

---

# *PRIDE Converter 2* User Guide

---

## Table of Contents

1. Introduction	3
2. <i>PRIDE Converter 2</i> Requirements	3
2.1. System Requirements	3
2.2. Additional Requirements	3
3. <i>PRIDE Converter 2</i> Tool Suite GUI User Guide	4
3.1. <i>PRIDE Converter 2</i>	5
3.1.1. Format Selection	5
3.1.2. File Selection	6
3.1.2.1. FASTA – First word	7
3.1.2.2. FASTA – Full ID line	7
3.1.2.3. FASTA – Match UniProt AC	8
3.1.2.3. FASTA – Match UniProt ID	8
3.1.3. Experimental Details	9
3.1.4. Sample Description	9
3.1.5. Protocol Description	10
3.1.6. Instrument Description	11
3.1.7. Software Processing	11
3.1.8. PTMs	12
3.1.9. Review	12
3.1.10. File Export	13
3.1.11. Validation Report	14
3.1.12. Helpful Tools	15
3.2. <i>PRIDE mzTab</i> Generator	15
3.2.1. Format Selection	15
3.2.2. File Selection	15
3.2.3. Generation Report	16
3.3. <i>PRIDE XML</i> Filter	16
3.3.1. Filter Information	16
3.3.2. File Selection	17
3.3.3. File Export	17

3.3.4. Filter Report	18
3.4. PRIDE XML Merger	18
3.4.1. Merger Information	18
3.4.2. File Selection	18
3.4.3. Master File Selection	19
3.4.4. Merger Report	19

# 1. Introduction

---

The PRIDE Converter 2 tool suite 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. These skeleton files require either manual or scripted editing to add quantitation and/or gel information.
- 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.

All tools have both a Graphical User Interface (GUI) and a Command-Line Interface (CLI). The GUIs have been designed to provide a rich, user-friendly interface while the CLIs have been developed mainly for tool and pipeline developers to be able to integrate the PRIDE Converter 2 tools in their own software to provide an efficient way to generate PRIDE XML from their own resources.

## 2. PRIDE Converter 2 Requirements

---

### 2.1. System Requirements

---

- Java: JRE 1.5 +
- CPU: 1 gigahertz (GHz) or faster 32-bit or 64-bit processor.
- Memory: 1 gigabyte (GB) RAM.
- Hard Disk: 55 MB available for installation, more if doing file conversions.
- Platform: Tested in Mac OS X, Linux, and Windows (XP, Vista, 7).

### 2.2. Additional Requirements

---

PRIDE Converter 2 requires Internet access for connection to the Ontology Lookup Service (OLS) web service and automatic access to PubMed records already published.

### 3. PRIDE Converter 2 Tool Suite GUI User Guide

To start the GUI, users must either double-click on the *PRIDE Converter 2* jar file or invoke the command-line without any arguments. The first screen that the user sees will be the tool selection screen. From there, the user can select which tool from the tool suite to launch (Figure 2).



Figure 2 – PRIDE Converter Tool Suite tool selection form

Each tool shares a common design framework to ensure a common and consistent look and feel across all applications (see Figure 3).

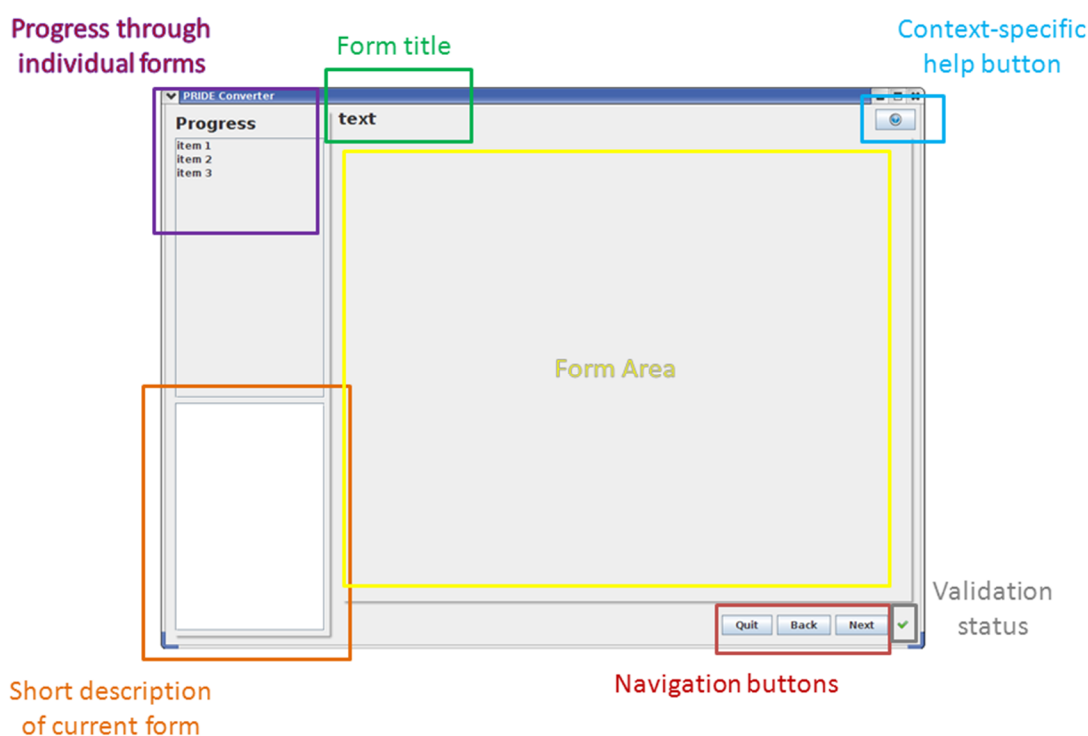


Figure 3. The Common GUI Navigation Panel.

Tools are made up of a series of forms, that can be navigated using the Next/Back/Quit buttons at the bottom right corner of the navigation frame. Each form will provide information that is used to update the GUI: a form title, a short description of its purpose and required fields, a context-specific help page and a mechanism to validate the form content to ensure that all required fields are filled in and contain correctly-formatted values.

The result of this validation are shown next to the navigation buttons (✅ if all fields are OK, ⚠️ on warnings, ❌ on errors). If errors are detected, the users will be shown an error dialog with more information on the validation rule that generated the error. Users will not be able to proceed with the conversion process until all errors are corrected.

