



Proteomics Data Collection (ProDaC) Work Package 5, Deliverable 5.2, Public Report M9

Bochum, June 29th, 2007

M. Eisenacher¹, L. Martens², M. Hamacher¹, T. Hardt¹, H. E. Meyer¹, C. Stephan¹

1 Medizinisches Proteom-Center, Ruhr-Universitaet Bochum, ZKF E.043, Universitaetsstr. 150, D-44801 Bochum, Germany

2 European Bioinformatics Institute, EMBL Outstation, Hinxton, Cambridge, UK

Corresponding Authors:

Christian Stephan, Christian.Stephan@ruhr-uni-bochum.de

Martin Eisenacher, Martin.Eisenacher@ruhr-uni-bochum.de

phone: +49 / 234 / 32 – 29275

fax: +49 / 234 / 32 – 14554

Contents

Proteomics Data Collection (ProDaC) Work Package 5, Deliverable 5.2, Public Report M9	1
Contents	1
Work Package 5: Data Flow Management – Short Summary	2
Objectives.....	2
Involved Partners	2
Deliverable W5.D2	2
Background	2
Introduction.....	2
Overview	3
Results stored in Microsoft Excel files (xls format)	4
Mascot mgf/dat and SEQUEST dta/out files	5
Ontology Lookup Service (OLS).....	6
Trans-Proteomic Pipeline.....	6
Local LIMS	6
Submission to PRIDE	7
Review process	8
Conclusions.....	9
Funding	9



ProDaC website: <http://www.fp6-prodac.eu/>

ProDaC mailing list: <http://lists.ruhr-uni-bochum.de/mailman/listinfo/prodac>

Abbreviations

MPC: Medizinisches Proteom-Center, Germany

EBI: European Bioinformatics Institute, United Kingdom

ETH: Federal Inst. of Technology ETH Zurich, Switzerland

VIB: Department of Medical Protein Research, University of Ghent, Belgium

UCD, Conway Institute of Biomolecular and Biomedical Research University College Dublin, Ireland

LundU: Lund Swegene, Bioinformatics Facility, Lund University, Sweden

GB: GeneBio, Switzerland

BDAL: Bruker Daltonics GmbH, Germany

MS: MatrixScience, United Kingdom

About this Report

ProDaC consists of 7 work packages. At defined time points the work packages have to fulfill certain aims or to provide specific deliverables as for example reports. Some of the reports are confidential and only spread within the consortium; this report is a public one.

Work Package 5: Data Flow Management – Short Summary

Objectives

An optimal data workflow and management are mandatory for the progress and quality of ProDaC. Therefore, MPC will ensure the correct data flow between the participants and the data storage platform PRIDE [1, 2]. As it is planned that the submission of supplementary data will be obligatory for publications, the data submission system will be tested by extensive interaction with the journals.

Involved Partners

MPC (leader), EBI, ETH, VIB, UCD, LundU, GB, BDAL, MS

Deliverable W5.D2

“Elaboration of the data submission pipeline”

Background

According to the needs of the different partners, a data submission pipeline will be elaborated in an iterative process. The MPC will test the data submission pipeline within the own group. After successful implementation, the MPC will project the schema on the core partners and finally at the associated partners' sites. For this purpose, all new generated consortium communication platforms will be used (WP1).

Introduction

In the previous report of work package 5 (deliverable D5.1) it is derived, that the workflow of Proteomics data in the consortium has to consider two possibilities of data flow:

- i) A file-related branch, where files stored in the file system of instrument computers or at central file locations of an institute have to flow into the ProDaC repository. This branch includes input and output files from the Mascot™ and SEQUEST™ search engines, as well as files in the standard documentation formats mzData and mzXML. Furthermore files from the suite of proteomic applications “Trans-Proteomic Pipeline” established at the Institute

for Systems Biology have to be considered and result files stored in Microsoft Excel format. Last but not least, other xml formats and html are used for results at the partners' sites.

- ii) A database-related branch, where data from LIMS systems or existing databases have to flow into the ProDaC central repository. Examples for this branch are ProteinScape, IntelliMS, Proteios and Proline.

Regarding i) it is a crucial point, that spectra and results are stored into separate files, so before storing them into the ProDaC repository, a correct assignment of related spectra/result files has to be performed and the contents have to be merged.

