***Guide utilisation de la base de données Malina***

Le fichier Excel *MALINA\_MATRICE\_NO\_UNITS.xlsx*  est une matrice 363747 lignes X 2391 colonnes. Elle contient toutes les données recueillies et mises en ligne durant la mission Malina. Les colonnes de cette matrice correspondent aux différentes variables et les lignes correspondent aux observations. Chaque élément à l’intérieur de la matrice est unique et lorsqu’une observation n’existe pas pour une variable, on note NA. Un individu peut rechercher des données à même le fichier Excel s’il le désire mais vu la taille du fichier, certains ordinateurs s’exécuteront très lentement.

Heureusement, un programme R a été conçu pour aller chercher les données demandées à l’intérieur de la base de donnée. Ce programme s’exécute très rapidement et on obtient un objet de type *data.frame* qui devient utilisable dans R. Cet objet peut être modifié, utilisé pour programmer un graphique ou peut être exporter en un fichier quelconque (CSV, Excel, Text, etc.). Les données contenues dans le fichier Excel ont été transférées dans une base de donnée de type SQL. Cette base de données nommée Takuvik.sqlite a été réalisé grâce au logiciel SQLite. Elle est divisée en tables correspondant aux différents fichiers rendus disponible sur le site internet de Malina.

Pour utiliser le programme R, voici la procédure :

1. S’assurer que le fichier Takuvik.sqlite se retrouve dans votre répertoire courant. Pour modifier ce répertoire courant, utiliser la fonction setwd(). Par exemple,

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

Puisque mon fichier Takuvik.sqlite se retrouve dans le dossier CSV\_full\_test, je dois spécifier à R que je veux ce dossier comme répertoire courant.

1. S’assurer que le package RSQLite est installé sur votre ordinateur. Pour l’installer, tapez :

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

Lorsque le « package » est installé sur votre ordinateur, vous devez À CHAQUE DÉBUT DE SESSION DANS R lui dire que vous voulez travailler avec ce package.

> library(RSQLite)

1. Pour envoyer des requêtes SQL, vous devez vous connecter à la base de donnée. Pour y arriver,

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

À la fin de chaque session du programme R, veuillez vous déconnecter de la base de données pour éviter des erreurs futures.

> dbDisconnect(db)

1. Si vous désirez voir les tables présentes dans la base de données (pour avoir un indice de la table que vous aurez besoin pour aller chercher vos données par exemple), procédez comme suit :

> dbListTables(db)

Si vous désirez voir les noms des variables présentes dans une table, faites ceci :

> res <- dbSendQuery(db, ‘SELECT \* from mcgill’)

> data <- fetch(res, n=-1)

> names(data)

1. Pour aller chercher des données à l’intérieur des tables de la base de données, on doit envoyer des requêtes SQL. Pour ce faire, on procède avec un énoncé SELECT. Voici un exemple où l’on veut avoir toutes les variables contenues dans la table *aptot.*

> res <- dbSendQuery(db, ‘’SELECT \* from aptot’’)

> data<- fetch(res, n=-1)

> data

> dbClearResult(res)

L’énoncé db correspond à l’objet créer plus tôt qui établi une connexion avec la base

de données. On peut remplacer \* par les colonnes que l’on désire. Par exemple, si l’on veut seulement les données correspondant à AP784, on procèdera comme suit :

> res <- dbSendQuery(db, ‘’SELECT AP784 from aptot’’)

> data<- fetch(res, n=-1)

> data

> dbClearResult(res)

Si l’on veut les données AP784 mais seulement celles qui sont plus grande que 0.5, on procède comme suit en manipulant le *dataframe* créé :

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

Si on veut appliquer plus d’un critère, on ajoute chaque nouvelle condition une à la suite de l’autre séparé par « & ». Par exemple,

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

On note que le mot qui suit le signe de dollar correspond à la colonne (donc à la variable) dans le jeu de données.

