GOTIT 1.1 Database

documentation for users and developers

19/07/2019 - v 1.1

Florian Malard[1] | Philippe Grison[2] | Louis Duchemin[1] | Lara Konecny-Dupré[1] | Tristan Lefébure[1] | Nathanaëlle Saclier[1] | David Eme[3] | Chloé Martin[2] | Cécile Callou[2] | Christophe J. Douady[1]

(1) LEHNA : UMR CNRS 5023 Ecologie des Hydrosystèmes Naturels et Anthropisés, Université Lyon 1 , ENTPE, CNRS, Université Lyon

(2) BBEES - Unité Bases de données sur la Biodiversité, Écologie, Environnement et Sociétés (BBEES), Muséum national d'Histoire naturelle, CNRS

(3) New Zealand Inst. for Advanced Studies, School of Natural and Computational Sciences, Massey Univ., Auckland, New Zealand

Table of Content

I – Goal 2
II – Structure of the relational database 3
III – Requirements and technologies 6
IV – References 7
V – Tables and figures 8
VI – Supplementary materials 11

I – Goal

The database was developed to optimize the management and storage of species occurrence data that are produced within a laboratory either from morphological or molecular identification / delimitation. It is thus particularly suitable for designing and monitoring biodiversity projects that employ an integrative taxonomic approach combining morphology-based and DNA-based species occurrence data to document and explain species distribution patterns. Following an accepted terminology (de Queiroz, 1998; Hey *et al.*, 2003), we view species taxa, that is elementary units used in biodiversity research, as species hypotheses (SHs). Species taxa are either delimited using morphological or molecular criteria. Here, morphology-based species occurrence data are from specimens delimited using morphological criteria. DNA-based species occurrence data are from specimen sequences which are molecularly assigned to morphology-based species (i.e. molecular assignment). They can also be from specimen sequences which are molecularly clustered into groups of specimens considered as species hypotheses (i.e. molecular delimitation of species). We refer to the later as molecular operational taxonomic units (MOTU) but the diversity of molecular methods used to delimit these MOTU, and hence, the number of molecular taxonomic reference systems are fixed by the user. Individual specimens may thus be assigned to multiple SHs using distinct criteria such as morphological distinguishability and multiple DNA-based criteria. Changes in taxonomic patterns and species richness distribution patterns between any two sets of SHs can be further scrutinized, thereby paving the way towards understanding how the different properties of speciation (e.g. morphological distinguishability *vs*. genetic isolation) vary in strength relative to each other across taxa and geographic regions (Eme et al., 2018, Fišer *et al.*, 2018).

In addition to managing multi-criteria species occurrence data, the database offers a number of functionalities that are pivotal in optimizing biodiversity research within laboratories. First, it provides the necessary traceability for recovering the full set of methods and biological material linked to any species taxa occurrence data produced by the hosting laboratory. This is made possible because the database structure manages all steps of the data production process - from

sampling to DNA extraction and sequencing – and records the storage location of all vouchers produced during that process, including specimens and DNA extracts. The possibility for anyone to access biological material and replicate the methods used to produce any species occurrence data promotes scientific repeatability Second, the database accommodates species occurrence and DNA sequence data that are not produced by the hosting laboratory. This enables any user working on a particular species taxon or a set of taxa to dispose of all available data concerning those taxa into a single database. Third, although the database is designed to be deployed to fulfill a biodiversity laboratory data management needs, access to the hosting laboratory database can be granted to any users via a web application, including those outside the hosting laboratory. This promotes information sharing among laboratories while managing the permission to visualize, query and upload data for each user. At last, the database structure, web application and codes are openly distributed under the terms of GNU General Public License so that the tool can be adapted by advanced developers to fulfill laboratory's specific requirements.

II – Structure of the relational database

We provide in Figure 1 a simplified structure of the database: the full logical model is described in Figure 2. The full model has 15 core tables, 22 relational tables, 7 repository tables and one user table containing a total of 448 fields. Table 1 describes the content of the main tables and the technical features and content of fields within tables are provided in supplementary material S5.