PRIDE Converter 2 forms maintain a standard functionality scheme. Any field marked with a red asterisk (\*) is mandatory and the user cannot proceed to the next form until all mandatory fields are filled. The ➕ button can be used to add entries. When entries are selected from a table, the ✎ button can be used to edit those entries and the 🗑 button can be used to delete them.

### 3.1. PRIDE Converter 2

When launched in GUI mode, *PRIDE Converter 2* guides the user through a 12-step process to convert their search engine output files into well-annotated PRIDE XML files.

#### 3.1.1. Format Selection

Input format selection is the first choice that the user must make. When selecting a format, the user will be given information on the exact file type(s) required, as well as links to the vendor and/or format specification documentation (Figure 4). Please note that Mascot conversion is not available on the Mac OS platform, due to limitations imposed by the Mascot API itself. Please refer to Table 1 for the supported formats in *PRIDE Converter 2*.

Figure 4 – *PRIDE Converter 2* Format Selection form

**Table 1.** Supported Formats in *PRIDE Converter 2*.

Format Name	File Type	Data Content
Mascot	.dat	Spectra and Identifications
mzIdentML	.xml	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
mzML	.xml	Spectra Only
DTA	.dta	Spectra Only
MGF	.mgf	Spectra Only
mzData	.xml	Spectra Only
mzXML	.xml	Spectra Only
PKL	.pkl	Spectra Only

### 3.1.2. File Selection

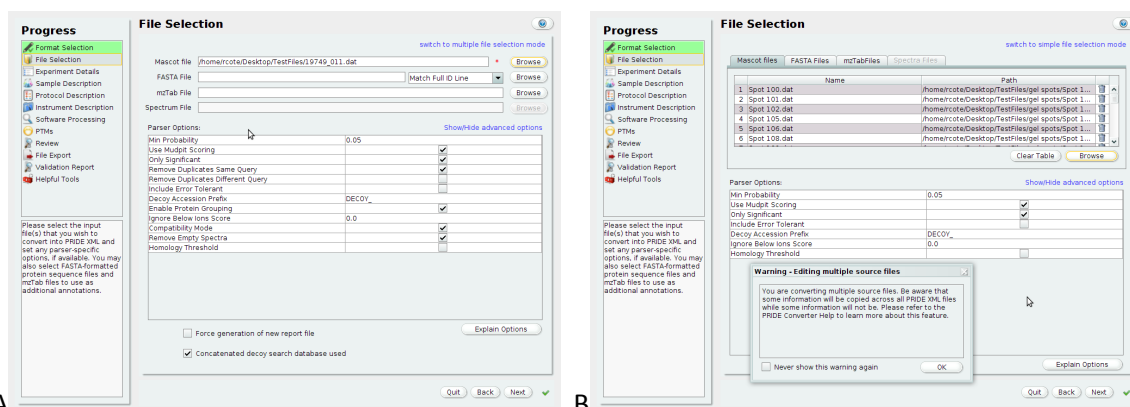
The next form is for file selection and setting DAO options, if applicable (Figure 5). Each format-specific DAO can have one or several custom options that can be set through the GUI (for example, the Mascot DAO, as shown in Figure 5). Sensible default options are always provided and only the basic required options are shown by default. Power users have the option to show all available options (if applicable) and those options will be stored in the report file such that the user can always review how the conversion process was configured. Tool tips will provide a description of the options and sensible/default values. Currently, only the peaklist-only DAOs do not have any options.

Some MS search result file formats link to spectra that are kept in external files (X!Tandem, mzIdentML and CRUX for example). For those DAOs, the user has the possibility to select a file containing the spectra to include in the conversion process. For the DAOs where externally-linked spectra are not applicable, the option to select spectra files is disabled.

By default, users will be shown a form that will only provide single-file selection dialogues but users who wish to do multiple file batch conversion can switch to an alternative view, which will allow them to select multiple files of each type (source, mzTab, spectra, etc).

When converting multiple files at once, a warning message will be displayed to the user indicating that some metadata will be copied across all files while some will not. The information that will be common across all files will include: sample annotations, post-translational modifications (PTM), protocol and instrumentation description and any other experiment-level annotations. Software processing, which can include file-specific software settings, score thresholds and false discovery rates will be taken from the individual report files associated with each input file to be converted.

Once the files have been selected, the *PRIDE Converter 2* runs in *prescan* mode to generate the report files, where the annotations that the user will input in the coming steps will be stored prior to actual conversion. Report file generation can take anywhere between seconds and several minutes and is directly proportional to the size and complexity of the input files. If several files are being converted at once, a report file needs to be generated for each input file, which will compound the time required.



**Figure 5 – PRIDE Converter 2 File Selection form.** Panel A shows the single-file selection mode while panel B shows the multi-file selection mode. Note that panel A shows the full set of options for the Mascot DAO while panel B only shows the default options.

Users have the possibility to link their search database protein sequences to the PRIDE XML files that they generate. This is particularly helpful in the case where custom-made protein sequence databases are used. The protein sequence database must be exported in FASTA format. From experience, we have observed that the protein accession that is assigned to an identification can differ significantly from the FASTA header, depending on if and how the search engine processes the protein sequence database. To solve this issue, PRIDE Converter 2 has several options on how the FASTA file should be processed during conversion.

