Proteomics and Bioinformatics group
Departement of Medical Protein Research
VIB and Faculty of Medicine and Health Sciences, Ghent University

http://www.proteomics.be

MS_LIMS MANUAL

Kenny Helsens
Niklaas Colaert
Steffi Wortelkamp
Lennart Martens

http://genesis.ugent.be/ms_lims

Contents

Introduction

Proteomics lims suite

Mass spectrometry based proteomics approaches produce large amounts of mass spectra that require processing, identification and possibly quantification before interpretation can be undertaken. High-throughput studies require automation of these various steps, as well as management of the data in association with the results obtained. We here present ms_lims, a freely available, open-source system based on a central database to automate data management and processing in mass spectrometry driven proteomics analyses.

ms_lims is mainly designed to automate data flow in the high-throughput proteomics lab. Taking spectrum files from a variety of *(pluggable)* fileformats (standard Micromass PKL file and Mascot Generic File support is provided), it transforms these to the Mascot Generic Format and stores them in the database, retaining LC information if present, and also allowing additional information to be stored for each individual LC run.

Another part allows the retrieval of the stored spectra in mergefiles of arbitrary size. These can then be submitted to a search engine, eg. *Mascot* from *Matrix Science*.

Subsequently, the results of these searches can be parsed and stored in a relational database structure for future reference.

ms_lims requires Java version 1.5 or above, which you can get here. Download the binaries here and a start-up script for your platform here.

Part I

Part one, the installation.

1
General

The installation of ms_lims cannot be done by a typical installer. The lims system is dependent from multiple instances that you have to process one by one.
The installation takes four steps.

Step one covers the installation of a Java Runtime Environment(JRE). Ms_lims was created with Java Development Kit(JDK) 1.5 and therefore needs a JRE starting version 1.5 (also known as Java 5) or later.

Step two *(shortly)* covers the installation of a database system. Ms_lims works around a central Relational Database Management System (RDBMS) to control the proteomics data. By preference, the freely available open source MySQL RDBMS is used for storage and manipulation of the proteomics data.

Step three covers the building of a bridge between Mascot and ms_lims.

Step four covers the installation of ms_lims itself.

2

Java

Since ms_lims was developed in Java, it runs on every operating system that has an up-to-date Java installation.

Previous java installation?

Most probably, Java is already installed on your computer due to the widespread use of Java nowadays.

If you do not know for sure if Java 1.5 or higher is installed, do the following:

- Open the shell in linux or the command window in windows

- Enter `java -version`

If Java is already installed, you will see something as below where x stands for the version.

```
{java version "1.x.0_01"}
```

If yours is beyond version 1.5, then you must upgrade Java. If nothing meaningful appears, then Java is not yet installed. In both ways, you have to install a new Java version.

New Java installation

The installation of a Java is quite straightforward.

- Goto http://java.com

- Follow the main download link and download the installer

- When finished, open the installer and follow the instructions

Java should be properly installed by now. Proceed to the next step.

3

Relational Database Management System (RDBMS)

3.1 Why a RDBMS?

> A short definition of a RDBMS may be a DBMS in which data is stored in the form of tables and the relationship among the data is also stored in the form of tables.
> *Wikipedia*

As for a proteomics oriented lims, whether you want to store fragmentation spectra or retrieve peptide identifications - the relation must be saved between the fragmentation spectrum and its peptide identifications. At all times, the database has a central role in the lims for storing, managing and delivering this proteomics data.

Multiple RDMBS systems are availlable: MySQL, Oracle, Firebird, PostgreSQL, MS SQL Server are the most common examples. Even though there are all distinct RDBMS, they are similar as they are all SQL implementations. SQL means Structured Query Language and serves as a language for humans to communicate to the database holding all the proteomics data. All of these SQL implementing databases can be used with ms_lims by using different drivers. Hence, some of these are commercial while other are free open-source driven efforts. We prefer to use the popular open source database MySQL by default and will therefore orient this manual towards this database.

3.2 MySQL

About MySQL

> The MySQL RDBMS has become the world's most popular open source database because of its consistent fast performance, high reliability and ease of use.
> *mysql.com*

For this reason amongst others, the developers of the lims system have chosen MySQL as the RDMBS of preference. A list of short instructions on the installation of MySQL follows.

Getting MySQL

First, we will download the installer from the MySQL website.

- Goto the MySQL website at <http://mysql.com>

- Click on the **download** tab in the top

- Choose the **MySQL community server** to continue

- Select the essential installer of your operating system and download it by the link at the right

After downloading has finished, proceed to the installation.

Installing MySQL

- Open the installer

- Select the typical installation and proceed

- Click the install button to start the intallation of the MySQL server

- Wait for the installation to complete. After completion, enable the **configure now** checkbox and to proceed to configure the database

Now the installation has completed, the MySQL database needs some extra configurations regarding performance and security.

Configuring MySQL

- Verify you are now in the configuration window titled **'MySQL server instance configuration wizard'**

- Select the standard configuration

- Install as a windows service and name it MySQL and make the service launch automatically each time the computer starts

- Modify the security settings and enter a root pasword. Consider this as the 'Master' password for fully controlling the MySQL database. Obviously this is powerful and should therefore not be known by all users.

- Enable root access from remote machines

- Disable the creation of an anonymous account

- Execute!

The MySQL server is up and running now. We can now proceed installing ms_lims itself.

4

Mass Spectrometry LIMS(ms_lims)

Getting ms_lims

An installer is supplied at http://genesis.ugent.be/download/ for your convenience. The installer runs on both Windows or *NIX platforms, provided they run the Java virtual Machine version 1.5 or above.

Installing ms_lims

Open the downloaded installer. The wizard guides you through a few familiar steps such as defining an installation target and license information. After the installation has finished, go to the installation folder. During installation, all required files are copied here. This includes the Java libraries in the core folder, the configuration files in the conf folder. An uninstaller is supplied as well in the corresponding folder.

In the root folder, you find two startup scripts (.bat for windows and .sh for *nix operating systems)

ms_lims starts the main user interface to work with ms_lims

configuration starts the configuration interface to setup the ms_lims database or to add users, instruments and protocols into an existing ms_lims database system

Upon installing ms_lims, we are interested in the configuration interface. Hence, we will come back to the main user interface in the next chapter.