The database structure describes a species-occurrence data-production process as classically followed by many biodiversity institutions, including research laboratories and museums. We refer to this data-production process as "internal" in that sense that data are produced internally by the institution hosting the data base (i.e. the hosting institution). As part of this process, a site is visited (table 'Site' in Fig. 1) and sampled ('Sampling'). Then, samples are sorted to provide specimen lots ('Internal biological material') containing individual specimens that supposedly belong to a single species taxa. Some specimens are isolated from a lot ('Specimen') for

morphological and/or molecular analysis. Isolated specimens or part of them are used for dissecting and mounting on microscope slides ('Slide') and/or producing DNA sequences ('DNA', 'PCR', 'Chromatogram', and 'Internal sequence'). This procedure generates biological material - specimen lots, specimen slides and DNA extracts - that is physically-stored at identified locations (Storage). Then, we incorporated external data into the modelling process, either species occurrence data ('External biological material') or DNA sequence data ('External sequence'). External data are not produced by the institution hosting the database. Rather, they arise from the literature (e.g. species occurrence data) or from data bases (e.g. DNA sequence data from GenBank). Giving the possibility to bring together internal and external data into the same data base enables any user working on a particular species taxon or a set of taxa to dispose of all available data concerning those taxa. Biological materials and sequences (either internal or external) can be linked to literature references ('Source'). Species names are assigned to biological material, specimens and sequences and the identification criterion used for assignment is specified in the table 'Identified species'. At last, MOTUs ('MOTU number') are assigned to DNA sequences, specifying well-defined delimitation methods and sequence data sets used to delimit MOTUs ('MOTU'). 'Repositories' act as dictionaries containing terms used to fill in properties of tables and the 'user' table enables to set up the role and privileges of the database users.

We emphasize three characteristic features of the database. First, the data base provides the necessary traceability for recovering the set of methods (from sampling to molecular analysis) and biological material (specimen lot, specimen slide and specimen DNA extract) linked to any species taxa occurrence data produced by the hosting institution. Giving the possibility to anyone to check the scientific veracity of the identification and distribution of species taxa is a desirable, albeit sometimes overlooked, property of taxonomy. Second, the database allows managing simultaneously geographic occurrence data for species taxa delimited using any criteria. The same occurrence data can be linked to two distinct taxonomic reference systems. The first reference system established by the user (table 'Taxon') holds a list of binomial species names which correspond mostly to species taxa that have historically been delimited and described based on morphological criteria. Specimen lots, single specimens and sequences are linked to that

reference system via the relational table 'identified species' which specifies the identification criterion used while assigning the lots, specimens and sequences to a species name. The list of possible criteria is set up by the user in the 'vocabulary' table but it may typically include criteria such as "morphology" and "molecular assignment". A user may thus assign a DNA sequence, hence a specimen, to a species name based on the similarity between that sequence and a reference sequence assigned to that species taxa. This is typically the approach involved in DNA barcoding *sensu stricto* (i.e. the molecular identification of individuals of already known species; Hebert *et al.*, 2003), but the database keeps trace of the criterion used during the attribution process. The database also integrates the historicity of species attribution since a lot, a specimen and a sequence may be attributed several times either to the same species (changing the identification criterion) or different species. We let the user decide whether the same species should necessarily be attributed to a DNA sequence, the specimen to which that sequence belongs and the lot containing that specimen. Indeed, we believe that some flexibility is needed here. For example, a specimen lot for which some specimens are identified morphologically as "genus X species Y" may be attributed to that species. Yet, a juvenile specimen from that lot may be attributed morphologically to "genus X" because juveniles cannot be identified morphologically to species level. Sequencing that juvenile specimen may yield a sequence matching a different species (i.e. the lot contains in fact two species) to which the sequence will be molecularly assigned.

