Database procedures

Download PDF version

The Coral Trait Database is an open source research initiative that aims to make all observations and measurements of corals accessible in order to more rapidly advance coral reef science. Anyone collecting coral trait data (e.g., collected in field and laboratory studies, extracted from the literature, or by other means) can join and contribute to the growing data compilation. Contributors have control over the privacy of their data and greatly benefit from being able to download complementary public data from the database in a standard format for use in their analyses. We hope that private data will become public once the contributor has published them, which will subsequently be cited when their data are used in analyses by other people. We have carefully designed a citation system that ensures full transparency about the origin of each individual data point as well as larger data compilations and compilations of other peoples' data (such as data extracted from literature for meta-analyses). By becoming a data contributor, you become a member of the Coral Trait Database Consortium.

> The Coral Trait Database Consortium (2014) The Coral Trait Database. http://coraltraits.org

Contents

1. Mission
2. Governance

   - Administrator
   - Managerial Board
   - Editorial Board
   - Taxonomy Advisory Board
   - Database programmers

3. Traits

   - What information can be contributed
   - Observations and measurements

4. Data submission

   - Accepted data
   - Small submissions
   - Bulk submissions (recommended)
   - Back-end imports
   - Submitting other people's data

---

1. Mission

To be useful, big data need to be organised in a big way.

- Advance coral reef science through the gathering and dissemination of data resources
- Provide an open access portal providing unrestricted interactive and automated access to information about corals
- Avoid redundant data gathering efforts
- Fairly recognize those who collected original data, as well as those who compile larger data compilations
- Provide a platform for community engagement in data quality control
- Facilitate reproducible science
- Do the above with a sustainable future for the database in mind, in terms of management and longevity

Top

---

2. Governance

Other than part-time programming and administrative help, the database is run and maintained by volunteers.

Administrator

- Toni Mizerek (part-time)

Managerial Board

For decisions requiring a vote, at least three of the Managerial Board should agree.

- Joshua Madin
- Andrew Baird
- Daniel Falster
- Emily Darling

Editorial Board

Editors listed *here* are responsible for quality control of trait data.

Taxonomy Advisory Board

- Andrew Baird
- Danwei Huang

Technical Advisory Board

The database relies on external people with breadth of experience and technical know-how, and who appreciate the subtle difficulties of managing large data compilations.

- Matt Kosnik

People to potentially approach:

- Matt Jones
- Mark Schildhauer
- John Alroy
- Ethan White
- Scott Chamberlain or Karthik Ram (ropensci)

Database programmers

- Joshua Madin
- Surendra Shrestha (part-time)

---

3. Traits

The Coral Traits Consortium is flexible about the definition of trait.

A true trait is a heritable quality of an organism. However, including species-level characteristics, such as geographical range size and conservation status, in the same database makes analyses of true traits much more effective. In addition, the database captures contextual (ancillary) data about the environment in which a trait measurement was recorded (e.g., "traits" of habitat, seawater or an experiment). Contextual "traits" are important for understanding true traits (e.g., as predictors or co-variates). Individual-level (true) traits, species-level characteristics and contextual data are all considered traits in the database.

The current list of traits can be found *here*.

What information can be contributed

The database currently accepts individual-level and species-level measurements of both zooxanthellate and azooxanthellate Schleractinain corals.

Public data must be citable.

Unpublished can also be imported into the database if it is kept private. Private data can be made public once published and associated with a resource. Benefits of accepting unpublished data include:

- Being able to easily download the data alongside useful public data for analysis
- Being more likely to contribute the data at end of project and improve its longevity
- Identify duplicate research efforts

4

`Individual-level` measurements typically include other measurements of the environment that provide useful context. One (or more) trait measurements of a coral and other contextual measurements are bound together as an observation. For example, an measurement of larval swimming speed was measured within the context of swimming direction and water temperature, both of which provide important information about that particular swimming speed measurement.

`Species-level` measurements are either traits that do not vary among individuals over ecological time-scales (such as sexual system and if zooxanthellate or not) or emergent characteristics of individuals within a species (such as geographical range size, conservation status, wave exposure preference and lower depth).