Furthermore the sample and protocol information may be missing, especially in the file-related branch. In both branches the correct usage of standardized controlled vocabulary or ontology terms as demanded by the standard formats is not expected yet.

Overview

The idea of the data submission pipeline as derived from the ProDaC project description and the aims of work package 5 is pointed out in figure 1.

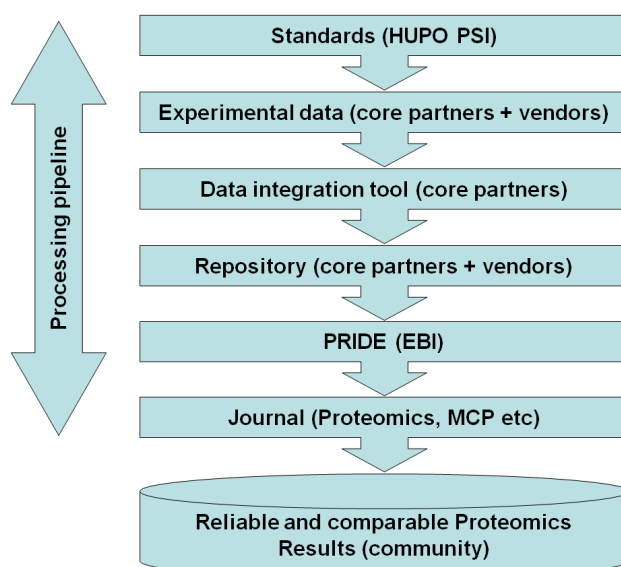


Figure 1: Idea of the ProDaC Data Submission Pipeline

A more detailed incorporation of this idea after the analysis of the relevant tools used in the ProDaC consortium (work package 3) and the analysis of storage solutions (work package 5, deliverable D5.1) is given in figure 2. The different possibilities of data flow are described in the next sections.