The second referential established by the user (tables 'MOTU number' and 'MOTU') holds a list of numbers which correspond to operational taxonomic units that are delimited using molecular methods. This referential is part of the field of DNA taxonomy which consists in using DNA to delineate species boundaries (Tautz *et al.*, 2003; Fontaneto *et al.,* 2015). The characteristic feature of MOTU referential is that it is renewed much more often than binomial species name referential. MOTUs are regularly generated by scientists by applying several molecular delimitation methods to sets of DNA sequences belonging to multiple specimens and multiple morphologically-distinguishable species taxa (see for example Dellicour & Flot 2015; Fišer *et al.*, 2018 for a review of molecular delimitation methods). The outcome of MOTU delimitation (i.e. the number of MOTU and their boundaries) is sensitive to the delimitation method and sequence

data set used. Therefore, the GOTIT database is conceived to accommodate several MOTU reference systems while specifying the delimitation methods and sequence data set used in each referential. Internal and external sequences are linked to MOTU numbers stored in the table 'MOTU number' and the table specifies the molecular delimitation methods used in each referential. The list of methods is set up by the user in the 'vocabulary' table and the sequence data set used in each referential is described in the table 'MOTU'.

Third, the database structure accommodates data arising from external sources including the literature, species occurrence data bases and nucleotide sequence databases. External data were introduced into the logical data model as two separate tables – 'External biological material' and 'External sequences' - because they often lack the necessary level of description to be included as part of the internal data-production process. Yet, the two tables are linked to the referential 'Taxon' and the table 'External sequence' is linked to the MOTU referential following the same procedures as described above for the internal specimen lots and sequences. At last, the two tables 'External biological material' and 'External sequences' are linked via the table 'Sampling' to the table 'Site' containing geographic coordinates, thereby enabling to map simultaneously species occurrence data and MOTUs arising from external and internal sources.

III – Requirements and technologies

The GOTIT data base was developed using the open source object-relational database system postgreSQL (https://www.postgresql.org/) and its graphical user interface administration tool pgAdmin (https://www.pgadmin.org/). Users can access any subset of data they need from the database using PL/pgSQL queries in pgAdmin. A number of query examples are provided in Appendix S6.

IV – References

de Queiroz, K. 1998. The general lineage concept of species, species criteria, and the process of speciation and terminological recommendations. In D. J. Howard, & S. H. Berlocher (Eds.), Endless forms: Species and speciation (pp. 57–75). Oxford, UK: Oxford University Press.

Dellicour S., Flot J.-F. 2015. Delimiting species-poor data sets using single molecular markers: A study of barcode gaps, haplowebs and GMYC. Systematic Biology, 64(6), 900–908.

Eme D., Zagmajster M., Delić T., Fišer C., Flot J.-F., Konecny-Dupré L., Pálsson S., Stoch F., Zakšek V., Douady C. J., Malard F., 2018. Do cryptic species matter in macroecology? Sequencing European groundwater crustaceans yields smaller ranges but does not challenge biodiversity determinants. Ecography 41: 424–436.

Fišer C., Robinson C.T., Malard F. 2018. Cryptic species as a window into the paradigm shift of the species concept. Molecular Ecology 27: 613–635.

Fontaneto D., Flot J.-F., Tang C.Q. 2015. Guidelines for DNA taxonomy, with a focus on the meiofauna. Marine Biodiversity, 45, 433–451.

Hebert P.D.N., Cywinska A., Ball S.L., DeWaard J.R. 2003. Biological identifications through DNA barcodes. Proceedings of the Royal Society B, 270(1512): 313–321.

Hey J., Waples R.S., Arnold M.L., Butlin R.K., Harrison R.G. 2003. Understanding and confronting species uncertainty in biology and conservation. Trends in Ecology and Evolution 18(11): 597–603.

Tautz D., Arctander P., Minelli A., Thomas R.H., Vogler A.P. 2003. A plea for DNA taxonomy.

Trends Ecol. Evol. 2003; 18 (DOI: 10.1016/S0169-5347(02)00041-1)

V – Tables and figures

Table 1. Content of main tables in the GOTIT database. Abbreviations for table type are as follows: C: core table; R: repository table; L: relational table.