Observations and measurements

Observations bind related measurements. For example, observing the same coral and measuring its height and weight results in one observation with two measurement (each corresponding with a different trait of the coral). If water temperature was also measured, then this also belongs to the same observation, but as a contextual trait.

`Observation-level` data include the coral species, location and resource. These data are the same for all measurements corresponding to the observation. When entering or importing trait data, the following is required.

- User
- Access (public or private)
- Coral species
- Location
- *Resource

*Resource can be left blank for unpublished data, but data must be kept private.

`Measurement-level` data include the trait, value, standard (unit), methodology, and estimates of precision (if applicable). When entering or importing trait data, the following is required.

- Trait
- Value
- Standard

Top

---

4. Data submission

The Coral Traits Database is a research tool, not a meta-data catalogue.

A meta-data repository captures high-level information about your data set, so that people can easily find it. Examples of meta-data repositories include DRYAD, Ecological Archives and Nature Scientific Data. You are encouraged to submit data sets to meta-data repository to help ensure their longevity and the reproducibility of the results for whcih the data were originally collected.

The Coral Traits Database captures data-level information so that measurements from multiple data sets can be integrated, extracted, compared and analysed.

One way to think of The Coral Traits Database is a very large data-set being cobbled together by the coral reef community for everyone to use, avoiding redundant efforts and speeding up science.

Accepted data

- Individual-level raw data you collected yourself that has been published
- Individual-level data extracted from existing publications (typically not raw; e.g., means from tables and figures)
- Species-level model-derived data that has been published
- Species-level expert or group opinion data that has been published
- Unpublished data if kept private until published

To contribute data, you need to make a database account (*here*) and then contact the database Administrator. Signing up alone does not enable you to contribute data.

Having a primary, peer-reviewed resource is essential for maintaining data quality, contributor recognition and scientific rigour

Small submissions

If you are entering one or a few observations, you can use the Add Observation link in the menu bar. The bulk import instructions will help you understand the general data entry protocol. For instance, you need to enter the locations and resources, and also ensure the traits your require exist, before you can save an observation.

Bulk submissions (recommended)

Even if only entering small amounts of data, the bulk submission process is recommended. Information about importing large amount of data is *here*.

> Bulk imports retain a copy of the import spreadsheet that allows you to easily try again if errors are made.

Back-end imports

If your data is well-managed, you can ask one of the database programmers to upload it for you. The data will be associated with your name and made private. You are required to make the data public yourself (if desired).

Submitting other people's data

Entering published data not already in the database in strongly encouraged to improve the data's longevity and augment data analysis. A case in which this might occur is a meta-analysis. The data enterer can keep the data they submit private until their study is published.

> The key ojective is to extract the data in such a way as to avoid people ever needing to go back to the primary resource again.

For example, extracting only the mean value of a trait from a paper, without extracting any measure of variation or the context in which the trait was measured, will mean that the data may not be useful for other purposes. Someone else might need to go back and extract the information again, and there is a chance your initial efforts won't be cited.

- **Primary resources only.** Often people enter data from summary tables in papers that come from other (primary) resources. It is important to enter the data from the primary resource for two reasons: (1) so that the primary resource's author is credited for their work, and (2) to avoid data duplication, where the same data are entered from both the primary and secondary resource. Secondary resources, such as meta-analyses, can be credited for accummulating data by using database citation manager.

- **Careful extraction.** Copy values from tables carefully and double check. Extracting data from figures can be done with software like ImageJ, where a scale can be set based on axis values and measurements of plotted data made, including error bars.

- **Gather important context.** Enter contextual data as well, which might require reading the methods. This might be as simple as the depth or habitat in which corals were measured, and potentially the same for all observations. Such information is very useful. However, contextual information can get complicated quickly. For example, when the planar area (size) of a colony is measured each year for 10 years, context will include an individual identifier to capture that the same colony was measured, as well as the year to determine the order in which measurements were made. Examples of capturing a range of contextual situations can be found on the Imports page, which will be added to as new situations arise.

Quality control

There are three levels of data review.

