Guidelines for handling ProteomeXchange reprocessed datasets

Version 1.0.2, September 2019.

Table of contents

1.1 A proposed change in terminology

As of December 2017, ProteomeXchange (PX) uses the term "version" to refer to updated PX XML messages that revise the previous PX XML files corresponding to a particular dataset (e.g., to add the PubMed ID). From now on, the term "version" SHALL be deprecated. Henceforth, the term "revision" SHALL refer to technical updates of a submitted dataset and metadata that do not alter the original intent of the submission. Henceforth, the term "reanalysis" SHALL refer to a reprocessing of the original raw data to achieve a new result, either by the same authors, or more commonly by a new group of actors. Each new reprocessing of the same original dataset represents a new "reanalysis". A "revision" implies that the previous instance is obsolete. A "reanalysis" does NOT imply that earlier "reanalyses" are obsolete (Figure 1).

1.2 General principles

For the public announcement of reprocessed datasets, there are two main cases supported: (i) data reprocessing by (or through) PX members (Section 1.2.1), and (ii) data reprocessing by third parties (Section 1.2.2).

A reprocessed PX dataset (RPXD) can correspond to the reprocessing of an individual dataset, a subset of an individual dataset, or a group of datasets (complete or subset). In terms of traceability, it is recommended that all RPXDs reprocess the entire original dataset(s), since an entire dataset is easier to trace back than subset lists of raw files. However, it is indeed permitted that only parts of the original dataset/s are reprocessed. From now on in the text, even if not specified, when the term "groups of datasets" is mentioned, it also refers to "groups of entire or subset datasets", meaning a group of input raw files.

When the term “Raw data” is used, it refers to the files produced originally by the mass spectrometer, or a simple format conversion thereof. Please notice that different types of mass spectrometry files (including the raw data) can be available for a given dataset.

1.2.1 Data reprocessing by (or through) PX members

This first scenario includes those cases when a PX resource directly performs the reanalysis of the datasets (e.g. PeptideAtlas), or when the reanalysis is performed by community members using the infrastructure of a resource (e.g. MassIVE).

In this scenario, each new reprocessed dataset (an individual one or any unique group/combination of specific datasets) gets a unique RPXD identifier, assigned by ProteomeCentral (PC). These RPXD identifiers can only be issued to PX members. The integer number sequence generator associated with PXDs and RPXDs is shared, so an RPXD will never have the same number as a PXD dataset. For instance, it is not possible to have two datasets, one being PXD001000 and the other RPXD001000.

The PX XML message for each RPXD MUST explicitly provide a link to the originally submitted dataset (or group of datasets) used in the reanalysis. This allows easy reference to the original datasets. However, the original PX XML messages for the source datasets available in PC will need to be updated as well. The PC database will maintain the reverse associations, and will generate a PX XML message “on request” so that for any PXD dataset, a user can see in PC or request via an Application Programming Interface (API) all RPXDs that list the original PXD as a source. Therefore, the original PX XML files will not be updated by the receiving repositories (those hosting the original versions of the submitted datasets).

Origin of reprocessed datasets

RPXD identifiers can be generated from previous PX datasets (the ideal situation), but also from datasets not originally submitted or available in any of the PX repositories. Therefore, there can be different scenarios with regard to the origin of the reprocessed datasets (either individually or as a part of a group of datasets), as follows:
a)- Ideally, the original dataset SHOULD have a PXD identifier, in which case it MUST be indicated as the dataset origin.
b)- If no PXD identifier is available, an alternative identifier from any of the PX resources MAY be used, for instance a PeptideAtlas, MassIVE, jPOST, iProX or PRIDE identifier. This could potentially happen in two cases:

- Recently submitted datasets that do not comply with all PX requirements.
- Older datasets submitted before PX was started (before 2012).

c)- If no alternative dataset identifier is available from any of the PX resources, the original raw files need to be re-uploaded to a PX resource. Two options are possible:

(i)-Create a PXD dataset, indicating the original data producers. In this case, the original authors should be contacted to obtain the original processed result files as well. All reanalyses of the dataset would then be in scenario (a) as indicated above.
(ii) - Create a RPXD dataset containing the spectrum files, clearly indicating the source of the data but listing as main contacts the people that perform the reanalysis. The processed results will be those corresponding to the reprocessing. One example of this case would be those datasets produced by the CPTAC, that are available in the CPTAC portal.

In both cases, if available, the URL from which the files of the original dataset(s) were obtained should be reported and any related PubMed IDs should also be included.

Requirements for generating and storing reprocessed datasets

A graphical summary of the guidelines can be seen in Figure 1. The following rules apply:
1- Processed new results and raw data MUST be available as part of the reprocessed dataset (as for any PXD), and they MUST be stored in the PX resource responsible for the reanalysis.

2- However, raw files do not need to be duplicated in the PX member repository file system (e.g. both in the original and reprocessed dataset/s). Symbolic or hard links MAY be used.