Table	Table type	Content
Site	C	Coding of site, features and geographic location, habitat classification
Country	R	List of countries in which the sites are located
Municipality	R	List of municipalities in which the sites are located
Sampling		Coding of sample, taxa targeted during sampling, sampling method and effort, measurements, sample status (e.g. to be processed)
Program	R	Description of programs which fund sampling: program title, coordinators, funding agency …
Internal biological material	C	Coding and content of specimen lots collected by the hosting laboratory including number and features of specimens (e.g. sex, stages, pigmentation, eyes).
Composition of internal biological material	C	
Specimen	C	Coding of single specimens isolated for morphological and/or molecular analysis, features of specimen, coding of specimen-containing tubes
Specimen slide	C	Coding and description of microscope slides containing dissected specimens. Link to photos of microscope slides.
DNA	C	Coding of DNA extracts, extraction method, and DNA quality and concentration.
PCR	C	Coding of PCR, targeted gene, forward and reverse primer, PCR quality, and PCR characteristics (e.g. semi-nested)
Chromatogram	C	Coding of chromatogram, primer and quality of chromatogram
Internal sequence	C	Coding of sequence, coded link towards sequence alignment, sequence status (e.g. validated, in progress), accession number (if any).
External biological material	C	Coding and content of specimen lots arising from literature sources including number and features of specimens (e.g. sex, stages, pigmentation, eyes).
External sequence	C	Coding of sequences arising from literature sources, coded link towards sequence alignment, sequence origin, sequence status (e.g. validated, pseudo gene), accession number, gene, taxon associated to the sequence in the source
Identified species	L	Relational table linking internal and external biological materials, specimens, internal and external sequences to a species taxa in the taxon table. Each link has an attribute corresponding to the criterion used for identification (e.g. morphology, molecular assignment)
Taxon	R	Taxonomic reference list containing taxon ranks and names
MOTU number	L	Table containing MOTU numbers attributed to sequences as well as information on the MOTU data set and MOTU generating methods.
MOTU	C	
Source	C	Coding and description of literature references: year, title
Box	C	Coding and description of storing boxes containing internal biological materials (including single specimens isolated for morphological and/or molecular analysis), specimen slides, and DNA extracts.
Vocabulary	R	Terms used to fill in properties of entities (e.g. habitat types, primers, etc…)
Person	R	List of persons (names, full names) involved in the data building process (sampling, DNA extraction etc…)
Institution	R	List of institution names to which the persons belongs

FIGURE 1. Simplified schematic model of GOTIT database. Numbered and colorized lines correspond to distinct pathways. 1) red: internal data production; 2) yellow: storage; 3) orange: external data; 4) black: data sources; 5) blue: species assignment.