La procédure pour sélectionner plus d’une colonne dans une table est la suivante :

> res <- dbSendQuery(db, statement = paste("SELECT STATION,AP784", "from aptot"))

> data <- fetch(res, n= -1)

> data

> dbClearResult(res)

1. Si l’on veut mettre à jour une table dans la base de données, on utilise la fonction dbWriteTable(). En effet, si l’auteur d’un des fichiers de données Malina apporte des changements quelconque à son fichier, il est facile de le mettre à jour dans la base de données. Cette fonction va réécrire la table donc il est important de lui donner le même nom. Par exemple, si l’auteur du fichier *anap\_database\_malina\_V2.xls* renvoie une copie de son fichier avec des données qu’il a recalculées, on a qu’à réimporter ce fichier dans R et utiliser la fonction SQL. On procède ainsi :

-D’abord on converti le fichier Excel en fichier csv

> 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. Par défaut, lorsqu’on envoie des requêtes SQLite dans R, on obtient un *data frame* que l’on peut modifier dans R. Si l’on veut exporter ces données dans un format qui sera lu avec Excel, on procède ainsi :

\*\*\* Notre data frame se nomme data dans cet exemple.

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

Le fichier csv data\_anap.csv se retrouvera dans le dossier qui correspond à votre répertoire courant dans R. Pour importer un fichier csv en Excel, sélectionnez la fonction importer et suivre les étapes. Vous devrez chochez le délimiteur point-virgule seulement.

1. Pour ajouter un nouveau jeu de données, on utilise la même fonction que pour mettre à jour une table, i.e. dbWriteTable(). On s’assure que le nom que l’on donne à notre nouvelle table n’est pas déjà utilisé dans la base de données et que le nom de la table représente assez bien le jeu de données.

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

l’argument conn réfère au nom de la connection de la base de données SQL. J’ai décidé de l’appeler db. L’argument name est le nom de la table dans la base de données qui doit être entre guillemètes et l’argument value est le nom de l’objet R qui contient les données. Ceci veut dire qu’on doit avoir importé dans R le fichier csv voulu et on lui assigne un certain nom.

1. RÉFÉRENCES

Chaque personne qui désire utiliser les données d’une certaines table se doit d’aller vérifier les spécifications de l’auteur. En effet, l’auteur d’un jeu de données a presque toujours un fichier README d’associé à ses données ou bien les informations se retrouvent dans une feuille Excel. Que ce soit pour connaitre les unités des variables ou toute autre information, il est très important d’aller vérifier s’il y a certaines informations qui n’ont pas pu être intégré à la base de données. Vous trouverez une liste des tables de la base de données et les liens correspondent en annexe.

Il est important de noter que certains fichiers Excel ont été modifiés au niveau de leur format puisqu’il n’était pas conforme à un format « base de données ». En effet, pour qu’un fichier puisse être compatible avec le format d’une base de données, les observations doivent se retrouver sur des lignes et les noms de variables dans des colonnes. Toute information qui ne peut s’ajuster à un format « tableau » ne peut être ajoutée. De plus, on ne peut pas retrouver des symboles autres que : \_ / \. Lorsqu’on veut ajouter un nouveau fichier, on doit toujours s’assurer que le fichier soit conforme sinon on doit le modifier. Vous recevrez 2 dossiers contenant tous les fichiers qui ont été utilisés et modifiés pour la base de données, un avec les fichiers en format Excel et l’autre les fichiers en format CSV.

Il est aussi important de noter qu’il se peut que certaines données ne correspondent pas toujours avec les variables voulues. Ce phénomène est causé par les différences dans l’enregistrement des données. Par exemple, certaines bouteilles ont été enregistrées à différentes profondeurs et on marque 2.9m au lieu de 3m. De plus, il se peut que vous recherchiez des données de profondeur dans la matrice mais que la variable est été enregistrée sous WATER\_DEPTH au lieu de DEPTH. Ceci est dû au fait que je ne me suis pas permis de changer les noms des variables des auteurs. Il est donc parfois nécessaire d’aller voir le nom des variables contenues dans les tables. Pour y arriver, voir la section 4.

Bonne chance!

Claude-Anne Blouin

claudeanne.blouin.1@gmail.com

**ANNEXE I**

RÉFÉRENCES

\*\* Si vous ne connaissez pas le mot de passe pour vous authentifier, essayer de copier-coller le lien directement dans votre navigateur.

|  |  |  |
| --- | --- | --- |
| **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. |