Configuring ms_lims

The configuration tool of ms_lims helps to tie up ms_lims with the MySQL database that was installed before. Different configuration steps are separated in different tabs on the left side.

summary Gives an brief overview of the active setup

database Creates the ms_lims relational database scheme into your MySQL database. A relational schema specifies a collection of table definitions, providing structure for the mass spectrometry data stored in the database instance. Just as fragmentation spectra are stored in one table, peptide identifications are stored in different table. Hence, a connection between both tables is maintained. More, both are connected to another table that reflects which instrument was used. As such, you can eventually retrieve all spectra from a given instrument or all peptides containing a particular sequence.

- **Connect** to the MySQL database system you previously installed by using the 'root' username and password . Fill in the hostname of the computer that has the MySQL database installed. If the MySQL server is installed on this system, fill in **'localhost'** to refer to this system. Otherwise, fill in the name of the computer as it exists in the network.
- **Create SQL Datase** to create a new database on the MySQL database server. Fill in an appropriate name such as **'projects'** to identify the database.
- **Set SQL Scheme for ms_lims** to the (ex.) 'projects' database.

Now the database is structured, it is ready to store mass spectrometry data in a relational manner.

users Add or remove ms_lims users from the database system. An ms_lims user can be a mass spectrometrist storing fragmentation spectra or a informatician storing peptide identification results.

protocols Add or remove ms_lims protocols from the database system.

instrument Add or remove ms_lims instruments from the database system. ms_lims is independent from mass spectrometer vendors by storing fragmentation spectra in a uniform manner. Hence, as different vendors have different forms of output, .pkl files versus .xml or single versus merged files, ms_lims has different engines to store each fragmentation spectrum as a single entity independent from vendors. In this panel you are ought to define which instrument is used in order to store the data into ms_lims appropriately.
If your instrument of choice is not listed, please contact the authors

Ok, the MySQL database is now running a relational database scheme capable for managing proteomics data. Only a few minor steps to go!

5

Tie up the loose ends

Two more issues remain, both are about tying up the loose ends between ms_lims and other software. First is about allowing ms_lims users accessing the MySQL database. Second is about telling ms_lims where to find the peptides identified from fragmentation spectra by Mascot.

Adding MySQL users

It is important to distinguish two user identities. First there is an ms_lims identity for users in the lab environment, being the researchers working on the bench or at the mass spectrometer. These identities are created by configuring ms_lims. Second there is a MySQL identity to access the database itself. While the former is a simple identifier who the project or data belongs to, the latter comes with a password and purely serves to interact with the MySQL database. Currently, there is only one MySQL user: the 'root' user. As you might remember, we named this the 'master' user as it comes with full control. This user can create other users with equal or minor permissions. We will now add new users to the database to tie up ms_lims to the MySQL database.To add new users to the MySQL system, we prefer to use the **MySQL Administrator Tool**.

- Goto the MySQL website at http://mysql.com

- Click on the **download** tab in the top

- Choose the **GUI Tools** on the left to continue

- Select the essential installer of your operating system and proceed by the **'mirror'** link at the right

- To avoid registration, click **'No thanks, just take me to the downloads!'**

13

- Select the HTTP or FTP link from a location near to you to download the GUI Tools

- Open the installer and folow the straightforward installation instructions

After the installation of the GUI Tools has finished, locate and start the MySQL Administrator tool. Before the actual Administrator tool starts, we must establish a connection to the MySQL database we want to configure. In our case, this is the MySQL database we have just installed.

- Fill in the hostname of the computer that has the MySQL database installed. If the MySQL server is installed on this system, fill in **'localhost'** to refer to this system. Otherwise, fill in the name of the computer as it exists in the network.

- Enter **'root'** as the username

- Enter the password you entered while modifying the security settings for the 'root' user (we also referred this as the 'Master' password)

- Establish a connection to the MySQL database

This MySQL administrator tool allows you to configure all types of settings of the MySQL system. All of these are thoroughly documented at the website but *we recommend the default settings*. We will now use the Administrator tool to create new users in the database.

- Select the **'user administration'** in the left

- Select the 'add new user' button in the bottom

- Fill in the name for the MySQL user and a password (you are free to fill in more personal information)

- Select the **Scheme Privileges'** tab in the top

- Select the ms_lims database beyond the Schemata header **Scheme Privileges'** tab in the top

- Enable the 'SELECT', 'INSERT' and 'UPDATE' prilege into the Assigned Privileges by clicking the single arrow buttons

- Save the new user by clicking 'Apply Changes' in the bottom

- ..

- Repeat this procedure untill every user has access to the MySQL database

Ok, by now we have a ms_lims database scheme running on a MySQL server that can be accessed by multiple users. One last issue remains: peptide identifications by Mascot.

Mascot Daemon Task Database

ms lims does not provide a tool to start Mascot database searches, but relies on well the established Mascot daemon tool instead. All information concerning the searches exerted by Mascot Daemon (e.g., start time, duration, source spectra, etc..) are stored in a compact Microsoft access database file. By default, this Mascot daemon database file is named *TaskDB.mdb* and can be found in the default Mascot Daemon installation folder. If the appropriate location of this database is provided to ms lims, then the searches that have been run by Mascot daemon and the resulting peptide identifications can be directly imported into ms lims.

The last step in the installation process therefore involves to give ms lims access to the Mascot daemon Task database. On a Windows platform, this can be done easily be creating a system data source named TaskDB which links any request to the Mascot daemon Task database.

Follow the instructions below to create a new system data source:

- Go to **Control Panel** (via the start menu)

- Go to **Administrative Tools**

- Go to **Data Sources**

- Go to the tab named **System DSN**

- Click on the **Add...** button to add a new System data source

- Select (one of) the **Microsoft Access drivers (.mdb)**

- Enter *TaskDB* in the **Data Source Field** *mind that ms lims will only accept a database named 'TaskDB'!!*

- Enter *Mascot deamon task database* in the **Description** field.

- Click **Database** and browse to the MascotDaemon installation folder to select the *TaskDB.mdb*. The default setting thereto is *C://program files/mascotdaemon/*.

- Click **OK** to finish the creation of the system data source.

From now on, ms lims can request the "TaskDB" data source to the operating system, and is subsequently able to read all Mascot daemon searches and import the required searches into ms lims.

