Data Descriptor: A database of marine phytoplankton abundance, biomass and species composition in Australian waters

Claire H. Davies et al.*

There have been many individual phytoplankton datasets collected across Australia since the mid 1900s, but most are unavailable to the research community. We have searched archives, contacted researchers, and scanned the primary and grey literature to collate 3,621,847 records of marine phytoplankton species from Australian waters from 1844 to the present. Many of these are small datasets collected for local questions, but combined they provide over 170 years of data on phytoplankton communities in Australian waters. Units and taxonomy have been standardised, obviously erroneous data removed, and all metadata included. We have lodged this dataset with the Australian Ocean Data Network (http://portal.aodn.org.au/) allowing public access. The Australian Phytoplankton Database will be invaluable for global change studies, as it allows analysis of ecological indicators of climate change and eutrophication (e.g., changes in distribution; diatom:dinoflagellate ratios). In addition, the standardised conversion of abundance records to biomass provides modellers with quantifiable data to initialise and validate ecosystem models of lower marine trophic levels.

<table>
<thead>
<tr>
<th>Design Type(s)</th>
<th>data integration objective ● database creation objective</th>
</tr>
</thead>
<tbody>
<tr>
<td>Measurement Type(s)</td>
<td>phytoplankton ecology</td>
</tr>
<tr>
<td>Technology Type(s)</td>
<td>data item extraction from journal article</td>
</tr>
<tr>
<td>Factor Type(s)</td>
<td></td>
</tr>
<tr>
<td>Sample Characteristic(s)</td>
<td></td>
</tr>
</tbody>
</table>

Correspondence and requests for materials should be addressed to C.D. (email: claire.davies@csiro.au). #A full list of authors and their affiliations appears at the end of the paper.
Phytoplankton are microalgae at the base of the food web and directly or indirectly support all marine life. As highly efficient primary producers they are critical to maintaining biodiversity and supporting fisheries throughout the ocean. Due to their high turnover rates and sensitivity to changes in environmental conditions phytoplankton are useful indicators of changing oceanographic conditions, climate change, and deterioration in water quality. Some phytoplankton produce toxins which may be accumulated by filter feeding shellfish, causing irritation, serious illness or death to animals and humans.

A litre of seawater may contain up to one million algal cells, representing at least 100–150 different species. These include the larger phytoplankton, dominated by the diatoms and dinoflagellates, but also include flagellates and the coccoid picoplanktonic forms. Currently, 537 intraspecific dinoflagellate and 938 diatom taxa are known to inhabit coastal and oceanic waters around Australia. The fractions of nanoplankton flagellates and coccoid picoplankton, although smaller in size, can account for up to 90% of the total phytoplankton chlorophyll under low biomass scenarios in offshore waters. The latter are difficult, or impossible, to identify with light microscopy. Including a reliable estimate of biomass, along with cell abundance will provide more realistic information about the phytoplankton community structure at a particular point in time.

Many researchers have phytoplankton data sitting around on paper records or in spreadsheets, some published and some unpublished, which may eventually be lost or misplaced. Consultancies often hold data archives from several research projects from which data are released only to the client with no time and incentive to publish. These data, archived thoroughly, standardised and freely available to a wider audience, are an invaluable resource for the research community. Small, individual datasets with limited stand alone impact, can collectively provide valuable additions to large scale spatial and temporal studies. Many of these data have previously been published in some form. Journal articles, theses and reports and especially older publications, rarely include the relevant dataset so the data are not available for use after publication unless the author releases the data to a publically accessible platform. Even when data are released, they are often only presence data and are not explicit as they could be.

The Australian Phytoplankton Database has been collated from literature, active and retired researchers, consultancies, archives and databases. Only data with the relevant corresponding metadata about collection location, date and methods has been included. The Australian Phytoplankton Database contains data on marine phytoplankton abundance, biomass and species composition. It can be used to:

- develop distribution or biogeographic maps;
- determine community structures and range changes over time or oceanographic conditions;

Data availability scale

<table>
<thead>
<tr>
<th>Publication options</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thesis</td>
</tr>
<tr>
<td>Journal report</td>
</tr>
<tr>
<td>Client report</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data storage options</th>
</tr>
</thead>
<tbody>
<tr>
<td>Notepad</td>
</tr>
<tr>
<td>Personal computer</td>
</tr>
<tr>
<td>Networked database</td>
</tr>
<tr>
<td>Webpage</td>
</tr>
</tbody>
</table>

Figure 1. Flow diagram of data collation, verification and release to AODN.
understand dynamics of harmful algal species to help inform the aquaculture industry\textsuperscript{11};

- develop inputs to climate, ecosystem and fisheries models to inform management about resources\textsuperscript{12}.

