***Malina database user guide***

The following Excel file *MALINA\_MATRICE\_NO\_UNITS.xslx*  is a 363747 by 2391 matrix. All the collected data available on the Malina web database is stored in this matrix. The variables correspond to the matrix’s columns and the observations are stored on each line. Each element inside the matrix is unique and if an observation does not exist for a particular variable, one should read NA. The user will be able to get his data directly from the Excel matrix only if his computer permits it since the file is extremely large. Some computers may allow it but they will run very slowly.

Fortunately, a program written in R exists so that one could get the data he wants faster. This program speaks to an SQL database to retrieve the data one is looking for. SQL is designed to store, fetch, and manipulate data very efficiently therefore being faster than Excel. The software R is the interface used on top of SQL because it is easier to work that way instead of directly into the computer terminal. R has also a lot of interesting features such as its ability to convert the selected data into a simple 2-dimensionial data frame object. Once the user has all the data he wanted into an R object, it is easy to modify this object containing the data into a graphic or to export it as a CSV, Excel or text file. The SQL database is named Takuvik.sqlite and this file can be moved around on your computer or it can be stored on the Internet so that many other scientists could use it. The database is broken down into tables, which corresponds to the files available on the Malina web site.

In order to use the R program please follow the steps:

1. The file Takuvik.sqlite has to be in the folder corresponding to your current directory. To modify your current directory, use the function setwd(). For example,

* setwd("/Users/claude-anne/Desktop/CSV\_full\_test")

Since my Takuvik.sqlite file is in the folder CSV\_full\_test, I need to specify R that I want this folder as current directory.

1. Make sure the RSQLite package is installed on your computer. To install it type

* install.packages(‘’RSQLite’’, dependencies = TRUE)

When the package is installed, you will need to load it before every R session. You need to tell the software that you want to work with this package EVERY TIME you open it by typing :

* library(RSQLite)

1. To send SQL queries, you need to be connected to the database. To do so, you need to type :

* db <- dbConnect(SQLite(), dbname = ‘’Takuvik.sqlite’’)

Before leaving your R session, you need to tell SQL that you want to be disconnected from the database. This step is very important to prevent bugs. You will need to type :

* dbDisconnect(db)

1. Since you probably don’t know every table stored in the database, it would be a good idea to look at what is available to you. The table name will give you a hint on what kind of data is in each table. To do so :

* dbListTables(db)

If you are still not sure the table contains what you are looking for, you can view the variables (or the columns) of the table by doing so :

* res <- dbSendQuery(db, ‘SELECT \* from mcgill’)
* data <- fetch(res, n=-1)
* names(data)

1. In order to retrieve the specific data you want inside tables in the databse, you need to send SQL queries. To achieve this task, you need a SELECT statement. The following example shows you how to retrieve all the data from table *aptot*.

* res <- dbSendQuery(db, ‘’SELECT \* from aptot’’)
* data<- fetch(res, n=-1)
* data
* dbClearResult(res)

the keyword *db is* the R object in which the database connection is stored. You can replace \* with the desired columns if you do not want all the data from the table. For example, if you only want the observations for the variable AP784, you can type :

* res <- dbSendQuery(db, ‘’SELECT AP784 from aptot’’)
* data<- fetch(res, n=-1)
* data
* dbClearResult(res)

The object *data* stores the data you just got from the database in a data frame format that you can manipulate easily. You could only want to keep the observations from the column AP784 that are greater than 0.5 by typing :

* newdata <- data[which(data$AP784 > 0.50)]

If you want to apply multiple criteria to your data set, simply add the new conditions one after the other and separate them with &. For example,

* newdata <- data[which(data$AP784 > 0.05 & data$STATION == 235]

Please note that the word following the dollar sign corresponds to the data set column (or variable).

The steps to select more than one column from a table are the following :

* res <- dbSendQuery(db, statement = paste("SELECT STATION,AP784", "from aptot"))
* data <- fetch(res, n= -1)
* data
* dbClearResult(res)

1. If you have to update a particular table in the database, you need the function *dbWriteTable().* You will need to do so if one of a Malina fle author update his document. This function will rewrite an existing table so it is very important to give the updated table the same name as before. For example, if the author of the file *anap\_database\_malina\_V2.xls* sends another copy with data he has recalulated, you need to import the updated file into R and use the SQL function. Before importing the file into R, you need to convert the Excel file into a CSV file.

