Data Submission Guidelines for the ProteomeXchange Consortium

This document aims to provide detailed guidelines for the users to submit mass spectrometry (MS) derived proteomics data to the ProteomeXchange (PX) Consortium of proteomics resources (1) (http://www.proteomexchange.org).

Table of contents

1 Types of dataset submissions

The PX resources support two types of dataset submissions, depending on the different proteomics data workflows and the data formats available.

a) **Complete submission:** A complete (also known as "supported") submission ensures that the identification results and the corresponding mass spectra (see definitions of data types in Appendix I) can be parsed, integrated and visualised by the PX resource and/or in free-to-use stand-alone tools such as PRIDE Inspector (available at https://github.com/PRIDE-Toolsuite/pride-inspector/releases). To achieve that, processed identification results need to be provided in a standard format (e.g. mzIdentML (2), mzTab (3)), and optionally as well in a different open data format (e.g. PRIDE XML). In addition, all the submitted files are made available to download.

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

2 Proteomics data resources in ProteomeXchange

In the ProteomeXchange Consortium there are currently two types of proteomics data resources (Figure 1):

a) ***Archival*** **resources**: Their main mission is to store MS based proteomics data. There are two types:
   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 data dependent acquisition (DDA) workflows). The current examples in the Consortium are PRIDE Archive, MassIVE and jPOST (see Section 2.1).
   b. *Focused* resources: They support specifically one type of data workflow and will not store data from other proteomics approaches. An example is the PASSEL component of PeptideAtlas, which is the representative for Selected Reaction Monitoring (SRM) approaches (see Section 2.2).

b) ***Secondary data*** **resources**: These ones build upon the primary data provided by submitters, which are stored in the *Archival* resources. There are two representative resources: PeptideAtlas and MassIVE (see Section 2.3). MassIVE is then both an *Archival* and a *Secondary data* resource.

2.1 List of Universal Archival resources

Currently, there are two *universal* Archival resources available:

1- PRIDE Archive (http://www.ebi.ac.uk/pride/archive, EMBL-European Bioinformatics Institute, Cambridge, UK). Data submission documentation is available here or in this publication (4).

2- MassIVE (https://massive.ucsd.edu/, University of California San Diego (UCSD), San Diego, CA, US). Data submission documentation is available at https://massive.ucsd.edu/ProteoSAFe/help.jsp.

3- jPOST (http://jpost.org/, jPOST Project Team, Japan). Data submission documentation is available at https://repository.jpostdb.org/help.

2.2 List of Focused Archival resources

1- PASSEL (Institute for Systems Biology, Seattle, WA, USA) is the only *focused* resource at present. Data submission documentation available at http://www.peptideatlas.org/passel/.

2.3 List of Secondary Data Resources

PeptideAtlas (http://www.peptideatlas.org/, Institute for Systems Biology, Seattle, WA, USA) is the only secondary data resource at present. Documentation is available at http://www.peptideatlas.org/.

MassIVE (https://massive.ucsd.edu/, University of California San Diego (UCSD), San Diego, CA, USA).

2.4 ProteomeCentral: the common Portal for PX datasets

ProteomeCentral (available at http://proteomecentral.proteomexchange.org) is the portal for 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.

3 Data workflow for original datasets

The overall ProteomeXchange data workflow is summarized in Figure 1.