### 3.1.2.1. FASTA – First word

In this scenario, the search engine assigns the first word of the FASTA ID line as the identification accession. Given a FASTA ID line that looks like the following:

```
>P30443 Full=HLA class I histocompatibility antigen, A-1 alpha chain;
```

The report file identifications, after being enriched with the FASTA information, will look like this:

```
<Identification>
  <Accession>P30443</Accession>
  ...
  <FastaSequenceReference>8</FastaSequenceReference>
</Identification>
```

The *FastaSequenceReference* points to a Sequence element, as illustrated below.

```
<Sequence id="8" accession="P30443">
MAVMAPRTL L L L L L S G A L A L T Q T W A G S H S M R Y F F T S V S R P G R G E P R F I A V G Y V D D T Q F V R F D S D A A S Q K M E P R A P W I E Q E P E Y W D Q E T R N M K A
H S Q T D R A N L G T L R G Y N Q S E D G S H T I Q I M Y G C D V G P D G R F L R G Y R Q D A Y D G K D Y I A L N E D L R S W T A A D M A A Q I T K R K W E A V H A A E Q R R V Y L E G
R C V D G L R R Y L E N G E T L Q R T D P P K T H M T H P I S D H E A T L R C W A L G F Y P A E I T L T W Q R D G E D Q T Q D T E L V E T R P A G D G T F Q K W A A V V P S G E E Q
R Y T C H V Q H E G L P K P L T R W E L S S Q P T I P V G I I A G L V L L G A V I T G A V V A A V M W R R K S S D R K G G S Y Q A A S S D S A Q G S D V S L T A C K V
</Sequence>
```

### 3.1.2.2. FASTA – Full ID line

In this scenario, the search engine assigns the complete ID line FASTA ID line as the identification accession. This occurs when the search engine does not process or index the source FASTA database, and is especially useful when custom databases or custom identifiers are used.

The report file identifications, after being enriched with the FASTA information, will look like this:

```
<Identification>
  <Accession>Custom Fasta Header Protein 1</Accession>
  ...
  <FastaSequenceReference>12</FastaSequenceReference>
</Identification>

...

<Sequence id="12" accession="Custom Fasta Header Protein 1">
MGDREQLLQRLARLAEQAERYDDMASAMKAVTELNEPLSNEDRNLLSVAYKNVVGARRSSWRVISSIEQKTMADGNEKKLEKVKAYREKIEKEL
ETVCNDVLSLLDKFLIKNCNDFQYESKVFYLLKMGDYRYLAEVASGEKKNSVVEASEAAYKEAFEISKEQMOPHTPIRLGLALNFSVFYYEII
QNAPEQACLLAKQAFDDAIAELDTLNEDSYKDSTLIMQLLRDNLTLWTSDDQDEEAGEGN
</Sequence>
```

### 3.1.2.3. FASTA – Match UniProt AC

---

If the FASTA file used contains ID lines that are generated according to the UniProtKB specification, they should look something like the following:

```
>sw|P31946|1433B_HUMAN RecName: Full=14-3-3 protein beta/alpha;AltName: Full=Protein
1054;AltName: Full=Protein kinase C inhibitor protein 1;Short=KCIP-1;Contains:RecName:
Full=14-3-3 protein beta/alpha, N-terminally processed;
```

Some search engines will process these FASTA ID lines and only assign the UniProtKB protein accession. The report file identifications, after being enriched with the FASTA information, will look like this:

```
<Identification>
  <Accession>P31946</Accession>
  ...
  <FastaSequenceReference>9</FastaSequenceReference>
</Identification>

...

<Sequence id="9" accession="P31946">
MTMDKSELVQKAKLAEQAERYDDMAAMKAVTEQGHLSNEERNLLSVAYKNVVGARRSSWRVISSIEQKTERNEKKQMGKEYREKIEAELQ
DICNDVLELLDKYLIPNATQPESKVFYLLKMGDYFRYLSEVASGDNKQTTVSNSQAYQEAFAEISKKEMQPTHTPIRLGLALNFSVFYYEILNS
PEKACSLAKTAFDEAIAELDTLNESYKDSTLIMQLLRDNLTLWTSNQDGEDGAGEGN
</Sequence>
```

### 3.1.2.3. FASTA – Match UniProt ID

---

Using the same FASTA ID line as in the preceding section, some search engines will assign the UniProtKB ID. The report file identifications, after being enriched with the FASTA information, will look like this:

```
<Identification>
  <Accession>1433B_HUMAN</Accession>
  <CuratedAccession>P31946</CuratedAccession>
  ...
  <FastaSequenceReference>9</FastaSequenceReference>
</Identification>

...

<Sequence id="9" accession="P31946">
MTMDKSELVQKAKLAEQAERYDDMAAMKAVTEQGHLSNEERNLLSVAYKNVVGARRSSWRVISSIEQKTERNEKKQMGKEYREKIEAELQ
DICNDVLELLDKYLIPNATQPESKVFYLLKMGDYFRYLSEVASGDNKQTTVSNSQAYQEAFAEISKKEMQPTHTPIRLGLALNFSVFYYEILNS
PEKACSLAKTAFDEAIAELDTLNESYKDSTLIMQLLRDNLTLWTSNQDGEDGAGEGN
</Sequence>
```

Note, in this last case, the presence of a *CuratedAccession* element where the correct UniProtKB accession has been linked to the UniProtKB identifier. Users are strongly urged to use UniProt accessions instead of identifiers, as entry identifiers are notoriously changeable, making data comparisons and downstream analysis more difficult and cumbersome.



### 3.1.3. Experimental Details

The first annotation form deals with experimental and project details (Figure 6). Users must provide a project name, an experiment title and a short label to describe the file being converted. In the case where multiple files are being converted at once, the user needs to provide a project name in this form but will have the opportunity to provide individual experiment titles and short labels in a special form which is only displayed for this purpose. Users need to provide contact information for at least one person – usually the data submitter or the primary investigator. A full name, valid e-mail address and institute contact details are required for each contact. Contacts can be saved as templates to avoid the necessity to re-enter the information at each run of the application. Users also have the option to provide a reference that supports the data being submitted to PRIDE. References can either be provided as free text or use an automatic PubMed ID lookup to retrieve the full citation, if indexed and available online. An Internet connection is required for this.