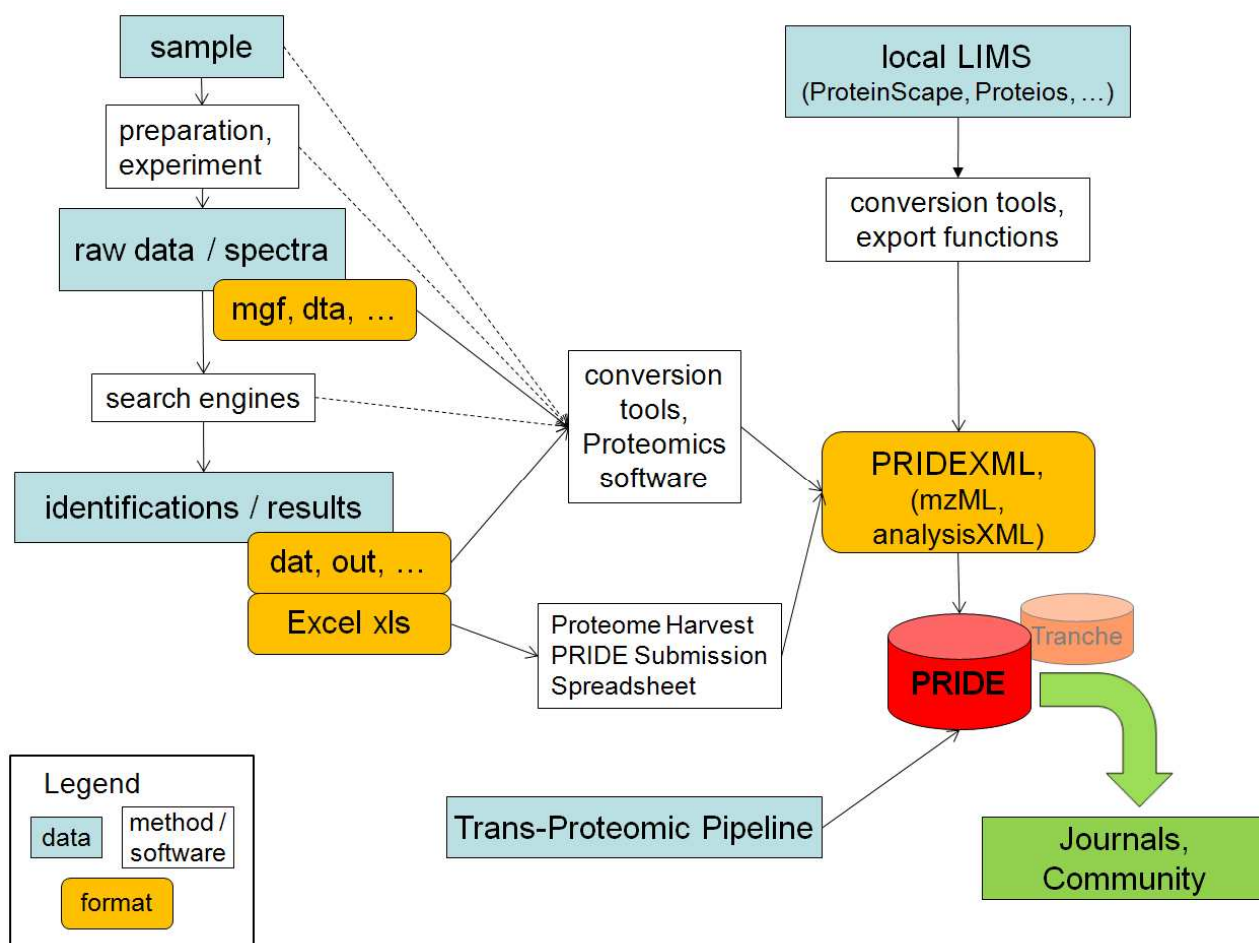


Figure 2: Details of Data and Information Flow

Results stored in Microsoft Excel files (xls format)

For protein identification results stored in Excel, a simple possibility is required to upload them into PRIDE. The EBI has already set up an interactive Excel spreadsheet (called Proteome Harvest PRIDE Submission Spreadsheet) to provide laboratories with the ability to prepare PRIDE 2.1 XML and therefore to submit data to PRIDE.

The current version of the spreadsheet allows the generation of a complete PRIDE XML file but does not allow the import of mass spectrometry data (e.g. peak list data) unless this is in mzData format (Version 1.05).

As well as generating valid XML, this sheet also includes direct access to the Ontology Lookup Service at the EBI. This allows to search for appropriate controlled vocabulary and ontology terms without needing to leave the spreadsheet. The only requirement to use this functionality is internet access when working with the spreadsheet. Figure 3 shows an excerpt of the worksheet.

Once a valid XML file is created, it can be submitted to PRIDE (see below). The first sheet of the Microsoft Excel workbook (entitled 'Introduction' on the tab at the bottom) includes complete instructions for populating the spreadsheet.

See <http://www.ebi.ac.uk/pride> for details and comprehensive Adobe Flash tutorials.

This sheet contains information related to the instrument used for Mass Spectrometry. Note that at as a minimum you must supply at least one parameter describing the source, the analyzer components and the detector.						
Parameters Describing the Source (At least one parameter is mandatory)						
CV Params				User Params		↓ Status
↓ CV Name	↓ Accession	↓ Term Name	↓ Value (optional)	Double Click to activate buttons below		
PSI	PSI:1000057	ElectrosprayInlet		Search for CV Term	Add Param	cvParam - OK
		Diameter	50nM	Search for CV Term	Add Param	userParam - OK
Parameters Describing the Analyzer Components (At least one analyzer component is mandatory)						
Note that each component may be described by one or more CV / User parameter terms. To indicate the ordering and grouping of these terms, please enter an integer in column A below.						
Parameters				User Params		↓ Status
CV Params				User Params		
↓ Analyzer Component Number	↓ CV Name	↓ Accession	↓ Term Name	↓ Value (optional)	Double Click to activate buttons below	
1	PSI	PSI:1000081	Quadrupole		Search for CV Term	
1			Volume	3 mL	Search for CV Term	Add Param
2	PSI	PSI:1000083	RadialEjectionLinearIonTrap		Search for CV Term	Add Param
2			Length	4 cm	Search for CV Term	Add Param
3	PSI	PSI:1000078	AxialEjectionLinearIonTrap		Search for CV Term	Add Param

Figure 3: Excerpt of the instrument details input section of the Proteome Harvest PRIDE Submission Spreadsheet

Mascot mgf/dat and SEQUEST dta/out files

Mascot peak lists (mgf format) and result files (dat format) and SEQUEST peak lists (dta format) and result files (out format) are the most frequently used file formats in the consortium. They will be converted into standard formats like PRIDE XML, mzML or analysisXML by early conversion tools (developed e.g in work package 3) and later by implementations in the various Proteomics software systems.

