The PRoteomics IDEntification (PRIDE) Converter 2 Framework: An Improved Suite of Tools to Facilitate Data Submission to the PRIDE Database and the ProteomeXchange Consortium*[S]

Richard G. Côté‡‖‖, Johannes Griss‡‖‖, José A. Dianes‡, Rui Wang‡, James C. Wright§, Henk W.P. van den Toorn¶, Bas van Breukelen¶, Albert J. R. Heck¶, Niels Hulstaert‖, Lennart Martens‖**, Florian Reisinger‡, Attila Csordas‡, David Ovelleiro‡, Yasset Perez-Rivevol‡ ‡‡, Harald Barsnes§§, Henning Hermjakob‡, and Juan Antonio Vizcaíno‡¶¶**

The original PRIDE Converter tool greatly simplified the process of submitting mass spectrometry (MS)-based proteomics data to the PRIDE database. However, after much user feedback, it was noted that the tool had some limitations and could not handle several user requirements that were now becoming commonplace. This prompted us to design and implement a whole new suite of tools that would build on the successes of the original PRIDE Converter and allow users to generate submission-ready, well-annotated PRIDE XML files. The *PRIDE Converter 2* tool suite allows users to convert search result files into PRIDE XML (the format needed for performing submissions to the PRIDE database), generate mzTab skeleton files that can be used as a basis to submit quantitative and gel-based MS data, and post-process PRIDE XML files by filtering out contaminants and empty spectra, or by merging several PRIDE XML files together. All the tools have both a graphical user interface that provides a dialog-based, user-friendly way to convert and prepare files for submission, as well as a command-line interface that can be used to integrate the tools into existing or novel pipelines, for batch processing and power users. The *PRIDE Converter 2* tool suite will thus become a cornerstone in the submission process to PRIDE and, by extension, to the ProteomeXchange consortium of MS-proteomics data repositories. *Molecular & Cellular Proteomics 11: 10.1074/mcp.O112.021543, 1682–1689, 2012.*

The sharing of biological data in the public domain is generally considered to be good scientific practice. This concept of data sharing has gained substantial traction in the field of MS-based proteomics, in which the PRIDE[1] (PRoteomics IDEntifications) database (http://www.ebi.ac.uk/pride) at the European Bioinformatics Institute (EBI, Cambridge, UK) is one of the most prominent public data repositories (1). PRIDE stores MS and MS/MS spectra, the derived peptide and protein identifications and expression values if available (the processed experimental results), and any associated metadata. It is important to highlight that data stored in PRIDE is not reprocessed after submission. PRIDE, in its current form, represents the submitter's view of the data. PRIDE is also a founding member of the ProteomeXchange (PX) consortium (http://www.proteomexchange.org) (2). The PX members, led

From the ‡Proteomics Services Team, EMBL Outstation, European Bioinformatics Institute (EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, UK; §Proteomic Mass Spectrometry, Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK; ¶Biomolecular Mass Spectrometry and Proteomics, Bijvoet Center for Biomolecular Research and Utrecht Institute for Pharmaceutical Sciences, Utrecht University, The Netherlands, and Netherlands Proteomics Centre; ‖Department of Medical Protein Research, VIB, B-9000 Ghent, Belgium; **Department of Biochemistry, Ghent University, B-9000 Ghent, Belgium; ‡‡Department of Proteomics, Center for Genetic Engineering and Biotechnology, Havana, Cuba; §§Proteomics Unit, Department of Biomedicine, University of Bergen, Norway

Received July 9, 2012, and in revised form, August 31, 2012
⌘ *Author's Choice*—Final version full access.
Published, MCP Papers in Press, September 4, 2012, DOI 10.1074/mcp.O112.021543

[1] The abbreviations used are: API, Application Programming Interface; BBSRC, Biotechnology and Biological Science Research Council; CLI, Command-Line Interface; CV, controlled vocabulary; DAO, data access object; EBI, European Bioinformatics Institute; GUI, graphical user interface; JAR, java archive; LIMS, Laboratory Information Management System; MIAPE, Minimum Information About a Proteomics Experiment; NIH, National Institutes of Health; OLS, Ontology Lookup Service; PMF, peptide mass fingerprinting; PRIDE, PRoteomics IDEntifications database; PSI, Proteomics Standards Initiative; PTM, post-translational modification; PX, ProteomeXchange; UniProtKB, UniProt KnowledgeBase; XML, eXtensible Markup Language.

by PRIDE and PeptideAtlas (3), are currently working toward the implementation of a system that enables the automated and standardized sharing of MS-based proteomics data between the main proteomics repositories. In this framework, PRIDE is the initial submission point for tandem MS data. Currently, the first pilot PX submissions (containing raw data and processed results) have already been carried out (http://proteomecentral.proteomexchange.org) and the system is now starting to accept regular submissions. At present, submissions to PRIDE are performed using a publicly available XML data format called PRIDE XML, which is built around the mzData data standard format (4).

Several scientific journals (*e.g. Molecular and Cellular Proteomics, Proteomics,* and *Nature* Publishing Group journals) are supporting a gradual move toward mandating public deposition of MS data to support the publication of related manuscripts. In parallel, several funding agencies (such as The Wellcome Trust, NIH, and BBSRC) are also enforcing the public availability of experimental data in the context of their funded projects. Despite these efforts, the field of MS proteomics is still lagging behind other more mature "omics" disciplines in terms of public data availability (5). In practical terms, a major contribution to this public data-sharing policy trend is provided by the availability of reliable and user-friendly submission tools. Such tools must be able to capture properly the experimental data and any supporting technical and biological metadata. In addition, to encourage MS data deposition the submission process has to be as easy as possible.

