data in their own raw binary format, which most commonly require a vendor specific software suite or programming interface to open. These raw MS data formats may be converted to intermediate "peak list" formats, where again a number of text based formats have been developed by different instrument or software vendors (e.g. MGF, pkl and dta) – for a broader discussion and references see [1]. In the identification space, each search engine has tended to export results in their own text-based or XML-based format e.g. Mascot dat (text), X!Tandem XML, OMSSA OMX (XML), and so on. The TPP (Trans Proteomic Pipeline) has also had its own suite of open source formats, mzXML (for raw data and peak lists), pepXML for peptide identifications and protXML for protein identification. In the quantitation domain, most software tools tend to export results as CSV (Comma Separated Values) or HTML reports for end user visualisation or store results in their own internal XML-based formats.

The HUPO-PSI (Proteomics Standards Initiative) has been working in this area for a number of years, acting as a forum for collaboration from different academic and industry groups to agree on a set of common interchange formats. A key aspect of PX has been to contribute actively to the PSI, and drive the completion of formats and the development of tool sets. Recent developments include the release and maintenance of several XML standards, including mzML for storing raw and processed MS data[2], TraML for storing the input transitions in SRM studies[3] and mzIdentML for storing peptide and protein identification data[4]. A key deliverable of PX is to improve capabilities for public deposition of quantitative data. PX members have taken a leading role in formulating a detailed XML-based standard called mzQuantML, capable of storing a complete trace of metadata and data, used to arrive at relative or absolute quantitation values for peptides or proteins[5]. It was also recognised that, particularly for quantitative data, there is a need for an end-user focussed format that could be more easily loaded into spreadsheet software or statistical analysis software. This requirement led to the development of the mzTab specification [Griss *et al.,* submitted] (http://code.google.com/p/mztab/). A number of software tools and programming interfaces have been developed with input from the PX consortium, enabling easier end user analysis and developer uptake for integration of formats into software pipeline.

As described in the following section, the current PX tool set supports a number of existing formats, converting identification and spectral data to the PRIDE XML format, and also now to mzIdentML (identification results) plus the corresponding spectra in a text or XML-based file (ideally mzML), thus simplifying future re-analysis of data sets. As quantification formats stabilise and tools become mature, support for PSI standards will be developed by the consortium.

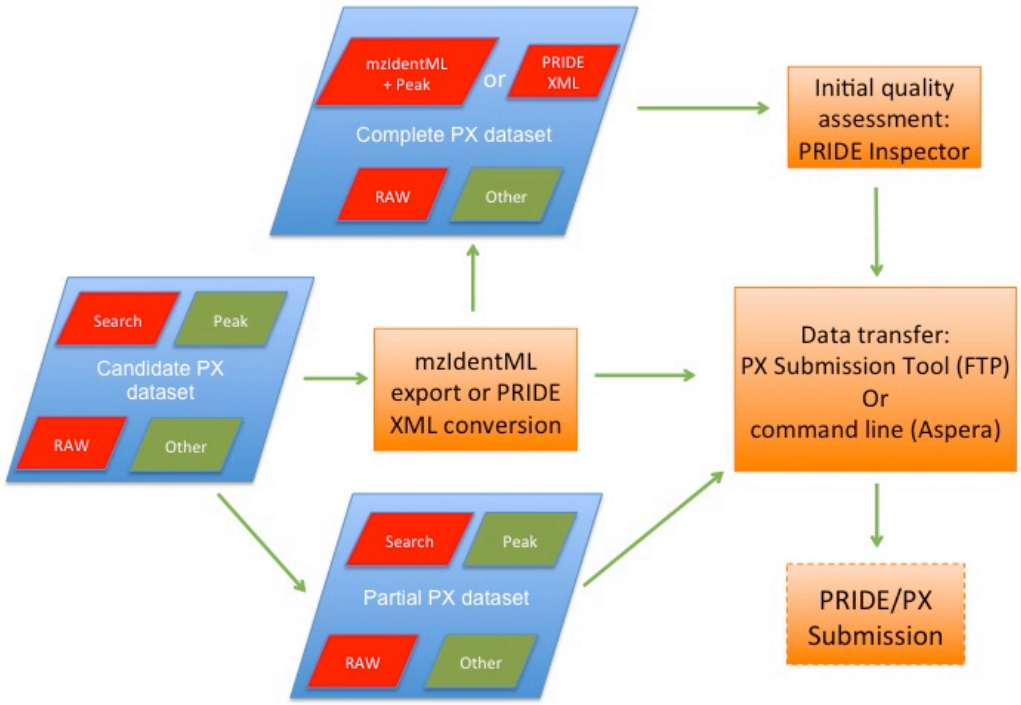
The PSI has also developed a number of MIAPE (Minimum Information About a Proteomics Experiment) guidelines[6] (http://www.psidev.info/miape). The PX consortium encourages

submitters to provide as much metadata annotation as possible at the moment of submission. To make this possible, in the PX context, mzML and mzIdentML semantic validators have been developed[7]. It is important to highlight that the data workflow provided by the ProteoRed *MIAPE Extractor* (http://www.proteored.org/MIAPEExtractor) provides fully MIAPE compliant PRIDE XML files prepared for a PX submission, including a data compilation/integration step, and a data curation step consisting, for example applying an FDR threshold (see http://proteo.cnb.csic.es/trac/attachment/wiki/WikiStart/figuraAbstract.png).

5. Tools available and ways to submit data to PX

5.1. MS/MS data submissions to PRIDE

An overview of the file types and available tools to submit MS/MS data to ProteomeXchange *via* PRIDE is provided in the figure below.



Overview of the types of files and tools available to perform a MS/MS submission to PX *via* PRIDE. See sections 5.1.1-5.1.3 for more details.

5.1.1. Creation of supported files for "Complete" submissions

5.1.1.1. PRIDE XML

PRIDE XML is the internal format supported by PRIDE since its inception.

A) Tools developed by the PRIDE team

PRIDE Converter 2[8] (http://code.google.com/p/pride-converter-2/) is the most recent conversion tool developed by the team. It can work in batch mode and it can be integrated into automatic pipelines due to its modular software architecture. It is composed of 4 independent applications:

-The *PRIDE Converter 2* application will convert MS search result files containing identification and spectra into PRIDE XML.

-The *PRIDE mzTab Generator* will produce skeleton mzTab files from MS search results files. At present, these skeleton files require either manual or scripted editing to add quantitation and/or gel information, but will be updated for automated insertion of quantitation results from different community file formats when the mzTab format is finalised.

-The *PRIDE XML Filter* will remove identifications or spectra from PRIDE XML files based on a series of configurable filters.

- The *PRIDE XML Merger* will combine several PRIDE XML files into a single one.

List of the formats supported by PRIDE Converter 2 by August 2013 (table below).

Format Name	File Type	Data Content
Mascot	.dat	Spectra and Identifications
X!Tandem	.xml	Spectra and Identifications
OMSSA	.csv	Spectra and Identifications
SpectraST	.txt	Spectra and Identifications
CRUX	.txt	Spectra and Identifications
MSGF	.txt	Spectra and Identifications
Proteome Discoverer	.msf	Spectra and Identifications
DTA	.dta	Spectra Only
MGF	.mgf	Spectra Only
mzData	.xml	Spectra Only
mzXML	.xml	Spectra Only
PKL	.pkl	Spectra Only

List of formats supported by PRIDE Converter 2.

Tutorials for general users and developers are available at the PRIDE Converter 2 Google Code page (http://code.google.com/p/pride-converter-2/).

B) Other tools developed by collaborators with capability for direct submission to PRIDE

1- PeptideShaker (peptide-shaker.googlecode.com/). It can use as input Mascot .dat, X!Tandem XML and OMSSA .omx files.

2- ProteinLynx Global Server (PLGS, Waters Corporation). It has an exporter to PRIDE XML from version 2.4. Improved support from version 3.0.

3- OmicsHub Proteomics (Integromics, https://www.integromics.com/products/proteomics/).

4- hEIDI (http://biodev.extra.cea.fr/docs/heidi). Local LIMS system.

5- Proteios[9] (http://www.proteios.org/). LIMS system.

6- EasyProt[10] (http://easyprot.unige.ch/). Software platform for the analysis of MS/MS data.

7- ProteinScape (Bruker).

8- The ProteoRed MIAPE Extractor tool (http://www.proteored.org/MIAPEExtractor). It is able to generate fully MIAPE compliant (MS-MSI) PRIDE XML files containing much more detailed metadata than the minimal required by PX submission.