Similar to the Proteome Harvest PRIDE Submission Spreadsheet in a first step the spectra and result files will be collected and converted. Where necessary, missing sample information and information about experiment and preparation procedures will be queried semi-interactively from the user incorporating the appropriate CV / ontology terms.

A prototype of a conversion tool developed in work package 3 – named ProCon (Proteomics Conversion Tool) is shown in figure 4.

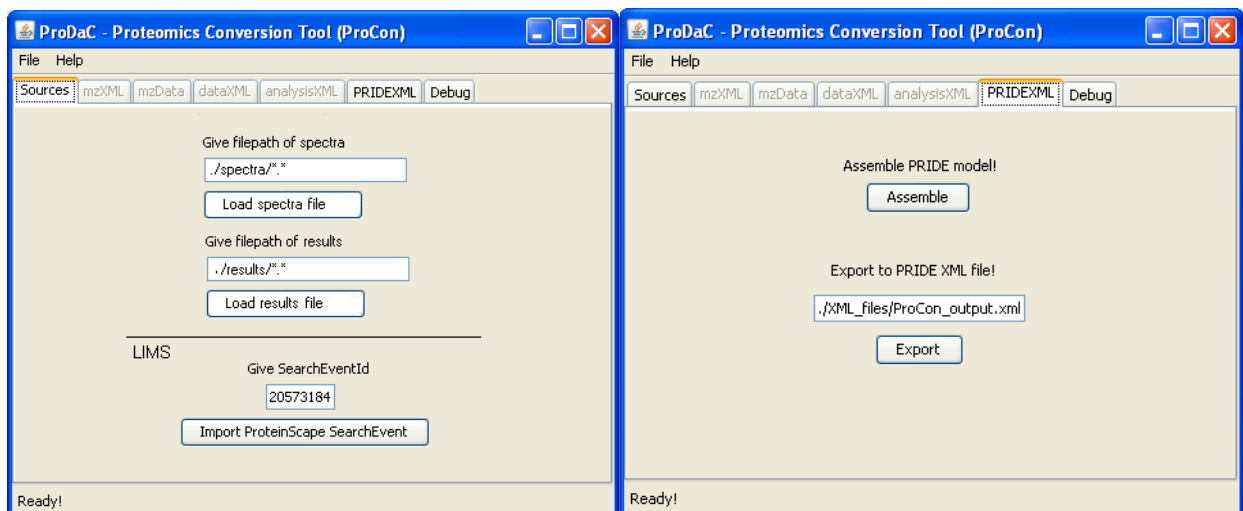


Figure 4: The sources tab and the PRIDEXML tab of the ProCon prototype

Ontology Lookup Service (OLS)

To support non-Bioinformatics scientists with an easy-to-use interface for browsing and searching ontologies (e.g. PSI CVs), the EBI has implemented an Ontology Lookup Service (OLS, <http://www.ebi.ac.uk/ols>, [3]). In a specified ontology (e.g. PSI-MS) a term or accession number can be searched or the ontology as a whole can be browsed or visualized in a tree-like view (see figure 5).