* anap <- read.csv(‘anap\_database\_malina\_V2.csv’, sep=’;’, header = TRUE, dec=’,’ , stringsAsFactors=FALSE)
* dbWriteTable(conn = db, name = ‘anap’, value = anap, row.names=FALSE, header = TRUE, overwrite = TRUE)

1. By default, when you are sending SQL queries, R will return the data in an R data frame object. If you do not like R or that you feel more comfortable working with an other format, you can export the R data frame in Excel, text file, csv file and more. Here’s an example to export your data in an Excel format :

\*\*\* Here the data frame is named *data.*

* write.table(data, file = ‘data\_anap.csv’ , sep=’;’, col.names=NA, qmethod = ‘double’)

The csv file *data\_anap.csv* will appear in your current directory folder. The *file* keyword corresponds to the name YOU want your file to be named on your computer. On Mac OSX, R will not let you export your data directly in a .xls format. You need to export it in a csv format and then in Excel you select the import function and then you select semi-column only for the delimiters.

1. To add a new table in the database, you need the same function used to update a table which is *dbWriteTable().* Make sure that you give the new table a name that illustrates a bit what kind of data it contains. Also, the new table needs to be named with a title that has not been previously used in the database. Also, the file has to be imported in R before proceeding. For example, if you import a file into R and name it *eclairement* and you also want the table to be named this way, you will type :

* dbWriteTable(conn = db, name = ‘eclairement’, value = eclairement, row.names = FALSE, header= TRUE)

The keyword *conn =* refers to the database connection name (we used *db* throughout this guide). Then, *name =* is the table name for the database (it needs to be inside quotation marks) and the keyword *value =* is the name of the file you imported into R. It is easier to always import your data in a csv format.

1. REFERENCES

Each and every person who wishes working with the database has the duty to check if the author has written any specifications for his or her data. Each table comes from a certain file available on the malina web database and almost each file has a README associated with its data or an Excel worksheet. It could be units speficitations or comments on the experiment for example. You will find at the end of this document a list of every table in the database with its web link where you can find all the information needed.

Also, it is important to note that many original Excel files were modified to fit a matrix format. In order to be read into a database, a file has to fit in a matrix. That means allo the observation should corresponds to the rows of a matrix and the variables should be the columns of the matrix. Any comments or specifications should be placed into a column called remarks or comments. Nothing can be placed outside of the matrix. One can ONLY use these following symbols : \_ / \.

When one wants to add a new file, it needs to fit into the matrix format otherwise it needs to be modify before importing it into a database table. There are 2 folders available to you with all the files used to populate the database, one folder with Excel formats and the other with csv formats.

It is also important to note that certain observations will not match exactly with the wanted variables. This will likely occur when a recorded observation has been entered differently in one’s file. For example, some bottles were recored at a 2.9m depth electronically but the observation was entered as 3m in the Excel spreadsheet. It could also be explained by the fact that the tables in the databse are not connected with each other. What I mean is that the observations recorded by the scientists do not match exactly with the information electronically recorded. Also, some people decided to name their variables differently but they mean the same thing. For example, some file will have DEPTH as a variable and another file will have WATER\_DEPTH but they mean the same thing. Unfortunately, I was not authorized to change anything except for the format of the different files so it would be a great idea to check for the right spelling of a table column. To do so, please read section 4 of this document.

Good luck!

