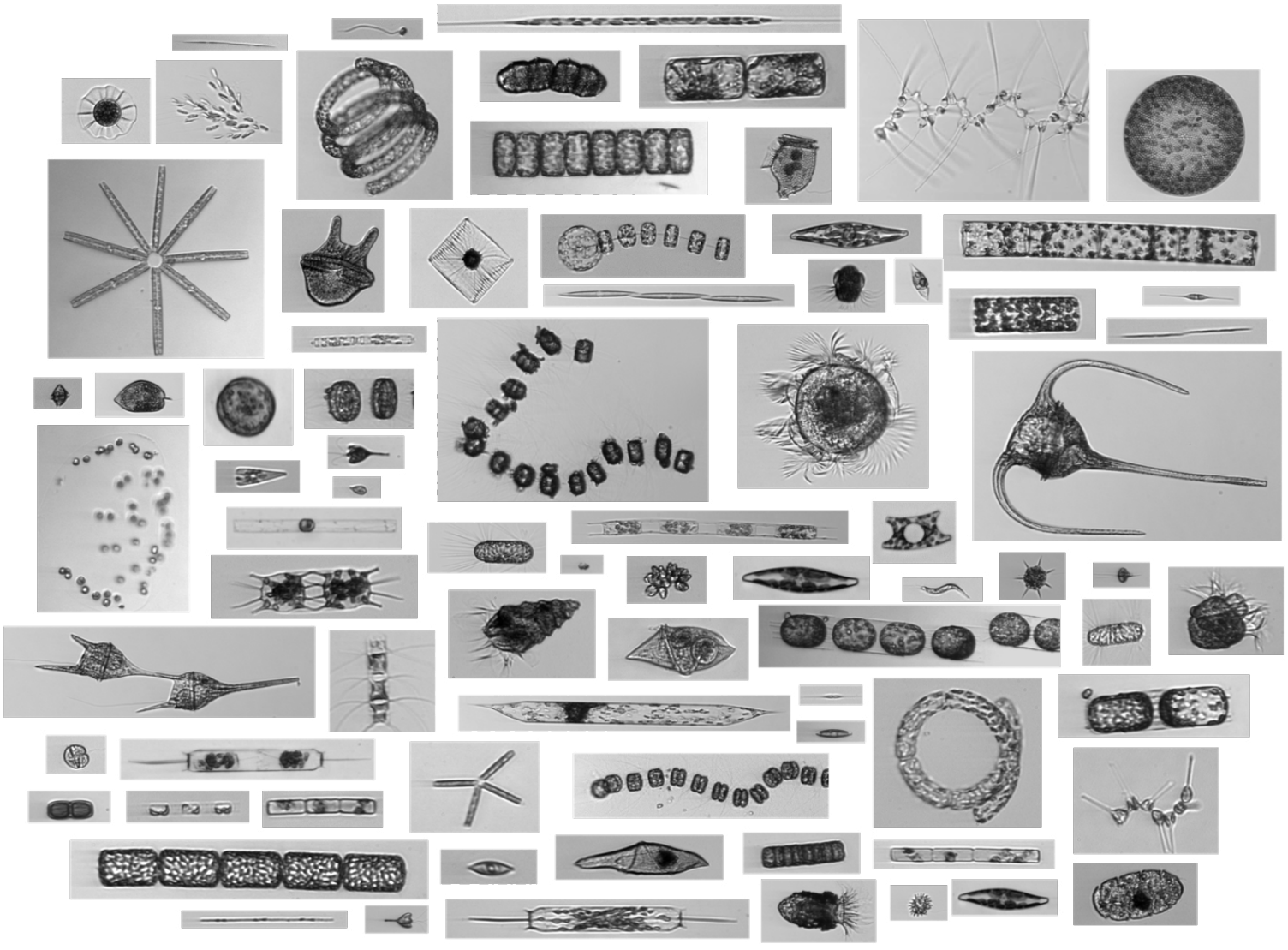
Standards and practices for reporting plankton and other particle observations from images
Technical Manual



AUTHORS

Aimee Neeley
NASA Goddard Space Flight Center/Science Systems and Applications Inc.

Stace Beaulieu
Woods Hole Oceanographic Institution

Chris Proctor
NASA Goddard Space Flight Center/Science Systems and Applications Inc.