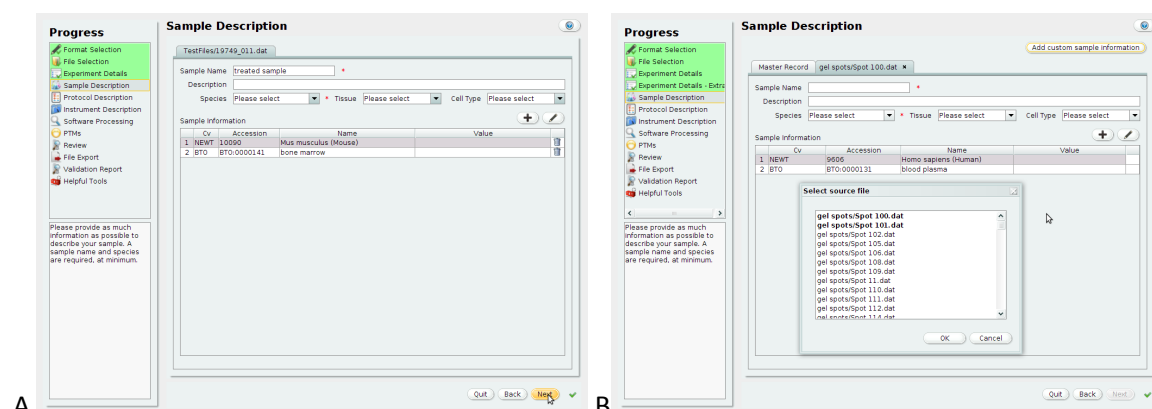
Figure 6 consists of two panels, A and B, showing the 'Experiment Details' form in PRIDE Converter 2. Panel A (left) is for single-file conversion and includes fields for Project Name, Experiment Title, Short Label, and Description. It also has a 'Contacts' table with columns for Name, E-Mail, and Institution, and a 'Reference' table with columns for Reference Line, PubMed ID, and DOI ID. Panel B (right) is for multi-file conversion and features a table with columns for Source File, Experiment Title, and Short Label. An 'Auto-generate' button is present at the top right of the table. Both panels have a 'Progress' sidebar on the left with options like Formal Selection, File Selection, Experiment Details, Sample Description, Protocol Description, Instrument Description, Software Processing, PTMs, Review, File Export, Validation Report, and Helpful Tools.

**Figure 6** – PRIDE Converter 2 Experiment Details form. Panel A shows the single-file conversion mode while panel B shows the additional step that is required to provide experiment titles and short labels for each source file to be converted.

If multiple files are being converted, the project name will be identical across all files, but the combination of experiment title and short label must be unique to each file. Users can auto-generate placeholder data, based on the project name, but are encouraged to fill in these fields with meaningful information.

### 3.1.4. Sample Description

The sample description form (Figure 7) allows the user to provide taxonomy, tissue type and cell type annotations using pull-down menus that contain the most commonly used values already in PRIDE. A sample name and species are required. Users still have the possibility to use the comprehensive Ontology Lookup Service (OLS) to look up alternative terms that are not already provided. If multiple files are being converted, the information in the “master report” will be automatically copied over to each file. Users will have the possibility to add file-specific controlled vocabulary (CV) terms by clicking on the “add custom sample information” button and selecting the proper source file. Files with custom annotations will be shown in bold. Any annotation coming from the master report will not be editable or deletable from the individual files’ panel, but users will be able to add new terms. Any terms added or deleted from the master report view will always be propagated to all files.



**Figure 7 – PRIDE Converter 2 Sample Description form.** Panel A shows the single-file conversion mode while panel B shows the how annotation of multiple files. Note that, in panel B, the entries for the selected individual file cannot be deleted, as they were set in the ‘Master Record’ tab.

### 3.1.5. Protocol Description

In the protocol description form (Figure 8), *PRIDE Converter 2* requires a protocol name to be provided, at minimum. Protocols can have multiple steps, and each step can be described with one or several CV terms. Please look at the templates provided with the tool for examples. The “Load” button can be used to load existing templates. A popup window will be shown, where users can select the desired template from a list. To save a template once all of the information is correctly entered, users simply need to enter a name (in this case, a protocol name) and click on the “Save” button. A file will be saved to the templates folder in the installation directory of *PRIDE Converter 2*. Should there already be a file with the same name, the user will be prompted to either rename the template or allow the existing one to be overwritten.

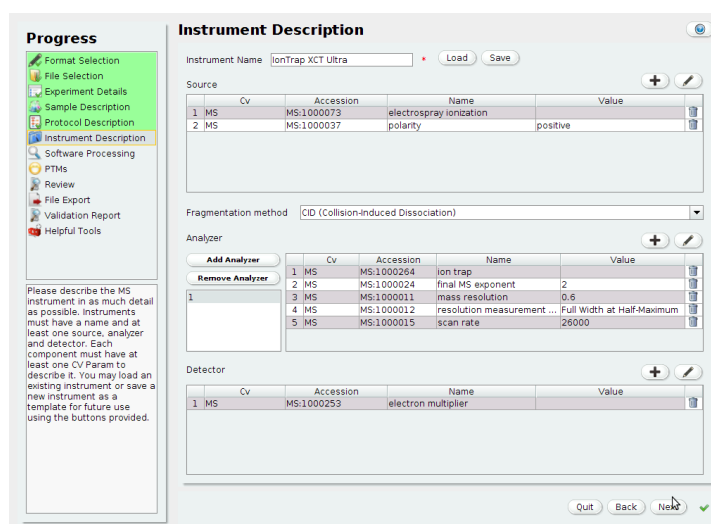
*PRIDE Converter 2* provides the possibility to save commonly-used annotations such as sample, instrumentation and protocols as templates which can be reused in subsequent conversions. A set of basic templates is provided with the tool suite. Users can load and update templates to better suit their own requirements.

Figure 8 shows the 'Protocol Description' form in PRIDE Converter 2. It has a 'Progress' sidebar on the left with the same options as Figure 7. The main area is titled 'Protocol Description' and includes a 'Protocol Name' field (containing 'SDS-PAGE') and 'Load' and 'Save' buttons. Below these are 'Add Step' and 'Remove Step' buttons. A list on the left shows 'Step 1', 'Step 2', 'Step 3', and 'Step 4', with 'Step 3' selected. The main table, 'Step Description', has columns 'Cv', 'Accession', 'Name', and 'Value'. It contains one entry: '1 SEP sep:00142 enzyme digestion Trypsin'. At the bottom are 'Quit', 'Back', and 'Next' buttons.

**Figure 8 – PRIDE Converter 2 Protocol Description form.**

### 3.1.6. Instrument Description

The instrument description form (Figure 9) is another form that can use templates, as in most cases, the instrumentation available to a user will not change dramatically over time. “Source”, “Analyzer” and “Detector” elements need to be annotated, preferably with CV terms from the Proteomics Standards Initiative (PSI) Mass Spectrometry (MS) controlled vocabulary (<http://psidev.cvs.sourceforge.net/viewvc/psidev/psi/psi-ms/mzML/controlledVocabulary/psi-ms.obo>).