Claude-Anne Blouin

claudeanne.blouin.1@gmail.com

REFERENCES

|  |  |  |
| --- | --- | --- |
| **TABLE** | **LIEN INTERNET** | **AUTEURS** |
| A | http://www.obs-vlfr.fr/proof/ftpv/malina/basic\_files/ctd/ | basic file |
| anap | http://www.obs-vlfr.fr/proof/ftpv/malina/data/p2557-2560/Absorption\_Particulaire/Absorption\_Particulaire/ | Belanger S. |
| ap | http://www.obs-vlfr.fr/proof/ftpv/malina/data/p2557-2560/Absorption\_Particulaire/Absorption\_Particulaire/ | Belanger S. |
| aphy | http://www.obs-vlfr.fr/proof/ftpv/malina/data/p2557-2560/Absorption\_Particulaire/Absorption\_Particulaire/ | Belanger S. |
| aptot | http://www.obs-vlfr.fr/proof/ftpv/malina/data/p2557-2560/Absorption\_Particulaire/Absorption\_Particulaire/ | Belanger S. |
| bacterial | http://www.obs-vlfr.fr/proof/ftpv/malina/data/p30151\_2479\_Joux\_bact\_prod/ | Joux F/Jeffrey W. |
| bbsiop | http://www.obs-vlfr.fr/proof/ftpv/malina/data/p2450-2449/bb\_siop\_Malina/bb\_siop\_Malina/ | Reynolds R. |
| chla | http://www.obs-vlfr.fr/proof/ftpv/malina/data/p2355/ | Gosselin M./Bélanger S. |
| mcgill | http://www.obs-vlfr.fr/proof/ftpv/malina/data/lansard/ | Mucci A./Lansard B. |
| novembre | http://www.obs-vlfr.fr/proof/ftpv/malina/data/p\_raimbault/ | Tremblay J.E/Raimbault P. |
| ce600 | http://www.obs-vlfr.fr/proof/ftpv/malina/data/p2536-2545\_leymarie/ | Antoine D./Leymarie E. |
| ctd | http://www.obs-vlfr.fr/proof/ftpv/malina/data/p2360\_2363\_30259\_ron\_benner/ | Benner R. |
| zodiac | http://www.obs-vlfr.fr/proof/ftpv/malina/data/p2359\_2361\_2364\_ron\_benner/ | Benner R. |
| stnlipids | http://www.obs-vlfr.fr/proof/ftpv/malina/data/p2408\_2426/ | Bouloubassi I. |
| stncasq | http://www.obs-vlfr.fr/proof/ftpv/malina/data/p2409\_2427/ | Bouloubassi I. |
| ultrapath | http://www.obs-vlfr.fr/proof/ftpv/malina/data/p2136-2137-2138/ | Bricaud A. |
| res3 | http://www.obs-vlfr.fr/proof/ftpv/malina/data/p\_IOP\_doxaran/ | Doxaran D. |
| spmpoc | http://www.obs-vlfr.fr/proof/ftpv/malina/data/p\_spm\_poc\_doxaran/ | Doxaran D./Ehn J./Babin M. |
| downcast | http://www.obs-vlfr.fr/proof/ftpv/malina/data/p2338\_2341\_2354\_ac9\_bb9\_ehn/ | Ehn J. |
| stnxrf | http://www.obs-vlfr.fr/proof/ftpv/malina/data/p2437/ | Giraudeaux |
| phytosylvie | http://www.obs-vlfr.fr/proof/ftpv/malina/data/p2376\_2377/ | Gosselin M./Bélanger S. |
| scamp | http://www.obs-vlfr.fr/proof/ftpv/malina/data/p2107\_scamp/ | Gratton Y. |
| salinitylog | http://www.obs-vlfr.fr/proof/ftpv/malina/data/p2529-2530\_salinity\_/ | Gratton Y./Prieur L. |
| hookerbarge | http://www.obs-vlfr.fr/proof/ftpv/malina/data/p2110\_c-ops-barge/ | Hooker |
| hookermalina | http://www.obs-vlfr.fr/proof/ftpv/malina/data/p2123\_hooker/ | Hooker |
| photoparam | http://www.obs-vlfr.fr/proof/ftpv/malina/data/p30091/ | Huot Y. |
| abundance | http://www.obs-vlfr.fr/proof/ftpv/malina/data/bacteria\_AAP\_jeanthon/ | Jeantohn C./Bœuf D. |