Ivona Cetinić
NASA Goddard Space Flight Center/Universities Space Research Association

Joe Futrelle
Woods Hole Oceanographic Institution

Inia Soto Ramos
NASA Goddard Space Flight Center/Universities Space Research Association

Heidi Sosik
Woods Hole Oceanographic Institution

Emmanuel Devred
Fisheries and Oceans Canada

Lee Karp-Boss
University of Maine

Marc Picheral
French National Centre for Scientific Research

Nicole Poulton
Bigelow Laboratory for Ocean Sciences

Collin Roesler
Bowdoin College

Adam Shepherd
Woods Hole Oceanographic Institution

EDITORS

Heather Benway
Woods Hole Oceanographic Institution

Mai Maheigan
Wood Hole Oceanographic Institution

BIBLIOGRAPHIC CITATION:

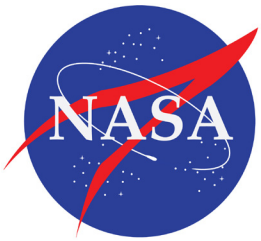
Neeley, A., Beaulieu, S., Proctor, C., Cetinić, I., Futrelle, J., Soto Ramos, I., Sosik, H., Devred, E., Karp-Boss, L., Picheral, M., Poulton, N., Roesler, C., and Shepherd, A.. 2021: Standards and practices for reporting plankton and other particle observations from images. 38pp. DOI: 10.1575/1912/27377.

ACKNOWLEDGMENTS

This report was an outcome of a working group supported by the Ocean Carbon and Biogeochemistry (OCB) project office, which is funded by the US National Science Foundation (OCE1558412) and the National Aeronautics and Space Administration (NNX17AB17G). AN, SB, and CP conceived and drafted the document. IC, IST, JF and HS contributed to the main body of the document as well as the example files. All members of the working group contributed to the content of the document, including the conceptualization of the data table and metadata format. We would also like thank the external reviewers Cecile Rousseaux (NASA GSFC), Susanne Menden-Deuer (URI) Frank Muller-Karger (USF), and Abigail Benson (USGS) for their valuable

COVER IMAGE

Phytoplankton FlowCytobot image collage by Heidi Sosik, WHOI.



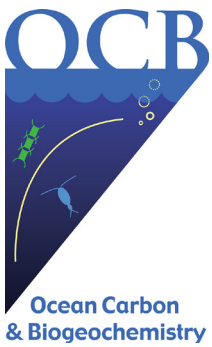


Table of Contents

Abstract 1

1. Introduction..... 2

2. Creating a data file from image data..... 5

2.1 Required provenance for plankton and particle observations from images..... 6

Figure 1. IFCB workflow overview chart..... 7

2.2 Steps to standardizing the data table..... 8

2.2.1 General workflow to create a taxonomic lookup table..... 9

2.2.2 Detailed workflow to create a taxonomic lookup table..... 10

Figure 2. Workflow to create a lookup table..... 11

2.3 Metadata headers specific to plankton and other particle data..... 12

2.4 The data table..... 13

2.4.1 Imperfect matches between data provider category and taxonomy.. 15

2.4.2 Non-conforming ROIs..... 15

2.4.3 Defining non-conforming ROIs..... 16

2.4.4 Optional supplemental definitions of non-conforming ROIs..... 16

2.4.5 Submitting data to OBIS..... 16

3. Creating and submitting a file to SeaBASS..... 18

3.1 The file format..... 18

3.2 Submission of images..... 19

3.3 Required documentation for data submission..... 19

3.3.1 Protocol documentation..... 20
3.3.2 The checklist..... 20

4. Appendix A: The YAML File..... 22

5. Appendix B: SeaBASS Fieldnames..... 23

6. Appendix C: Single Sample Data File Example..... 25

7. Appendix D: Example Protocol Document..... 28

8. Appendix E: Example Checklist..... 31

9. References..... 32

Abstract

This technical manual guides the user through the process of creating a data table for the submission of taxonomic and morphological information for plankton and other particles from images to a repository. Guidance is provided to produce documentation that should accompany the submission of plankton and other particle data to a repository, describes data collection and processing techniques, and outlines the creation of a data file. Field names include `scientificName` that represents the lowest level taxonomic classification (e.g., genus if not certain of species, family if not certain of genus) and `scientificNameID`, the unique identifier from a reference database such as the World Register of Marine Species or AlgaeBase. The data table described here includes the field names `associatedMedia`, `scientificName`/ `scientificNameID` for both automated and manual identification, `biovolume`, `area_cross_section`, `length_representation` and `width_representation`. Additional steps that instruct the user on how to format their data for a submission to the Ocean Biodiversity Information System (OBIS) are also included. Examples of documentation and data files are provided for the user to follow. The documentation requirements and data table format are approved by both NASA's SeaWiFS Bio-optical Archive and Storage System (SeaBASS) and the National Science Foundation's Biological and Chemical Oceanography Data Management Office (BCO-DMO).