The Instrument Description form is divided into two main sections: Progress and Instrument Description. The Progress section on the left contains a list of steps: Format Selection, File Selection, Experiment Details, Sample Description, Protocol Description, Instrument Description (highlighted), Software Processing, PTMs, Review, File Export, Validation Report, and Helpful Tools. Below this list is a text box with instructions: "Please describe the MS instrument in as much detail as possible. Instruments must have a name and at least one source, analyzer and detector. Each component must have at least one CV Param to describe it. You may load an existing instrument or save a new instrument as a template for future use using the buttons provided."

The Instrument Description section on the right has a title bar "Instrument Description" and a "Load" button. It contains three tables: Source, Analyzer, and Detector. The Source table has columns: Cv, Accession, Name, and Value. The Analyzer table has columns: Cv, Accession, Name, and Value. The Detector table has columns: Cv, Accession, Name, and Value. Below the tables are buttons for "Add Analyzer" and "Remove Analyzer". At the bottom are "Quit", "Back", and "Next" buttons.

Cv	Accession	Name	Value
1 MS	MS:1000073	electrospray ionization	
2 MS	MS:1000037	polarity	positive

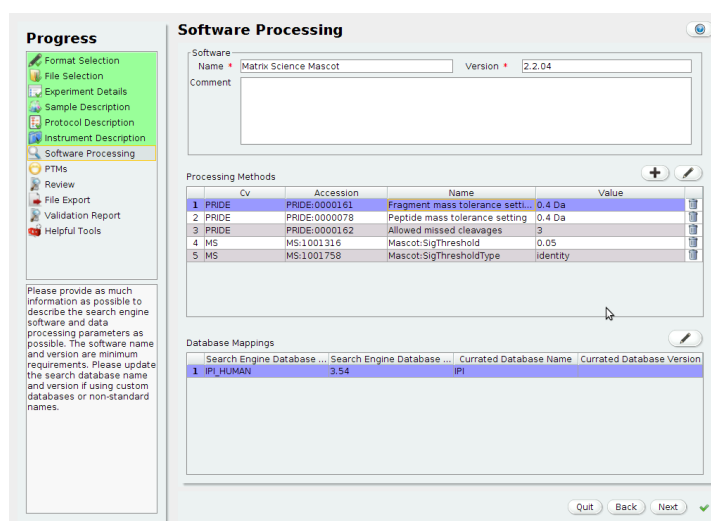
Cv	Accession	Name	Value
1 MS	MS:1000264	ion trap	
2 MS	MS:1000024	final MS exponent	2
3 MS	MS:1000011	mass resolution	0.6
4 MS	MS:1000012	resolution measurement ...	Full Width at Half-Maximum
5 MS	MS:1000015	scan rate	26000

Cv	Accession	Name	Value
1 MS	MS:1000253	electron multiplier	

Figure 9 – PRIDE Converter 2 Instrument Description form.

### 3.1.7. Software Processing

The software processing form (Figure 10) report information on the software settings and the search database used. The “software” and “processing method” information will more often than not be reported directly by the DAO and should not be modified. **Users should be aware that if converting multiple files**, these two sections will be taken directly from the source file and any additions made to the GUI will not be carried across all converted files.



The Software Processing form is divided into two main sections: Progress and Software Processing. The Progress section on the left contains a list of steps: Format Selection, File Selection, Experiment Details, Sample Description, Protocol Description, Instrument Description, Software Processing (highlighted), PTMs, Review, File Export, Validation Report, and Helpful Tools. Below this list is a text box with instructions: "Please provide as much information as possible to describe the search engine software and data processing parameters as possible. The software name and version are minimum requirements. Please update the search database name and version if using custom databases or non-standard names."

The Software Processing section on the right has a title bar "Software Processing" and a "Version" field. It contains two tables: Processing Methods and Database Mappings. The Processing Methods table has columns: Cv, Accession, Name, and Value. The Database Mappings table has columns: Search Engine Database, Search Engine Database, Curated Database Name, and Curated Database Version. At the bottom are "Quit", "Back", and "Next" buttons.

Cv	Accession	Name	Value
1 PRIDE	PRIDE:0000161	Fragment mass tolerance setti...	0.4 Da
2 PRIDE	PRIDE:0000078	Peptide mass tolerance setting	0.4 Da
3 PRIDE	PRIDE:0000162	Allowed missed cleavages	3
4 MS	MS:1001316	Mascot:SigThreshold	0.05
5 MS	MS:1001758	Mascot:SigThresholdType	identity

Search Engine Database	Search Engine Database	Curated Database Name	Curated Database Version
1 IPI_HUMAN	3.54	IPI	

Figure 10 – PRIDE Converter 2 Software Processing form.

The “search database” mappings will help standardize the names of the sources of the protein sequences that have been used to generate the search databases. The most common sources of protein sequences have been included (UniProtKB/Swiss-Prot, UniProtKB/TrEMBL, International Protein Index (IPI), Ensembl, RefSeq, NCBI nr database) and the users are encouraged to provide either a release number or the date the search database was built. Users can also manually enter a custom database name, if the supplied options are not appropriate. The “curated database name” is mandatory, but the “curated database version”, while desired, is not. A warning will be generated if the database version is omitted however.

### 3.1.8. PTMs

The next step is the review and/or correction of PTM assignments (Figure 11). Please refer to the section entitled “PRIDE Converter Automatic PTM Assignment” for more information on how PTMs are assigned and how they can be manually corrected or annotated.

	Search Engine	Accession	Name	Database	Residues	Mono Delta
1	Phospho (S)	MOD:00696	phosphorylated ...	MOD	S	79.966324
2	Oxidation (M)	MOD:00425	monohydroxylat...	MOD	M	15.994919
3	Carbamidometh...	MOD:00397	iodoacetamide ...	MOD	C	57.021465
4	Acetyl (Protein ...	MOD:00394	acetylated resid...	MOD	N-term	42.010559
5	iTRAQ (K)	MOD:01499	iTRAQ4plex-116 ...	MOD	K	144.102066
6	Phospho (T)	MOD:00696	phosphorylated ...	MOD	T	79.966324
7	Phospho (Y)	MOD:00696	phosphorylated ...	MOD	Y	79.966324
8	iTRAQ (N-term)	MOD:01499	iTRAQ4plex-116 ...	MOD	N-term	144.102066

Figure 11 – PRIDE Converter 2 PTM form.