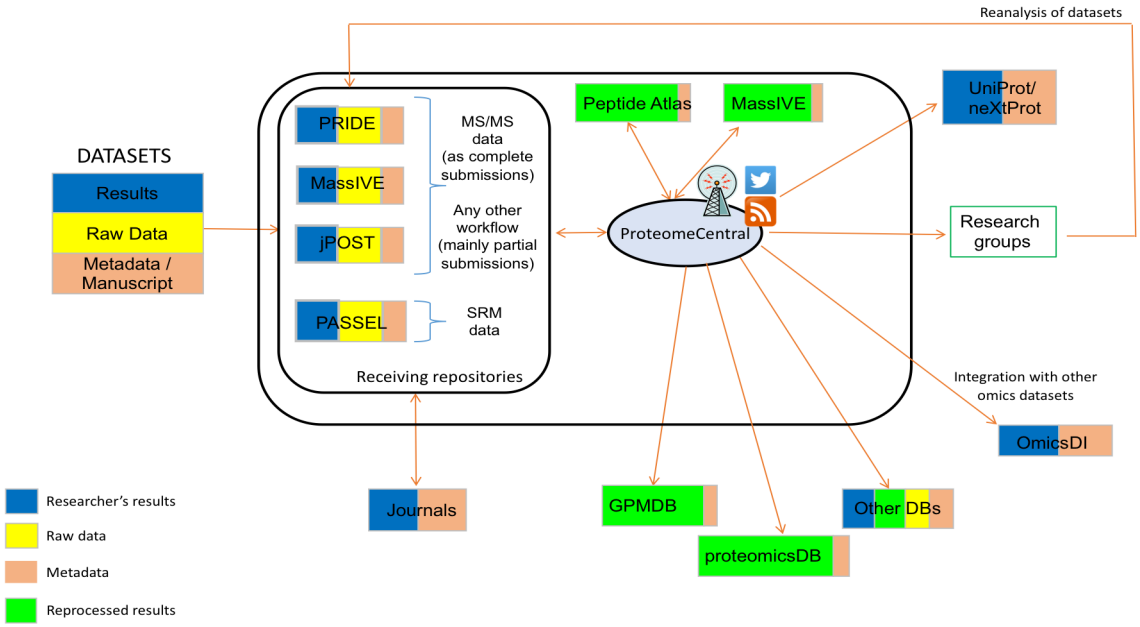

Figure 1: Overview of the ProteomeXchange data flow.

Original datasets coming from any proteomics data workflow can be submitted to any of the *universal* Archival Resources (PRIDE Archive, MassIVE or jPOST). Examples of data workflows are shot-gun (bottom-up) proteomics (Data Dependent Acquisition, DDA), top down, or Data Independent Acquisition (DIA) approaches (e.g. SWATH-MS), among many others. All of the submitted datasets will get a unique PXD identifier (see details at http://www.ebi.ac.uk/miriam/main/collections/MIR:00000513).

However, it is highly RECOMMENDED that datasets from data workflows explicitly supported by existing *focused* archival resources, other than shot-gun proteomics (the most universal and widely used approach), are submitted to that resource, and not to any of the *universal archival* resources. At present, SRM/MRM datasets should be submitted to PASSEL (the only PX resource of this type). The same recommendation will be implemented for additional proteomics approaches if other *focused resources* are included in the Consortium in the future.

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.

In any case, for each submitted PX dataset it is mandatory to include the following data types:

(i) Mass spectrometer output files (see Appendix I).
(ii) Protein/ peptide identifications. Depending on the type of submission, 'supported identification results' (e.g. mzIdentML) will be needed for complete submissions. In the case of Partial submissions, any type of search engine output files are supported.
(iii) Processed peak list spectra formats. These files are needed to enable the connection between the identifications and the mass spectra. In the case of complete submissions performed with mzIdentML, these files are mandatory (since peak lists are not included in mzIdentML *per se*). These files are optional in the case of Partial submissions (since mass spectrometer output files are available anyway).
(iv) Metadata: Related biological and technological metadata provide the experimental context. Different resources have different metadata requirements (see individual documentation for each resource), but at least information needs to be provided to be able to generate the PX XML format (used by the ProteomeCentral resource, see Appendix II).

Other optional data types can also be included in a submitted dataset, for instance:
(i) Quantification software output files: Quantification results.
(ii) Gel images.
(iii) Files used to perform the mass spectral searches, either sequence database files or spectral library files.
(iv) Any other data type (e.g. scripts, pdf files, etc).

In addition, a mechanism to submit mass spectrometry imaging data (as a Partial submission) has been described in this publication (5). See Table 1 below for more details.

Table 1. Summary of submission guidelines for each PX resource, depending on the data workflow involved.