1 Introduction

Over the last 10 years, the number of satellite algorithms for deriving phytoplankton community composition (PCC) and size classes (PSCs) has grown exponentially (e.g., Mouw et al. 2017; Sathyendranath 2014), and these parameters have been used for various applications, from assessing climate change impacts on marine ecosystems to understanding the mechanisms that regulate global biogeochemical cycles (e.g., Mouw et al. 2016; Le Quere et al. 2005). Several global climate models, including the National Aeronautics and Space Administration's (NASA) Ocean Biogeochemical Model and the Darwin Project hosted by Massachusetts Institute of Technology, have been developed to better understand and predict phytoplankton community composition and community dynamics (Dutkiewicz 2020). These models include multiple phytoplankton types and inherent optical properties, thus increasing the demand for coincident measurements of PCC and optical properties, which can then be used to parametrize and validate the models. Plankton diversity and abundance have been shown to be sensitive to climate variability (Behrenfeld 2014; Gobler 2020). Moreover, the identification and enumeration of nonliving particles, such as detritus and fecal pellets, are important for estimating carbon export to the ocean interior (Siegel et al. 2016; Boyd et al. 2019). Marine ecological time series that include Essential Ocean Variables (EOVs, Global Ocean Observing System (GOOS)) and Essential Biodiversity Variables (EBVs, Group on Earth Observations Biodiversity Observation Network (GEO BON), Muller-Karger et al. 2018) are critical to understanding large-scale environmental and ecosystem variability over longer time scales (Boss et al. 2020). In this document, we provide recommendations for EBV-usable and EBV-ready data sets (defined by Kissling et al. 2018) to align with Findability, Accessibility, Interoperability and Reusability (FAIR) Data Principles for data management (Wilkinson et al. 2016).

New algorithms, models, and other applications increasingly require more detailed information about phytoplankton community and particle composition. For example, NASA's upcoming ocean color mission Plankton, Aerosols, Clouds, ocean Ecosystem (PACE) will collect hyperspectral ocean color data that will provide the capability to resolve different spectral signatures of phytoplankton (Werdell et al. 2019). Advances in the field require greater availability of phytoplankton taxonomic information across multiple taxonomic levels.

Concurrent with increasing demand for this kind of information, there have been technological advances in phytoplankton detection, from microscopy to conventional flow cytometry and, most recently, automated imaging-in-flow cytometry, such as the Imaging FlowCytobot (IFCB; Sosik and Olson 2007). With these new observational tools producing datasets of potentially high spatial, temporal, taxonomic and morphological resolution, it is imperative that we develop adaptable informatics solutions to ensure that these data sets continue to serve the evolving needs of a broad range of users. As such, there is a critical need for high-quality ground-truth data sets to aid in the development and evaluation of satellite-derived ocean color products, validation and parameterization of ecosystem and global models, and hypothesis-driven research.

Agencies that fund basic and applied oceanographic and limnological research, including but not limited to NASA, the National Science Foundation (NSF), the National Oceanic and Atmospheric Administration (NOAA), and the European Organization for the Exploitation of Meteorological Satellites (EUMETSAT), require repositories of high-quality *in situ* environmental, optical, and phytoplankton properties for science and modeling studies, algorithm development, and product validation. NASA's SeaWiFS Bio-optical Archive and Storage System (SeaBASS) and NSF's Biological and Chemical Oceanography Data Management Office (BCO-DMO) are leading examples of such community resources. These publicly available data repositories are routinely used in hypothesis-driven research, as well as for model parameterization and validation. To date, phytoplankton properties archived in SeaBASS have been limited to pigment concentrations, absorption coefficients, growth rates and grazer mortality. A small but growing number of datasets containing phytoplankton information can be found in BCO-DMO, including some that provide taxonomic information (example).

Access to high-resolution plankton and other particle data informs future needs in algorithm development, validation, and calibration efforts. The objectives of individual studies will determine what parameters (e.g., taxonomic-level cell or particle concentrations, biovolume, cellular carbon) are needed for development and validation of models and algorithms. To facilitate access to such *in situ* plankton and particle data products, the Ocean Carbon & Biogeochemistry (OCB) Phytoplankton Taxonomy Working Group (PTWG) brought phytoplankton ecologists, taxonomists, and algorithm developers together with data and informatics specialists, including members of the BCO-DMO and SeaBASS teams, to establish best practices for providing taxon and morphologically resolved plankton and other particle observations using international data standards. Two meetings of the Working Group were held at the Woods Hole Oceanographic Institution (WHOI) in Woods Hole, Massachusetts over two days each in June 2017 and 2018. The group discussed the key needs (e.g., standardized data reporting and terminology) and challenges (large and complex data sets and metadata) associated with the provision of taxon-resolved plankton data (e.g., cell counts, biomass, size distributions, etc.). Working Group members formulated provenance requirements to enable users to trace data products back to raw data, including documentation of data processing steps. By developing a set of common practices around provenance for plankton and particle observations, we enable users of the data to (1) make informed decisions about which products can be integrated or compared across datasets, instruments, etc., and (2) reproduce or reprocess products to standardize them across datasets or instruments in such a way that products can also be updated if processing approaches improve or become more standardized. This document takes into account the many needs and challenges identified by this working group. The objectives of this document are to:

- Prescribe a standardized data table format for *in situ* data contributors that addresses the need for machine-readable taxonomic and morphological information from plankton and other particles, including ingestion of essential provenance information for reusability
- Develop vocabulary important for interoperability and reusability of data
- Explore options for how a primary repository could:
    - ◊ Store and serve the contributed data and metadata
    - ◊ Produce, store, and serve higher-level data products

The following chapters give a detailed description of the necessary procedures and data submis-

sion requirements. Chapter 2 describes the reporting and documentation standards for the data processing steps, including metadata headers, field (data table column) names and formatting the data table. Chapter 3 provides the instructions to create a data file for submission to NASA's SeaBASS repository from imaging-in-flow cytometric methods. Incoming submissions to BCO-DMO will be profiled by the data managers to determine if the files are compliant to the data table defined in this document. The BCO-DMO data managers will work with submitters to ensure that data are published and comply with the NSF Ocean Sciences Sample and Data Policy.

To illustrate the requirements for the data files containing imaging-in-flow cytometry and submersible microscope observations to a public repository, we provide examples of a data file and documentation. The examples are based on data files published by the NSF's Environmental Data Initiative (EDI) repository to show that these guidelines would be applicable for a data submission to other repositories (Sosik et al. 2020). Additionally, this protocol may be applied to image data collected by various imaging-in-flow cytometric and submersible microscope instruments (e.g., Walcutt et al. 2020, Menden-Deuer et al. 2020).

The standards and practices defined in this document for reporting plankton and other particle observations from imaging-in-flow cytometric methods build a foundation for other taxon and morphology resolving methods, such as classical microscopy and standard flow cytometry. However, the data resolved by the aforementioned techniques do not completely conform to the data file structure described here. Therefore, separate, additional instructions will follow this document to address these data types: "*Data Standards and Practices for plankton observations from classical microscopy*" as well as "*Data Standards and Practices for plankton observations from standard flow cytometry*".

2 Creating a data file from image data

It is essential that a data file submitted to any repository include documentation that provides detailed information associated with image collection (e.g., instrument settings and water sample collection method), and image processing methods (e.g., manual and automated classification methods, biovolume computation method). The expected data file prepared for submission to a repository must, at minimum, specify individual level counts with automatic and/or manual classifications, biovolume and size parameters for the target (particle or living organism) in each region of interest (ROI; a rectangular subset of pixels), which is considered Level 1b in the sample processing scheme (Figure 1). To this end, the Working Group has developed a data file format that fulfills Level 1b data requirements of known repositories for ROIs in which targets have been classified by an automatic and/or manual classifier. While this document focuses on submissions to SeaBASS, the structure of the data table and metadata headers developed during this activity for plankton and other particle data should apply to any repository.

In the following subchapters, we define the process leading to the development of such a file. Subchapter 2.1 lists the specific requirements to be met for Level 1b ROI data sets and describes the importance of preserving and thoroughly documenting the methods of sample collection and image processing, or provenance, for the observations of plankton and other particles. Subchapter 2.2 details the process set in place to standardize the taxonomic information and morphological characteristics (e.g., size, shape) into a data table. Next, the framework of a complete data file is defined and includes: (1) 'header fields', a series of metadata lines placed before the data table to provide critical information about the data such as 'who, what, when, where, how' the samples and data were collected and processed (Subchapter 2.3); (2) 'field names' that define the columns of the data table, and (3) formatting of the data table (Subchapter 2.4). Although the structure of the data table can apply to any repository, the format of the header section may vary between repositories. In this document, we provide an example data file that meets the requirements of NASA's SeaBASS repository; however, the structure of the 'field names' or data column headers provides a path to submitting Level 1b plankton and particle imaging data to other repositories. For reusability, it is essential that all information described below be included with each data file. Lastly, we describe the protocol documentation and checklists that are required to accompany a data set, using a submission to SeaBASS as an example. The protocol documentation must provide a detailed account of sample collection methodology and data post-processing procedures.