1. `Contributor-level` review at time of submission, Once submitted, data are tagged as *pending*.
2. `Editor-level` review. The relevant Editor/s for traits in your submission are automatically notified by email. The contributor may be contacted by the Editor if there are any issues with the submission. The Editor will *approve* the submission once satisfied.
3. `User-level` review. Anyone signed-in as a database user can report an issue with an observation record, and the submitter and the Editor will be notified by email.

Error checking

Basic error checking will ensure data submissions fit into the database. Error checking will improve as different issues arise. Measurement records with the same coral species, location, resource and value will be flagged as potential duplicates.

Top

---

5. Downloading data

Information about downloading data is *here*.

Top

Primary and secondary resources

If you publish a study using data from the Coral Traits Database, it is your responibility to cite the data correctly (see License). There are two levels of resources.

`Primary resources` (should) exist for every observation in the database. When data are downloaded, you are also sent a list of these resources for citing in your work (see Downloads).

`Secondary resources` exist when someone has gone to the trouble to compile other people's data, and have published this compilation (e.g., a meta-analysis). This collection of observations will have a secondary resource.

Not complete, because still working on citation management system.

Publishing with data from the database

If you publish a study using data from the Coral Traits Database, it is your responsibility to submit that snapshot of data to the journal or a meta-data repository to ensure your results are reproducible.

> The database is constantly growing and changing.

The database keeps track of changes through time, but does not currently have a mechanism for downloading historical snapshots. Instead, downloaded files have the date appended to them for reference.

People have expressed concern about the proliferation of duplicate coral trait data sets. Our hope is that the Coral Traits Database will remain the primary data resource. Data snapshots submitted to meta-data repositories are to ensure the reproducibility of the published results.

------

6. Database design

The Coral Trait Database has been developed according to the Observation and Measurement Ontology (OBOE) that was developed at the National Center for Ecological Analysis and Synthesis (see references below). The system has been designed with reef corals in mind, but is a generic system for data in which entities are observed, and then traits of these entities are measured. The key distinction between OBOE and other observational models is that trait-entity relationships and observation context are made explicit. In other words,

the database preserves meta-data about data values as well as the standard meta-data about data sets.

Several contextual constraints have been implemented in the Coral Trait Database to simplify the structure and improve speed. For example, we model only five observed entities: User, Resource, Location, Time and Coral. Measurements of "traits" can be taken of all five entities; however, we constrain the measureable traits for the first four of these entities. Measurements of "traits" for Coral are highly flexible. Also for simplicity, many contextual entities, such as "habitat" or "plot", are measured at the Coral entity level. For example, a Coral "planar area" (i.e., colony size) can be measured as expected, but so can traits for contextual entities such as "water temperature", "depth" and "habitat". Such constraints on the model can be relaxed if necessary, but are in place to improve the performance of the model for current purposes.

Database design references

- Madin JS, Bowers S, Schildhauer M and Jones M (2008) Advancing ecological research with ontologies. Trends in Ecology and Evolution. 23:159-168.
- Bowers S, Madin JS and Schildhauer M (2008) A conceptual modeling framework for expressing observational data semantics. Lecture Notes in Computer Science.
- Madin JS, Bowers S, Schildhauer M, Krivov S, Pennington D and Villa F (2007) An ontology for describing and synthesizing ecological observational data. Ecological Informatics 2:279-296.
- The foundation for this web application (e.g., session and user models) was developed using Michael Hartl's Ruby on Rails tutorial.

Database usage

Website and data download activity are tracked using Google Analytics.

Top

---

7. Database code

The database was developed using Ruby on Rails, is open source, and can be found at Github.

Top

---

8. License

Basically, if you enter data into the database and make it public, the data can be reused by others if they cite it correctly.

Similarly, if you download and use data in an analysis to be published, you must cite primary (and, if applicable, secondary) resources correctly.

A couple of key points:

- Content must be associated with a published scientific or other scholarly research document
- Content may only be entered by an individual who represents and warrants that s/he is the creator and owner of the content or otherwise has sufficient rights to be able to make the content available under a CC0 Waiver
- The primary language of the content must be English.
- Once CC0 is applied by making data public and associating it with a published research Document, the data can not be made private again.

Top