This was the philosophy that drove the development of the original PRIDE Converter (6) (http://pride-converter.googlecode.com), an open source and platform-independent software tool for the submission of proteomics data to PRIDE. PRIDE Converter can convert input data from a large variety of popular MS proteomics formats into PRIDE XML, guiding the user through the process by a graphical user interface (GUI). As a result, PRIDE Converter made the submission of MS data a much easier and more straightforward process, especially for researchers without bioinformatics support. PRIDE Converter has definitely been a key factor in the huge growth in data content in PRIDE since 2008 (7) and has become the *de facto* submission tool to PRIDE for most researchers. PRIDE Converter has been regularly updated and more than 30 different releases have been made publicly available. However, after receiving extensive feedback from users, it became apparent that the original PRIDE Converter had some limitations mainly in terms of software architecture, memory requirements, difficulties to extend the supported formats, and a lack of functionality for performing batch conversions (a frequent request). In addition, new use cases needed to be supported, such as support for quantitative information and the ability to easily post-process the large XML files generated during the conversion process. To overcome these limitations, we decided to design a new submission tool from the ground up, which would be suitable to the evolving needs of our submitters.

TABLE I
Tools in the PRIDE Converter 2 tool suite

Tool name	Function
PRIDE Converter 2	Converts search engine output files into valid, well-annotated PRIDE XML files ready for submission.
PRIDE mzTab Generator	Generates skeleton mzTab files where the user can add quantitative and/or gel data.
PRIDE XML Merger	Merges several PRIDE XML files together, while maintaining internal consistency in spectra and peptide links.
PRIDE XML Filter	Post-processes PRIDE XML files according to filter rules to remove contaminants, empty spectra and/or update the protein inference assignments.

In this manuscript we describe the PRIDE Converter 2 framework, including all of its new features and supported use cases. We are certain that future submitters to PRIDE and to the PX consortium will benefit immensely from the availability of this new submission tool.

EXPERIMENTAL PROCEDURES

The *PRIDE Converter 2* tool suite is developed in Java and all the source code is available online (http://pride-converter-2.googlecode.com). It is distributed as open source under the very permissive Apache License, Version 2.0.

The development of *PRIDE Converter 2* had several goals:

- Provide a series of tools dedicated to specific tasks.
- Each tool should be accessible through a command-line interface (CLI) for integration into (third-party) PRIDE XML generation and annotation pipelines.
- Each tool should be accessible through a GUI to provide a rich, user-friendly experience.
- Each tool has to be as efficient as possible in its use of resources to keep a low memory profile.
- Support as many input formats as possible by reusing existing Application Programming Interfaces (APIs) and code libraries.
- Keep the GUI as consistent as possible across applications in the tool suite by reusing components, where possible.
- Improve on the original PRIDE Converter tool and support new use cases required by our users.

To achieve these goals, the *PRIDE Converter 2* tool suite took inspiration from the Linux toolchain approach, in which small applications that are dedicated to a single purpose can be chained together to perform powerful operations with minimal resource overhead. As such, the PRIDE Converter 2 tool suite consists of four different applications: *PRIDE Converter 2, PRIDE mzTab Generator, PRIDE XML Merger,* and *PRIDE XML Filter* (Fig. 1, Table I). All of these tools are bundled within a single executable JAR file for convenience. Users can launch the GUI either by double-clicking on the JAR file or by invoking it without arguments from the command-line. If arguments are provided, the CLI is launched, allowing batch processing and, as a key point, integration into existing and newly built pipelines.

Supplemental Files S1 and S2 are provided as a *PRIDE Converter 2* guide for general users, and developers, respectively. More details about the technical implementation can be found in supplemental File S2, section 1.

Support for Novel Use-cases—The original PRIDE Converter had some functional and practical limitations. Driving the work behind the development of the *PRIDE Converter 2* tool suite was the desire to not only overcome the shortcomings of the original tool but also add functionality that had been repeatedly requested by our users.

As such, the *PRIDE Converter 2* tool suite has added conversion support for a number of new data formats (see Table II) and other formats will likely follow over time. Moreover, support for existing formats has also been improved. For example, it is now possible to submit peptide mass fingerprint (PMF) data generated by Mascot (http://www.matrixscience.com). Also, the addition of quantitative data to PRIDE XML files has been greatly improved by integrating support for mzTab files.

The mzTab format is meant to be a light-weight, standard tab-delimited file for MS-based proteomics data, developed by the Proteomics Standards Initiative (PSI). Designed to be easy to parse, it contains only the minimal information required to evaluate the results of a proteomics experiment (http://mztab.googlecode.com). Users can generate skeleton mzTab files using the *PRIDE mzTab Generator* and then use the produced mzTab files as a basis to provide quantitative information as part of the conversion process in *PRIDE Converter 2*. Gel and spot-related information can also be added to the mzTab files, making the capture of gel-associated information much more straightforward (supplemental File S2, section 4). Users can now also provide their original search databases in FASTA format (supplemental File S1, section 3). This is essential to maintain data provenance for nonstandard protein databases and makes it easier to map the identified proteins across all protein databases, a process that is performed as a matter of course in the PRIDE database to maximize search capabilities (8).