2.1 Required provenance for plankton and particle observations from images

In order to ensure transparency and facilitate postprocessing, each step of image collection and processing (Figure 1) must be diligently recorded and included in the data reporting. In this workflow diagram, we indicate the activities (ovals) and some entities (rectangles) that must be described to enable reusability of data products derived from plankton and other particle images. Metadata for these activities and entities represent essential provenance (i.e., to describe the derivation of a data product from its raw data source). Although Figure 1 is specific to the Imaging FlowCytobot (IFCB), a similar flow chart could be constructed for other instruments (e.g., FlowCAM, 4-Deep, CytoBuoy). Briefly, documentation is required for any data set and must include a detailed description of the following (from Figure 1):

- Instrument settings for data acquisition (e.g., trigger thresholds and image resolution)
- Image processing methods and versions (e.g., software package information, method of computing particle biovolume, size features extracted)
- Manual annotation and automated classification methods (e.g., machine learning software information and its version number)
- Taxonomic or morphological assignment (e.g., selection of "winning" category and matching to machine-readable identifiers provided by the World Register of Marine Species (WoRMS) or AlgaeBase taxonomic databases)
- Grouping individual classified particles into higher-level groups (e.g., diatoms, dinoflagellates) or size classes (micro, nano and picoplankton); these data are considered Level 2. (Level 2 data will not be discussed in this document.)

The data processing levels are described as follows:

Level 0: Raw images collected by the imaging-in-flow cytometer

Level 1a: Automated classification by an algorithm (automated annotation) and/or manual annotation

Level 1b: Individual level counts with automatic (including interpretation of class scores or probabilities) and manual classifications, and biovolume and size parameters for each ROI

Level 2: Summary data for sample e.g., taxonomic groupings

Each step of data reporting will be covered in detail in the following chapters and subchapters. Detailed information should be included within an accompanying protocol document, with specific items also provided as headers in each data file.