5.1.1.2. mzIdentML

As mentioned in the previous section, mzIdentML[4] 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 and PX. 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 is the list of software that can export mzIdentML v1.1 (see an updated list at http://www.psidev.info/tools-implementing-mzidentml):

1- Mascot (Matrix Science, http://www.matrixscience.com/). From version 2.4.

2- MS-GF+ (http://proteomics.ucsd.edu/Software/MSGFPlus.html#pubs).

3- Phenyx (GeneBio, http://www.genebio.com/products/phenyx/).

4- 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).

5- TPP (pep.xml and prot.xml files): The idConvert tool from can be downloaded from ProteoWizard[11], or is bundled with the TPP directly starting with version 4.6.3.

6- X!Tandem and OMSSA: Using the mzidLibrary[7] (https://code.google.com/p/mzidentml-lib/).

7- Scaffold (Proteome Software, http://www.proteomesoftware.com/products/scaffold/). From version 4.0.

8- OpenMS[12].

9- MIAPE MSI Extractor (http://proteored.org/miape/, ProteoRed, Madrid)

10- PAnalyzer[13]: Tool to perform protein inference analysis (https://code.google.com/p/ehu-bio/wiki/PAnalyzer).

11- Tools from D. Tabb lab: Myrimatch[14], Pepitome (spectral library search)[15] , TagRecon[16] and IDPicker[17].

5.1.2. Checking the files before submission (initial quality assessment)

A) Tool developed by the PRIDE team

PRIDE Inspector[18] (http://code.google.com/p/pride-toolsuite/wiki/PRIDEInspector). This is an open source rich client application for inspecting MS-based proteomics data. Experiments can be examined based on different views emphasising 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 PRIDE XML. 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.

B) Other tools developed by collaborators

1- PRIDE Viewer[19] (http://proteo.cnb.csic.es/prideviewer/). It can visualize PRIDE XML files.

2- 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.

3- mzIdentML validator (http://psi-pi.googlecode.com/svn/trunk/validator/trunk/mzid-validator.html): a Java-based tool to validate semantics and MIAPE compliance of mzIdentML files[7].

4- ProteoRed MIAPE Extractor tool workflow (http://www.proteored.org/MIAPEExtractor): After the MIAPE information, data can be integrated, inspected and validated before the PRIDE XML creation.

5- ProteoIDViewer (https://code.google.com/p/mzidentml-viewer/)[7]. Java based viewer optimized for mzIdentML files, with some analysis functionality as well.

5.1.3. File submission to PRIDE: the PX submission tool

5.1.3.1. General Information

System Requirements:

- Java: JRE 1.6 +
- CPU: 1 gigahertz (GHz) or faster 32-bit or 64-bit processor
- Memory: 1 gigabyte (GB) RAM
- Hard Disk: 50 MB available
- Platform: Tested on Mac OS X, Linux, and Windows

Additional Requirements:

- Internet access is needed to connect to the PX web services for user login and submission.
- FTP should also be enabled for large file upload on the submitter's LAN firewall.

Source Code:

http://proteomexchange.googlecode.com/svn/px-submission-tool

5.1.3.2. Functionality, Design and Implementation Details

The PX submission tool (current version is 2.0.0) is a standalone graphical user interface (GUI) application written in Java and released under the Apache 2 open source license.

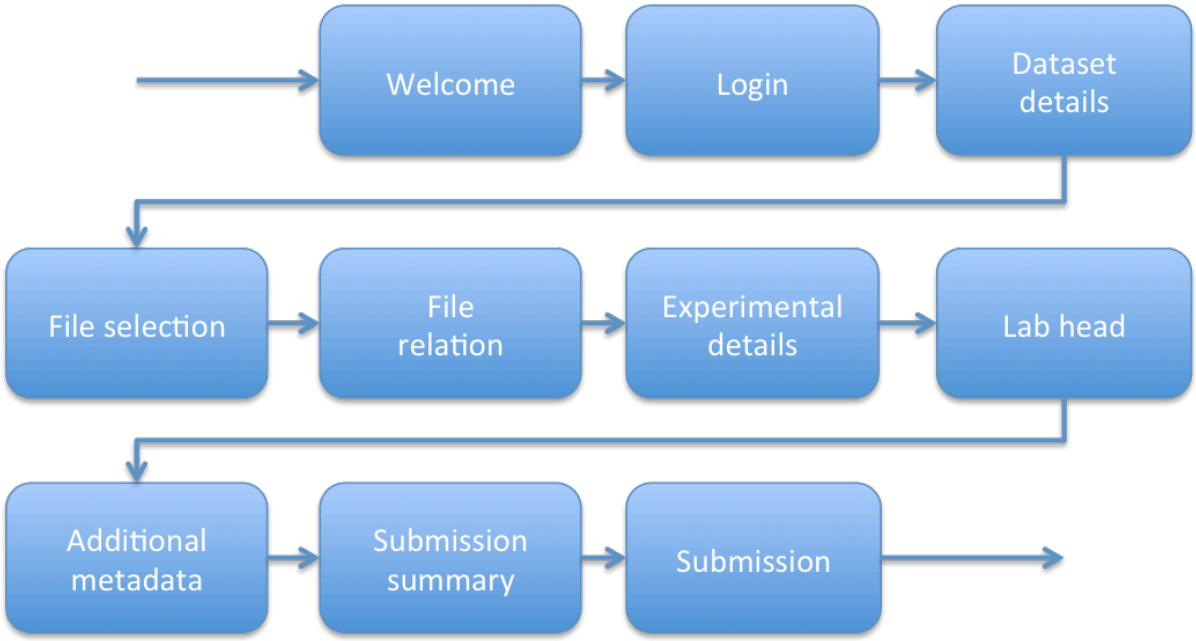
The "PX submission tool" can: (i) select all the files needed for submission; (ii) interactively group related different types of files (for instance, link the corresponding mass spectrometer output and results files); (iii) ensure a minimum level of metadata (especially for 'partial submissions'); and (iv) send the files to the EBI *via* FTP. Large datasets can be submitted conveniently using a batch system. With that aim, a tab delimited file format was designed (details about the file format specification at http://www.proteomexchange.org/sites/proteomexchange.org/files/documents/proteomexchange_submission_summary_file_format.pdf).

The architecture of the tool consists of two loosely coupled modules: the data access module and the GUI module. The data access module is a library called *px-submission-core*, for reading, writing and validating the custom PX submission files. In the submission tool, the

library is acting as an intermediate data model for storing and validating the user's input before a data submission is completed.

The data access module can also be used as a third party library in other applications or pipelines, and is particularly useful for bulk data submission where it is impractical to capture many relationships between files. A script using this library can be used to generate a PX submission file, which in turn can be loaded into the submission tool with all the fields being populated automatically.

The GUI module of the submission tool performs the submission process by guiding users through a series of steps, where each step focuses on a particular aspect of the submission, as shown in figure below (see section 7 Supp. Notes and figures therein).



PX Submission tool workflow

1. 'Welcome' step: the entry step to the submission tool, currently it allows users to choose two main types of submissions: "Complete" and "Partial". For each type of submission, it also reminds the submitter about the information required to complete a submission. Furthermore, the 'Welcome' screen provides options for both resubmission and bulk submission.

2. 'Login' step: this step requires the submitter to login to PRIDE as a registered user, so the ownership and contact details can be assigned to the submitted dataset.

3. 'Dataset details' step: the form asks the user to provide some general metadata about the dataset, such as: project title, description, sample processing protocol, data processing protocol and experiment type.

4. 'File selection' step: the user specifies all the project related files that need to be submitted together. It also tries to assign a file type automatically, which can be overwritten by the users if needed. Currently, there are seven different file types: 'RESULT' for processed results, 'SEARCH' for search engine output, 'PEAK' for peak list files, 'RAW' for MS instrument mass spectrometer output files, 'QUANT', for quantification results, "GEL", for gel images, and 'OTHER' for any other file types.

5. 'Files relation' step: this step groups the files that belong to the same experiment. It is mandatory that each result file needs to have at least one 'RAW' file mapped to it directly.

6. 'Experimental details' step: the form asks the user to provide some details about the experiment, such as: species, tissue, cell type, disease, instrument, modification, quantification and experimental factor. The details needed are different depending on the type of submission ("Complete" or "Partial"). For "Complete" submissions, each 'RESULT' file needs to be annotated, whereas for "Partial" submissions, only general annotations at the dataset are needed.

7. 'Lab head' step: this step requires submitter to provide the contact details of the lab head or the principal investigator. This information will be used as a secondary contact point and to group datasets in the future.

8. 'Additional metadata' step: this step is an optional step for capturing additional metadata about the dataset, such as: parent project, PubMed ID if published, previous ProteomeXchange accession if the dataset is a reanalysis of previous data and links to other Omics datasets, in case the submission is part of a multiomics study and data from other techniques have been submitted to other resources.