Another user requirement fulfilled by the *PRIDE Converter 2* tool suite is the ability to post-process the initially generated PRIDE XML files. For example, users can now use the *PRIDE XML Filter* tool to remove contaminants and empty spectra prior to submission. Finally, in the case of gel-based proteomics experiments in which each gel spot produces one MS experiment, the original PRIDE Converter tool would generate one PRIDE XML file per spot. This meant that a single project could cover several dozens, if not hundreds of PRIDE experiment accession numbers. The *PRIDE XML Merger* can now merge together an arbitrarily large number of PRIDE XML files into a single file, while keeping the links between identified peptides and their underlying spectra consistent. This means that users will be able to obtain a single PRIDE accession number to refer to their collated experimental data.

RESULTS

PRIDE Converter 2—Most users will be best served by the tool's user-friendly GUI. This interface has the benefit of a comprehensive context-sensitive help module, and provides instant feedback on the fields and annotations that are required at each step of the wizard-like conversion process. Each dialog the user edits will be validated before moving on to the next step. If the form contains any errors (such as empty mandatory fields or data entered in incorrect formats) the user is immediately informed and the conversion process is blocked until the error is fixed by the user. Therefore, once a dialog is filled out the user can be sure that the required information is present. The file validation process can also generate warnings, which would not block the process but should still be taken under consideration to generate optimally annotated PRIDE XML files.

The command line interface is mainly geared toward power users who have the capacity to parallelize batch conversions and/or those who already have mechanisms to programmatically provide all of the necessary metadata required to produce valid report files, and ultimately PRIDE XML files. When using the command line interface, the *PRIDE Converter 2* tool must be invoked in two modes: *prescan* and *output* (refer to Fig. 1). The *prescan* mode needs to be run first and will take the results files obtained from an MS experiment (*i.e.* spectra with or without accompanying peptide and protein identifications) and will generate an intermediary report file from these. Two additional types of files can optionally be provided in the *prescan* mode to enrich the report: the protein sequence search database used in the proteomics experiment (in FASTA format) and mzTab files providing quantitative data and gel/spot information. Once the report file has been properly annotated, the *PRIDE Converter 2* tool is then run in *output* mode to generate a well-formed PRIDE XML file.

PRIDE Converter 2 currently supports the formats shown in Table II. New formats can easily be supported simply by implementing the Java *DAO* (Data Access Object) interface. This interface provides methods to access and retrieve information on metadata, spectra, peptides, proteins and post-translational modifications (PTMs) from the source files. The DAOs try to extract as much information as possible from the source files to provide a sensible starting point for the annotation process. A full developer's guide to programming custom DAOs is available at http://code.google.com/p/pride-converter-2/wiki/HowToWriteADao and in the supplemental File S2 (section 2). To avoid having to (re-)create parsers for the various formats, *PRIDE Converter 2* makes use of best-of-breed existing reusable APIs where available (Table II). When this was not possible new parsers were developed.

The report file generated by the *prescan* will contain all of the reported protein and peptide identifications, and will furthermore serve as the basis for all subsequent controlled vocabulary annotation that will make its way into the final PRIDE XML file, including, but not limited to, details on contacts, protocols, instrumentation and software processing, journal references, search database annotations, protein sequences, and PTMs. If provided, the report file will also contain any quantitative and gel-based information. *PRIDE Converter 2* will attempt to automatically curate protein accessions coming from various sources into the preferred PRIDE format for that source as part of the *prescan* process (for example, the submitted protein identifier "*sp P29375 KDM5A_HUMAN Lysine-specific demethylase 5A OS = Homo sapiens GN = KDM5A PE = 1 SV = 3*" will be cleaned up to P29375, which is the default protein identifier format in PRIDE for UniProtKB entries).