Part II

The End User Tools

6

Overview of end-user tools in ms-lims

After ms-lims has been successfully installed, the end-users can start making use of the client tools that work the database. Some tools serve to insert information in the database, while others assist to extract and interpret the information that has been stored in the database. The hierarchy of the tools is similar to the time course of an experiment. First, a project must be created to encompass all information involved in a single experiment, and this can be done via the ProjectManager tool (6.2). Second, the MS/MS spectra must be stored in the ms-lims database by the SpectrumStorage tool (6.3), and the MS/MS spectra can be extracted from the database subsequently by MergerGUI (6.4) for further analysis by external database search engines. Once the Ms/MS spectra interpretation into peptide identifications by Mascot has been completed, the searches can be stored into ms-lims by IdentificationGUI (6.6). Furthermore, if peptide quantification was performed in parallel by Mascot, then this can be persisted into the database as well by the QuantitationGUI (6.7). Following these storage tools, all information can be further analyzed via standardized project reports of the ProjectAnalyzer tool (6.9) or via custom SQL queries in the GenericQuery tool (6.8). Finally, the results can be validated in depth on the level of peptide identifications by Peptizer (6.10) or peptide quantitations by Rover (6.11). Each of these tools are explained in the following sections.

The appearance of the main GUI for ms_lims is a menu bar and list of buttons. Here you can choose which tool you would like to start. The buttons from up to down follow the workflow which you should meet when working with ms_lims. You can use shortcuts to start the tools, e.g. `Alt+1` starts the ProjectManager. Just try out the pulldown menu in the menu bar. Now we will take a closer look at the tools.
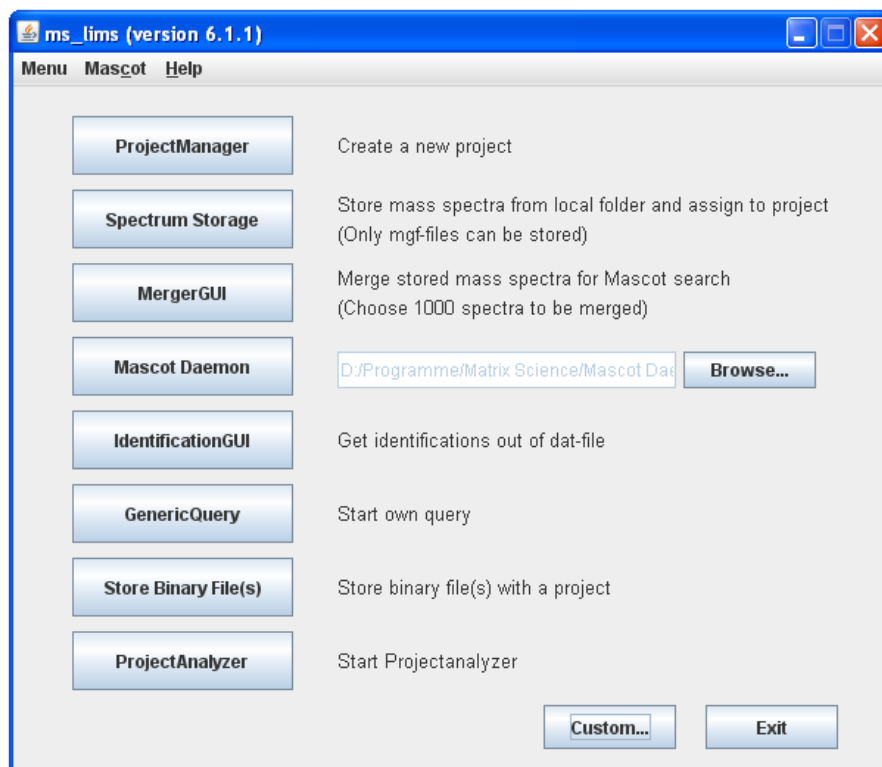
Figure 6.1: The ms_lims menu consists of several buttons. Just click on the buttons to start the corresponding tool.

6.1 Starting ms-lims

The main tools of ms_lims are provided by an application GUI, which can be started by double clicking the ms-lims jar executable file. When you start up ms_lims, the database connection dialog will appear. Enter your username and password here to connect to the ms-lims database.

6.2 Managing projects by the ProjectManager

The ProjectManager simply organize the projects which you can create or modify here. A project includes the mass spectra and mascot searches of an experiment. To get a better overview of your projects, you can check the box *Sort projects alphabetically*.
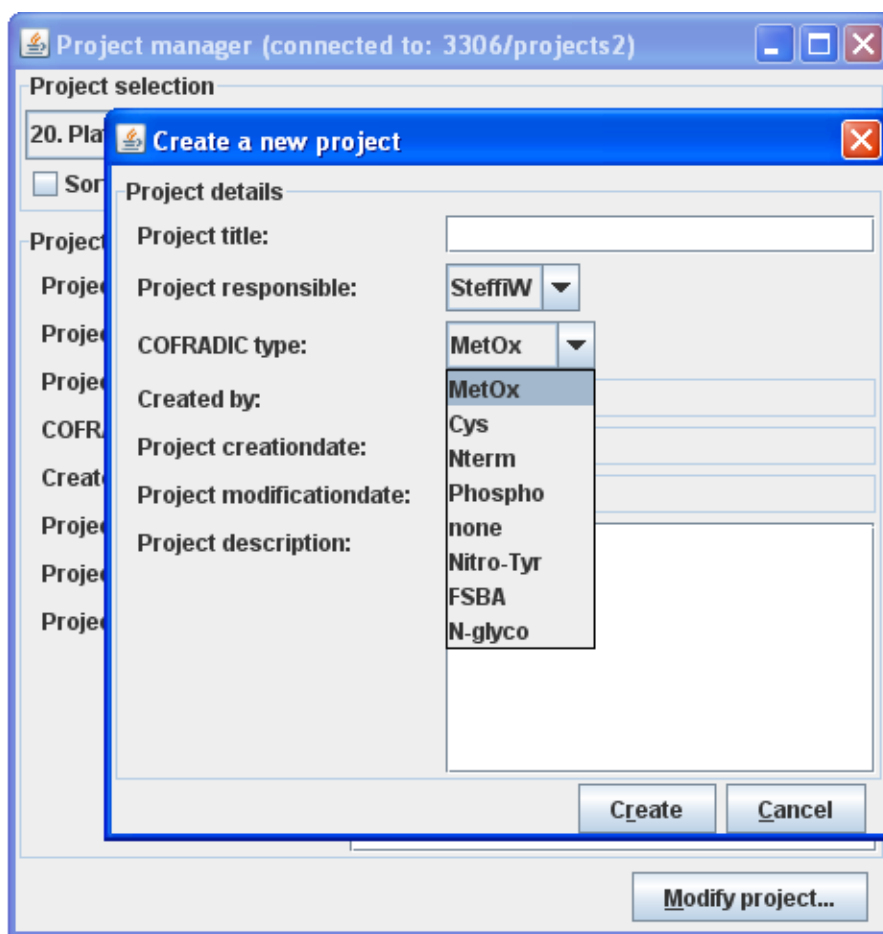
Figure 6.2: The ProjectManager allows you to create new projects.

