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

2022-02-28

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).

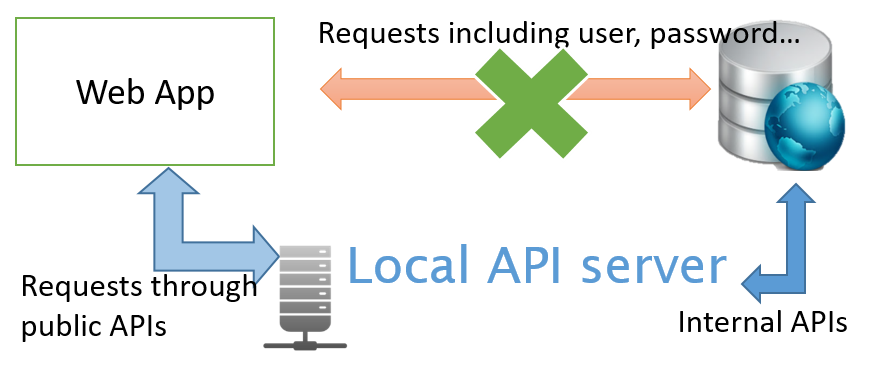


Fig. . Simplified schematic for web API of the bio-database

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)

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 |

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

## 3.1 OpenCPU server configuration

This is a reference to table .

Table  Dataset demo

| mpg | cyl | disp | hp | drat |
| --- | --- | --- | --- | --- |
| 21.0 | 6 | 160 | 110 | 3.90 |
| 21.0 | 6 | 160 | 110 | 3.90 |
| 22.8 | 4 | 108 | 93 | 3.85 |
| 21.4 | 6 | 258 | 110 | 3.08 |
| 18.7 | 8 | 360 | 175 | 3.15 |

## 3.2 Internal API package: odbapi