The Australian Phytoplankton Database is available through the Australian Ocean Data Network (AODN: http://portal.aodn.org.au) portal. This portal is the main repository for marine data in Australia. The Phytoplankton Database will be maintained in the CSIRO data centre and can be updated with new records, which will automatically upload to the AODN. Researchers wishing to submit new data should contact the corresponding author or the AODN. A snapshot of the Australian Phytoplankton Database as it is at the time of this publication has been assigned a DOI and will be maintained in perpetuity by the AODN (Data Citation 1).

The Australian Phytoplankton Database has been built with ease of use and minimising user error in mind. Therefore it provides the clean data at a level that requires minimal interpretation. The CSIRO database holds all the raw data, for example original species names and ambiguous records, from these datasets, and researchers can request further information from the corresponding author if required.

### Methods

Samples were mainly collected from Niskin bottles, net drops or tows, and the Continuous Plankton Recorder (CPR). These are all standard methods of collecting phytoplankton samples\textsuperscript{13,14} and many are still largely reliant on a phytoplankton manual written in 1978 (ref. 15). A few samples were collected using automated samplers on moorings\textsuperscript{16}. The sampling was done via research vessels, container ships and small boats by experienced researchers, students and volunteers. The majority of the samples were preserved using Lugol’s solution, although formalin, paraformaldehyde and glutaraldehyde have also been used. Different methods of preservation can affect the condition of the sample and which taxa are well preserved\textsuperscript{4,17}. Samples were analysed with standard methods including light microscopy, transmission or scanning electron microscopy which are described in Hallegraeff, et al.\textsuperscript{4} All methodological variations within our phytoplankton database are detailed in the metadata where available and recorded for each data entry. Where available a citation is referenced for each project, which gives details on methodologies and limitations from that project (Table 1 (available online only)).

There were three stages in data gathering. The first stage was to conduct a literature review of Australian phytoplankton data. Any literature that contained abundance or presence data was digitised and uploaded into the CSIRO maintained Oracle database. The second stage was to scan already existing databases, such as the CSIRO data trawler, the Ocean Biogeographic Ocean System (OBIS) and the Atlas of Living Australia (AOLA). These repositories only store presence records. Relevant data were selected and uploaded into the database. The third stage was to ask researchers to contribute any other data sets that they had. All data were organised into a standard format and uploaded into the database. Data were then served to and hosted by the AODN.

All taxa have been verified as accepted species and given the currently accepted name as defined by the World Register of Marine Species database (WoRMS—http://www.marinespecies.org/aphia.php?p=...
Identifi cation of the smaller phytoplankton is often to a coarser taxonomic level as many cannot be distinguished to species using light microscopy. In some studies electron microscopy has been used to determine species, but in other studies functional groups have been identifi ed. This data set does not include accessory pigment data which can help resolve these smaller taxa18 although it can be thought of as a complementary dataset. Over 20 years of pigment data are available in Australia via the AEsOP database (http://aesop.csiro.au/).

Cell biovolume is calculated as per Hillebrand, et al.19 following the suggested shape factors for each genera. Size parameters were estimated from measurements taken from Australian samples or Australian references where available8,20 and other sources where not21–23. In some data sets, direct measurements of size classes (e.g., P599 the Australian Continuous Plankton Recorder Survey and P597 the IMOS National Reference Stations), were used in preference to literature values. For some taxa there was insuffi cient information available to estimate a biovolume, these were generally the rare taxa. Rather than estimate a size class without any information available, these have been left blank.

