Infrastructure and Web Applications for Application Programming Interface (API) of Bio-database, Ocean Data Bank (ODB) 海洋資料庫生物資料庫應用程式介面之基礎建構與網際網路應用

Chiyu Weng 翁其羽

2022-03-04

Table of Contents

# 1 Introduction

## 1.1 An infrastructure for information services of ODB’s bio-database

How can we design an infrastructure that help people easier, more safer, more convenient for adding ideas to use the bio-database of Ocean Data Bank (ODB)? ODB is dedicated to help academia to curate and use the databases whose data is mainly collected through marine research vessels which are supported by Ministry of Science and Technology (MOST), Taiwan. The data of ODB is in privacy under the restriction of the data release policies of MOST, and ODB compiles raw data after reviewing someone’s application. This restriction limits open usage of the information services we deliver. To bridge ODB’s information services between open usage and databases is a prerequisite, whereas a well-defined application programming interface (API) is often the answer ([Box 1.1](#box1-1)). In addition, ODB’s information services are usually not for implementing new theory or algorithm on scientific researches, but for helping academia to use data, check data patterns, and create data modeling in their researches. These considerations form the basis for how to construct the infrastructure of information services to use the bio-database of ODB:

Box1.1

✓ API is a set of interfacing specifications for machine-to-machine communications. Here we focus on web API, i.e., data transport upon HTTP request/response structure. Most popular web API protocol is REST (Representational State Transfer), and increasingly used, GraphQL API. Commonly used data formats are JSON and XML.

✓ API can provide a secure way to access the bio-database (internal APIs, Fig. ) with authority management. On the other hand, API can also provide open-access methods for public information compiled from the bio-database (public APIs).

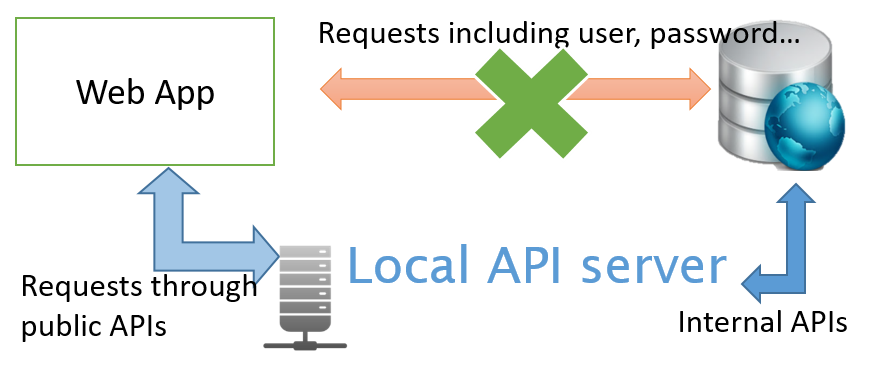


Fig. . Simplified schematic for web API of the bio-database

1. Don’t reinvent the wheel: ODB’s information services should be focused on problem solving. “Don’t reinvent the wheel” here means use existing, proven, and open source software (OSS) to get solutions. There are many OSS packages in data manipulation, data visualization, database-engine wrapper, Geographic Information System (GIS) processing, and statistics which are the major fields of data science that used for ecological applications. To be able to add these OSS packages upon the infrastructure in a more flexible way facilitate us to bring ideas into practice.
2. Faster development and deployment: It’s practical to design the infrastructure of ODB’s information services under the premise that most of the researchers are not programmers and no need to have background for database when using the data. But the researchers usually have to demonstrate their ideas, to check data patterns, and to modeling their data more efficiently. That’s why we need to have an infrastructure that is suitable for faster development and deployment.
3. Scalability, reproducibility, and consistency: To be easier to combine user data with ODB’s data so that the researchers can check data patterns in different scales. It’s also necessary to make the results be reproducible and consistent.

We present an infrastructure which may fulfill these requirements for constructing ODB’s information services upon (Fig. 1.2). We use R ()

# 2 **Re-structuring bio-database**

## 2.1 *Mutating a new bio-database on PostgreSQL*

The bio-database of ODB had been constructed since 2009 and completely set-up on Microsoft SQL Server since 2015. The data curation is basically done by a Microsoft C# program. These protocols of data curation was documented in 圖輯. For developing an open cross-platform framework for web APIs of bio-database, first, we mutated a bio-database on PostgreSQL, i.e., exported from SQL Server, and then re-imported to a new database by using PostgreSQL.