9. 'Submission summary' step: Provides a summary view of the submission for the user to review before the file upload begins.

10. 'Submission' step: Submits the complete dataset to the ProteomeXchange consortium using the FTP protocol.

The ProteomeXchange submission tool was built with fault tolerance in mind. It can stop and resume an existing submission if needed. Also if the application exits unexpectedly, the submission tool keeps a local record, and the user can resume the submission by just restarting the application.

The PX submission tool has been designed and developed in a way that it could be potentially used as well by other resources in the future. As highlighted before, the 'RESULT' files format supported for performing a "Complete" submission are PRIDE XML and mzIdentML version 1.1.0. Once the dataset is finalized, the submission is reviewed by a curator, and then loaded into the PRIDE database.

5.1.3.3. New open source libraries made available with PX submission tool

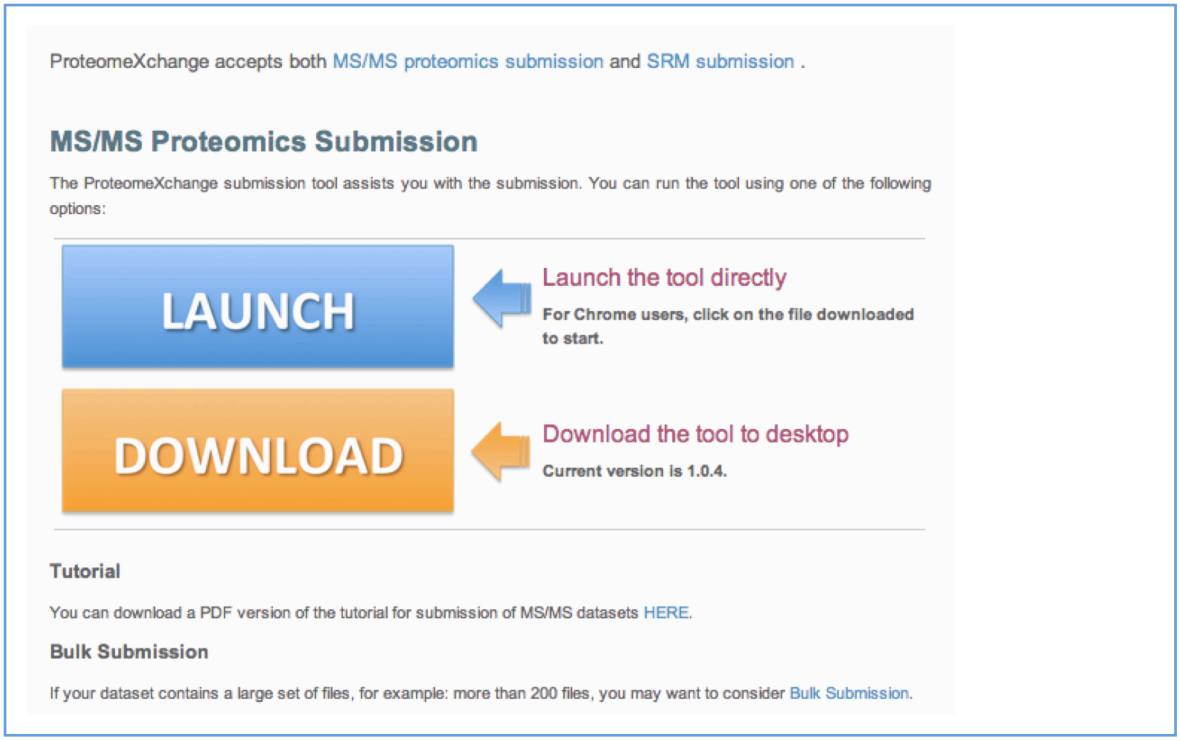
px-submission-core	
Website & Source Code	http://proteomexchange.googlecode.com/svn/px-submission-core
Description	The *px-submission-core* library is a sub-module of the PX Submission tool and is also designed to be an independent library. Therefore, it can be easily integrated into other pipelines or tools for creating, reading and validating a submission summary file. It offers the following key features: • Full object modelling of the submission summary file • Parsing PX submission summary file • Writing PX submission summary file from memory • Validate PX submission summary file
License	Apache 2 open source license
Language	Java

px-submission-core library [seems to be two Table 2s]

5.1.3.4. PX Submission Tool Java Web Start

In addition to the desktop application, the PX submission tool has also been integrated into the PX website (http://www.proteomexchange.org/submission), in the form of a Java Web Start application (http://www.oracle.com/technetwork/java/javase/tech/index-jsp-136112.html). The benefits of this alternative are:

1. It does not require manual installation, and users can launch it directly from the PX submission page (see figure below).
2. It also enables one click experience. In addition, the users will always get the latest version of the tool.



Screenshot of the PX submission page where it is possible to launch the Web Start version of the PX Submission Tool (http://www.proteomexchange.org/submission).

5.1.4. File submission to PRIDE: Command line support using Aspera

The FTP protocol (used in the PX submission tool) is sometimes not ideal, depending on the total size of the files to be submitted, or on the location of the submitter. At the moment, there is an alternative way to upload data to PX *via* PRIDE, using Aspera file transfer protocol (http://asperasoft.com/). However, at the moment of writing, only command line support is available so some bioinformatics expertise is required for the submitter before using this service. In the future, we plan to offer this functionality in the PX submission tool as well.

Instructions about how to use Aspera are available at http://www.ebi.ac.uk/pride/help/archive/aspera.

5.1.5. Examples of Partial submissions to PRIDE

The list of fully supported/unsupported files will be dynamic since new converters/exporters are frequently under development connecting proteomics tools to specific repositories.

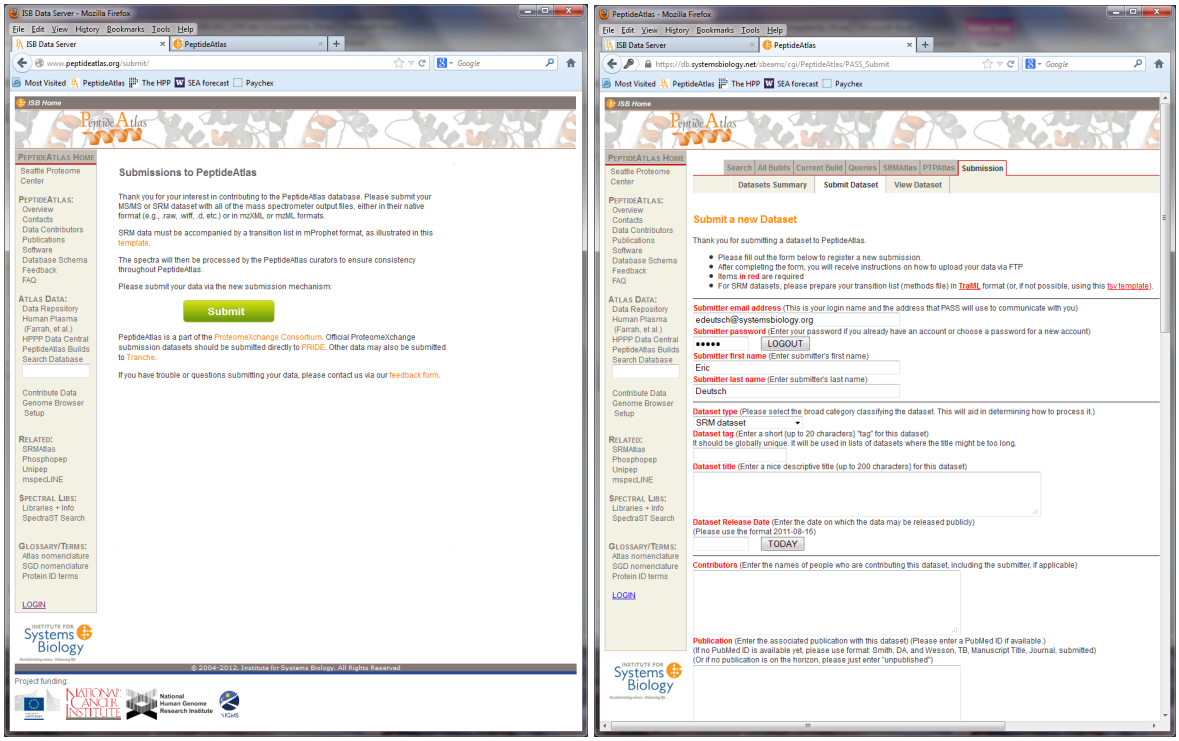
Partial submissions are allowed for workflows whose output files cannot be converted into PRIDE XML or mzIdentML due to lack of a suitable working converter/exporter. These submissions will not get a DOI and the files will be available to download but not fully integrated in PRIDE.