	PRIDE	**PASSEL**	**MassIVE**	**jPOST**
DDA MS/MS				
Partial	Yes	No	Yes	Yes
Complete: mzIdentML	Yes	No	Yes	Yes
Complete: mzTab	No	No	Yes	Yes
Complete: TSV	No	No	Yes	No
Complete: PRIDE XML	Yes	No	No	No
Other workflows				
Targeted SRM/MRM	Partial only	Partial and complete	Partial only	Partial only
DIA MS/MS	Partial only	No	Partial and complete	Partial only
Top-down	Partial only	No	Partial only	Partial only
Mass spectrometry imaging	Partial only	No	Partial only	Partial only

3.1 Submission workflow for Selected Reaction Monitoring (SRM) datasets

New datasets acquired *via* SRM should be submitted to PASSEL, as the only focused Archival resource currently supporting this type of approaches.

For such submitted datasets, 3 main items are required:

1. Mass spectrometer output files, preferably raw files (Appendix I).
2. Transition list describing the peptides that the instrument targeted.
3. Analysis results.

Once submissions are received, they are checked by a curator, run through the PASSEL pipeline, and then loaded into the PASSEL database.

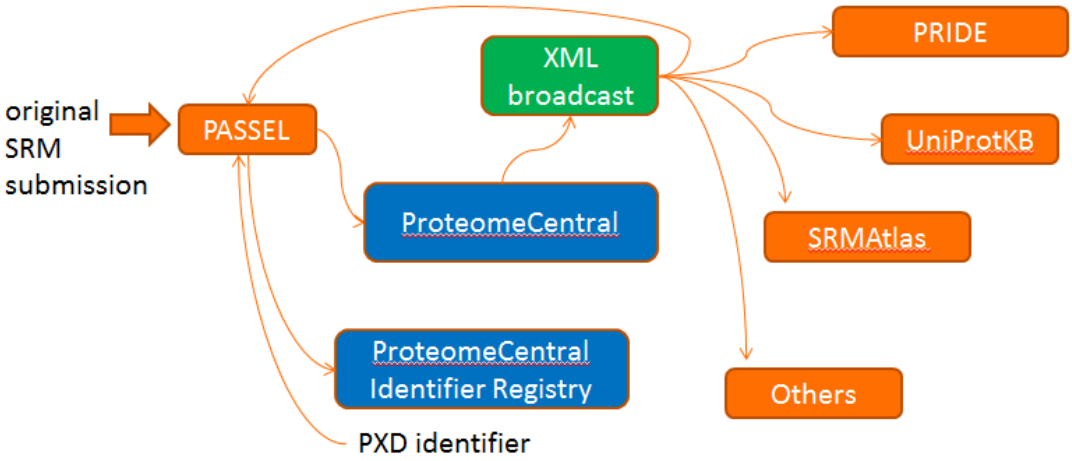

Figure 2. Workflow for original SRM data submissions to PASSEL.

4 Workflow for reprocessed datasets

The workflow for reprocessed datasets starts when any *secondary data* resource of the PX Consortium (at present PeptideAtlas and MassIVE) make a reinterpretation of existing data in any of the *Archival* resources. A new ProteomeXchange identifier will be obtained from ProteomeCentral (it is a RPXD identifier instead of the standard PXD identifier). As an example, see dataset RPXD000665 in ProteomeCentral: http://proteomecentral.proteomexchange.org/cgi/GetDataset?ID=RPXD000665).

However, the original PX accession number is retained in the PX XML message to allow coordinated search for different views of data from one given submission. This ensures that a simple one-time submission from a contributor is automatically distributed to all PX repositories with sufficient information. When the reanalysis is done by a PX member a XML broadcast will be produced, which will include the new PXD identifier, but also the old one. All the relevant information about the connection between the datasets will be stored in ProteomeCentral. Three main situations may arise when a PX dataset is reanalysed:

a) If the data reinterpretation gets published in a separate publication as 'independent' findings:
- Data must go to a *universal Archival resource* (e.g. as any other new MS/MS dataset).
- It is not mandatory to re-upload the raw data (references to URLs are allowed in this case, if this case is supported by the Archival resource).

b) The reinterpretation does not get published as 'independent' new findings. In this case, data can be kept in a *Secondary data resource*. For instance, this applies to all new PeptideAtlas builds that get published.

c) In the case of a mixture of new and reprocessed data in one given dataset, they should be considered to be a new dataset, so the dataset should be submitted to the corresponding *universal Archival resource*.