Box2.1

Why not just use original Microsoft SQL Server as backend database for developing APIs of bio-database?

✓ SQL Server is a commercial software with annual license fees.

✓ SQL Server is suited for the Microsoft Server based framework, but not for an open cross-platform framework.

Why choose PostgreSQL?

✓ PostgreSQL is a powerful open-source object-relational database systems.

✓ PostgreSQL can be well integrated in Geographic Information System (GIS) applications by its PostGIS extension. - It’s an important reason why we choose PostgreSQL. Most of APIs in ODB are used to develop GIS related applications.

As a backend database for API in production, we do not need some columns of variables originally in SQL Server (Fig. 2a) that used to remark some quality control (QC) issues. So when mutating the bio-database on PostgreSQL, we also did some simplifications for database schema (Fig. 2b). Roughly speaking, we left the “remark\_xxx” or “flag\_xxx” columns in SQL Server only, and in PostgreSQL, re-arranged the tables to make frequently-used variables more easily be accessed in table “cast\_site” (site information) and “taxa\_data” (abundance data of taxonomic groups). Another important modification is to construct classification of taxonomic groups in table ’taxon\_group” that can help us in analytical statistics among taxonomic groups when using bio-database data. The major parts of schema for bio-database on PostgreSQL include: taxon\_group:

1. cast\_site:
2. c00\_cast\_tbl:
3. t00\_taxa\_tbl:
4. taxa\_data:

# NOW on Ubuntu, this driver not support old SQL Server 2008-R2. It only works on Windows  
# On Ubuntu, the driver setting is in /etc/odbcinst.ini, e.g. driver={ODBC Driver 18 for SQL Server}  
library(RODBC) #R interface of ODBC driver to connect SQL Server  
library(sqldf) #runing SQL statements on R data frames  
library(data.table) #data.table is superb fast in R data manipulation  
library(magrittr) #pipe function  
ms\_conn <-   
 paste0('driver={SQL Server};server=', sqlServerHost,  
 ';database=', sqlServerDB,  
 ';uid=', sqlServerUser,  
 ';pwd=', sqlServerPass) %>%  
 odbcDriverConnect()  
  
# Query all taxonomic abundance in SQL server  
taxa\_data <- sqlQuery(ms\_conn, 'select \* from dbo.Taxa\_record') %>% setDT()  
  
# Query site  
cast\_site <- sqlQuery(ms\_conn, 'select \* from dbo.Cast') %>% setDT()  
  
close(ms\_conn) # Close database connection

Table  Taxaonomic abundance in bio-database

| taxarec\_id | cast\_id | taxonomic\_name | taxon\_count | original\_unit |
| --- | --- | --- | --- | --- |
| T0000000001 | C00000013 | Acartia bifilosa | 18.9960 | per m3 |
| T0000000002 | C00000014 | Acartia negligens | 31.9352 | per m3 |
| T0000000003 | C00000015 | Acartia negligens | 14.4690 | per m3 |
| T0000000004 | C00000001 | Acartia negligens | 48.5928 | per m3 |
| T0000000005 | C00000002 | Acartia negligens | 39.0825 | per m3 |
| T0000000006 | C00000003 | Acartia negligens | 47.9880 | per m3 |

Table  Casting sites in bio-database

| cast\_id | station\_id | date | depth\_lower\_bound | mesh\_size | gear\_type |
| --- | --- | --- | --- | --- | --- |
| C00000001 | S000001 | 1999-08-16 | 163 | 150 | NORPAC Net |
| C00000002 | S000002 | 1999-08-16 | 150 | 150 | NORPAC Net |
| C00000003 | S000003 | 1999-08-16 | 124 | 150 | NORPAC Net |
| C00000004 | S000004 | 1999-08-16 | 132 | 150 | NORPAC Net |
| C00000005 | S000005 | 1999-08-18 | 75 | 150 | NORPAC Net |
| C00000006 | S000006 | 1999-08-18 | 68 | 150 | NORPAC Net |