Some examples are at present: MaxQuant[20] or ProteinPilot™ (AB SCIEX), among others. In this case, the mandatory elements for performing a submission are:

- Search engine output files.
- Raw files (mass spectrometer output files).

5.2. SRM data submissions *via* PASSEL

Submission to PASSEL (PeptideAtlas SRM Experiment Library) is performed *via* the PeptideAtlas Submission System (PASS), accessible at http://www.peptideatlas.org/submit/. Basic requirements are described at this page (figure below, a). These requirements are essentially the same as for MS/MS datasets as described above with the important addition of the SRM transitions lists, which are the instructions given to the mass spec for data acquisition. Clicking the [SUBMIT] button leads the user to a web form requesting basic information about the submission (figure below, b).



Screenshots from the PASSEL/PASS submission process. (a) left: Submission overview and summary page describing the process and providing links to more information. (b) right: Data annotation submission format, wherein users register/login and then provide basic metadata about the dataset they are submitting.

Once the basic metadata has been satisfactorily provided, an FTP account is automatically created, and the submitter is invited to upload their mass spectrometer output files, transition lists, interpreted results, and additional metadata to the FTP account. The user may then finalize the submission, so that it cannot be edited any more, although the user does have the option of reverting the dataset back into an editable mode later if necessary.

Once the dataset is finalized, the submission is reviewed by a curator, and then loaded into the PASSEL database, wherein the data may be browsed interactively.

6. List of PX partners and stakeholders

6.1. Partners of the PX EU FP7 grant

People representing the partners of the EU FP7 grant 'ProteomeXchange' who participated in any of the PX meetings in Heidelberg (2011), San Diego (2012) and/or Liverpool (2013):

Name	Surname	Organization and Country
Juan Pablo	Albar	ProteoRed, CNB-CSIC, Madrid, Spain
Conrad	Bessant	Queen Mary University of London, UK
Philip	Andrews	University of Michigan, USA
Harald	Barsnes	University of Bergen, Norway
Pierre-Alain	Binz	GeneBio, Geneva, Switzerland; SIB Swiss Institute of Bioinformatics, Geneva, Switzerland
Alan	Bridge	SIB Swiss Institute of Bioinformatics, Geneva, Switzerland
David	Campbell	Institute for Systems Biology, Seattle, USA
Alex	Campos	Integromics SL, Madrid, Spain
Niklaas	Colaert	University of Ghent / VIB, Belgium
Attila	Csordas	EMBL-EBI, Cambridge, UK
Eric	Deutsch	ISB, Seattle, USA
Martin	Eisenacher	University of Bochum, Germany
Lucia	Espona Pernas	ETH Zürich, Switzerland
Eduardo	Gonzalez-Couto	Integromics SL, Madrid, Spain
Johannes	Griss	EMBL-EBI, Cambridge, UK
Kenny	Helsens	University of Ghent / VIB, Belgium
Henning	Hermjakob	EMBL-EBI, Cambridge, UK
Niels	Hulstaert	University of Ghent / VIB, Belgium
Andrew	Jones	University of Liverpool, UK
Hans-Joachim	Kraus	Wiley-VCH Verlag GmbH, Germany / *Proteomics*
Lennart	Martens	University of Ghent / VIB, Belgium
Salvador	Martínez de Bartolomé	ProteoRed, CNB-CSIC, Madrid, Spain
Gerhard	Mayer	University of Bochum, Germany
David	Ovelleiro	EMBL-EBI, Cambridge, UK
Da	Qi	University of Liverpool, UK
Florian	Reisinger	EMBL-EBI, Cambridge, UK
An	Staes	University of Ghent / VIB, Belgium
Christian	Stephan	University of Bochum, Germany
Julian	Uszkoreit	University of Bochum, Germany
Juan Antonio	Vizcaíno	EMBL-EBI, Cambridge, UK
Ioannis	Xenarios	SIB Swiss Institute of Bioinformatics, Switzerland

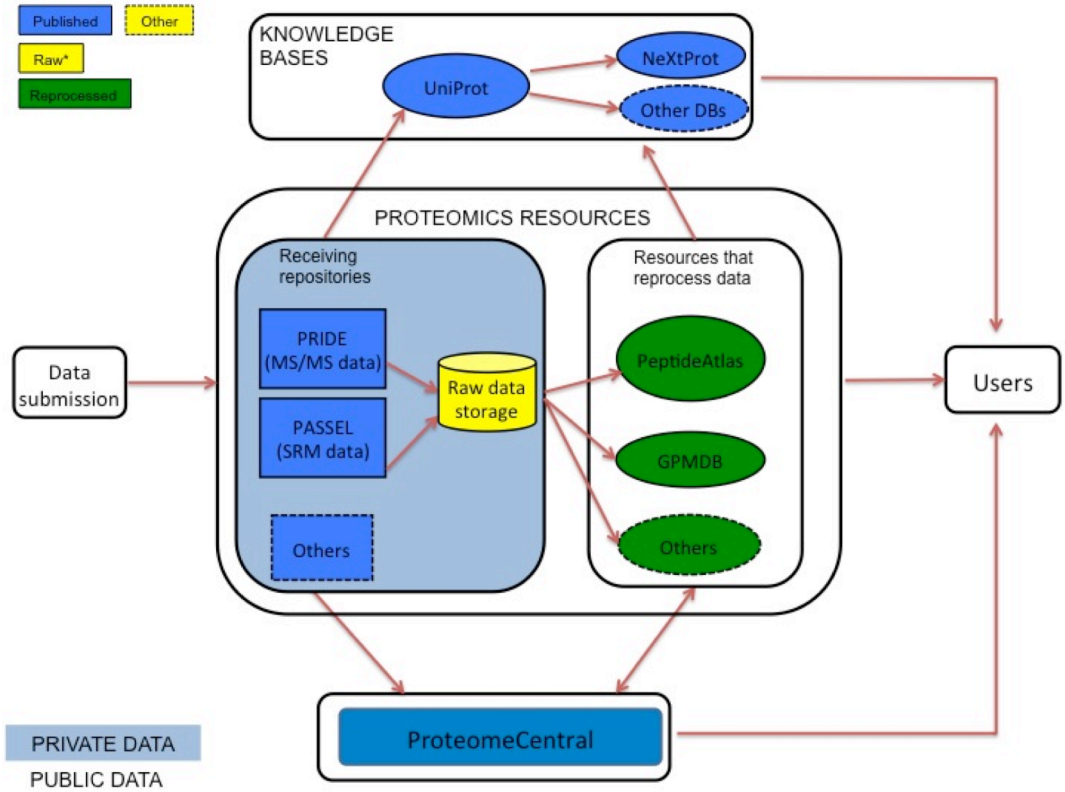
6.2. PX Stakeholders

People representing prominent proteomics groups or scientific journals, who also participated in the PX meetings in Heidelberg (2011), San Diego (2012) and/or Liverpool (2013):

Name	Surname	Organization and Country
Yasmeen	Ahmad	Centre for Gene Regulation & Expression, University of Dundee, UK
Nuno	Bandeira	University of California, San Diego, USA
Conrad	Bessant	Queen Mary University of London, UK
Christoph	Borchers	University of Victoria, Genome BC Proteomics Centre, Canada
Ralph	Bradshaw	University of California, San Francisco / *Molecular and Cellular Proteomics*
Robert	Chalkley	University of California, San Francisco / *Molecular and Cellular Proteomics*
Matt	Chambers	Vanderbilt University Medical Center, USA
Andrew	Dowsey	University of Manchester, UK
Markus	Elsner	*Nature Biotechnology*, Germany
Laurent	Gatto	University of Cambridge, UK
Christopher	Gerner	Medical University of Vienna, Austria
William	Hancock	Northeastern University, Boston, USA/ *Journal of Proteome Research*
Seul-Ki	Jeong	Yonsei Proteome Research Center, Seoul, South Korea
Sangtae	Kim	University of California, San Diego, USA
Ruth	McNally	ESRC Cesagen, Lancaster, UK
Robert	Moulder	University of Turku, Finland
Markus	Muller	SIB Swiss Institute of Bioinformatics, Geneva, Switzerland
Alexey	Nesvizhskii	University of Michigan, USA
Gilbert S.	Omenn	University of Michigan, USA
Jan	Sklenar	The Sainsbury Laboratory, Norwich, UK
Shin	Kawano	Database Center for Life Science, Research Organization of Information and Systems, Bunkyo-ku, Tokyo, Japan
Robert	Petryszak	EMBL-EBI, Functional Genomics Group
Serhiy	Souchelnytskyi	Karolinska Institutet, Stockholm, Sweden
Tao	Xu	The Scripps Research Institute, San Diego, USA
Weimin	Zhu	Taicang Institute of Life Sciences Information, Suzhou, China
Nobel	Zong	University of California, Los Angeles, USA

7. Supplementary Figures

Supp. Figure 1. Detailed PX data workflow. "Published" means data as published in journals. "Raw" means the mass spectrometer output files. "Private" refers to the time while data remains private, during the manuscript review process. Once the manuscript is accepted for publication, the data becomes "Public".