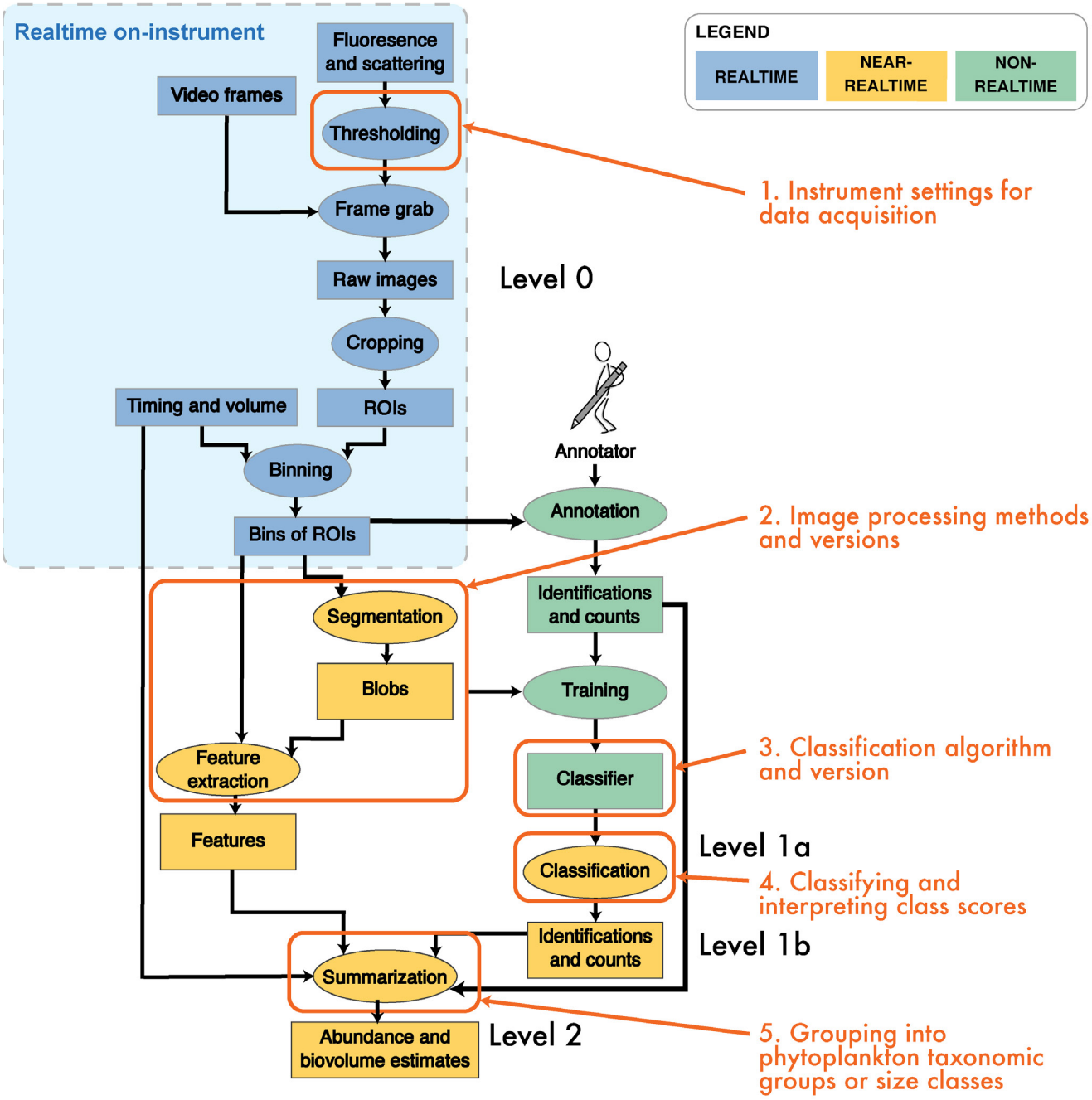


Figure 1. IFCB workflow overview chart. Arrows in the direction of the workflow, starting with raw images at the top and leading to data products at the bottom. Activities and entities are represented by ovals and rectangles, respectively. For more information and access to code, please visit https://github.com/hsosik/ifcb-analysis. (diagram credit: H. M. Sosik and J. Futrelle)

2.2 Steps to standardizing the data table

To standardize plankton taxonomic and morphological information in the data file, the data provider must create a lookup table that matches their morphological categories to the lowest level, with the accepted scientific names paired to identifiers in an authoritative taxonomic database. This lookup table is essential to provide machine-readable taxonomic information for each ROI. To provide machine-readable taxonomic identifiers for living organisms, we recommend WoRMS, as it is the official taxonomic reference list for the Ocean Biodiversity Information System (OBIS). OBIS requires the Darwin Core international standard terms of the full scientific name (scientificName) that provides the lowest taxonomic rank that can be identified, whether it be at the species, genus, class, or a higher rank. An external, machine-readable and resolvable identifier, or object number, that returns nomenclatural (not taxonomic) details of a name (scientificNameID) should also be included for each ROI. OBIS can harvest a `scientificNameID` containing a Life Science Identifier (LSID), a persistent globally unique identifier for biological objects in a database, from WoRMS. The use of LSIDs from a referenced database permits machine-readable taxonomic identifications. An LSID used in a taxonomic database consists of a uniform resource name (urn) that contains the following information in order: network identifier, a root DNS name of the reference database, a namespace, and the object number or taxon identifier (a living product) that is unique to the biological object defined by that referenced database. The namespace defines the type of information provided by the object number. As such, the status of taxonomic identifiers can change (e.g., from accepted to unaccepted) when new information becomes available. It would be untenable for a data repository manager to routinely modify archived data files with revised/updated taxonomic names. Therefore, a traceable ID linked back to a taxonomic reference database that is regularly updated is necessary.

An example of a `scientificNameID` would be:

In this example, 'urn:lsid' indicates the ID that is specific to life science data and is used for all files, marinespecies.org is the url for the reference database WoRMS, and the namespace 'taxname' informs the user that the following number represents a unique numerical identifier or taxon identifier in WoRMS. In the above example, 233015 represents the taxon identifier (AphiaID) in WoRMS for the dinoflagellate species *Karenia brevis*. WoRMS provides a continuously updated and comprehensive list of marine organisms, with species names and synonyms and information regarding higher classification and parent taxon for each organism.

AlgaeBase is another example of a reference database that provides unique LSIDs. WoRMS currently integrates Aphia with taxonomic information from AlgaeBase. A different example LSID using AlgaeBase would be:

where www.algaebase.org is the URL for the database and 86701 is the unique identifier for Eukaryota. LSIDs should be included in the data table and associated with each ROI. The use of referenced and traceable taxonomic identification for living organisms will be required for data submission to public repositories, such as SeaBASS and BCO-DMO.

2.2.1 General workflow to create a taxonomic lookup table

A taxonomic lookup table is essential to ensure the accurate pairing of data provider categories, the categories used by the data provider to name the organism or particle for an automated classification (not necessarily a scientific name, e.g., pennate or detritus), to their `scientificName` and `scientificNameID`. In this section, we describe the steps necessary to create a lookup table. Starting with the data provider's categories for automated and/or manual classifications, the `scientificName`/`scientificNameID` pairs can be determined manually by searching WoRMS or automatically using web services with a script or with the WoRMS Taxon Match Graphical User Interface (GUI).

When `scientificName`/`scientificNameID` pairs have been determined manually, we recommend confirming that each `scientificName`/`scientificNameID` pair is accepted in WoRMS either using the GUI or by using an automated workflow in a script. When using web services to determine the `scientificName`/`scientificNameID` pairs, some manual cleanup may be required to ensure the correct `scientificName`/`scientificNameID` pairs are provided. Using web services can also correct a misspelled `scientificName` and retrieve hierarchical ranks. There may be instances when different automated scripts provide contrasting results. For example, the R package 'worrms' can yield different results than the R package 'taxize' (either of which may be correct depending on the case). Moreover, automated services will match names exactly, and there may be cases in which one wants to provide a `scientificName`/`scientificNameID` pair at a higher taxonomic rank. It is preferable to generate `scientificName`/`scientificNameID` pairs for all the names being considered for either the automated classifier ('automated') or manual annotations ('manual'). Providing a list of all `scientificName`/`scientificNameID` pairs assessed by the automated classifier with the data submission enables the determination of both presence and absence of annotations in the Level 1b file. This topic will be explored further in Subchapter 2.3. Next, detailed instructions to create a lookup table are provided (Figure 2).