### 3.1.9. Review

Once all of the required annotations are provided, the GUI displays a quick review screen before proceeding to copy the metadata across all report files (Figure 12). Any warnings generated by the automatic form validation that occurs at each step will be displayed here. If multiple files are being converted at once, the total number of input files will also be shown. Users can add additional experiment-level CV terms that would reflect the experiment as a whole but are not appropriate in any of the previous sections. The DAOs and PRIDE Converter 2 itself will automatically add params for ‘XML generation software’ and ‘Original MS data file format’.

	CV	Accession	Name	Value
1	PRIDE	PRIDE:0000175	XML generation software	PRIDE Converter Toolsuite 2.0...
2	PRIDE	PRIDE:0000218	Original MS data file format	Mascot dat file
3	PRIDE	PRIDE:0000219	Date of search	2009-08-08 14:02
4	PRIDE	PRIDE:0000175	XML generation software	PRIDE Converter Toolsuite 2.0...
5	MS	MS:1001083	msims search	

Messages:  
The information you have entered in the previous forms has generated 5 warning messages. Please review them before generating the PRIDE XML files.

**Figure 12** – *PRIDE Converter 2* Review form.

Updating the report file (if performing a single conversion) or copying the data across all report files (if converting multiple files) can take anywhere between seconds and several minutes and is directly proportional to the number of files that need to be updated.

### 3.1.10. File Export

The next step is the generation of the PRIDE XML files (Figure 13). Alternatively, the graphical conversion process can be halted at this point, as all of the report files are now complete and validated and the conversion process can be scripted and batched using the Command Line Interface (CLI). This is generally only practical for users who need to convert a significantly large number of files and who have access to a cluster of computers where the conversions can be parallelized. In most cases, a single desktop machine with sufficient memory and disk space will be more than sufficient.

*PRIDE Converter 2* can generate compressed files using the gzip algorithm to save on disk space. This is enabled by default, as is the option to remove any temporary work files, such as report files, once the application exits. *PRIDE Converter 2* will always include all spectra in the generated PRIDE XML file unless the “include only identified” spectra option is selected. If the experiment contains interaction data that would benefit from being submitted to the IntAct interaction database (<http://www.ebi.ac.uk/intact>) at the EBI, users are encouraged to select the relevant option in the export dialog. This will add a CV term in the generated XML file that will be recognized by the submission process and the person whose contact details have been captured will be contacted by an IntAct curator for further discussion.

**Figure 13 – PRIDE Converter 2 File Export form.**

This panel also sets the various filtering options, which are discussed in greater detail in the section entitled “PRIDE XML Filter”.

Generating the PRIDE XML files can take anywhere between minutes and hours. This is proportional to the size, number and complexity of the source files. Users should ensure that they have sufficient disk space. If filtering is enabled, the full PRIDE XML file is first written to disk, then written again in its filtered form after it has been processed. The report file and the intermediate PRIDE XML file are only deleted when *PRIDE Converter 2* exits if the “remove temporary files” option is selected.

### 3.1.11. Validation Report

Once the PRIDE XML files are written, a status report is shown, indicating how many PRIDE XML files have been generated (Figure 14).

**Figure 14 – PRIDE Converter 2 Validation Report form.**

### 3.1.12. Helpful Tools

The final screen (Figure 15) of the GUI invites users to review the generated PRIDE XML files using the PRIDE Inspector tool (<http://code.google.com/p/pride-toolsuite/wiki/PRIDEInspector>) and to submit their data *via* the ProteomeXchange (PX) consortium (<http://www.proteomexchange.org>).

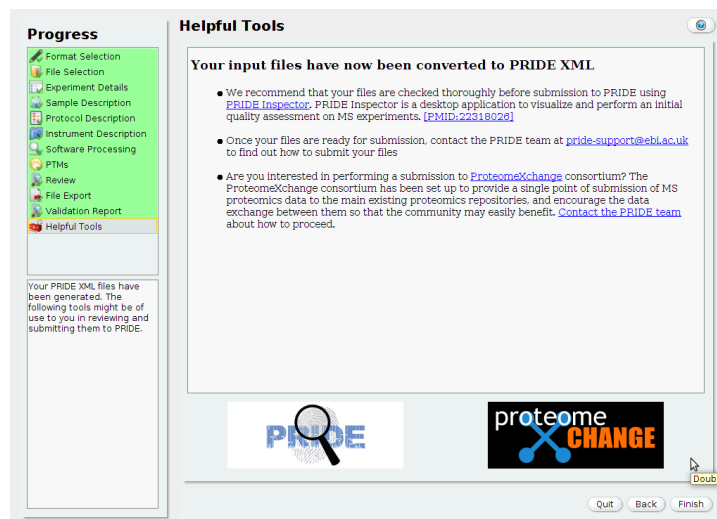


Figure 15 – PRIDE Converter 2 Helpful Tools form.

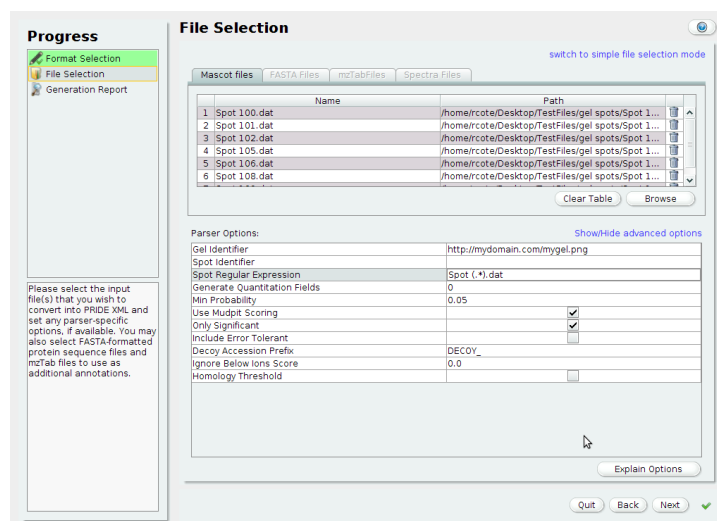
## 3.2. PRIDE mzTab Generator

### 3.2.1. Format Selection

The input format selection form for the PRIDE mzTab Generator is identical to the one used in PRIDE Converter 2 (Figure 4), with the exception that peaklist-only formats are disabled and cannot be selected.