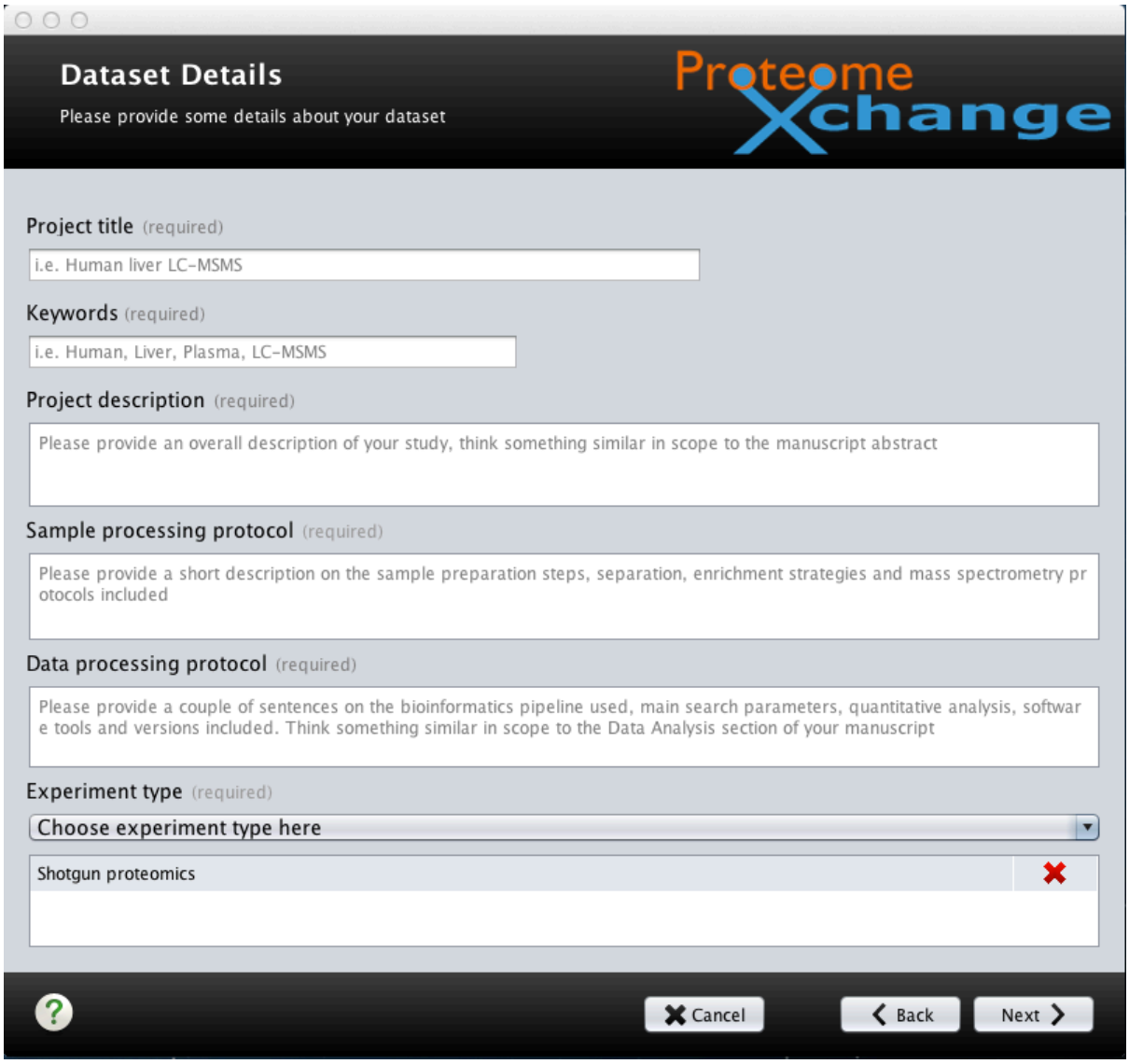


The following screenshots have been generated from the PX submission tool version 2.0.0 (the tool is available at http://www.proteomexchange.org/submission).

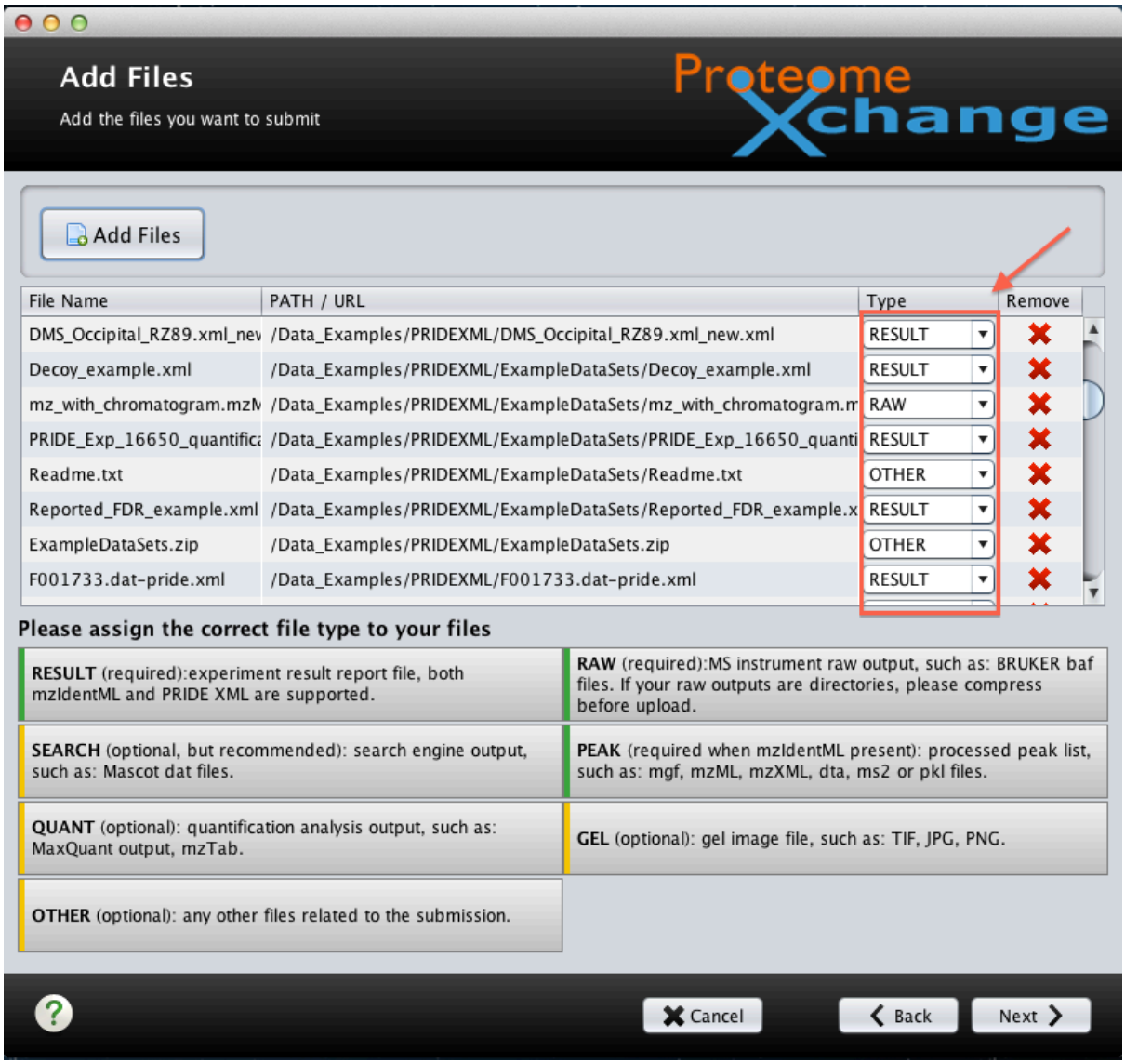
Supp. Figure 2. 'Welcome' screen of the PX Submission tool, highlighting the two default submission types: complete and partial. In this first screen, it is also possible to select the 'Bulk submission' mode.

Supp. Figure 3. 'Login' screen of the PX submission tool. Login to PRIDE, to be identified as the owner of the dataset.