2.2.2 Detailed workflow to create a taxonomic lookup table

1. The user must begin with a list of unique names or data provider categories (and IDs if determined manually). The IDs could be the AphiaID or the full scientificNameID.
2. Decide whether to try an automated script such as WoRMS_verify.Rmd or taxonomyCleanr; otherwise, skip ahead to step 5 to start with WoRMS Taxon Match GUI.
3. Script: If IDs are available, generate a scientificName from the IDs provided. The user must perform a Boolean check (true or false logic) of the provided name with the generated scientificName determined from the ID. The Boolean check will need to be a fuzzy match, not a strict match, owing to possible misspellings, lowercase letters, etc. If all returns are true, then check that the scientificName/scientificNameID pairs are accepted. Replace unaccepted scientificName/scientificNameID pairs with the accepted pairs and skip to step 5.
4. Script continued: If IDs are not available, generate a `scientificName` from the provided name (i.e., resolve provided names), then generate the ID from the resolved `scientificName`. It is important that the script also generates the classification (or at least a higher rank, such as Phylum) for the resolved `scientificName` to account for cases when the same name is in very different places on the tree of life. Inspect for appropriate higher rank, then check if the `scientificName`/`scientificNameID` pairs are accepted. Replace unaccepted `scientificName`/`scientificNameID` pairs with the accepted pairs.
5. Regardless of whether a script was used, the user should run the names as provided through the WoRMS Taxon Match GUI as recommended by the Marine Biodiversity Observation Network (MBON). The user should select the LSID, classification, and 'taxon status' to be included in the output. 'Taxon status' shows whether the `scientificNameIDs` are accepted by the authority. In some cases, the GUI output is ambiguous, prompting the user to select from a list (this step may require manually browsing WoRMS to assist in the selection of the correct `scientificNameID`). Occasionally the output from the GUI will provide a better `scientificName`/`scientificNameID` pair than an automated script. For example, we found that the R package 'taxize' could not accommodate brackish phytoplankton taxa.
6. A cleaning script with hard coding is usually necessary whether starting with an automated script or the GUI. For names that are possible to match with web services, the GUI has its own issues, e.g., it cannot output the taxon status "accepted, alternate representation." Some names cannot be matched with web services, e.g., a category representing morphologically similar taxa on different branches of the tree of life (e.g., pennate diatom). Additionally, some names represent objects that are not organisms (e.g., detritus) and, therefore, cannot be matched with web services.
7. A cleaning script with hard coding is also important to accommodate case-by-case decisions whether to standardize the machine-readable `scientificName`/`scientificID` pair to a higher rank. A best practice is to retain the original provided name (`data_provider_category`) in the data table along with the higher rank accepted `scientificName`/`scientificID` pair.
8. To confirm the accuracy of the LSIDs, manually paste the LSIDs into an online LSID

Resolver. As of April 2020, WoRMS LSIDs resolve to human-readable HTTP webpages but Algaebase LSIDs do not.