3. The level of metadata annotation for each RPXD dataset must comply with the PX XML requirements in effect at the time of submission of the RPXD. The use of terms like “Information not available” (or analogous ones) for the existing mandatory fields in the PX XML IS NOT allowed. This may require the collection of additional metadata over what was available in the PXD record.

4. As indicated above, it is encouraged that all RPXDs reprocess the entire original dataset(s). In that case, the following CV term can be used (“Reprocessed complete dataset”, MS:1002861). However, it is permitted that only part of the original dataset is reprocessed. If this is done, then the dataset must be annotated with the CV term “Reprocessed subset dataset” (MS:1002862) in the PX XML file.

5- Any subsequent “reanalysis” of a dataset performed or submitted through the same PX member (e.g. due to a more recent reference protein sequence database) will be considered as a new “reanalysis”. A new RPXD identifier SHOULD NOT be generated in that case. See below for how this scenario should be implemented.

6- However, if a PXD (or group of PXDs) is reprocessed by different PX members, a different RPXD identifier MUST be used by each member. The original PXD will be then linked to each RPXD.

7- If different groups of raw files are reprocessed and included under the same RPXD identifier (e.g. in the case of different “subsets of datasets”), each “Reanalysis” MUST clearly indicate the exact list of raw files used. In any case, it is obvious then all the raw files included in a given “Reanalysis” MUST have been used in a previous “Reanalysis” instances.

8- In the current implementation of PC, it is only easy to access the latest version of each PX XML document. The PC interface SHALL be changed to enable the linking and access to all previously announced PX XML document instances for any PXD or RPXD dataset.

9- The same concept of “complete” or “partial” submissions used for original PXD datasets apply. The term “Complete” dataset applies when it is possible for the resource, to parse and link the identification results with the originating mass spectra. The “Partial” submission term applies when identification results and mass spectra cannot be linked, so files are just made available to download. It is discouraged, but explicitly permitted that an RPXD be a “partial” submission (as of December 2017, this would be applicable e.g. for MaxQuant and/or OpenMS).

Handling and versioning of reprocessed datasets

Each “reanalysis” will be modeled as a member of a container. Each container represents all the “reanalyses” of a unique set of original datasets (either an individual one or a group of them), performed by a specific PX member repository to which the RPXD accession is assigned. See Figure 1 for more details.

Each “reanalysis” of the datasets referenced by the container will be assigned a unique “reanalysis number”, which allows individual “reanalyses” to be uniquely identified. Individual “reanalyses” should be identified as follows:

RPXD000001.x (where *RPXD00001* is the container and *x* is the reanalysis number)

In this way, it is possible to refer to both each specific “reanalysis”, as well as to all “reanalyses” of the datasets in the container submitted through a PX resource. If the “reanalysis” number is not used or simply lost due to parsing issues, access to the container will list all attached individual “reanalyses”.

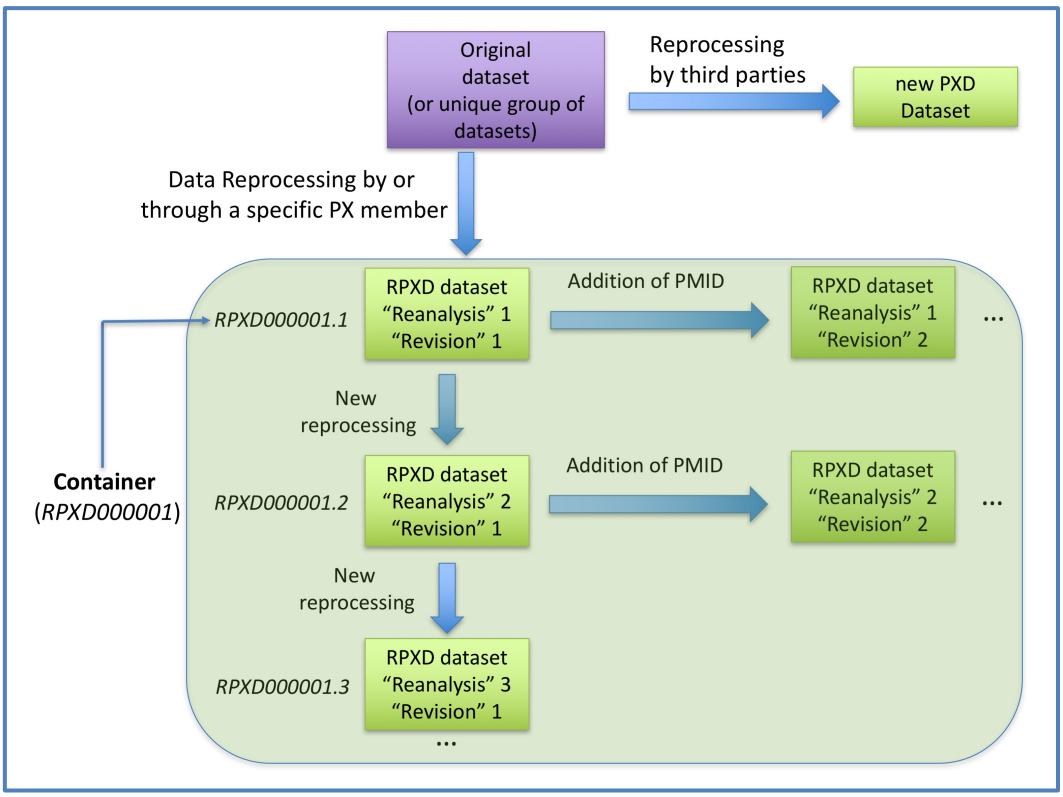

Figure 1. Graphical summary of the guidelines for reprocessed datasets.