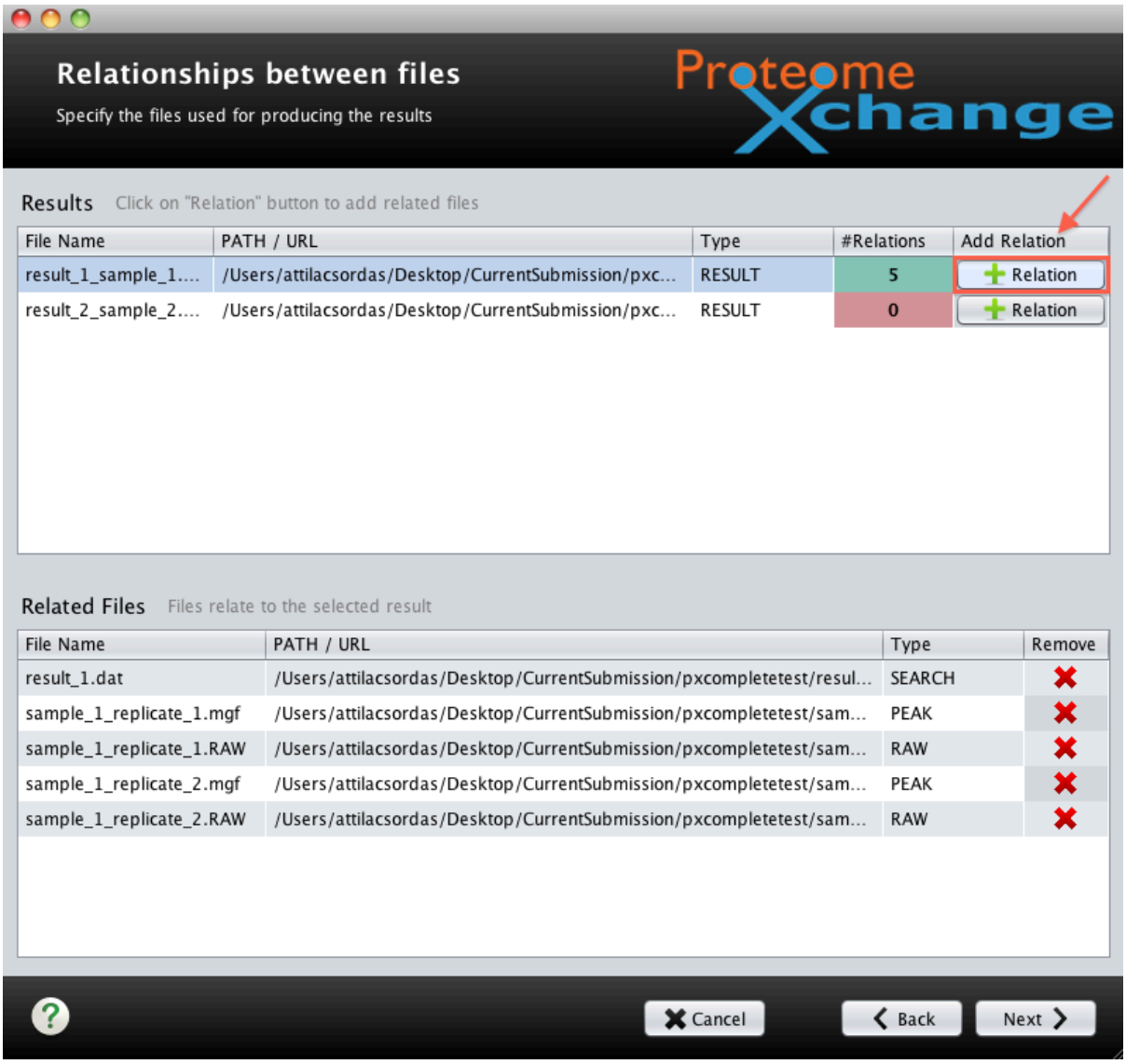
Supp. Figure 4. 'Dataset Details' screen of the PX Submission tool. The submitter needs to provide extra metadata information about the dataset.



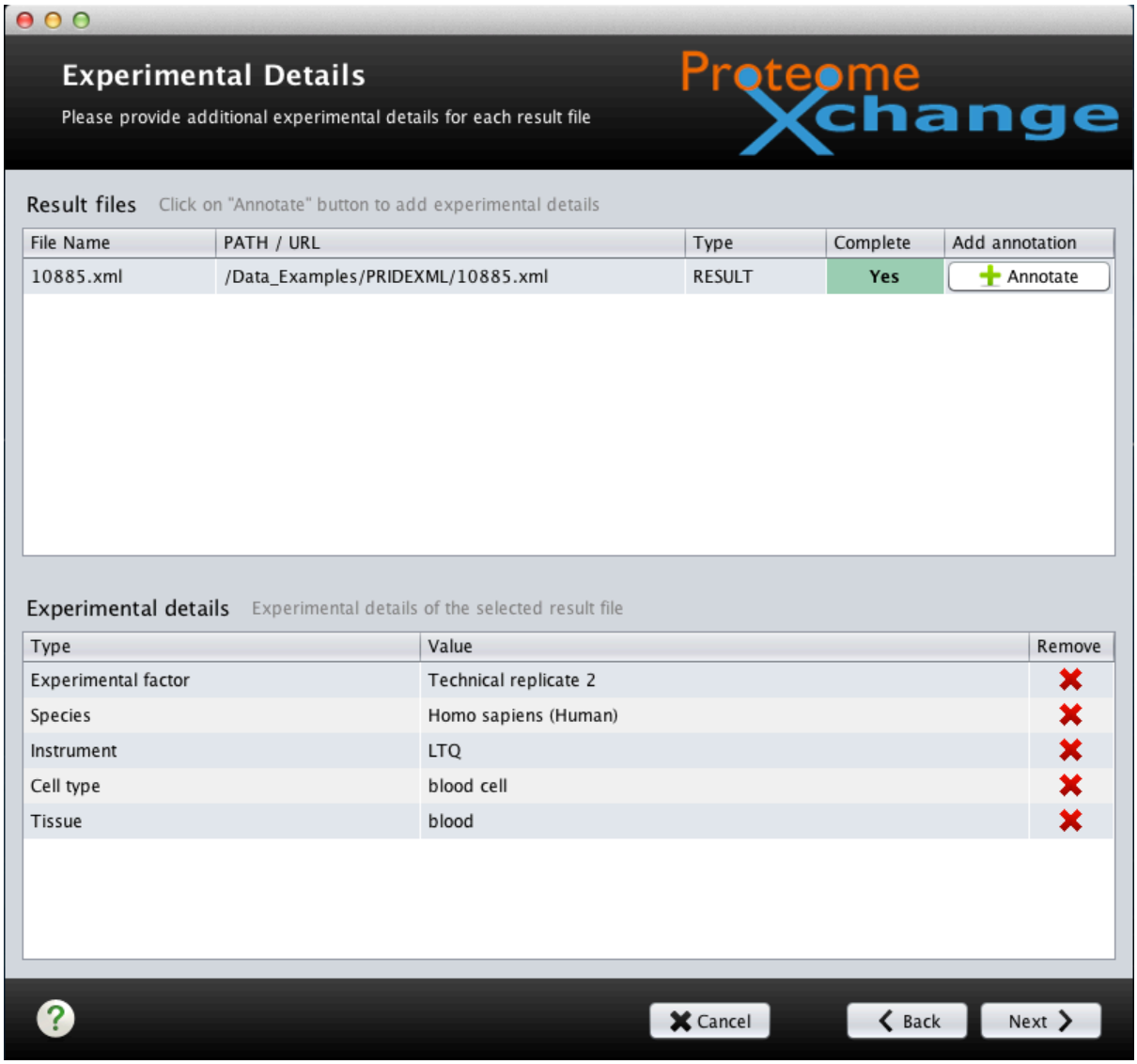
Supp. Figure 5. 'Add files' screen of the PX Submission tool. Different files need to be selected and assigned a 'Type'. All the files that are part of the submission need to be selected at this stage of the submission.



Supp. Figure 6. 'Relationships between files' screen of the PX Submission tool. At this stage, the 'RESULT' files need to be linked to the other types of files included in the submission (at least 'RAW' files will also be present).