To create a new project, click on the button *Create new project*. Enter a meaningful title for your project. In the next steps you will identify your project only by this title. Choose the project responsible person then and select the COFRADIC type if you did a COFRADIC experiment. If you did a other kind of experiment, please select *none* (See figure 6.2). The project description allows you e.g. to make a note of the experimental setup of the project. When you have finished select *Create*. A project number will be assigned automatically to your project. If you just want to edit an existing project, click on the button *Modify project*. Don't forget to save the changes on your project. To exit this tool just close the tool window.

6.3 Storing MS/MS spectra by the SpectrumStorageGUI

This tool allows you to store spectra assigned to a project in the database. When you start the tool, there's a dialog where you can choose the mass spectrometer instrument where the spectra were derived from. Please note that the instrument table has to be set up in the database where

the correct storage engine class has to be defined for the relevant instrument.
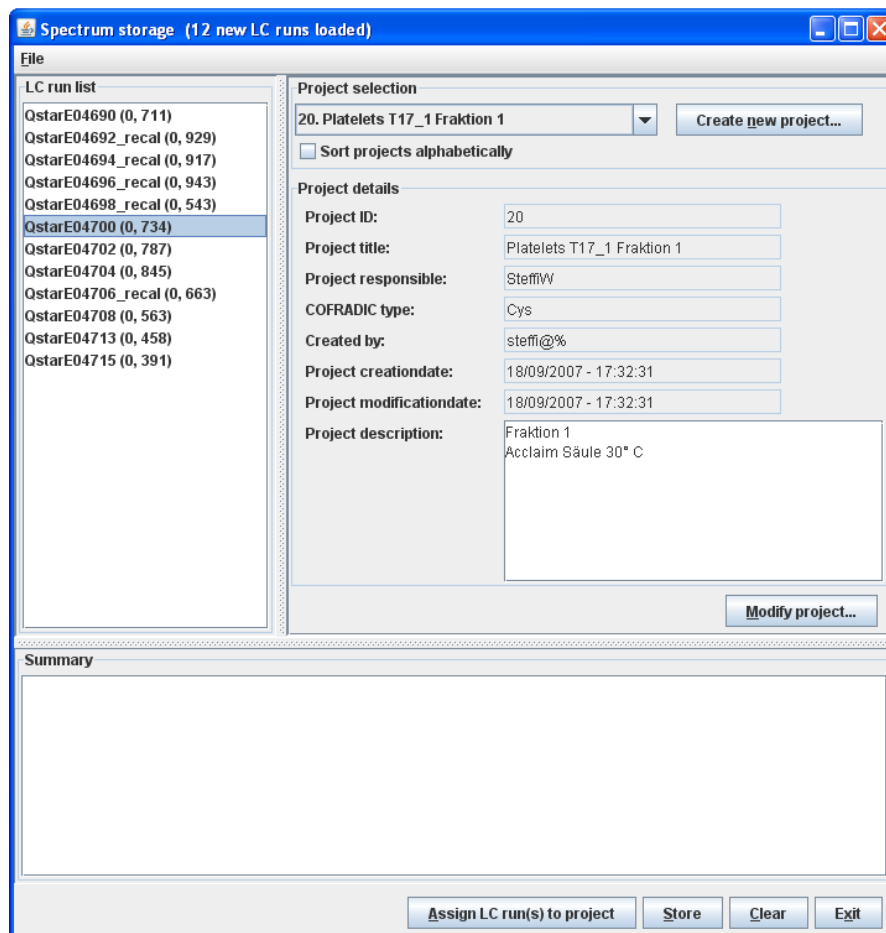
Figure 6.3: Here you can assign mass spectra to a project and store them in the database.

The window of the SpectrumStorageGUI consists of three parts: left, right and lower window.

In the left window (LC run list) you'll see the spectra files you just loaded. By doubleclicking on a file in the LC run list you can add a comment concerning that LC run.

Choose the project to which the spectra belong to in the right window *Project selection*. You also can create or modify the project here. Mark all spectra you want to add to the choosen project in the LC run list. Click on the button *Assign LC run(s) to project*.
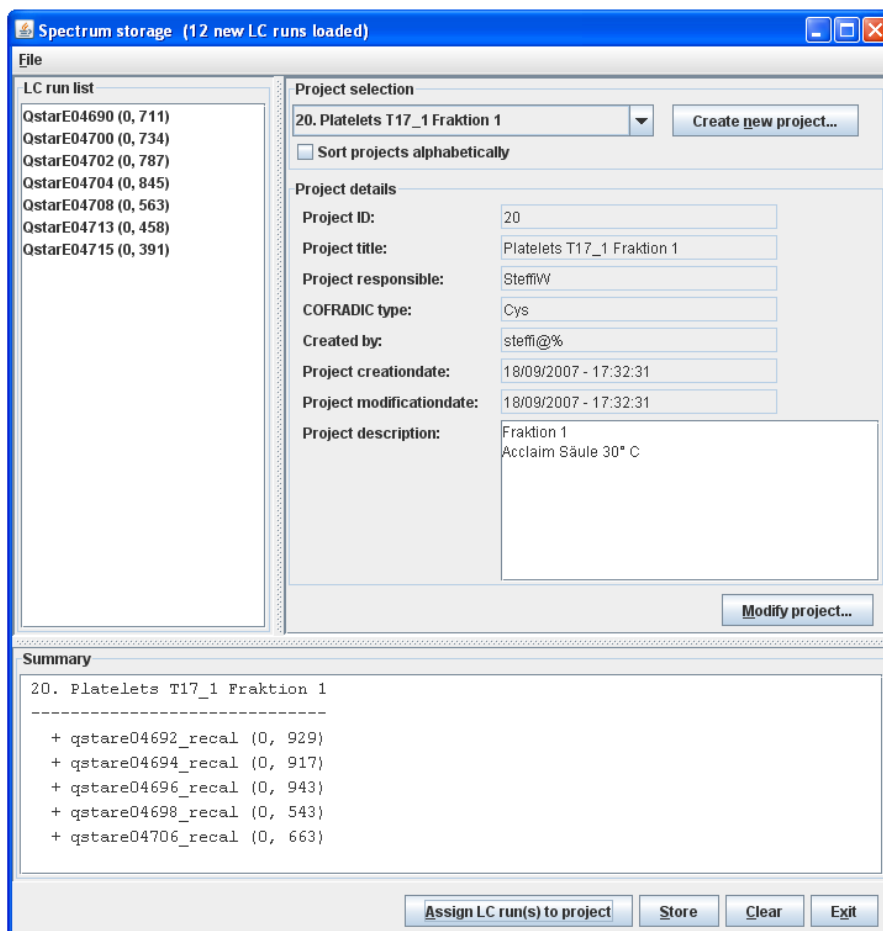
Figure 6.4: Mark the spectra you want to assign to a project in the upper window.