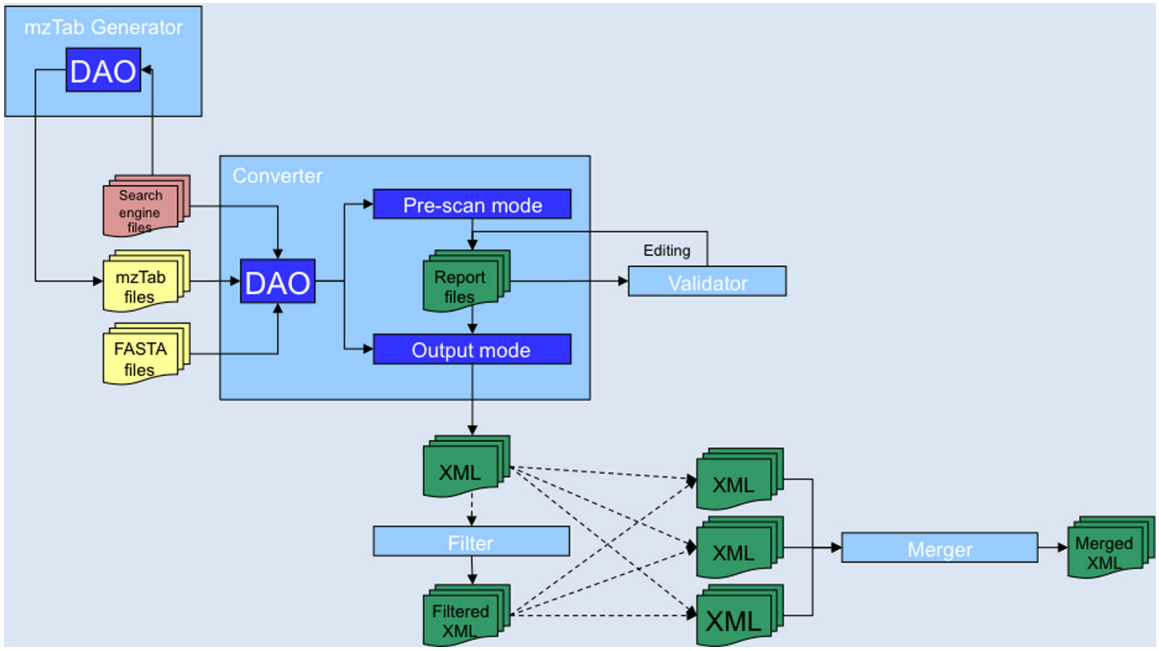

FIG. 1. **Schematic overview of the workflow and interactions between the tools in the *PRIDE Converter 2* tool suite.** The *PRIDE mzTab Generator* parses the search result files and generates skeleton mzTab files that can be used as input files to *PRIDE Converter 2*. The PRIDE XML files generated by *PRIDE Converter 2* can be filtered using the *PRIDE XML Filter* tool and/or merged into a single PRIDE XML file using the *PRIDE XML Merger* tool.

TABLE II
Supported formats in PRIDE Converter 2

Format name	File type	Data content	New in *PRIDE Converter 2*	Used APIs
Mascot	.dat	Spectra and Identifications	No	Mascot API (19)
mzIdentML	.xml	Spectra and Identifications	Yes	jmzIdentML (20)
X!Tandem	.xml	Spectra and Identifications	No	xtandem-parser (21)
OMSSA	.csv	Spectra and Identifications	No	New
SpectraST	.txt	Spectra and Identifications	Yes	New
CRUX	.txt	Spectra and Identifications	Yes	New
MSGF	.txt	Spectra and Identifications	Yes	New
Proteome Discoverer	.msf	Spectra and Identifications	Yes	Thermo MSF Parser (14)
mzML	.xml	Spectra Only	Yes	jmzML (22)
DTA	.dta	Spectra Only	No	jmzReader (23)
MGF	.mgf	Spectra Only	No	jmzReader (23)
mzData	.xml	Spectra Only	No	jmzReader (23)
mzXML	.xml	Spectra Only	No	jmzReader (23)
PKL	.pkl	Spectra Only	No	jmzReader (23)

The report file XML schema is well-defined and annotated (http://pride-converter-2.googlecode.com/svn/trunk/report-api/src/main/resources/reportfile.xsd), and a Java API has been provided to generate report files, making it easy to integrate this functionality into an existing proteomics LIMS as a first step in exporting data to PRIDE XML. Once the report file has been generated and annotated, either manually, programmatically, or by using the *PRIDE Converter 2* GUI, the *output* mode of PRIDE Converter is invoked to generate a submission-ready PRIDE XML file.

PRIDE Converter 2 Overview—When launched in GUI mode (Fig. 2), *PRIDE Converter 2* guides the user through a 12-step process to convert their search engine output files into well-annotated PRIDE XML files (Fig. 3, supplemental File S1, section 3). Input format selection is the first choice that the user must make. Each format-specific DAO can have one or several custom options that can subsequently be set through the GUI. 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 the choices for these options will be stored in the report file such that the user can always review how the conversion process was configured.