| ectoenzyme | http://www.obs-vlfr.fr/proof/ftpv/malina/data/p30092/ | Joux F./Ortega E. |
| respiration | http://www.obs-vlfr.fr/proof/ftpv/malina/data/p2390/ | Joux F./Ortega E. |
| prad | http://www.obs-vlfr.fr/proof/ftpv/malina/data/RAD\_papakyriakou/ | Papakyriakou |
| uvp5odv | http://www.obs-vlfr.fr/proof/ftpv/malina/data/p2351\_30016\_uvp5\_/ | Picheral M. |
| uvp5zoo | http://www.obs-vlfr.fr/proof/ftpv/malina/data/p2351\_30016\_uvp5\_/ | Picheral M./Marec C. |
| psdcoulter | http://www.obs-vlfr.fr/proof/ftpv/malina/data/p2371\_30021\_rick/coulter\_malina\_reynolds/ | Reynolds R. |
| eclairement | http://www.obs-vlfr.fr/proof/ftpv/malina/data/p\_eclairementsurf\_LMGEM/ | Sempere R. |
| doc | http://www.obs-vlfr.fr/proof/ftpv/malina/data/p\_30259\_2359\_DOC\_LMGEM/ | Sempere R. |
| raimbault | http://www.obs-vlfr.fr/proof/ftpv/malina/data/p\_raimbault\_incub/ | Tremblay J.E/Raimbault P. |
| cytometry | http://www.obs-vlfr.fr/proof/ftpv/malina/data/cyto\_FCM\_vaulot/ | Vaulot D. |
| trflp | http://www.obs-vlfr.fr/proof/ftpv/malina/data/p2322\_2407\_vaulot\_/ | Vaulot D. |
| hookerpigment | http://www.obs-vlfr.fr/proof/ftpv/malina/data/p2375/ | Wright V./ Hooker S. |
| xie | http://www.obs-vlfr.fr/proof/ftpv/malina/data/p2477\_2478-xie\_huixiang/ | Xie H. |
| tameamund | http://www.obs-vlfr.fr/proof/ftpv/malina/basic\_files/logbook/ctd-rosette-btl/ | basic file |
| station | Malina\_stations.xls (le lien ne fonctionne pas) | basic file |
| shipoperation | ship\_oepration\_log\_\_2010\_02\_24 (le lien ne fonctionne pas) | basic file |
| shiptrack | shiptrack\_time\_lat\_long.dat (le lien ne fonctionne pas) | basic file |
| meteo | http://www.obs-vlfr.fr/proof/ftpv/malina/basic\_files/logbook/meteo\_logbook/ | basic file |
| instruments | http://www.obs-vlfr.fr/proof/ftpv/malina/basic\_files/logbook/Rosette/ | basic file |
| niskin | http://www.obs-vlfr.fr/proof/ftpv/malina/basic\_files/logbook/Rosette/ | basic file |
| pufM | http://www.obs-vlfr.fr/proof/ftpv/malina/data/p2402\_2403\_2404/ | Jeanthon C./ Bœuf D. |
| benthic\_sediment | http://www.obs-vlfr.fr/proof/ftpv/malina/data/p2413\_2420\_link/ | Link H./ Archambault P./ Chaillou G. |
| benthic\_flux | http://www.obs-vlfr.fr/proof/ftpv/malina/data/p2413\_2420\_link/ | Link H./ Archambault P./ Chaillou G. |
| pco2 | http://www.obs-vlfr.fr/proof/ftpv/malina/data/pCO2\_papakyriakou/ | Papakyriakou |
| procmet | http://www.obs-vlfr.fr/proof/ftpv/malina/data/METEO\_papakyriakou/ | Papakyriakou |
| lipidbiomark\_boxcore | http://www.obs-vlfr.fr/proof/ftpv/malina/data/p2428-2430/ | Tolosa I. |
| malina\_CE600\_release2\_output | on USB key | Leymarie E. |
| PR\_abundance\_qPCR | on USB key | Jeanthon C. |
| PR\_molecular\_diversity | on USB key | Jeanthon C. |
| blast\_unique\_PR\_sequences | on USB key | Jeanthon C. |
| PRB\_cultures | on USB key | Jeanthon C. |
| Lisst | on USB key | Reynolds R. |
| lipidbiomark\_waters | on USB key | Tolosa I. |