Then you can check your selection in the lower window *Summary*. By pressing *Clear* the selection is cancelled and can be done again. Finally, to commit the selected spectra to the database click the button *Store*. This process can take a while depending on the size and number of spectra.

6.4   Extracting MS/MS spectra from ms-lims by the MergerGUI

The job of the MergerGUI is in short, to merge your stored spectra to a set of files before you submit them to the Mascot search engine.
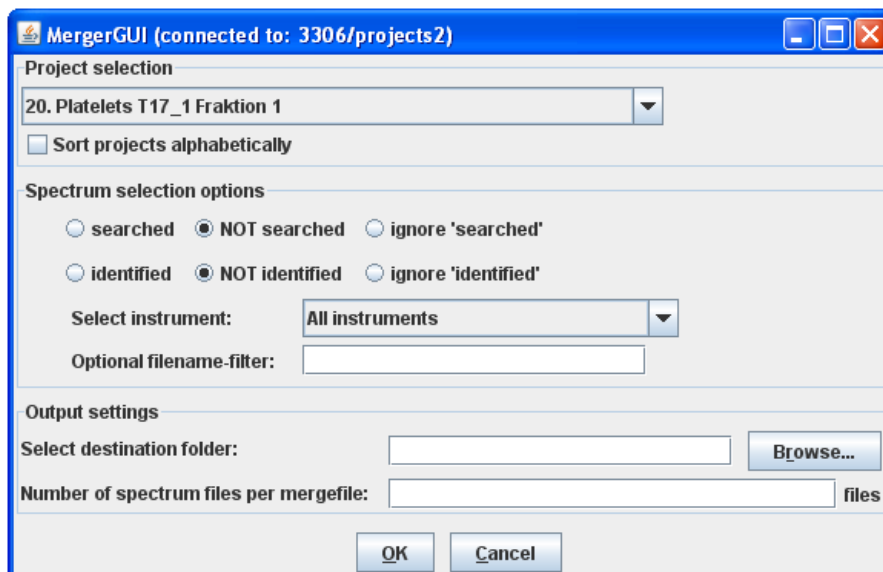
Figure 6.5: The MergerGUI organizes your spectra to a set of files

First choose the project whose spectra you want to be merged. When you merge the files for the first time, do not check the spectrum select options *NOT searched*, *NOT identified* in the section *Spectrum file options*. In the next step, select your instrument where the spectrafiles have been derived from and optional, choose a file filtername if a limit selection of files is needed, e.g. 'LTQ003*'. In the section *Output settings* pick or create a folder where the merged files can be stored. Enter the number of spectra files you want to be merged. A good starting point is a value of 1000 files. The next time if you want to redo a search, check the spectrum select options *searched* and *NOT identified* to get a subset of the spectra which need fitted search parameters for identification for example.

6.5 Identifying MS/MS spectra by Mascot Daemon

Up to now, ms_lims can only process datafiles from the Mascot search engine. This button will start your MascotDaemon application. You can store the local path to the MascotDaemon application permanently in the mascotdaemon.properties file like

```
MASCOTDAEMONFILE=
```

```
C:/Program files/Matrix Science/Mascot Daemon/Daemon.exe
```

or enter the path directly in the field beside the button.
Now select the merged files for searching with Mascot you have created before with the MergerGUI tool. Then start your Mascot search as usual. It is important to use the merged files for searching and not the original spectra files at this point, otherwise ms_lims will prompt an error message. The merged files are connected to the database and to your project.

6.6   Storing peptide identifications by the IdentificationGUI

This tool allows you to parse and view your Mascot search results directly from the Mascot dat files and finally store them in the database. IdentificationGUI connects with your task database from MascotDaemon, so make sure that ODBC coupling with the task database is present (see section XX). A tree list of your actual Mascot Daemon queries stored in the Mascot Task Database should appear, starting with the last task you have performed recently. Here you select the results you want to parse. Click on the tree nodes to see which files belong to you search. Select one or multiple searches from the tree list, then mark your desired result file(s) in the right window. The identity threshold score applied for the extraction of the selected files can be set at the bottom of the window. The default is set at the Mascot standard 95% confidence which allows a maximum of 5% false positive identifications.
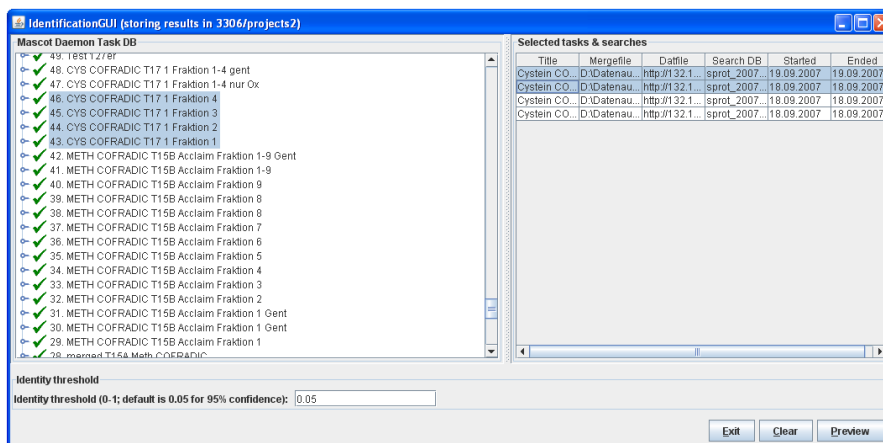
Figure 6.6: Choose the Mascot searches you want to parse

By selecting the *Preview* button the tool will start to retrieve the results and parse them. The procceding is indicated by a progressbar and according to the filesize this task can take a while. When the parsing of the results has finished, a table is displayed which contains the identified peak list from the selected search.