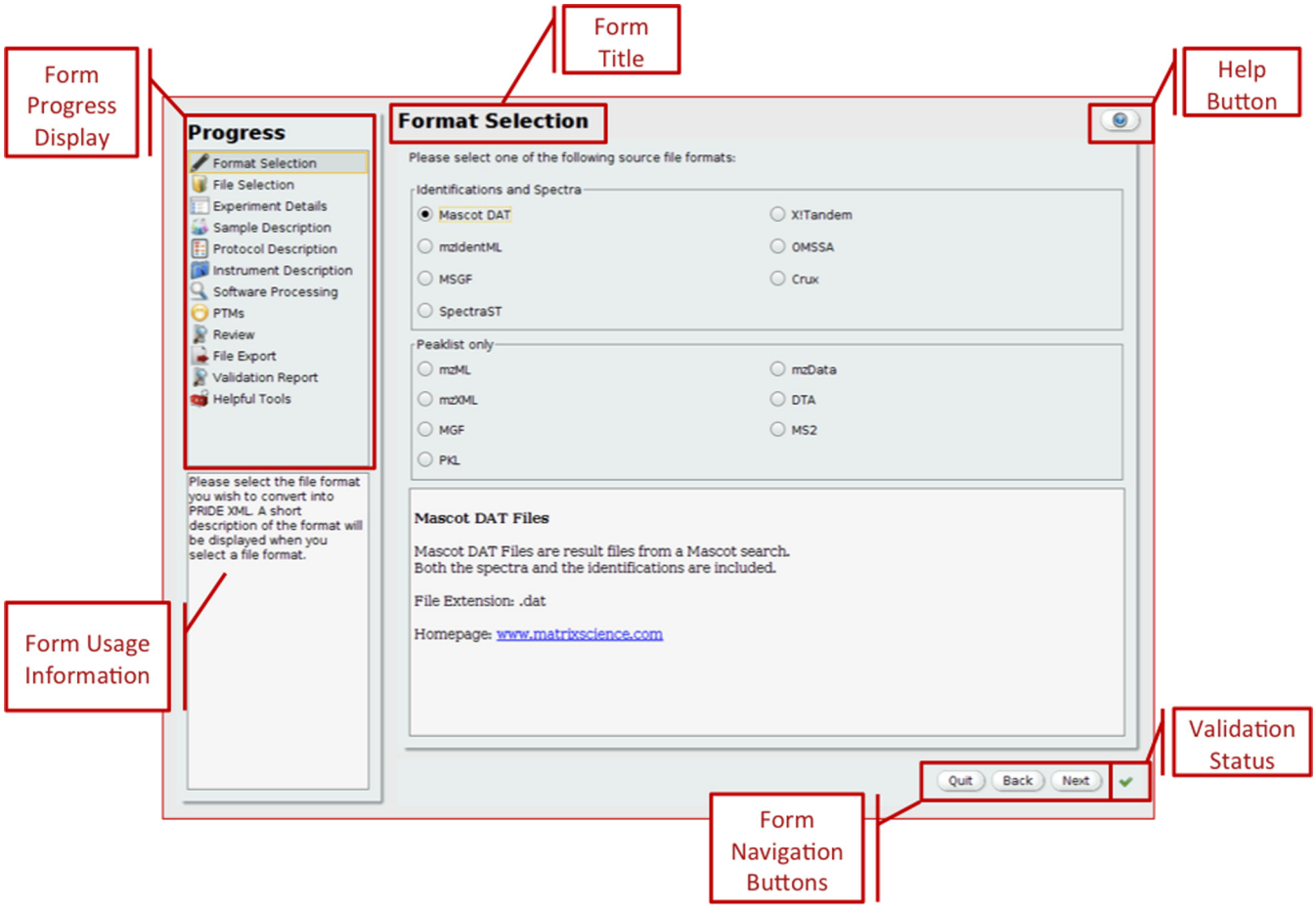

FIG. 2. **The *PRIDE Converter 2* Graphical User Interface.** The GUI for all tools shares common features, wherever possible, to improve usability and provide a consistent user experience. Each tool is composed of a series of forms that are presented in a wizard-like manner. Users can navigate through the forms using a series of buttons located on the lower right corner. A context-sensitive help button is always available, as is a short informative message on the role of the form and what information is expected. User input is always validated to ensure that all required fields are correctly filled-in and a graphical validation status is updated each time a navigation button is pressed.

The GUI allows users to convert multiple source files simultaneously while only having to enter the required annotations once. This can save a considerable amount of time, because annotations on contact information, sample details, instrumentation, protocols, software processing, PTMs, and search databases will typically be identical across related source files. *PRIDE Converter 2* also provides the possibility to save commonly used annotations such as instrumentation and protocols as templates, which can be reused in subsequent conversions. A set of basic templates is provided with the tool suite that users can update to better suit their own requirements. Sample annotation allows the user to provide taxonomy, tissue type and cell type annotations using pull-down menus that contain the most commonly used values found in PRIDE. Users still have the possibility to use the comprehensive Ontology Lookup Service (OLS) (9, 10) to look up alternative terms that are not already provided. If the user has included quantitative data *via* an mzTab file, the sample annotation form will be used to provide sample descriptions for the quantitation method.

Other useful features include the automatic mapping of PTMs most commonly observed in proteomics experiments to the appropriate controlled vocabulary (CV) terms in the PSI Protein Modification ontology (PSI-MOD) (11). Consistent PTM annotation has been a known issue in the past and a source of annotation errors in PRIDE data (7), as most search engines report PTMs in different ways using nonstandard terminology. *PRIDE Converter 2* attempts to assign a standardized PTM annotation based on a curated list of the most commonly observed PTMs and the mass delta as reported by the search engine. If a unique PTM can be assigned to a mass delta within a 0.1 Da mass tolerance, the annotation is automatically shown to the user. In cases where multiple PTMs can be assigned to a mass delta with a precision of 0.1 Da, *PRIDE Converter 2* will try to locate a unique PTM assignment to within 0.01 Da. If a unique match is found at the higher precision threshold, it will be assigned but the GUI will report the fact that multiple PTMs have been observed. In the case that multiple PTMs are still found at the higher precision threshold,