library(RPostgres) #R (using Rcpp) Interface to PostgreSQL  
library(odbapi) #Internal API package, which is explained in   
#e.g., A dataset with species Creseis conica, Abudefduf saxatilis, Labidocera gallensis, Metridia macrura, and M. asymmetrica is ready to be imported.  
test\_sp <- c("Creseis conica", "Abudefduf saxatilis",   
 "Labidocera gallensis", "Metridia macrura", "Metridia asymmetrica")  
  
pconn <- dbConnect(drv = RPostgres::Postgres(),  
 host = odbHost, port = odbPort, dbname = odbBioDB,  
 user = odbUser, password = odbPass)  
#Got taxonomy of bio-database on PostgreSQL  
taxonomy <- dbReadTable(pconn, name="taxon\_group") %>% setDT()  
#getSciName first retrieve query from existed taxonomy in "taxon\_group" table.  
#For those scientific names not in bio-database, retrieve them by web API of WORMS, GBIF, etc. Internally its magic is done through R package "taxize".  
new\_taxa <- odbapi::getSciName(dbuser = odbUser, dbhost = odbHost, dbname = odbBioDB,  
 taxon = test\_sp[!test\_sp %chin% taxonomy$show\_name],  
#Only the latter three species are new to “taxon\_group” table. This check can be ignored.  
 source = c("worms", "gbif")) #see Table 2.3

## [1] "connection to odbio database stage.."  
## [1] "fetch odbio taxon\_group.."  
## [1] "3 taxonomic names need to be queryed in taxize db. Please wait and check the interactive prompt..."  
## == 3 queries ===============  
## Error : (400) Bad Request - Labidocera gallensis  
## [1] "Warning: source not connect: worms"  
## == 3 queries ===============  
## v Found: Labidocera gallensis  
## v Found: Metridia macrura  
## v Found: Metridia asymmetrica  
## == Results =================  
##   
## \* Total: 3   
## \* Found: 3   
## \* Not Found: 0

Table  Update table of taxonomy (parts) in bio-database

| query | rank | taxon | class | order | family | genus | species |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Labidocera gallensis | species | Labidocera gallensis | Hexanauplia | Calanoida | Pontellidae | Labidocera | Labidocera gallensis |
| Metridia macrura | species | Metridia macrura | Hexanauplia | Calanoida | Metridinidae | Metridia | Metridia macrura |
| Metridia asymmetrica | species | Metridia asymmetrica | Hexanauplia | Calanoida | Metridinidae | Metridia | Metridia asymmetrica |

## 2.2 *Collaboration with larval fish database*

fconn <- dbConnect(drv = RPostgres::Postgres(),  
 host = larvaHost, port = odbPort, dbname = larvaDB,  
 user = odbUser, password = odbPass)  
# Similar schema to get the same tables from two different data source on PostgreSQL  
# Advantage: to design an integrated API to access them is more easier  
bio\_site <- dbReadTable(pconn, name="cast\_site") %>% setDT()  
larva\_site <- dbReadTable(fconn, name="cast\_site") %>% setDT()  
  
library(sf) # R package that handles geometries in GIS applications   
sites <- list(bio\_site[, source:="ODB"], larva\_site[, source:="CHIU"]) %>%   
 rbindlist(use.names = TRUE, fill = TRUE) %>%   
 .[, season := odbapi::datex\_season(date)] %>% # Internal API: convert date to season  
 .[, season := factor(season, levels = c(0,1,2,3),  
 labels=c("Spring","Summer","Autumn","Winter"))] %>%  
 st\_as\_sf(coords = c("longitude", "latitude"), crs = 4326)  
bbox <- st\_bbox(sites) # Bounding box, used in plotting Fig. 2.4  
dbDisconnect(pconn) # Close database connection  
dbDisconnect(fconn)