Figure 6.7: The Preview table shows a list of all identified peptides of the selected search

Here you can simply copy and paste the results to other applications as spreadsheet for example. *Hint:* Of course you can use the IdentificationGUI tool just to parse your Mascot searches without storing them to the database. You can adjust this preview table to your requirements and sort a certain column (descending or ascending), resize or move a column. If the *Column selection mode* is checked you can mark columns for copy and paste instead of marking the rows as normally. The colunms in the preview table presents following data:

Filename Refers to the original filename the spectra came from. Sometimes also some spectra header information can be seen here

Accession The protein accession as received from the search database. A left mouse button double click opens the entry the uniprot database

Sequence The peptide sequence as received from the search database

Modified sequence The peptide sequence is supplemented with the fixed or variable modifications if they occur. Note: a star marks the modifcation as fixed e.g. (Cmm*)

Ion Coverage Highlights the ionseries; for y-ions: red font colour, for b-ions: underlined

Start/End Start/End of the peptide within the protein sequence

Description The description as received from the search database

Title Refers to the search title in mascot

Score/Threshold Refers to the score/threshold from mascot

Confidence As set in IdentificationGUI before parsing

Calcutated/Experimental mass Refers to the calcutated/experimental mass from mascot

Isoforms Lists isoforms to that peptide

Precursor (m/z) The precursor mass of that peptide

Charge The charge of that peptide

Enzymatic Refers to the enzyme cleavage state: FE - correct enzymatic cleavage; NE - n-terminal correct; CE - c-terminal correct; EE - fully incorrect

Datfile The name of datfile which has been generated on this search

Search database filename Exact filename of the search database

Search DB Name of the search database

If you decided that the data in the preview table can be stored in the database, just press the *Store* button. Again you will see a progressbar and finally a small box informing you that all identifications have been stored in the database.

6.7   Storing peptide quantitations by the QuantiationGUI

Similar to the IdentificationGUI tool, the QuantitationGUI tool will first process all peptide quantitations associated with a set of peptide identifications from a Mascot search. If the peptide identifications are successfully mapped to peptide quantitations, then they are previewed prior to the final storage into the database.

6.8   Retrieving information from the database by custom SQL queries in the GenericQuery tool

GenericQuery allows you to do queries within your database using SQL (Structured Query Language). This feature is for advanced users, because you need a basic knowledge about how to perform a SQL query.
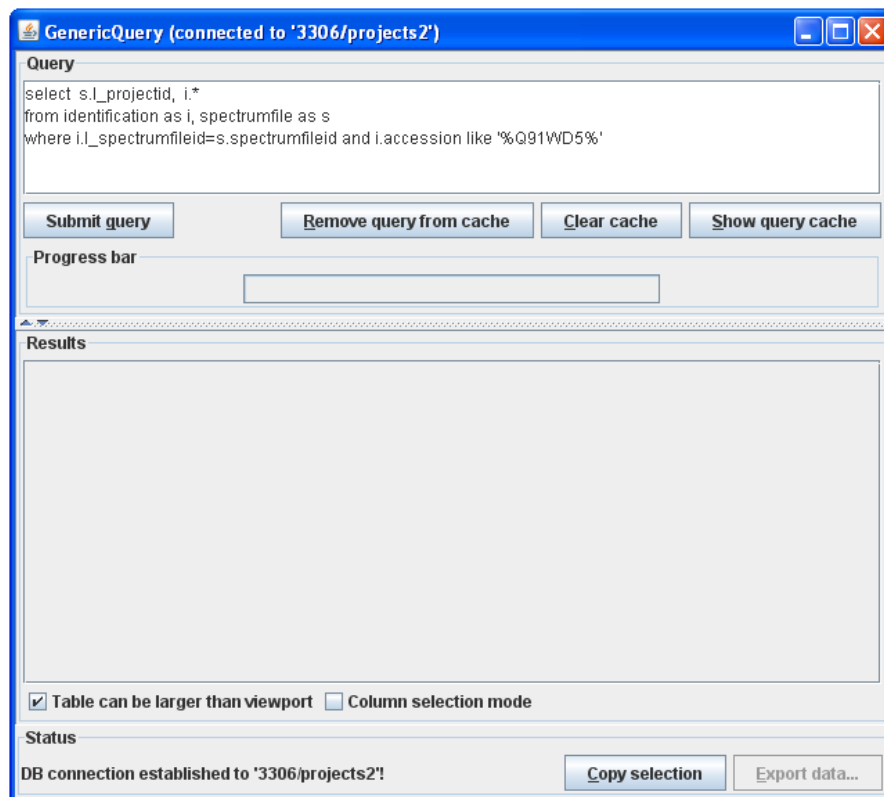
Figure 6.8: The GenericQueryGUI allows access to all tables and stored data by using SQL