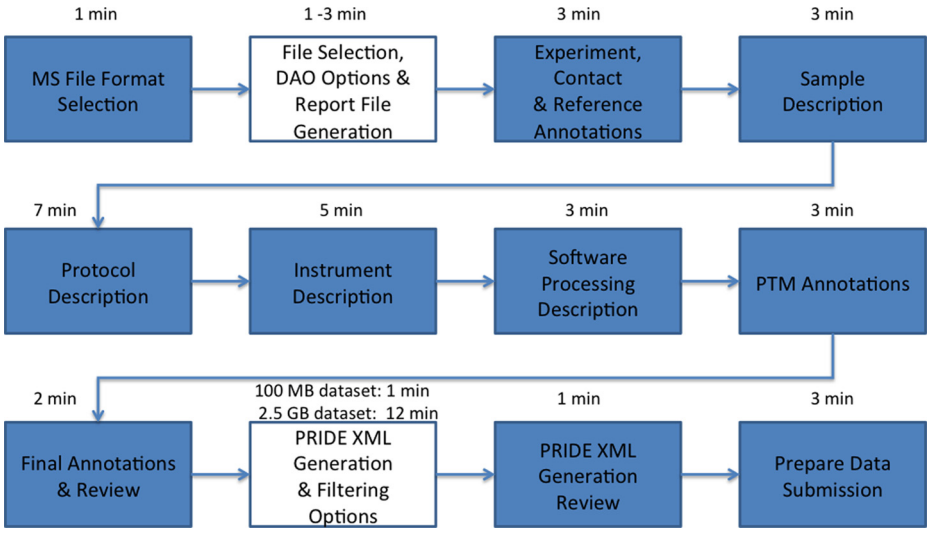

FIG. 3. **The 12-step process of converting search engine result files into well-annotated PRIDE XML files.** The approximate duration of the different steps in the conversion process are indicated to give an idea about the time required to do a submission. The boxes not filled are indicating that the duration of these steps depends on the size of the input files. The other steps are related to file selection and/or metadata annotation, and are independent of the size of the files. A Mac Book Pro laptop with 8 GBs of RAM running Mac OS X 10.6.8 was used to estimate the timings. To summarize, users select the appropriate format of their search engine files, then select the file(s) to convert and set any DAO-specific options, if applicable. The annotation process starts with contact, reference, and general project descriptions, then moves on to sample annotations, protocols, instrumentation details, and software processing details. The users are asked to review or complete the automatic PTM annotations and add any additional relevant experiment-level details. The report files are then finalized and the users can either stop the GUI process here or proceed to PRIDE XML file generation. This is where filtering options can also be set. Once the conversion process has completed, the users are invited to review their PRIDE XML files with the PRIDE Inspector tool and submit them to PRIDE and to the ProteomeXchange consortium.

no mapping is done. The GUI will report the fact that multiple PTMs have been observed within the mass tolerance window by highlighting the conflict in yellow. PTMs that have not been automatically assigned will be highlighted in red. The user must then edit the highlighted modifications manually to assign the correct PSI-MOD term either using the suggested PTMs, if available, or by searching the OLS for the correct PSI-MOD term (see supplemental File S1, section 6 for more details). Automatic PTM assignment is also performed in the CLI mode, but warnings highlighting multiple possible assignments are only shown in the console window and not the report file. It is therefore up to the user to confirm the proper assignment prior to the final conversion.

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. The next step is the generation of the PRIDE XML files. 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 CLI. This is generally only practical for users who need to convert a large number of files and have access to a computer cluster where the conversions can be parallelized. In most cases, a single desktop machine with average memory and disk space will be more than sufficient. The final screen of the GUI invites users to review the generated PRIDE XML files using the PRIDE Inspector tool (12) and to submit their data *via* the PX consortium. Please refer to sections 3 to 5 of the supplemental File S1 for a full user guide for all the tools in the *PRIDE Converter 2* tool suite.

PRIDE mzTab Generator—The *PRIDE mzTab generator* will generate skeleton mzTab files based on the same MS source files used by *PRIDE Converter 2*. The user has the same parser options as in *PRIDE Converter 2* and the generator settings are also stored in the produced mzTab file. This is very important as the mzTab files and the report files need to be generated with the same options as the conversion to produce sensible and consistent results. If mzTab files are used as part of the *PRIDE Converter 2 prescan*, the configuration settings of the mzTab files are read and if they do not match the configuration settings for the *prescan*, an error message is shown to the user and conversion is blocked until the differences are resolved.

The *PRIDE mzTab generator* has several parser options to handle gel and quantitative information. If the experiment contains quantitative information, it is possible for the *PRIDE mzTab generator* to automatically create placeholder annotations to describe the quantitation labels used in the experiment and add columns to the file for the quantitation values, that the users will be able to edit to add the quantitation values. If the experiment is gel-based, it is possible to link each identification to a specific spot on a specific gel (for example, "Spot 4" on "Gel A"). This information can also be automatically extracted from the filename if it is contained therein. All of this additional information will be stored in the

report file and will subsequently make its way into the final PRIDE XML file.

PRIDE XML Merger—It is a common scenario in MS-based proteomics experiments that several results files are produced from a single analysis. One example of this would be a gel-based MS experiment in which each spot typically yields a unique MS run and associated results file. It has already been mentioned that *PRIDE Converter 2* is able to load these input files and convert them in one batch, requiring only a single round of annotation for all of the source files. The *PRIDE XML Merger* is the next logical step in such a workflow, where all the individual spot files are merged into a single XML file ready for submission. Using the *PRIDE XML Merger*, it is possible to generate one PRIDE XML file per gel, which is a more convenient method than having one PRIDE XML file per single spot.

PRIDE XML Filter—The *PRIDE XML Filter* is designed to post-process the PRIDE XML files generated by *PRIDE Converter 2* and works at the level of protein identifications and spectra. To achieve this, the *PRIDE XML Filter* can remove empty spectra devoid of peaks or remove protein identifications that contain less than a specified number of peptides (useful to remove "one-hit wonders", for example). The *PRIDE XML Filter* can also take a list of contaminant protein identifications and use this as a blacklist to remove corresponding identifications from the XML file. The protein inference problem is one of the major challenges in reporting proteomics results (13). Unfortunately, the PRIDE XML format does not support properly the assignment of single peptides to multiple proteins, and grouping these into identification groups. Therefore, by default, *PRIDE Converter 2* reports all possible combinations of peptide to protein assignments making sure that no data is lost. This approach then significantly increases the number of proteins reported. Therefore, we have added a feature into the *PRIDE XML Filter,* which can take a whitelist of proteins generated using an external protein inference algorithm and then removes all not-fitting proteins from the generated PRIDE XML file. Although this is not an ideal solution we believe that it is a sensible compromise between the limitations of PRIDE XML and the incorporation of external protein inference results. Please refer to the section 7 of the supplemental File S1 for a more in-depth explanation on how the *DAO*s deal with protein inference.

DISCUSSION

The *PRIDE Converter 2* framework constitutes a big step forward compared with the original PRIDE Converter submission tool. The primary motivation behind the original tool has been however maintained here: the software must be as user-friendly as possible for biologists without much bioinformatics support. Going beyond that original goal, the framework now supports use cases that were absent from the original tool but were much in demand with users. As a result, *PRIDE Converter 2* can now be used by bioinformaticians/computer scientists to perform batch conversions, it can be integrated into pipelines to streamline submissions to PRIDE, and it supports PMF and quantitative data submissions. As of July 2012, *PRIDE Converter 2* has already been used to generate more than 2000 submitted PRIDE XML files, covering five different input file formats. We plan that the original PRIDE Converter application will be discontinued in the next few months.