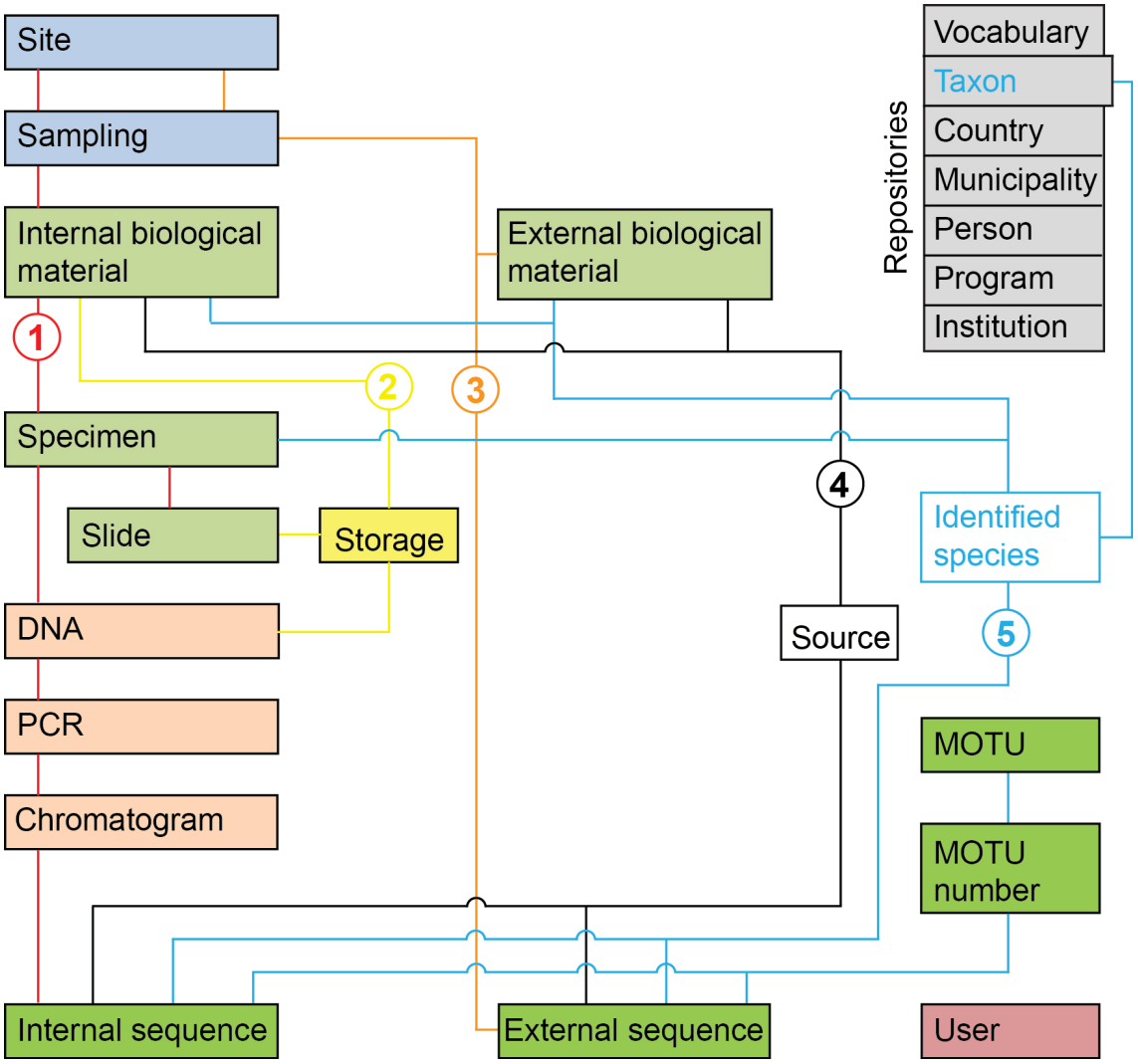
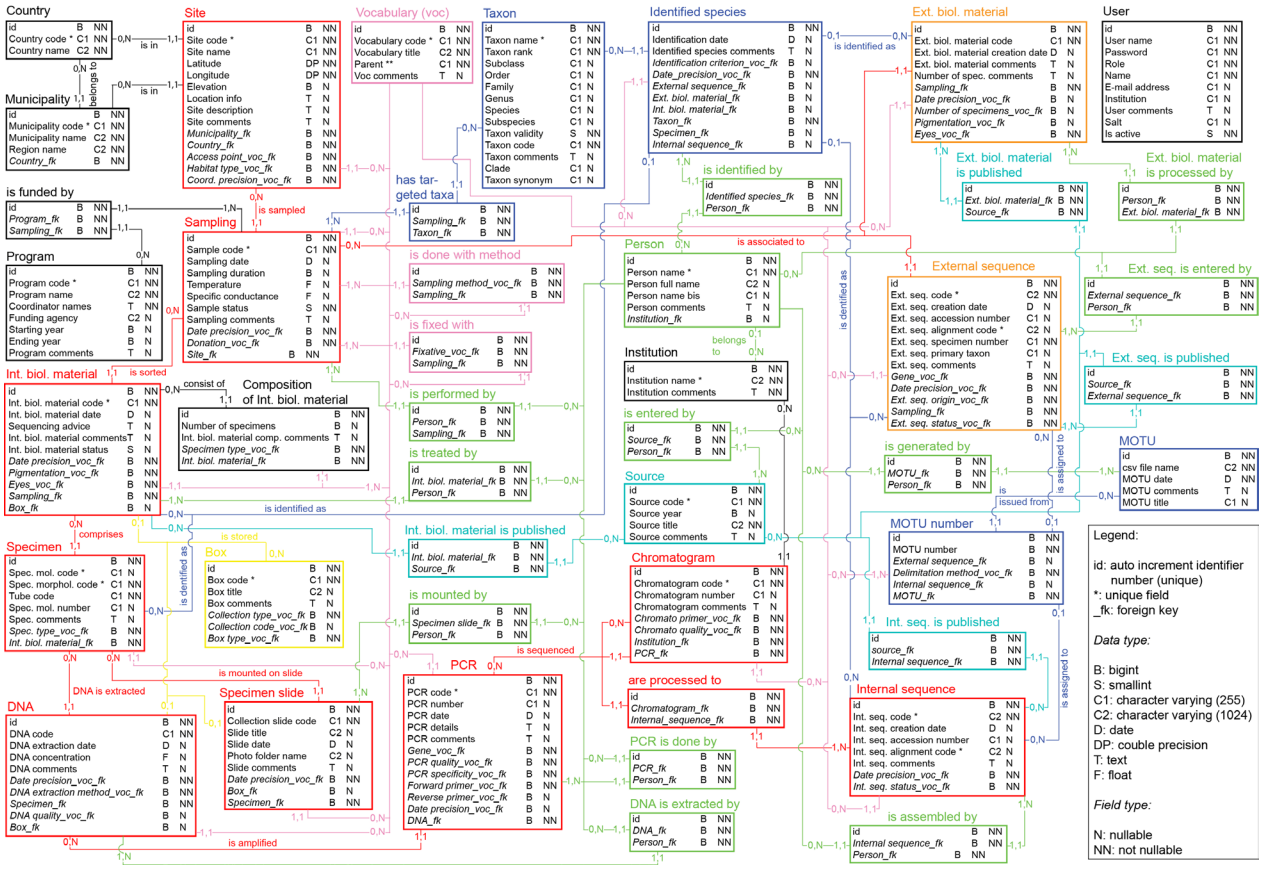