Data Records
Each data record represents the abundance or presence of a phytoplankton taxa at a certain point in space and time and has been given a unique record identifi cation number, P(project_id)_(sample_id)_ (record_id). Each data record belongs to a project, with each project having a unique identifi cation number, Pxxx. A project is de ned as a set of data records which have been collected together, usually as a cruise or study with the same sampling method and having the same person counting the samples. Metadata ascribed to a project relates to all the data records within that project. Details to identify each separate project are given in Table 1 (available online only). Each sample within that project has a unique sample_id. The sample_id has not been changed from the original data set to maintain traceability. So these may be duplicated between projects but P(project_id)_(sample_id) will be a unique entity in space and time. Species abundance records within the sample are given a unique record identi fi cation number, P(project_id)_(sample_id)_.

The majority of these projects have been uploaded as part of the collation of data for this database. None have been previously published as datasets but The IMOS National Reference Stations, P599, and Continuous Plankton Recorder Survey, P597, which together constitute half the data in this database, are continually updated and available through the AODN (https://portal.aodn.org.au/search?uuid=dfef238f-db69-3868-e043-08114f8c8a94 and https://portal.aodn.org.au/search?uuid=c1344979-701-0916-e044-00144f7bc0f4 respectively).

Table 1 (available online only) gives summary information on the project data sets, their space, time and taxonomic resolutions, numbers of samples and records available. Users can select data sets from this information and download as desired through the AODN.

Technical Validation
The Phytoplankton Database will provide an extensive resource for phytoplankton researchers, although there are some caveats due to the variety of the sampling and analysis protocols. The various sample collection methods infer that abundances might not always be directly comparable across projects. For example, quantitative methods such as bottle sampling (e.g., Project 599) will collect all but the rarest phytoplankton and will include the whole size spectrum, whereas semi-quantitative methods such as net sampling are selective and dependent on tow method, mesh size and the mix of species present in the water as some species may clog the net and trap smaller species that would otherwise go through the mesh (e.g., Project 509)24. By including all data collected using different methods and including this information as meta-data, researchers are able to analyse the relative abundance of each taxa within a project and compare across compatible projects. Metadata includes as much detail as is available about sampling methods and limitations and provides guidance to the users about the potential of each data set. Users should consider collection methods, preservation techniques and microscopic limitations when comparing datasets. All datasets have been standardised to taxa/m3 of water except P1070 where the units are taxa per gram of substrate measured. This project collected the phytoplankton by collecting substrates and analysing parts of the substrate. It is included here as the only data set on Gambierdiscus and associated benthic dinoflagellates from Australia which are important to the studies of the ciguatera.

All datasets submitted can be interpreted as con rm ing the presence of those species recorded. In some datasets, e.g., time series, it is possible also to infer absences, assuming that all species are looked for on each sampling occasion. Absences have been included in the data sets where the project information available allowed us the confi dence to interpret such absences correctly. The interpretation of absences from other projects is at the discretion of the user. We suggest that a project-by-project approach should be taken. If a taxa is not observed at all in a project, then the absence could be due to the taxa not being present, that taxa not being of interest to the analyst, or the inability to identify that taxa. Thus, a real absence should not be inferred. If a taxa is observed in some samples of a project, it can most likely be
assumed that the microscopist could identify the taxa, and that a real absence may be inferred in samples within that project where the taxa was not marked as present. Some data records were removed when there was ambiguity as to the identification of the taxa, i.e., when the taxonomic traceability, usually from older sources, was confused or when spelling mistakes make it unclear which one of two species was meant. Species known as freshwater species were removed

<table>
<thead>
<tr>
<th>Group</th>
<th>Equation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diatoms</td>
<td>$B = 0.288 \times V^{0.811}$</td>
</tr>
<tr>
<td>Dinoflagellates</td>
<td>$B = 0.76 \times V^{0.039}$</td>
</tr>
<tr>
<td>Tintinnids</td>
<td>$B = 444.5 + 0.053 \times V$</td>
</tr>
<tr>
<td>Ciliates</td>
<td>$B = 0.22 \times V^{0.080}$</td>
</tr>
<tr>
<td>Other protists (excluding diatoms)</td>
<td>$B = 0.216 \times V^{0.030}$</td>
</tr>
<tr>
<td>Silicoflagellates</td>
<td>as for diatoms</td>
</tr>
<tr>
<td>Phaeocystis antarctica</td>
<td>Use $B = 9 \text{ pg C cell}^{-1}$</td>
</tr>
</tbody>
</table>

Table 2. Information for converting biovolume $V$ ($\mu m^3$) to carbon biomass $B$ (pgC.cell$^{-1}$).