The modular software architecture, full documentation and free availability of the source code allow any third party to add support for a new format by simply providing a suitable implementation of the *DAO* interface. In practice, this has already happened in the last few months, as the module supporting data from Proteome Discoverer .msf results files was created independently, outside of the core PRIDE team, using the existing Thermo MSF Parser library (14), in the context of the EU FP7 project PRIME-XS (http://www.primexs.eu).

We expect that the new features we have added to the *PRIDE Converter 2* tool suite, such as possible integration into LIMS systems, batch conversion of files, and independent integration of new formats, will enable well-established proteomics groups to develop their own submission pipelines into PRIDE or integrate export to PRIDE XML files in other tools. Furthermore, given the well-documented *DAO* interface, we would encourage other groups who have in-depth expertise with data formats not currently supported to contribute conversion modules for *PRIDE Converter 2*.

We believe that the requirements for data availability by scientific journals and funding agencies can be addressed in a much more efficient and user-friendly way by this new framework. As well as support for new formats, we are hopeful that the open nature of *PRIDE Converter 2* will encourage third party validation schemes to automatically create a validation report for the generated PRIDE XML files. Support for the PSI validation framework (15) is already integrated into the *PRIDE Converter 2* framework and we have collaborations to develop semantic validation rules that would reflect various user needs, such as specific journal requirements and MIAPE guidelines.

Although the PRIDE database is still based on the PRIDE XML format, two modules for conversion of the mzIdentML v1.1 (16) and mzML v1.1 (17) formats, the two PSI standard formats for mass spectrometry data, and protein/peptide identifications, are provided by the *PRIDE Converter 2*. This less-than ideal solution is only an interim approach, because we are currently implementing native mzIdentML and mzML support in PRIDE. However, we will continue to support PRIDE XML as a valid submission format, at least for the medium term, for practical reasons. First, it will take some time before reliable and "easy to use" exporters for the new data standards become available for the many search engines and analysis pipelines. Second, there are several existing third-party pipelines that produce PRIDE XML files that we want to continue to support as well, at least until exporters to mzML/mzIdentML are developed by the groups maintaining

those pipelines. This is the case for the ProteinLynx Global Server (PLGS, Waters), hEIDI (http://biodev.extra.cea.fr/docs/heidi), OmicsHub Proteomics (Integromics), PeptideShaker (http://peptide-shaker.googlecode.com), and Proteios (18), among others.

One of the limitations of the PRIDE XML format is the limited support for protein inference. Protein groups can be reported but not in an ideal way (see the resulting file for Proteome Discoverer msf module). By default, all the peptide-to-protein mappings are reported in the PRIDE XML file. However, the user can still choose to report only the desired proteins by using the *PRIDE XML Filter* tool. Another use case that it is not ideally supported by the PRIDE XML format is the ambiguity for the position of PTMs. Nevertheless, several *DAO*s can report this information using a combination of several techniques. For more details about the approaches used, see section 7.2 in supplemental File S1.

Although the complexity and variation of proteomics workflows remains a major challenge, we expect the *PRIDE Converter 2* to be a major step forward in the user-friendly, comprehensive capture and reporting of proteomics data, and a key element in facilitating data submissions to the ProteomeXchange consortium.

Acknowledgments—We would like to thank Melih Birim for his input during the starting phase of the project.

* This work was supported by the Wellcome Trust [grant number WT085949MA] to J.G., R.G.C., J.A.D. and F.R. R.W. is supported by the BBSRC 'PRIDE Converter' grant [reference BB/I024204/1]. J.A.V. and N.H. are supported by the EU FP7 grant ProteomeXchange [grant number 260558]. J.A.V. is also supported by the EU FP7 grant LipidomicNet [grant number 202272]. H.B. is supported by the Research Council of Norway. H.W.P.vdT, B.vB, and A.J.R.H are supported by the Netherlands Proteomics Centre. H.W.P.vdT, B.vB, A.J.R.H and L.M. acknowledge support from the EU FP7 grant PRIME-XS [grant number 262067]. L.M. further acknowledges support from Ghent University (Multidisciplinary Research Partnership "Bioinformatics: from nucleotides to networks").

[S] This article contains supplemental Files S1 and S2.

¶¶ To whom correspondence should be addressed: Proteomics Services Team, EMBL Outstation, European Bioinformatics Institute (EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, UK. Tel.: + 44 (0) 1223 492686; E-mail: juan@ebi.ac.uk.

|||| Both authors contributed equally and should be considered joint first authors.