FIGURE 2. Logical data model of GOTIT database. The following abbreviations are used: Coord.: coordinates; Int. biol. mat.: internal biological material; Ext. biol. mat.: external biological material; Int. seq.: internal sequence; Ext. seq.: external sequence; Comp.: composition; Spec. : specimen; Morphol.: morphological; Mol.: molecular.



VI – Supplementary materials

We provide in the folder entitled "/doc/database" a list of resource files for managing and querying the database. Since the table and field titles in the database are in French, we provide for each resource file an English and a French version of it.

S1_Gotitdb_logical_model_en.jpg: Logical data model of GOTIT database (English version). The following abbreviations are used: Coord.: coordinates; Int. biol. mat.: internal biological material; Ext. biol. mat.: external biological material; Int. seq.: internal sequence; Ext. seq.: external sequence; Comp.: composition; Spec. : specimen; Morphol.: morphological; Mol.: molecular.

S2_Gotitdb_logical_model_fr.jpg: Logical data model of GOTIT database (French version).

S3_Gotitdb_conceptual_model_en.jpg: Conceptual data model of GOTIT database (English version). The following abbreviations are used: * : is ; ' : belongs to ; ^ : has ; ! : is obtained with ; ibm : internal biological material ; ebm : external biological material ; Int. seq. : internal sequence ; Ext. seq. : external sequence ; iden. criterion: Identification criterion ; nb of specimens : Number of specimens; Int. biol.mat.: internal biological material; Ext. biol. mat.: external biological material; Comp.: composition.

S4_Gotitdb_conceptual_model_fr.jpg: Conceptual data model of GOTIT database (French version). The following abbreviations are used: ': appartient à ; ass: assemblé(e); biomol: biologie moléculaire; c: °C (degré Celsius); chromato: chromatogramme; coll: collection; com.: commentaire; conc.: concentration; cond.: conductivite; echant.: echantillonnage; ext: externe; ind: individu; info: information; lat: latitude; long: longitude; micro_sie_cm: micro siemens par centimètre; morpho: morphologique; nb: nombre; num: numéro; sqc: séquençage / séquence;

S5_Gotitdb_tables_fields.ods: Description of tables, field characteristics and field contents in GOTIT database (English and French version).

S6_Gotitdb_sql_queries.pdf: A list of selected SQL queries to recover different subsets of data from GOTIT database.