![Figure 3](image-url)  

Figure 3. Demonstrated uses of phytoplankton data in the Australian Phytoplankton Database. (a) Map showing the range extension of *Noctiluca scintillans* over the past 150 years$^{28}$; (b) Parameterisation of an ecosystem model (Atlantis) of diatom abundance using satellite chlorophyll Phytoplankton Functional Type analysis and the Continuous Plankton Recorder species ratios; (c) Diatom diversity by latitude from the Continuous Plankton Recorder; (d) Distribution of potentially harmful algal bloom species *Dinophysis tripos*.  

assumed that the microscopist could identify the taxa, and that a real absence may be inferred in samples within that project where the taxa was not marked as present. Some data records were removed when there was ambiguity as to the identification of the taxa, i.e., when the taxonomic traceability, usually from older sources, was confused or when spelling mistakes make it unclear which one of two species was meant. Species known as freshwater species were removed
as the methods used to collect data were not aimed at freshwater environments and the inclusion of the odd records of these species would not be comprehensive or meaningful. Estuarine species were captured and the records kept. Data records with positions on land, with an uneven number for abundance or with impossible dates were also removed or converted to presence records.

Usage Notes
The database contains information on the functional group of the species, which can aid analysis. Functional groups include diatoms, dinoflagellates, flagellates, ciliates (including tintinnids), silicoflagellates, and cyanobacteria. Once downloaded and binned as required, data are suitable for use in the creation of ecological indicators. For example:

- Total diatoms, dinoflagellates
- Diatom:Dinoflagellate ratio
- Total phytoplankton abundance or biomass per degree square

Abundance (cells.L\(^{-1}\)) for the phytoplankton counts is given where it is available, providing more information about the productivity of an area than presence data alone. A low cell abundance may indicate a low level or production, but this may not be the case if these are large cells. The biomass data helps to show productivity of an area. The biovolume has been calculated for each cell count and when converted to biomass, is available for use by modellers and to assist in interpretation of an area’s productivity. An accepted method of converting biovolume to biomass is to assume that the cell has the density of water (1 mm\(^3\).L\(^{-1}\) = 1 mg.L\(^{-1}\)). Another useful conversion is to carbon biomass; full methods are readily found in the literature\(^{24-26}\). Table 2 gives details conversions of phytoplankton size data to carbon biomass.

Some of the data records are missing coordinates or dates. It was considered useful to keep these records as the presence of a taxa in a location may still be of value. The user may estimate coordinates from the location given and would then also be aware that these would not be the exact coordinates of the sample.

Data can be analysed in many different ways and in many software applications (e.g., R, Matlab). We include here some figures created in The R Project for Statistical Computing (https://www.r-project.org) to demonstrate some potential uses of the data (Fig. 3).

In some cases, notably project 599, the IMOS National Reference Stations, the phytoplankton component of the survey is only a part of the data available. Additional biogeochemical data are available for this data set via the AODN. Some of the projects, viz. P479, P599, P597, have corresponding zooplankton data freely available in The Australian Zooplankton Database\(^27\). These data sets can be matched by the project_id and the sample_id which are consistent across databases. The list of citations referenced in Table 1 (available online only) will also give users information as to how this data has been previously used from the discrete projects.

References


63. Hallegraeff, G. M. & Jeffrey, S. W. Annually recurrent diatom blooms in spring along the New South Wales coast of Australia. 

Data Citation

Acknowledgements
We acknowledge the contributions from all collaborators and their institutions. If data from multiple projects are used, please acknowledge this publication; if individual project data are used, please acknowledge use of data as per the custodian information. We would also encourage the data user to enter into collaboration with the researchers involved in the data they use—understanding the history of a project will add value to the research. The relevant acknowledgement data is available from the metadata files attached to the data.

If using data from Project 599 the National Reference Stations or Project 597 the Australian Continuous Plankton Recorder, please use the following acknowledgement:
'The data sourced from the Integrated Marine Observing System—IMOS is a national collaborative research infrastructure, supported by the Australian Government.'

Similarly for Projects 806 and 1066, please use the following acknowledgement:
'A part of the data included in the database was obtained with support from the Great Barrier Reef Marine Park Authority, through funding from the Australian Government Reef Program and from the Australian Institute of Marine Science'.