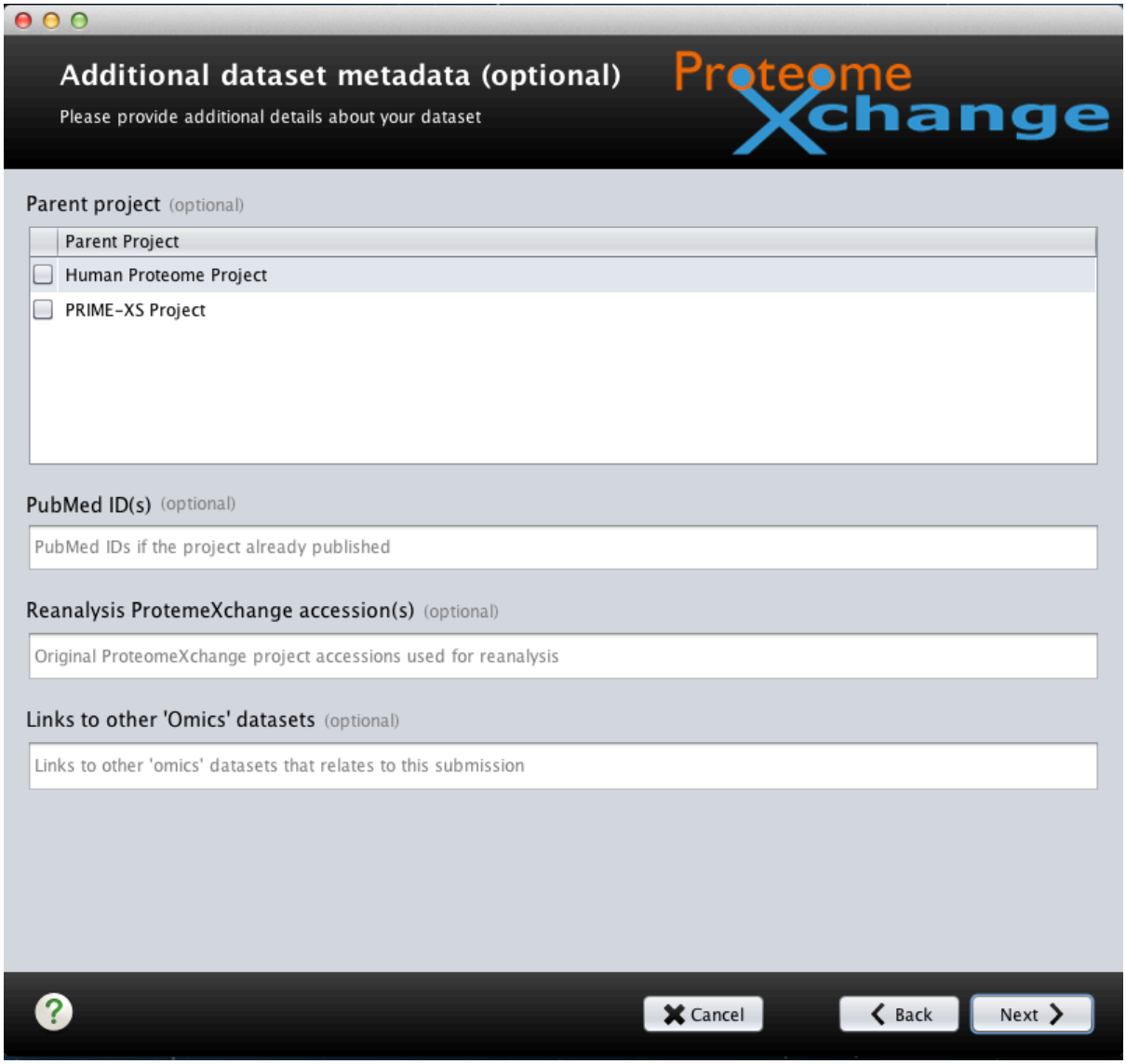


Supp. Figure 7. 'Experimental Details' screen of the PX Submission tool. The submitter needs to provide extra sample metadata information.

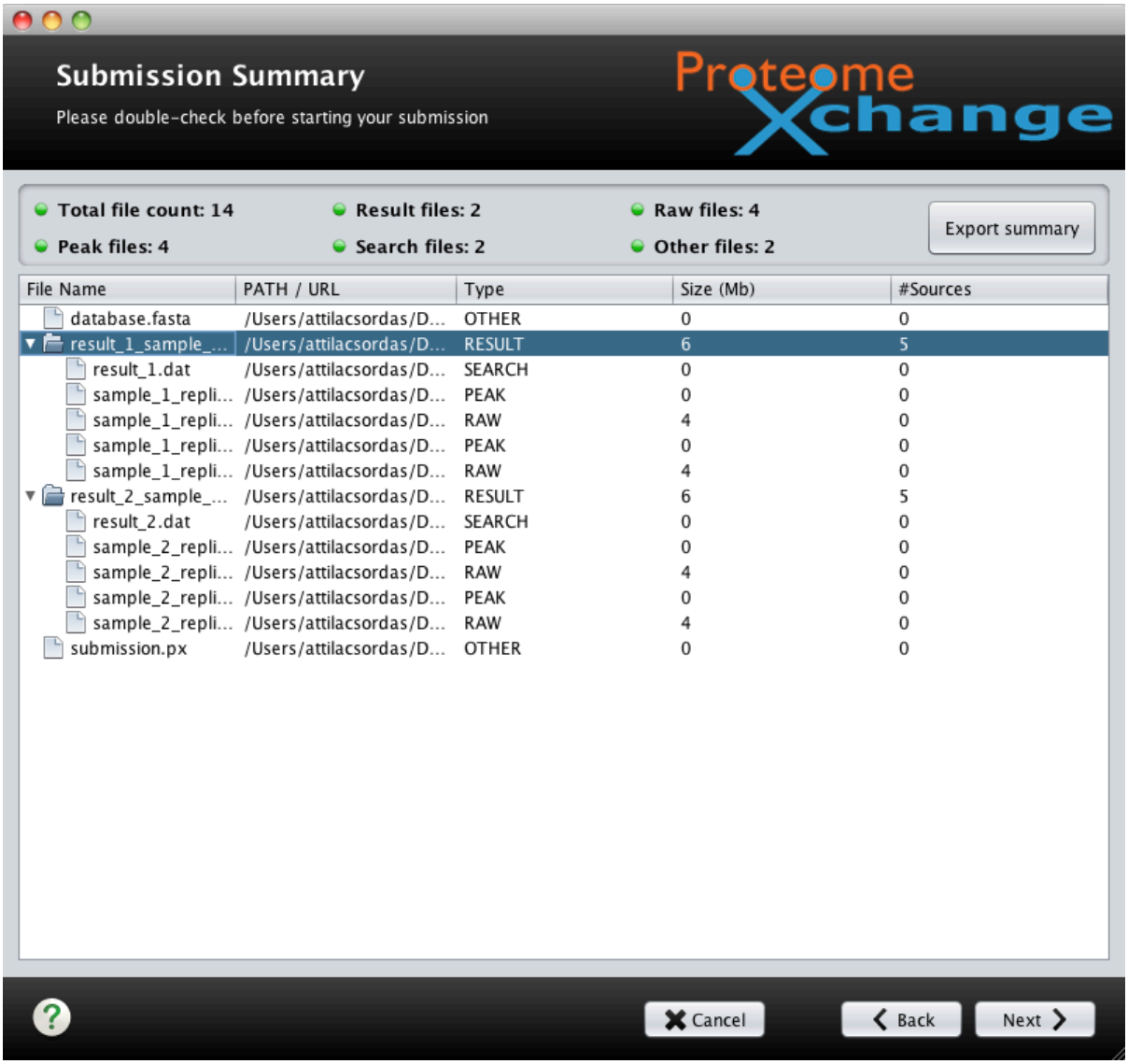


Supp. Figure 8. 'Lab head' screen of the PX Submission tool. It captures the contact details of the "lab head" or principal investigator.

Supp. Figure 9. 'Additional metadata' screen of the PX Submission tool. Optional step to capture details about parent project (if available), publication (in case the corresponding manuscript has been already published), reanalysis (if the submitted data is a reanalysis of a previous PX dataset) and other related 'omics' datasets.



Supp. Figure 10. 'Submission Summary' screen of the PX Submission tool. This is the last screen before the submission of the files actually starts. The submitter gets an overview of the files included in the submission.



8. Sustainability

In recent years, two major proteomics repositories, Peptidome (http://www.nlm.nih.gov/pubs/techbull/jf11/jf11_ncbi_reprint_sra.html) and Tranche[21], have ceased to operate, raising concerns about the long term sustainability of proteomics repositories like those forming the core of PX. Clearly data repositories require continuous funding for continuous operation, and PX partners are no exception.

While the past is not always a good predictor of the future, it often is the best we have. Both PRIDE and PeptideAtlas are well established community resources with a long track record, first publications in 2005[22, 23], and backing by strong organisations (EBI and ISB, respectively).