### 3.2.2. File Selection

The look and usability of the file selection form for the PRIDE mzTab Generator is almost identical to that described in the “File Selection” section of the PRIDE Converter 2 GUI user guide, with the addition of 4 mzTab-specific configuration options (Figure 16).



**Figure 16** – PRIDE mzTab Generator File Selection form.

The “gel identifier” option is used to set the gel identifier to be used for identifications in the generated mzTab file. This should generally be in the form of a URI. The “gel spot identifier” option sets the gel spot identifier to be used for identifications in the generated mzTab file. This option is ignored if the “gel spot regex” option is set. The “gel spot regex” option is used to extract the gel spot identifier based on the source file's name. The value of this option must correspond to a valid Java regular expression. The first matching group in the pattern will be used as a spot identifier. The “generate quant fields” option will add empty placeholder quantitative fields to the generated mzTab file for the number of specified reagents. These fields and columns will need to be manually or programmatically filled in with the proper values before the mzTab file is ready to be used by the PRIDE Converter 2.

The DAO options, if any, will be identical to those present in the PRIDE Converter 2 and will be stored in the mzTab file. In order to avoid unpredictable DAO behaviour, users **must** use the same settings when generating the mzTab files and when using the PRIDE Converter 2. As the options are stored in both the report file and the mzTab file, validity checks will be performed to ensure that both sets of options are identical. Any differences will cause an exception and the Conversion process will be blocked until all options match.

### 3.2.3. Generation Report

The final screen of the PRIDE mzTab Generator (Figure 17) will show the path of the newly generated mzTab file corresponding to each input file. If a gel identifier and/or a spot identifier or spot regex have been provided, the relevant information is also included in the report. This is particularly useful when a gel spot regular expression has been used to confirm that the correct spot identifier has been extracted from the input file name.

Input File	MzTab File	Gel ID	SPOT ID
gel spots/Spot 100.dat	gel spots/Spot 100.dat-mz...	http://mydomain.com/myge...	100
gel spots/Spot 101.dat	gel spots/Spot 101.dat-mz...	http://mydomain.com/myge...	101
gel spots/Spot 102.dat	gel spots/Spot 102.dat-mz...	http://mydomain.com/myge...	102
gel spots/Spot 105.dat	gel spots/Spot 105.dat-mz...	http://mydomain.com/myge...	105
gel spots/Spot 106.dat	gel spots/Spot 106.dat-mz...	http://mydomain.com/myge...	106
gel spots/Spot 108.dat	gel spots/Spot 108.dat-mz...	http://mydomain.com/myge...	108
gel spots/Spot 109.dat	gel spots/Spot 109.dat-mz...	http://mydomain.com/myge...	109

**Figure 17** – PRIDE mzTab Generator Generation Report form.

## 3.3. PRIDE XML Filter

### 3.3.1. Filter Information

The first screen of the PRIDE XML Filter provides an overview of the role of the tool and informs the user of several design assumptions and restrictions.



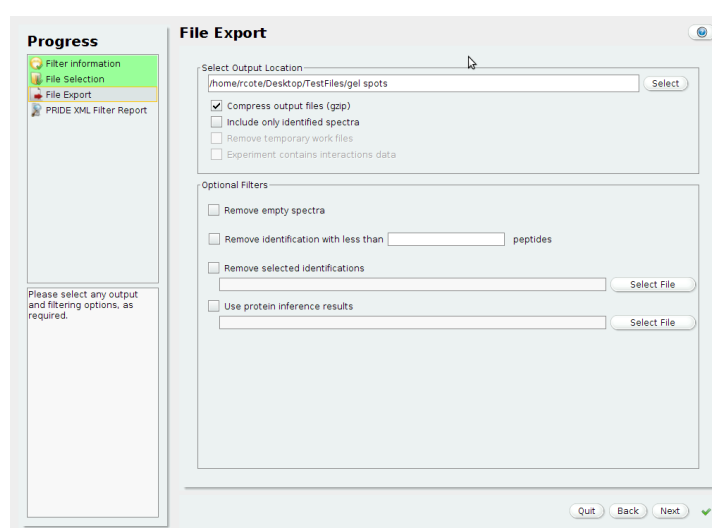
All the PRIDE XML files must be valid according to the PRIDE XML schema. Any number of filters can be selected in one go and all the filters will be applied to each selected input file.

### 3.3.2. File Selection

The look and usability of the file selection form for the PRIDE XML Filter is almost identical to that described in the “File Selection” section of the PRIDE Converter 2 GUI user guide, with the exception that all other file type selectors have been disabled and there are no configuration options to set. The PRIDE XML Filter can either take compressed (gzipped) or uncompressed PRIDE XML files as input, but all files need to be in the same state.

### 3.3.3. File Export

The look and usability of the file export form for the PRIDE XML Filter is almost identical to that described in the “File Export” section of the PRIDE Converter 2 GUI user guide, with the exception that some options have been disabled (Figure 18).



**Figure 18** – PRIDE XML Filter File Export form.

The “include only identified spectra” filter is the only filter that does not need to generate an intermediate PRIDE XML file, as this can be done at the DAO level. All other filters will first write a complete PRIDE XML file, as normal, then post-process this file to generate the fully-filtered PRIDE XML file. This will have consequences on the time and disk space required to fully produce the final file.

The “remove empty spectra” filter will remove any spectra that do not contain any peaks. The “remove identification with less than X peptides” filter will scan and remove any protein identification from the PRIDE XML file that contains less peptides than the number given as parameter.