The first instance of a "reanalysis" will be assigned "reanalysis" number 1. Therefore all RPXD containers will at least have one instance. Every new "reanalysis" of an RPXD will be submitted separately to PC with a new "reanalysis" number. The general order of operations will be as follows: a member repository will request an RPXD identifier. At this point, PC will only know that a RPXD identifier belongs to a specific PX member. When the member repository is ready, it submits its first PX XML file as "reanalysis" 1, thereby creating the container and the container's first "reanalysis" (Figure 1).

A "reanalysis" MAY also undergo a "revision" (like PXDs MAY, e.g. due to the addition of a PMID) to repair an error or insufficiency in the original document. However, as with PXDs, a "revision" only corrects or enriches the data or metadata without altering the intent of the "reanalysis" (Figure 1).

1.2.2 Data reprocessing by third parties

At present, reprocessed PX datasets will only get a RPXD identifier when the reprocessing is provided by one of the PX members. In case the reprocessed datasets are submitted by third parties, they must be submitted to one of the PX members following the default procedure for original datasets (Figure 1). We aim that in the near future the RPXD mechanism can also be used by third parties. However, before

this happens, these guidelines will need to be extended to cover the new possible scenarios when external parties to the Consortium are involved. Last but not least, PX members will need to implement these changes in their dataset submission software.

However, researchers submitting a reprocessed dataset MUST link explicitly to the originally submitted dataset (i.e. PX submission processes MUST allow submitters to reference a source PXD). The reprocessed datasets will get a PXD identifier and will be linked to the original dataset(s), as described in the previous section.

1.3 Guidelines for the annotation of the reprocessed datasets and for the generation of the PX XML messages

In the PX XML files generated coming from reprocessed datasets, the following information MUST be present:

1- Add information to the PX XML message about how the data was reprocessed.

2- Use as title of the dataset: “Reprocessed dataset: “, keeping the original title of the submission where the original raw data come from, in cases of a single source dataset. If there are multiple source datasets, or only part of a source dataset, then the original title may be altered to reflect the change. In case the reprocessing is also ‘Quantitative’, it is recommended to use instead “Reprocessed quantitative dataset:”.

3- In the <DatasetOrigin> element the following CV term must be used (PRIDE:0000397 or MS:1002863, “Data derived from previous dataset). There, the original datasets that were reprocessed must be included. Starting with PX XML version 1.3.0 there can be multiple elements <DatasetOrigin>. This enables the linking between original and reprocessed dataset(s) in PC. Also, it enables that two different identifiers for the same dataset can be used (for instance the corresponding PXD and MassIVE identifiers for a given dataset).

4- The detailed description of the reprocessing should be appended to the original dataset description with a clear marker “REPROCESSING METHODS:”. This may obsolete some information in the original part of the description, but this is okay as it may allow viewers to grasp how the “reanalysis” differs from the original or previous “reanalyses” (assuming both original and new descriptions are adequate).

5- The reference included in the RPXD dataset, if applicable, MUST be the one corresponding to a new publication, not the reference/s corresponding to the original dataset/s. If the reprocessed dataset is not published independently in a peer-reviewed article, a publication MUST NOT be specified. In addition, in the latter case, the <ReviewLevel> element in the PX XML file needs to be set to “Non-peer-review dataset”, irrespective of the review level of the original dataset.

6- The “lab head” and submitter related information will change to reflect the group responsible of the data reprocessing (not the original group who performed the submission of the dataset).

7- The use of terms like “Information not available” (or analogous ones) for the current existing mandatory fields in the PX XML file IS NOT permitted.

8- In case of multiple “reanalyses” associated with the same RPXD, the latest PX XML file MUST be displayed by default in PC, but all previous instances SHOULD also be available.

9- For PXD datasets, the “ChangeLog” field is used to explain what is different from the previous “revision”. In the case of RPXDs, “ChangeLog” should very briefly explain why the “reanalysis” was done (or ideally what it is unique about a particular “reanalysis”). For instance, a “ChangeLog” may simply be “Automated reprocessing with newer software and reference proteome” or “Reanalysis triggered by MassIVE user”. Detailed metadata about the analysis protocol (perhaps auto-generated text based on the reprocessing workflow) SHALL BE appended to the data description as described above, and not encoded in the “ChangeLog”. The “ChangeLogs” for the individual ‘Reanalyses’ SHOULD include the changes made only in that particular instance, not all the trail including the summary of the changes done in the “Container” as a whole. PC may be able to implement a mechanism to summarize and put together all the “ChangeLogs” for all ‘Reanalyses’.