Author Contributions

Additional Information
Table 1 is only available in the online version of this paper.

Competing financial interests: The authors declare no competing financial interests.


This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0

Metadata associated with this Data Descriptor is available at http://www.nature.com/sdata/ and is released under the CC0 waiver to maximize reuse.
Corrigendum: A database of marine phytoplankton abundance, biomass and species composition in Australian waters


Scientific Data 3:160043 doi:10.1038/sdata.2016.43 (2016); Published 21 Jun 2016; Updated 6 Dec 2016

A series of errors in our database were brought to our attention by readers, and have been corrected in an updated version of this database, which is accessible via the AODN at the following link: https://portal.aodn.org.au/search?uuid=75f4f1fc-bee3-4498-ab71-aa1ab29ab2c0.

The custodian details of several datasets were incorrect. These fields in the metadata table have been updated to correctly assign P744, P746, P748, and P778 to the Australian Antarctic Division, and P752 to the Royal Belgian Institute of Natural Sciences.

Species names and functional group assignments have been changed for a small number of records to fix identified errors. *Tripos brevis* and *Tripos arietinus* were spelt incorrectly, and have been duly corrected. Pedinellaceae was wrongly assigned to dinoflagellate as a functional group, and has now been re-assigned to flagellate. The 'Naked flagellate' group has been renamed 'Flagellate' as there is some inconsistency in the use of the term 'Naked flagellate' as well as the precise term that would be included. The functional group 'Other', has also been excluded as this contained data that was not necessarily phytoplankton but had been found in phytoplankton counts. The macroalgae *Murrayella australica, Cladophora* spp., *Chlorohormidium* sp., *Eudorina* spp., *Tribonema* spp., *Chlorohormidium* spp. were also removed.

In addition to these corrections, three datasets have been extended to include more recently acquired data: P 597 IMOS Australian Continuous Plankton Recorder survey (ongoing dataset, 59089 new records as of 2016-08-31); P599 IMOS National Reference Stations (ongoing dataset, 14669 new records as of 2016-08-31); and P1068 Great Barrier Reef Expedition 1928-29 (new dataset, 1340 new records). Table 1 provides a summary of the overall change in database contents.
This dataset will continue to grow and will be regularly updated with new data and any further corrections to the data. Users can email imos-plankton@csiro.au with any comments, which will be reviewed and included in future updates if applicable. The AODN portal will always direct the user to the most recent version, the original version will remain available at http://dx.doi.org/10.4225/69/56454b2ba2f79, and interim versions will be available on request.

This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0

© The Author(s) 2016

<table>
<thead>
<tr>
<th></th>
<th>Original</th>
<th>Version 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total records</td>
<td>3621847</td>
<td>3724221</td>
</tr>
<tr>
<td>Presence/Abundance records</td>
<td>222617</td>
<td>224262</td>
</tr>
<tr>
<td>Species level</td>
<td>123931</td>
<td>125959</td>
</tr>
<tr>
<td>Genera level</td>
<td>78153</td>
<td>79921</td>
</tr>
<tr>
<td>Higher level</td>
<td>20333</td>
<td>19382</td>
</tr>
<tr>
<td>Absence records</td>
<td>3399230</td>
<td>3499959</td>
</tr>
<tr>
<td>Total records</td>
<td>3621847</td>
<td>3724221</td>
</tr>
<tr>
<td>Taxa at species level</td>
<td>1499</td>
<td>1536</td>
</tr>
<tr>
<td>Taxa at genera level</td>
<td>522</td>
<td>512</td>
</tr>
<tr>
<td>Higher level taxa</td>
<td>107</td>
<td>89</td>
</tr>
</tbody>
</table>

Table 1. Summary of database contents.
Corrigendum: A database of marine phytoplankton abundance, biomass and species composition in Australian waters


Scientific Data 3:160043 doi:10.1038/sdata.2016.43 (2016); Published 21 Jun 2016; Updated 11 April 2017

The authors regret that Sarah A. Pausina was omitted in error from the author list of the original version of this Data Descriptor. This omission has now been corrected in the HTML and PDF versions of this Data Descriptor, as well as the accompanying Corrigendum.

This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0

© The Author(s) 2017