In the upper window you can type your query and submit it. Don't forget to connect to your database first (use projects e.g.). All submitted queries are stored in cache, so you can access recent queries (up to 40 entries). The lower window will show your result table, which one can simply copy and paste to spreadsheet applications or export the data in .html or .csv format.
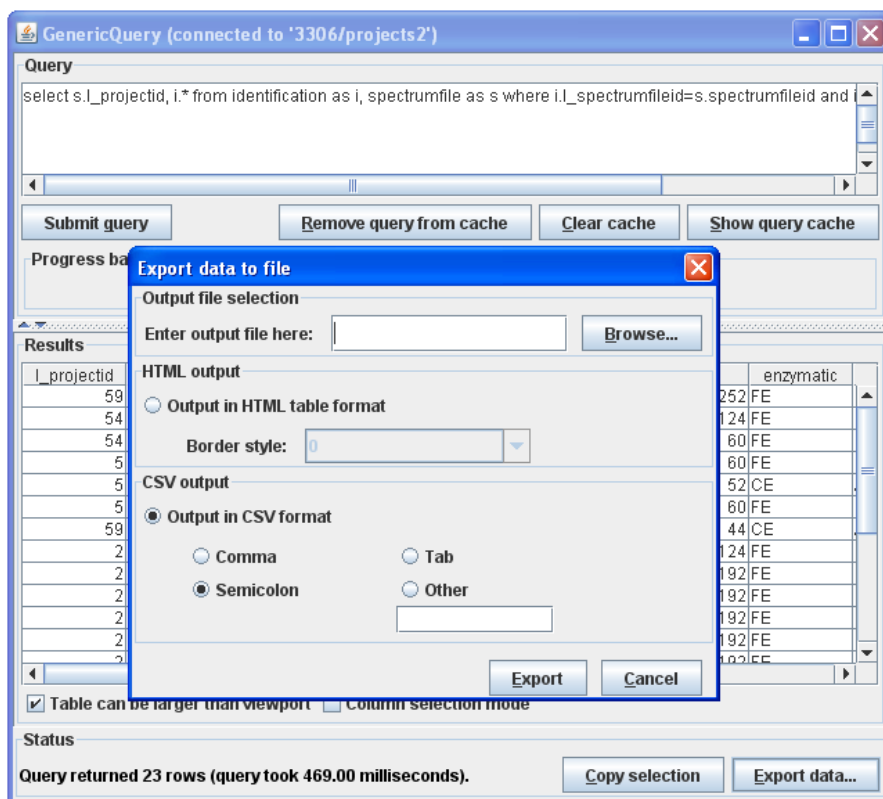
Figure 6.9: The GenericQueryGUI allows access to all tables and stored data by using SQL

Some useful queries for proposal are listed below:

select * from project where username like 'myname%' - List all my projects

select s.l_projectid, i.* from identification as i, spectrumfile as s

where i.l_spectrumfileid=s.spectrumfileid and i.accession like '%myAccession%' - List only projects and identifications where that protein accession has been identified

6.9 Retrieving standardized reports from the database by the ProjectAnalyzer tool

The ProjectAnalyzer tool provides a set of three tools for the data analysis to be performed on a selected project: *Binary file retriever tool*, *DescriptiveNumbersTool* and *Descriptive numbers tool*. Just select the tools from the drop-down menu.
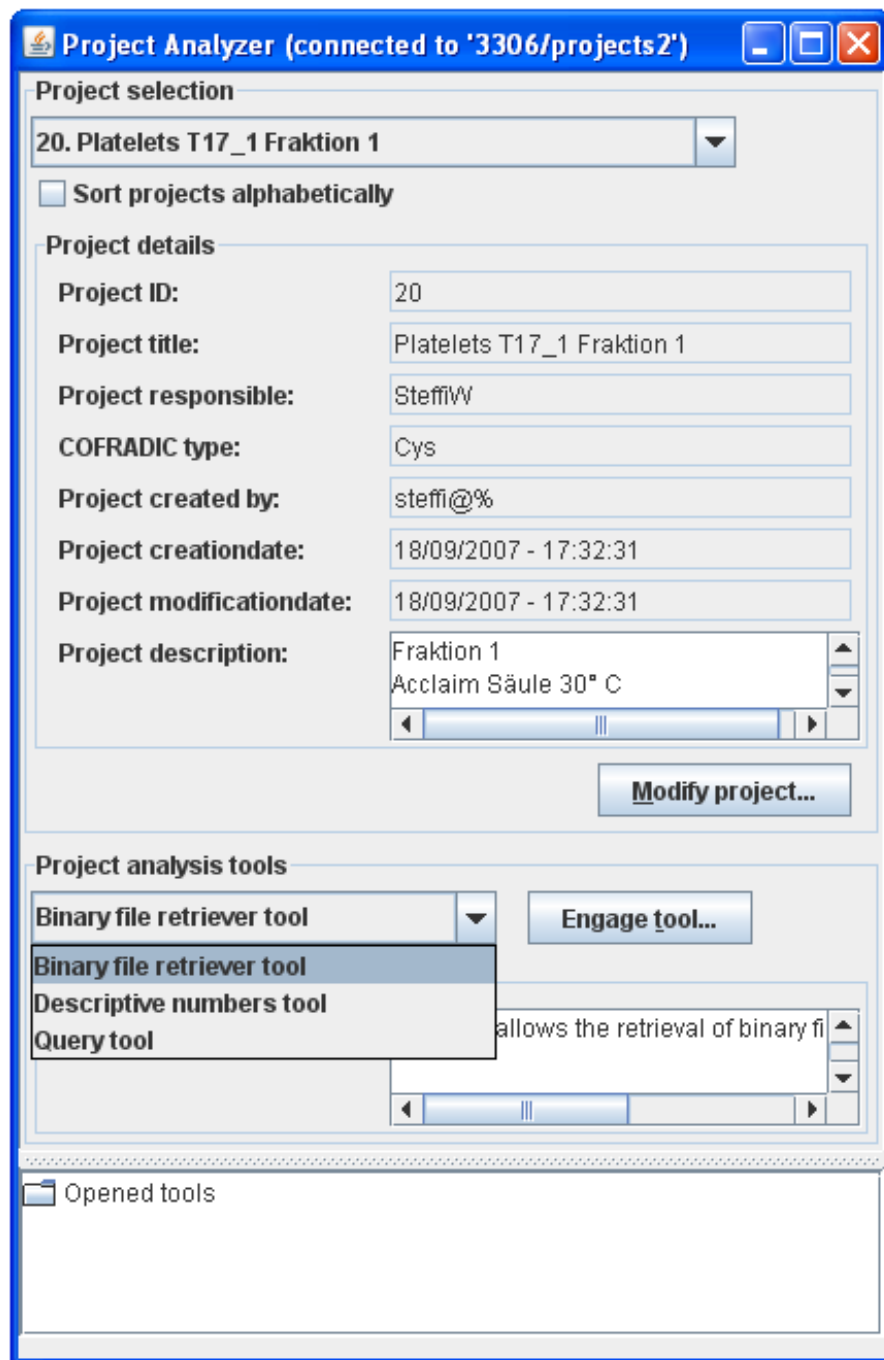
Figure 6.10: Three tools to analyze your projects