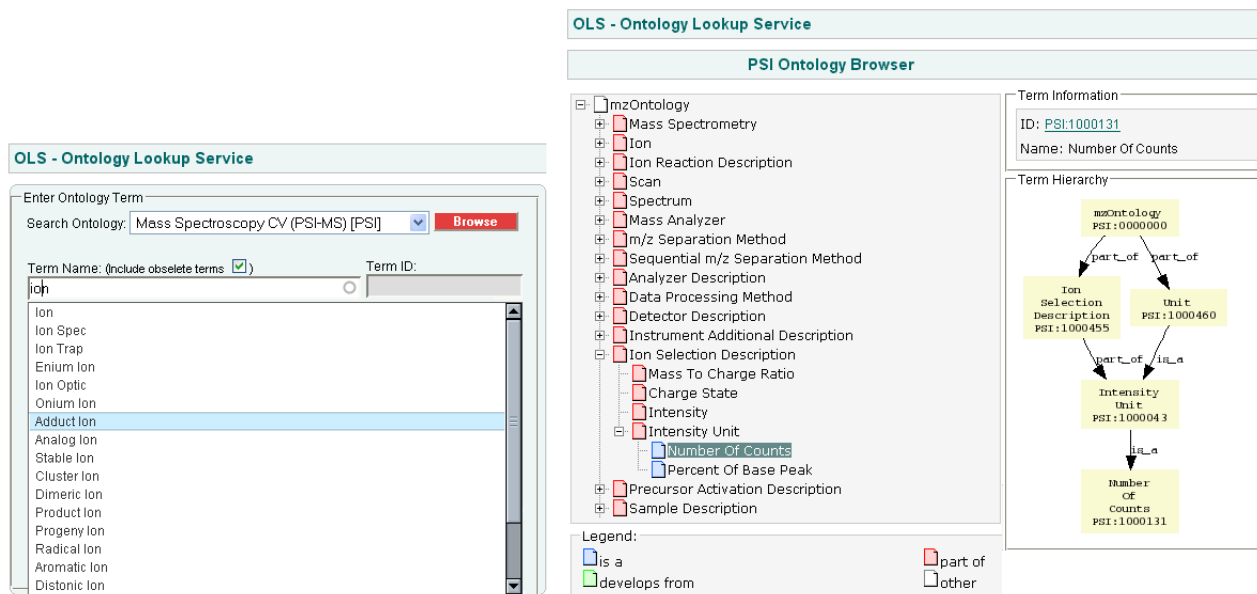


Figure 5: Ontology Lookup Service: Search (left) and Browse (right) functionalities

In addition to the “manual” interface to OLS there are possibilities to connect from own applications and developments. An alternative for own developments is the use of the appropriate OBO files (Open Biological Ontology, <http://obofoundry.org/>) of the OBI consortium (Ontologies for Biological Investigations, <http://obi.sourceforge.net/>). An OBO file contains ontology terms and the relations between them.

Local LIMS

A LIMS (Laboratory Information Management System) in the scope of ProDaC is defined as a database system containing information about samples, experiments, instruments, spectra and search engines. Whether it can perform identification searches or not is not relevant for this definition; important is the possibility to store performed identification runs and their parameters and results.

With this definition ProteinScope is a LIMS system, which may be locally installed at a partner site or which may be set up as a Data Collection Center like in the Human Brain Proteome Project consortium (HBPP, <http://www.hbpp.org/>, [4]).

In the data submission pipeline established in work package 5 conversion tools or export functionality will allow the data flow from ProteinScope to PRIDE XML. In the ProCon prototype first steps in that direction have been already implemented (see figure 4).

Trans-Proteomic Pipeline

The Trans-Proteomic Pipeline (TPP, <http://tools.proteomecenter.org/TPP.php>) developed at the Institute for Systems Biology (ISB) allows the conversion of raw mass spectrometry data to the mzXML peak list format by utilizing various tools (called sashimi tools, <http://sashimi.sourceforge.net/>). After identification of a set of mzXML files using a search engine such as SEQUEST, Mascot or X!Tandem, the results can be read by the post-processing tool PeptideProphet [5]. The results of this step are written to a single file in the pepXML format. These processed peptide identifications can then serve as an input to the protein assembly tool called ProteinProphet [6], which stores a protXML file as output.

The TPP can be pipelined to PRIDE formats by converting the mzXML files to the PSI standard mzData and pepXML / protXML to the identification section of PRIDE XML. The EBI has established a software library for that task and successfully uploaded several data sets from TPP to PRIDE (e.g., PRIDE experiment accession numbers 1755-1772, inclusive). An impression of the object model incorporated to realize this pipelining is shown in figure 6.