5 Data ownership

All ProteomeXchange resources do not assume editorial control or ownership over the submitted data; it maintains the original submitter as owner of these data. All ProteomeXchange resources require that a submitter is explicitly identified for each dataset. Upon public availability of the data, the original data ownership is maintained in the database, although obviously dissemination and reuse of the released data are no longer restricted at that point.

PASSEL and jPOST also do not assume editorial control of the submitted data. Users specify at submission time the date on which the data become publicly accessible. On this date, the data are automatically released. The data owner has the option of adjusting this date in case of review delays, etc.

6 Data privacy

All ProteomeXchange resources allow data to be kept private for any duration of time, until the owner of the data (as identified by the associated user account) gives explicit permission to release the data. However, a variant occurs when privately submitted data are associated with a manuscript submitted to a journal. Once the paper is published, the public availability of the corresponding submitted data will then be triggered without asking for permission to the submitters. In the particular case of PASSEL and jPOST, data become automatically available on the date that the submitter specifies.

All PX resources 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.

7 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. 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.
3. 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.
4. 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.
5. 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.
6. Pedrioli, P.G., Eng, J.K., Hubley, R., Vogelzang, M., Deutsch, E.W., Raught, B., Pratt, B., Nilsson, E., Angeletti, R.H., Apweiler, R. *et al.* (2004) A common open representation of mass spectrometry data and its application to proteomics research. *Nat Biotechnol*, **22**, 1459-1466.
7. 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.
8. 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.

8 Appendix I: Data types

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 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 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 (6) (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 (7) (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, 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, OMSSA xml, etc.
In addition to each specific format, a data standard format called mzIdentML (currently v1.1, http://www.psidev.info/mzidentml) (2) 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 (tab delimited file, http://www.psidev.info/mztab) (3) can represent both identification and quantification results.

- **Supported protein/peptide identification results:** This definition includes all protein/peptide identification processed data that can be fully represented by the receiving repository in ProteomeXchange. PRIDE Archive and MassIVE fully support mzIdentML, which can now be exported from a variety of tools (see updated list at http://www.psidev.info/tools-implementing-mzidentml). The PRIDE XML formal is also supported by PRIDE Archive (it was the original PRIDE data format), although it is not RECOMMENDED to use it if the same data can be represented in mzIdentML.
-
- **Quantification software output files**: the data and metadata generated by the software used for performing exclusively the quantification analysis of peptides and proteins. In addition to each specific format from each software tool, a data standard format called mzQuantML (currently v1.0, http://www.psidev.info/mzquantml) has been released by the PSI to represent this kind of information (8). As mentioned before, a second data format called mzTab (http://www.psidev.info/mztab) can also represent quantification results, although is currently not yet finished.

- **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.

9 Appendix II: Metadata and 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.

All elements in the schema are mandatory apart from the last ones (`ChangeLog`, `DatasetFileList`, `RepositoryRecordList` and `AdditionalInformation`). The corresponding .xsd file is available at https://raw.githubusercontent.com/proteomexchange/proteomecentral/master/lib/schemas/proteomeXchange-1.3.0.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.

10 Appendix III: 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.

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 three ways to do it:

1) One can follow the PX Twitter feed @proteomexchange.

2) 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.

3) One can subscribe to the following RSS feed:
http://groups.google.com/group/proteomexchange/feed/rss_v2_0_msgs.xml

11 Appendix IV: Membership in the ProteomeXchange Consortium

Applications for recognition as archival resources are welcome, and will be decided upon by the ProteomeXchange consortium based on the following key criteria:

1. Experience and funding level of resource.
2. Stability.
3. Availability of dedicated curation staff.
4. Ability to store and make accessible raw data, metadata, and interpretations.
5. Worldwide unrestrained availability of stored datasets for download.

The last version of the ProteomeXchange collaborative agreement (which can be found at http://www.proteomexchange.org/documents/proteomexchange-collaborative-agreement) describes the steps needed to become a member of the consortium.