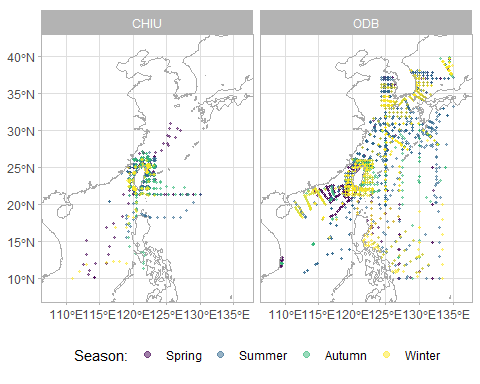


Fig. . Distribution of sampling sites in ODB bio-database

## 2.3 *Taxonomic composition of bio-database*

ODB bio-database collected a total of 137,819 records of abundance data, mostly zooplankton and larval fish, at 6700 sampling sites in 475 cruises. Fig. 2.5 shows the taxonomic composition in bio-database with the number of records and proportions. The largest proportion larval fish data, mainly comes from Dr. Chiu’s research, sums up to a 61,165 records with 673 species, 500 genera, and 190 families (Table 2.4).

## 2.4 *Eco-environment database in ODB*

This is a reference to table .

Table  Dataset demo

| name | data source | description | URL | header | x resolution | y resolution | records | longitude | latitude | years |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| nasa\_nobm\_v2017 | NASA/Goddard Space Flight Center | NASA Ocean Biogeochemical Model (NOBM) | https://gmao.gsfc.nasa.gov/reanalysis/MERRA-NOBM | longitude, latitude, date, season, mixlayer, seaice, chl, nitrate, iron, chlorophyte, diatom, coccolith, cyanobacteria | 1.25 | 0.667 | 658800 | 88.75 - 150.00 | 0.00 - 40.00 | 1998 - 2015 |
| nasa\_neo | NASA/Goddard Space Flight Center | NASA Earth Observations (NEO) | https://neo.gsfc.nasa.gov | longitude, latitude, date, season, sst, chl | 0.10 | 0.100 | 30927976 | 90.10 - 150.00 | 0.00 - 40.00 | 2002 - 2018 |
| ncep\_godas | NOAA/National Weather Service | NCEP Global Ocean Data Assimilation System (GODAS) | https://www.cpc.ncep.noaa.gov/products/GODAS | longitude, latitude, date, season, depth, salinity | 1.00 | 0.333 | 99118656 | 89.50 - 150.50 | -0.17 - 40.17 | 1980 - 2016 |
| noaa\_woa\_v2013 | NOAA/National Centers for Environmental Information | WORLD OCEAN ATLAS 2013 version 2 (WOA13) | https://www.nodc.noaa.gov/OC5/woa13 | longitude, latitude, month, season, depth, nitrate, phosphate, silicate | 1.00 | 1.000 | 116018 | 89.00 - 140.00 | 9.00 - 40.00 | - 2012 |
| svp\_pacific\_v2017\_025d | NOAA/Atlantic Oceanographic and Meteorological Laboratory (AOML) | Global Lagrangian Drifter Data | https://www.aoml.noaa.gov/phod/gdp | longitude, latitude, season, u, v | 0.25 | 0.250 | 58564 | 107.00 - 137.00 | 2.00 - 32.00 | - 2012 |
| ctd\_odb | ODB, IONTU | Marine hydrographic data from Conductivity-Temperature-Depth Profiler (CTD) | https://www.odb.ntu.edu.tw | longitude, latitude, date, season, cruise, depth, temperature, salinity, fluorescence, transmission, oxygen | NA | NA | 15076533 | 105.10 - 156.17 | 1.98 - 32.75 | 1985 - 2017 |
| depth\_odb\_v2011\_025d | ODB, IONTU | Bathymetric data in 0.5-degree compiled by ODB | https://www.odb.ntu.edu.tw | longitude, latitude, sea\_depth | 0.25 | 0.250 | 16093 | 105.00 - 135.00 | 2.00 - 35.00 | - 2011 |
| depth\_odb\_v2011\_ea1000m | ODB, IONTU | Bathymetric data about 1000m-gridded compiled by ODB | https://www.odb.ntu.edu.tw | longitude, latitude, sea\_depth | 0.01 | 0.010 | 9906301 | 105.00 - 135.00 | 2.00 - 35.00 | - 2011 |

# 3 OpenCPU framework for internal web APIs of bio-database

## 3.1 *OpenCPU server configuration*

## 3.2 *Internal API package: odbapi*