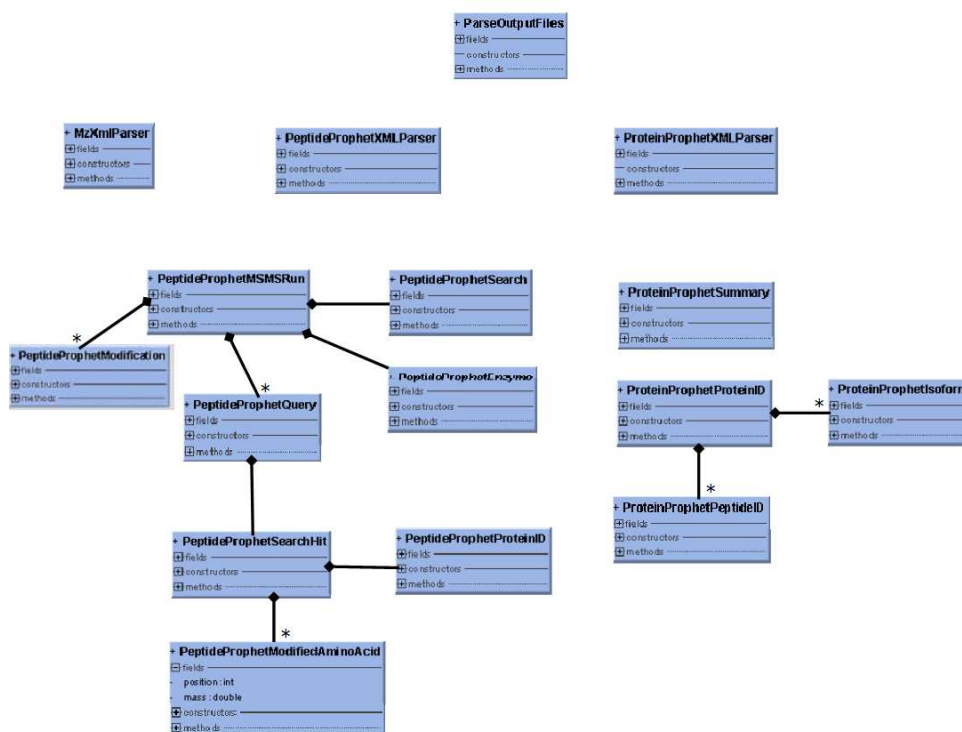


Figure 6: Objects used for pipelining TPP to PRIDE

Submission to PRIDE

Once a valid PRIDE XML is created, the submission to the PRIDE database is done with the help of a web-based submission form as shown in figure 7. As an additional functionality, an mzData 1.05 file can also be submitted as-is, allowing PRIDE to function as a pure mass spectrometry data repository as well. The submitted data set can be made “private” (explanation see next section) or can be made publicly available directly. A submission can replace a previous submission (controlled via check box) and accounts for reviewers can be automatically generated (also by simply selecting a checkbox).

Data Submission Form	
Check the 'Private Data?' check-box below if you wish to restrict access to the data you are uploading. <ul style="list-style-type: none"> • If unchecked, the data will be visible to all users as soon as it has been verified. • If checked, you will be given additional options to select a date when the data will become publicly available and to select collaborations that can view the data before it becomes public. 	
Private Data?	<input checked="" type="checkbox"/>
* Select a file to upload	<input type="text"/> <input type="button" value="Durchsuchen..."/>
* File Type	PRIDE 2.1 XML: <input checked="" type="radio"/> mzData 1.05 XML: <input type="radio"/>
Check the 'Replace Previous Submission?' check-box below if you are re-submitting data. Note that the contents of the 'ExperimentAccession' element for each experiment (integer value) must be the same as those of the experiment(s) that you are replacing. You will only be able to replace experiments that you originally submitted using your current login username. If you check the 'Replace Previous Submission?' check-box and the experiment accession number is not recognised, or has been submitted by somebody else, the submission will fail and no change will be made to the data held in PRIDE. (Submission success or failure will be reported back to you). Please note that you do not need to set the mzData accessionNumber element. This will automatically be set to the same value as the experiment accession number.	
Replace Previous Submission?	<input type="checkbox"/>
Check this box to automatically create accounts for reviewers if you are submitting data associated with a journal publication.	<input type="checkbox"/>
<input type="button" value="Upload Selected"/>	

Figure 7: PRIDE Data Submission Form

Review process

During the submission of a data set to PRIDE, it can be assigned a “private” status. “Private” data sets can be only seen and browsed by the owner and “collaboration” partners (see figure 8, left). New collaborations can be set up by registered PRIDE users. During the submission process an optional date can be specified when the data becomes available to the collaboration and another optional date, when the data set will be publicly available.