PRIDE has a limited amount of EBI core funding, ensuring at least basic operations. In addition, the project currently has a healthy mix of national and European support, including a Wellcome Trust Biomedical resource grant recently renewed for 2014-2017. Actual EBI data management infrastructure used by PRIDE (among other, larger resources) is supported by a UK Large Facilities Capital Fund grant with a runtime until 2021.

At least as important as actual funding prospects, PX comprises the principles of fully open data and mutual backup. All PX data is fully open access, which means that in case one partner ceases operations, any interested party can take over the data dissemination without hindrance by copyright. A commitment to transfer of data custodianship from one PX partner to the other in case of funding problems is agreed as the principle of mutual backup. In fact, this principle has already been informally implemented for Peptidome as an associated PX partner. While Peptidome data is still available from the NCBI FTP server, the complete dataset has been imported into PRIDE, reannotated, and is now a fully searchable part of PRIDE[24].

Based on a combination of stable funding and mutual support agreement among the PX partners, we believe any data submitted to PX is very likely to be available as long as required by the community, and we encourage all data producers to deposit their data in PX to increase discovery, reuse, citation, and public benefit of their data.

9. Membership to PX

The consortium is completely open to the participation of additional resources. Individual resources can join PX by implementing the PX data submission and dissemination guidelines, and metadata requirements. By August 2013, a list of format requirements for new members is being drafted and it is expected that it will be agreed in the near future, based on the existing one for IMEX, the consortium of molecular interactions databases (http://www.imexconsortium.org/).

However, it is important to highlight that other resources have already expressed a strong interest in joining the consortium. One example is the new MassIVE repository, led by Dr. N. Bandeira (included as an author in the manuscript) at University of California, San Diego.

10. Security model

The security model is implemented independently by each repository, both of which (PRIDE and PeptideAtlas/PASSEL) have an established track record in managing pre-publication, confidential data.

PRIDE allows data to be kept private for any duration of time, until the owner of the data (as identified by the associated PRIDE user account, see above) gives explicit permission to release the data. A variant occurs when privately submitted data are associated with a manuscript submitted to a journal. The public availability of the submitted data will then be coordinated with the publication of the associated article in correspondence with the journal editor. For PASSEL, data become automatically available on the date that the submitter specifies.

PRIDE and PASSSEL can automatically provide reviewer accounts for each submitted experiment, which can be communicated to journal editors and referees in a submitted manuscript, thus allowing confidential reviewing of the privately submitted data.

The date of submission, as well as the date of public release, is archived in the PRIDE database system. After public release of the data, the PRIDE experiment and the corresponding raw data files (and other files potentially included in the submission) will be made available to the general public without further reservations. The original ownership of the data will remain asserted in the PRIDE database, however. Any restrictions on data dissemination or reuse are obviously removed upon public availability of the data.

To summarize, datasets are private by default. At submission time, users are provided with a username and password that they can use to access the datasets during the manuscript review process. This is the default security implemented in any biological database with the exception of those dealing with personal information (e.g. the European Genotype Archive at the EBI, https://www.ebi.ac.uk/ega/).

11. List of relevant URLs

This is a list of URLs to the most updated documentation and resources:

A) General

- PX home page: http://www.proteomexchange.org

B) Data submission

- PX submission guidelines: http://www.proteomexchange.org/sites/proteomexchange.org/files/documents/px_general_guidelines.pdf

- PX submission tool: http://www.proteomexchange.org/submission

- PX submission tool tutorial:

http://www.proteomexchange.org/sites/proteomexchange.org/files/documents/px_submission_tutorial.pdf

- Web course "How to submit MS/MS data to PX *via* PRIDE" (EBI E-learning platform):

http://www.ebi.ac.uk/training/online/course/pride-submissions-proteomexchange

- How to do bulk submissions for MS/MS data:

http://www.proteomexchange.org/bulk-submission

- PX submission tool file format:

http://www.proteomexchange.org/sites/proteomexchange.org/files/documents/proteomexchange_submission_summary_file_format.pdf

- PASSEL submission form: https://db.systemsbiology.net/sbeams/cgi/PeptideAtlas/PASS_Submit?datasetType=SRM

C) Data access

- ProteomeCentral: proteomecentral.proteomexchange.org/

- PX XML schema: http://code.google.com/p/proteomexchange/source/browse/schema/proteomeXchange-1.1.0.xsd.

12. Abbreviations

CSV: Comma Separated Values

CV: Controlled Vocabulary

DOI: Digital Object Identifier

FDR: False Discovery Rate

GUI: Graphical User Interface

LIMS: Laboratory Information Management System

MIAPE: Minimum Information About a Proteomics Experiment

NCBI: National Center for Biotechnology Information

PASS: PeptideAtlas Submission System

PASSEL: PeptideAtlaS SRM Experiment Library

PRIDE: PRoteomics IDEntifications (database)

PX: ProteomeXchange

RSS: Rich Site Summary

TPP: Trans Proteomics Pipeline

TSV: Tab Separated Values

URL: Uniform Resource Locator

XSD: XML Schema Definition

13. References

1. Deutsch, E.W. File formats commonly used in mass spectrometry proteomics. *Mol Cell Proteomics* **11**, 1612-1621 (2012).
2. Martens, L. et al. mzML--a community standard for mass spectrometry data. *Mol Cell Proteomics* **10**, R110 000133 (2011).
3. Deutsch, E.W. et al. TraML--a standard format for exchange of selected reaction monitoring transition lists. *Mol Cell Proteomics* **11**, R111 015040 (2012).
4. Jones, A.R. et al. The mzIdentML data standard for mass spectrometry-based proteomics results. *Mol Cell Proteomics* **11**, M111 014381 (2012).
5. Walzer, M. et al. The mzQuantML data standard for mass spectrometry-based quantitative studies in proteomics. *Mol Cell Proteomics* (2013).
6. Taylor, C.F. et al. The minimum information about a proteomics experiment (MIAPE). *Nat Biotechnol* **25**, 887-893 (2007).
7. Ghali, F. et al. A toolkit for the mzIdentML standard: the ProteoIDViewer, the mzidLibrary and the mzidValidator. *Mol Cell Proteomics* (2013).
8. Cote, R.G. et al. 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 (2012).
9. Hakkinen, J., Vincic, G., Mansson, O., Warell, K. & Levander, F. The proteios software environment: an extensible multiuser platform for management and analysis of proteomics data. *J Proteome Res* **8**, 3037-3043 (2009).
10. Gluck, F. et al. EasyProt - An easy-to-use graphical platform for proteomics data analysis. *J Proteomics* **79C**, 146-160 (2012).
11. Chambers, M.C. et al. A cross-platform toolkit for mass spectrometry and proteomics. *Nat Biotechnol* **30**, 918-920 (2012).
12. Sturm, M. et al. OpenMS - an open-source software framework for mass spectrometry. *BMC bioinformatics* **9**, 163 (2008).
13. Prieto, G. et al. PAnalyzer: a software tool for protein inference in shotgun proteomics. *BMC bioinformatics* **13**, 288 (2012).
14. Tabb, D.L., Fernando, C.G. & Chambers, M.C. MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. *J Proteome Res* **6**, 654-661 (2007).
15. Dasari, S. et al. Pepitome: evaluating improved spectral library search for identification complementarity and quality assessment. *J Proteome Res* **11**, 1686-1695 (2012).
16. Dasari, S. et al. TagRecon: high-throughput mutation identification through sequence tagging. *J Proteome Res* **9**, 1716-1726 (2010).
17. Ma, Z.Q. et al. IDPicker 2.0: Improved protein assembly with high discrimination peptide identification filtering. *J Proteome Res* **8**, 3872-3881 (2009).
18. Cote, R.G. et al. The PRIDE Converter 2 framework: an improved suite of tools to facilitate data submission to the PRIDE database and the ProteomeXchange consortium. *Mol Cell Proteomics*, in press (2012).
19. Medina-Aunon, J.A., Carazo, J.M. & Albar, J.P. PRIDEViewer: a novel user-friendly interface to visualize PRIDE XML files. *Proteomics* **11**, 334-337 (2011).
20. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* **26**, 1367-1372 (2008).

21. Smith, B.E., Hill, J.A., Gjukich, M.A. & Andrews, P.C. Tranche distributed repository and ProteomeCommons.org. *Methods Mol Biol* **696**, 123-145 (2011).
22. Martens, L. et al. PRIDE: the proteomics identifications database. *Proteomics* **5**, 3537-3545 (2005).
23. Deutsch, E.W. et al. Human Plasma PeptideAtlas. *Proteomics* **5**, 3497-3500 (2005).
24. Csordas, A. et al. From Peptidome to PRIDE: public proteomics data migration at a large scale. *Proteomics* **13**, 1692-1695 (2013).