REFERENCES

1. Vizcaíno, J. A., Côté, R., Reisinger, F., Barsnes, H., Foster, J. M., Rameseder, J., Hermjakob, H., and Martens, L. (2010) The Proteomics Identifications database: 2010 update. *Nucleic Acids Res.* **38,** D736–742
2. Hermjakob, H., and Apweiler, R. (2006) The Proteomics Identifications Database (PRIDE) and the ProteomExchange Consortium: making proteomics data accessible. *Expert Rev. Proteomics* **3,** 1–3
3. Deutsch, E. W., Lam, H., and Aebersold, R. (2008) PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows. *EMBO Rep.* **9,** 429–434
4. Orchard, S., Montechi-Palazzi, L., Deutsch, E. W., Binz, P. A., Jones, A. R., Paton, N., Pizarro, A., Creasy, D. M., Wojcik, J., and Hermjakob, H. (2007) Five years of progress in the Standardization of Proteomics Data 4th Annual Spring Workshop of the HUPO-Proteomics Standards Initiative April 23–25, 2007 Ecole Nationale Superieure (ENS), Lyon, France. *Proteomics* **7,** 3436–3440
5. No authors listed. (2009) Credit where credit is overdue. *Nat. Biotechnol.* **27,** 579
6. Barsnes, H., Vizcaíno, J. A., Eidhammer, I., and Martens, L. (2009) PRIDE Converter: making proteomics data-sharing easy. *Nat. Biotechnol.* **27,** 598–599
7. Csordas, A., Ovelleiro, D., Wang, R., Foster, J. M., Ríos, D., Vizcaíno, J. A., and Hermjakob, H. (2012) PRIDE: quality control in a proteomics data repository. *Database* **2012**, bas004
8. Vizcaíno, J. A., Côté, R., Reisinger, F., Foster, J. M., Mueller, M., Rameseder, J., Hermjakob, H., and Martens, L. (2009) A guide to the Proteomics Identifications Database proteomics data repository. *Proteomics* **9,** 4276–4283
9. Côté, R., Reisinger, F., Martens, L., Barsnes, H., Vizcaino, J. A., and Hermjakob, H. (2010) The Ontology Lookup Service: bigger and better. *Nucleic Acids Res.* **38,** W155–160
10. Barsnes, H., Cote, R. G., Eidhammer, I., and Martens, L. (2010) OLS Dialog: An open-source front end to the Ontology Lookup Service. *BMC Bioinformatics* **11**, 34
11. Montecchi-Palazzi, L., Beavis, R., Binz, P. A., Chalkley, R. J., Cottrell, J., Creasy, D., Shofstahl, J., Seymour, S. L., and Garavelli, J. S. (2008) The PSI-MOD community standard for representation of protein modification data. *Nat. Biotechnol.* **26,** 864–866
12. Wang, R., Fabregat, A., Ríos, D., Ovelleiro, D., Foster, J. M., Côté, R. G., Griss, J., Csordas, A., Perez-Riverol, Y., Reisinger, F., Hermjakob, H., Martens, L., and Vizcaíno, J. A. (2012) PRIDE Inspector: a tool to visualize and validate MS proteomics data. *Nat. Biotechnol.* **30,** 135–137
13. Nesvizhskii, A. I., and Aebersold, R. (2005) Interpretation of shotgun proteomic data: the protein inference problem. *Mol. Cell. Proteomics* **4,** 1419–1440
14. Colaert, N., Barsnes, H., Vaudel, M., Helsens, K., Timmerman, E., Sickmann, A., Gevaert, K., and Martens, L. (2011) Thermo-msf-parser: an open source Java library to parse and visualize Thermo Proteome Discoverer msf files. *J. Proteome Res.* **10,** 3840–3843
15. Montecchi-Palazzi, L., Kerrien, S., Reisinger, F., Aranda, B., Jones, A. R., Martens, L., and Hermjakob, H. (2009) The PSI semantic validator: a framework to check MIAPE compliance of proteomics data. *Proteomics* **9,** 5112–5119
16. Jones, A. R., Eisenacher, M., Mayer, G., Kohlbacher, O., Siepen, J., Hubbard, S., Selley, J., Searle, B., Shofstahl, J., Seymour, S., Julian, R., Binz, P. A., Deutsch, E. W., Hermjakob, H., Reisinger, F., Griss, J., Vizcaino, J. A., Chambers, M., Pizarro, A., and Creasy, D. (2012) The mzIdentML data standard for mass spectrometry-based proteomics results. *Mol. Cell. Proteomics* **11,** M111.014381
17. Martens, L., Chambers, M., Sturm, M., Kessner, D., Levander, F., Shofstahl, J., Tang, W. H., Römpp, A., Neumann, S., Pizarro, A. D., Montecchi-Palazzi, L., Tasman, N., Coleman, M., Reisinger, F., Souda, P., Hermjakob, H., Binz, P. A., and Deutsch, E. W. (2011) mzML–a community standard for mass spectrometry data. *Mol. Cell. Proteomics* **10**, R110.000133
18. Häkkinen, J., Vincic, G., Månsson, O., Wårell, K., and Levander, F. (2009) The proteios software environment: an extensible multiuser platform for management and analysis of proteomics data. *J. Proteome Res.* **8,** 3037–3043
19. MatrixScience (2012) Mascot Parser API, http://www.matrixscience.com/msparser.html.
20. Reisinger, F., Krishna, R., Ghali, F., Ríos, D., Hermjakob, H., Antonio Vizcaíno, J., and Jones, A. R. (2012) jmzIdentML API: A Java interface to the mzIdentML standard for peptide and protein identification data. *Proteomics* **12,** 790–794
21. Muth, T., Vaudel, M., Barsnes, H., Martens, L., and Sickmann, A. (2010) XTandem Parser: an open-source library to parse and analyse X!Tandem MS/MS search results. *Proteomics* **10,** 1522–1524
22. Côté, R. G., Reisinger, F., and Martens, L. (2010) jmzML, an open-source Java API for mzML, the PSI standard for MS data. *Proteomics* **10,** 1332–1335
23. Griss, J., Reisinger, F., Hermjakob, H., and Vizcaíno, J. A. (2012) jmzReader: A Java parser library to process and visualize multiple text and XML-based mass spectrometry data formats. *Proteomics* **12,** 795–798