Note that this mechanism implicitly supports the peer review process (see figure 8, right): by checking the ‘create reviewer accounts’ checkbox (see figure 7), the submitter instructs PRIDE to set up a collaboration containing a PRIDE reviewer account. When submitting the manuscript to the journal, the author forwards the reviewer account to the Editor, or includes the login information in the submitted manuscript. The reviewers will then either receive the account from the Editor, or will find the details in the manuscript they are reviewing. They can then log in into PRIDE after which they will be able to review the private data set together with the manuscript. Once the manuscript is accepted, the PRIDE data set can obviously be made publicly available extremely easily.

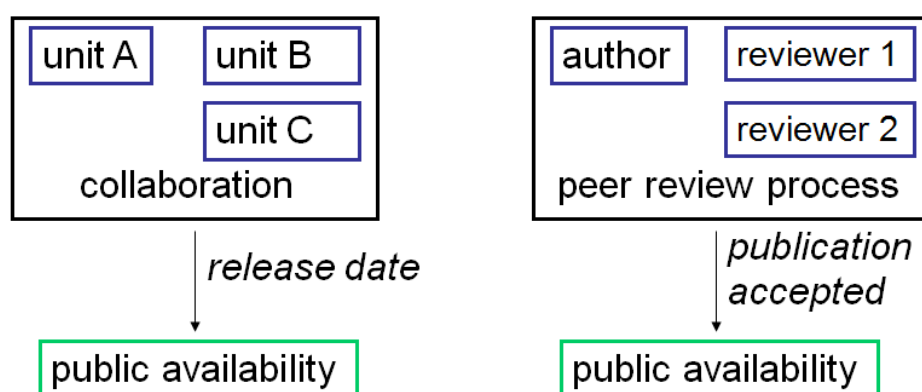


Figure 8: Privacy concept in PRIDE: collaboration view (left) and peer review view (right)

Conclusions

The establishment of the data submission pipeline is finished. Some of the flow possibilities are already implemented and can be used. A crucial task will be the finalization of the ProCon tool for conversion of Mascot mgf/dat and SEQUEST dta/out files into PRIDE XML and the import of peak lists and result data from ProteinScape. In both cases, a semi-interactive and semi-automatic way of sample and experiment annotation using CVs/ontologies will be an essential component.

Depending on the schedule some of the other file formats or storage solutions used in the consortium may be adopted to go through the ProDaC pipeline.

The pipeline will be tested inside the MPC (first during the development process, but afterwards also within the non-Bioinformatics working groups), then at the core partners' sites and finally at the associated partners' sites. The feedback of these tests will help to improve the tool.

Funding

ProDaC is funded by the European Commission as a Coordinated Action within the 6th European Union Research Framework Programme (Contract No.: 036814).

References

- [1] Martens L, Hermjakob H, Jones P, Adamski M, Taylor C, States D, Gevaert K, Vandekerckhove J, Apweiler R 2005 PRIDE: the proteomics identifications database. *Proteomics*. 2005 Aug;5(13):3537-45
- [2] Jones P, Côté RG, Martens L, Quinn AF, Taylor CF, Derache W, Hermjakob H, Apweiler R 2006 PRIDE: a public repository of protein and peptide identifications for the proteomics community. *Nucleic Acids Res*. 2006 Jan 1;34(Database issue):D659-63
- [3] Côté RG, Jones P, Apweiler R, Hermjakob H 2006, The Ontology Lookup Service, a lightweight cross-platform tool for controlled vocabulary queries. *BMC Bioinformatics* 2006; 7:97.
- [4] Hamacher M, Stephan C, Eisenacher M, Lewczuk P, Wiltfang J, Martens L, Vizcaíno JA, Kwon KH, Yoo JS, Park YM, Beckers J, Horsch M, de Angelis MH, Cho ZH, Apweiler R, Meyer HE 2007 High Performance Proteomics: 7(th) HUPO Brain Proteome Project Workshop March 7-9, 2007 Wellcome Trust Conference Centre, Hinxton, UK. *Proteomics*. 2007 Jul 3
- [5] Keller A, Nesvizhskii AI, Kolker E, Aebersold R 2002 Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal Chem*. 2002 Oct 15;74(20):5383-92
- [6] Nesvizhskii AI, Keller A, Kolker E, Aebersold R 2003 A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem*. 2003 Sep 1;75(17):4646-58