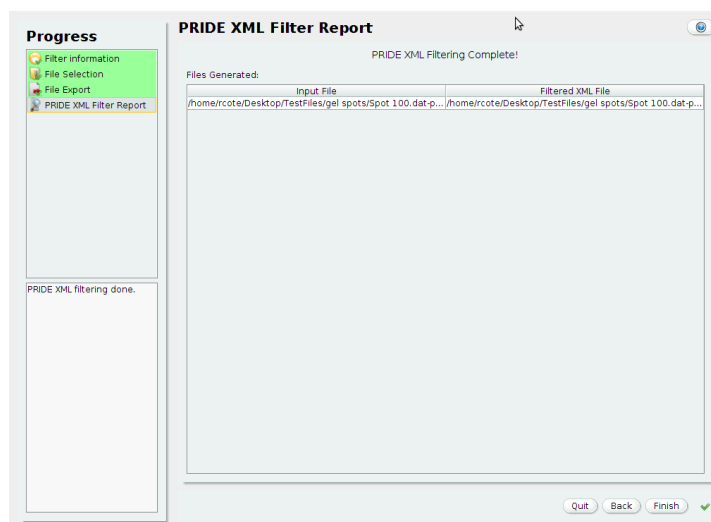
The “remove selected identifications” filter requires a list of contaminant protein identifications in the form of a text file with one protein accession per line. It will use this blacklist to remove corresponding identifications from the XML file where the submitted protein accession identically matches one of the blacklisted protein identifiers.

The “use protein inference results” filter also requires a list of protein identifications in the same manner as the previous filter but works in the complete opposite fashion. Only identifications found in the whitelist are retained in the final PRIDE XML file.

### 3.3.4. Filter Report

---

The final screen of the PRIDE XML Filter (Figure 19) will show the path of the newly generated filtered PRIDE XML file corresponding to each input file.



**Figure 19** – PRIDE XML Filter Report form.

## 3.4. PRIDE XML Merger

---

### 3.4.1. Merger Information

---

The first screen of the PRIDE XML Merger provides an overview of the role of the tool and informs the user of several design assumptions and restrictions. All the PRIDE XML files must be valid according to the PRIDE XML schema. A master file will be selected, which will serve as the basis for the final PRIDE XML file in terms of metadata. The identifications and spectra from the other files will be merged into the final output file. Users should note that the spectrum IDs from the source files will not necessarily be kept, but the internal references between peptide-spectrum\_refs will be maintained.

### 3.4.2. File Selection

---

The look and usability of the file selection form for the PRIDE XML Merger is almost identical to that described in the “File Selection” section of the PRIDE Converter 2 GUI user guide, with the exception that all other file type selectors have been disabled and the only DAO option available will be the option to compress the final output file (Figure 20). The PRIDE XML Merger can either take compressed (gzipped) or uncompressed PRIDE XML files as input, but all files need to be in the same state.

	Name	Path
1	Spot 100.dat-pride.xml	/home/rcote/Desktop/TestFiles/gel spots/Spot 1...
2	Spot 101.dat-pride.xml	/home/rcote/Desktop/TestFiles/gel spots/Spot 1...
3	Spot 102.dat-pride.xml	/home/rcote/Desktop/TestFiles/gel spots/Spot 1...
4	Spot 105.dat-pride.xml	/home/rcote/Desktop/TestFiles/gel spots/Spot 1...
5	Spot 106.dat-pride.xml	/home/rcote/Desktop/TestFiles/gel spots/Spot 1...
6	Spot 108.dat-pride.xml	/home/rcote/Desktop/TestFiles/gel spots/Spot 1...

**Figure 20** – PRIDE XML Merger File Selection form.

### 3.4.3. Master File Selection

A master file must be selected. All metadata annotations will be taken from this file to as a basis to generate the merged PRIDE XML file (Figure 21).

Please select the master file:

- gel spots/Spot 106.dat-pride.xml
- gel spots/Spot 100.dat-pride.xml
- gel spots/Spot 101.dat-pride.xml
- gel spots/Spot 102.dat-pride.xml
- gel spots/Spot 105.dat-pride.xml
- gel spots/Spot 106.dat-pride.xml
- gel spots/Spot 108.dat-pride.xml
- gel spots/Spot 109.dat-pride.xml
- gel spots/Spot 11.dat-pride.xml

**Figure 21** – PRIDE XML Merger Master File Selection form.

### 3.4.4. Merger Report

Once all PRIDE XML files have been merged, the user is shown a report form showing where the output file was written to and all the source PRIDE XML files which were used for the merge (Figure 22).

Progress

Merger Information

File Selection

Master File Selection

Merge Report

PRIDE XML merge done.

Merge Report

PRIDE XML Merging Complete!

Output file: /home/rcote/Desktop/TestFiles/gel spots/Spot 106.dat-pride.xml-merged.xml

Input Files

/home/rcote/Desktop/TestFiles/gel spots/Spot 100.dat-pride.xml

/home/rcote/Desktop/TestFiles/gel spots/Spot 101.dat-pride.xml

/home/rcote/Desktop/TestFiles/gel spots/Spot 102.dat-pride.xml

/home/rcote/Desktop/TestFiles/gel spots/Spot 105.dat-pride.xml

/home/rcote/Desktop/TestFiles/gel spots/Spot 106.dat-pride.xml

/home/rcote/Desktop/TestFiles/gel spots/Spot 108.dat-pride.xml

/home/rcote/Desktop/TestFiles/gel spots/Spot 109.dat-pride.xml

/home/rcote/Desktop/TestFiles/gel spots/Spot 11.dat-pride.xml

/home/rcote/Desktop/TestFiles/gel spots/Spot 110.dat-pride.xml

/home/rcote/Desktop/TestFiles/gel spots/Spot 111.dat-pride.xml

/home/rcote/Desktop/TestFiles/gel spots/Spot 112.dat-pride.xml

/home/rcote/Desktop/TestFiles/gel spots/Spot 114.dat-pride.xml

/home/rcote/Desktop/TestFiles/gel spots/Spot 115.dat-pride.xml

/home/rcote/Desktop/TestFiles/gel spots/Spot 117.dat-pride.xml

/home/rcote/Desktop/TestFiles/gel spots/Spot 118.dat-pride.xml

/home/rcote/Desktop/TestFiles/gel spots/Spot 119.dat-pride.xml

/home/rcote/Desktop/TestFiles/gel spots/Spot 121.dat-pride.xml

/home/rcote/Desktop/TestFiles/gel spots/Spot 122.dat-pride.xml

/home/rcote/Desktop/TestFiles/gel spots/Spot 123.dat-pride.xml

/home/rcote/Desktop/TestFiles/gel spots/Spot 125.dat-pride.xml

/home/rcote/Desktop/TestFiles/gel spots/Spot 126.dat-pride.xml

Quit

Back

Finish

Figure 22 – PRIDE XML Merger Report form.