6.9.1 Storing and retrieving Binary file(s) into ms-lims

Use this tool to append an informative protocol, image or text-file to a project. Define a descriptor as you like which specifies the binary file e.g. as text document, spreadsheet or picture.
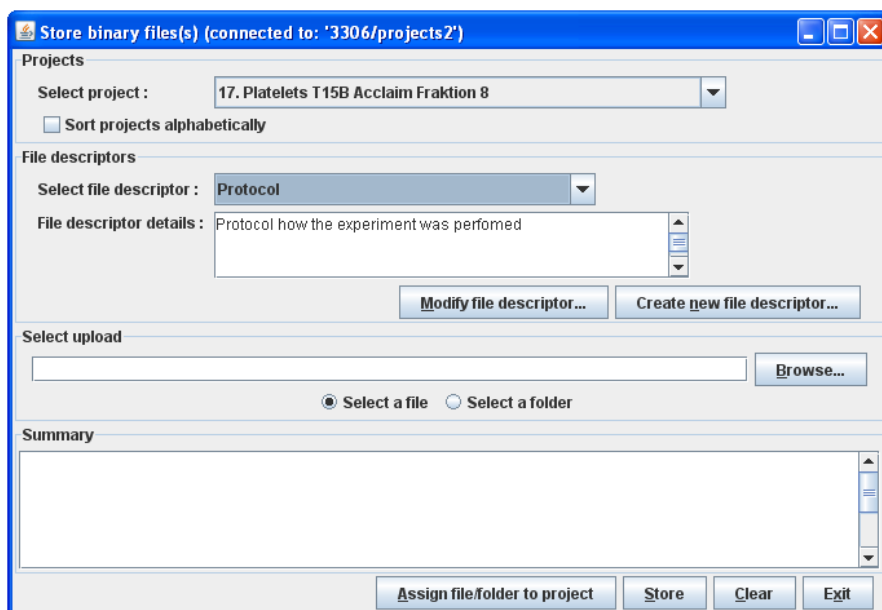
Figure 6.11: Assign a binary file to your project

To locate a binary file stored with a project, run the binary file retriever tool. Select your project in the upper window and then run the tool by hitting *Engage tool*. There'll be pop up a dialog where you can select the binary file and save it to any destination.

6.9.2 Descriptive numbers tool

This tool gives a result overview for COFRADIC experiments. If the experiment COFRADIC type *N-term*, *MetOx* or *Cys* has been selected when creating the project, this tool calculates some informative numbers. Some SQL queries will be performed then and generating the report therefore can take a while. A statusbar will inform you about the progress. The final report can be simply copied and pasted to any document.

6.9.3 Query tool

Here you can run a set of predefined SQL queries against the selected project. This option is very useful for beginners to start with analyzing the stored data. The queries are:
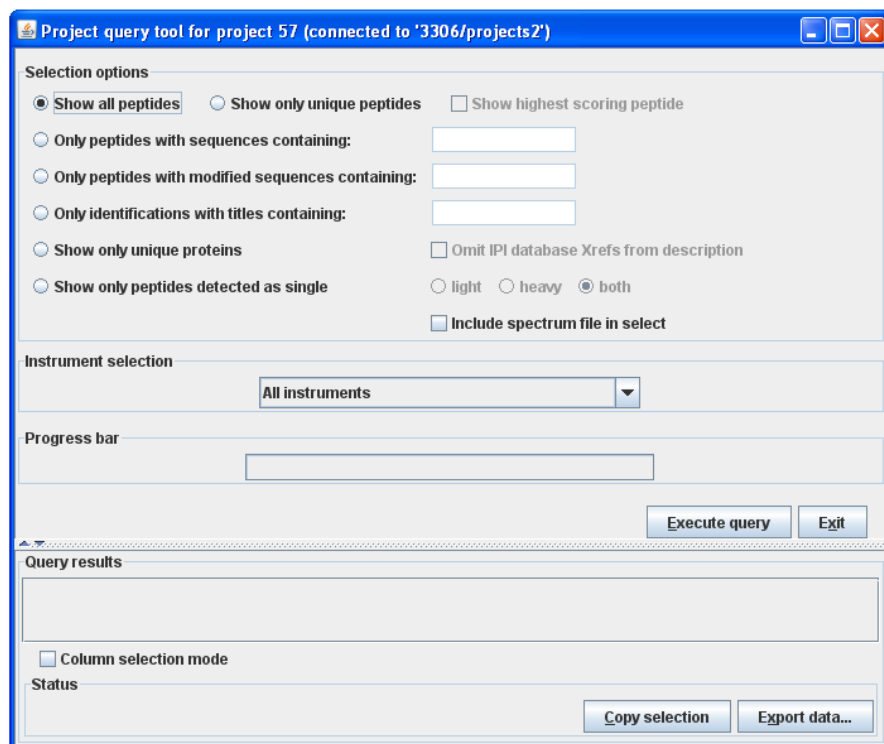
Figure 6.12: Query tool with predefined SQL queries

Show all identified peptides Simply list all identified peptides

Show only unique peptides Check the box if only unique peptides with the maximun score should be displayed

Only peptides with sequences containing Enter a searchstring use '%' as wildcard(s)

Only peptides with modified sequences containing Enter a searchstring; use '%' as wildcard(s)

Only identifications with title containing Enter a searchstring; use '%' as wildcard(s)

Show only unique proteins A list of all unique proteins - you can omit the appearance of cross references from the IPI database

Show only peptides detected as single Choose light, heavy or both kind of peptides

When you have selected a query, choose the instrument if required and press *Execute query*. It's useful to enable the checkbox *Include spectrumfile in select* so a spectrum viewer application opens when you right-click on a spectrumfile in the result table. Again, you can simply copy and past the results to a spreadsheet or export the data as explained in chapter 6.8.

Figure 6.13: The results table allows interactive exploration features

6.10 Validating peptide identifications by Peptizer

For manual validation of peptide identifications, ms-lims joins forces with the Peptizer tool. Peptizer supplies a platform allowing you to easily create an automated expert system to assure the quality of the peptide identifications.

The integration of ms-lims and peptizer is two-sided: Peptizer can inspect peptide identifications by their ms-lims derived identificationID, or by their association to a specific ms-lims project; and, when validation has been performed - the results can be persisted into ms-lims as well.

For more information on Peptizer, please visit the peptizer project site.

6.11 Validating peptide quantitation by Rover

For manual validation of peptide identifications, ms-lims joins forces with the Rover tool. Rover supplies a platform allowing you to easily filter the relevant outliers in your quantitative experiment. Furthermore, Rover provides an of quality related attributes in a protein centric view and thereby enables an optimal judgement on the reliability of the resulting peptide quantitation.

For more information on Rover, please visit the rover project site.