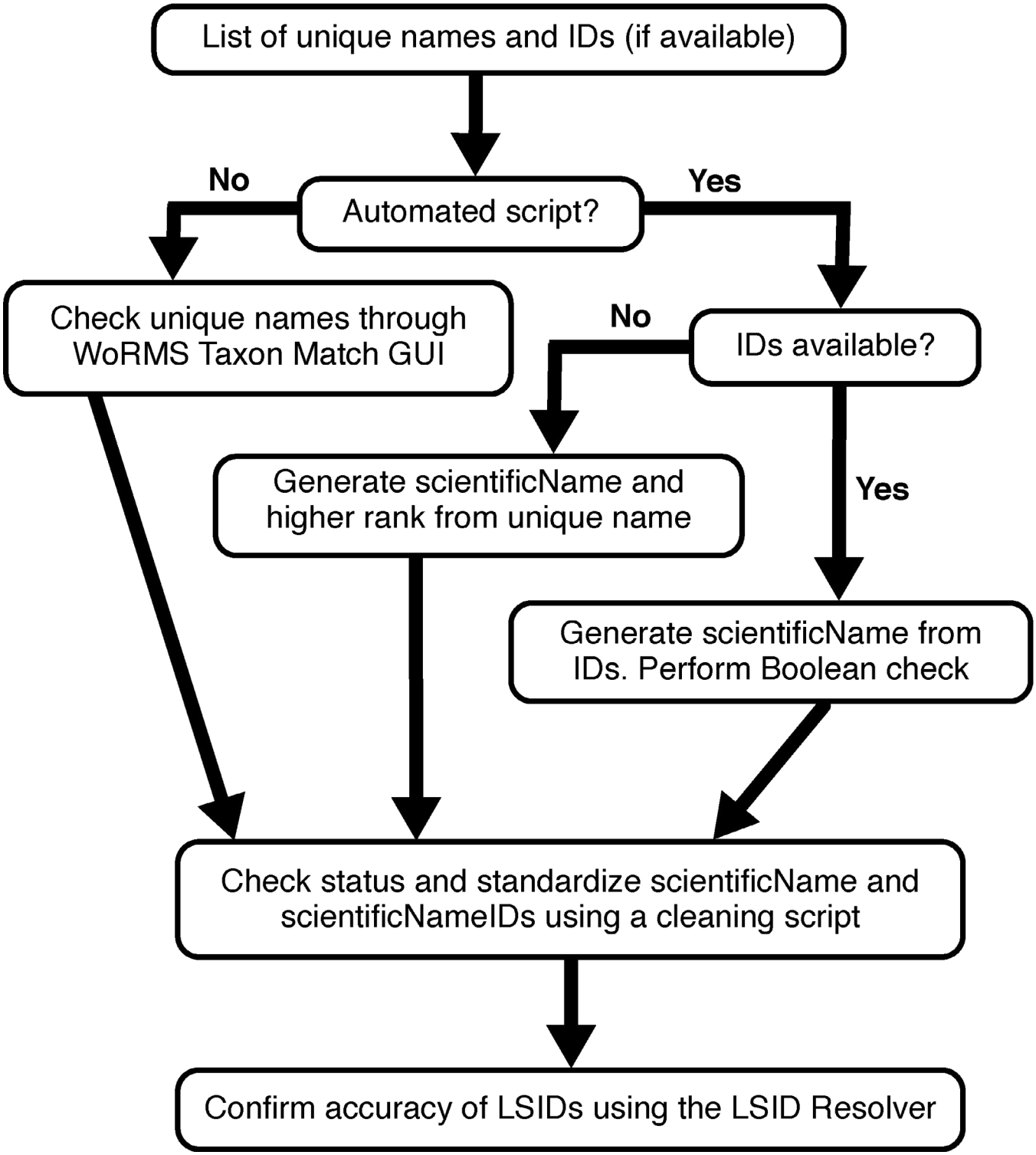


Figure 2. Workflow to create a lookup table.

2.3 Metadata headers specific to plankton and other particle data

Standard metadata headers that include data column labels, spatial and temporal information, and project information are required for data submission into any repository. For example, SeaBASS has a suite of required and optional self-describing metadata headers that consist of case insensitive keyword-value pairs using the form `/keyword=value`. Such standard metadata headers include information that identifies the institution of the data collector and the location at which a sample was collected (latitude, longitude, time, depth, etc.; Chapter 3). BCO-DMO and the EDI repository for NSF-funded Long-Term Ecological Research (LTER), also require submission of metadata forms that accompany a data submission. Through the activities of the Working Group, a new set of metadata headers were created specific to plankton and other particle image data that will be the focus of this subchapter. Note that additional requirements for SeaBASS can be found online or modeled from the data file examples found in Appendix C.

Each instrument and accompanying software package may use different terminological representations for particle size dimensions (i.e., length and width). For example, IFCB provides 'feret diameter' while the FlowCAM provides 'length' and 'width' for each ROI. Therefore, we created the following headers that accept different variable names for size dimensions:

1. `length_representation_instrument_varname`:
   (required, if applicable) the instrument's variable name equivalent to '`length_representation`' (e.g., `maxFeretDiameter`). Replace any spaces with underscores.
2. `width_representation_instrument_varname`:
   (required, if applicable) the instrument's variable name equivalent to 'width_representation' (e.g., `minFeretDiameter`). Replace any spaces with underscores.

Additional metadata headers that contain plankton and other particle information are discussed below and must be included in the data files:

1. `volume_sampled_ml`:
   (required) original volume of sample collected in units of milliliters (ml)
2. `volume_imaged_ml`:
   (required) subset of `volume_sampled_ml` that was imaged in units of milliliters (ml)
3. `pixel_per_um`:
   (required) number of pixels per unit length in units of micrometers (um; ASCII SeaBASS files use the letter u instead of the Greek letter μ)
4. `associatedMedia_source`:
   (optional, and may not be applicable) a unique persistent URL pointing to the landing page for a water sample from which multiple ROIs are derived
5. `eventID`:
   (required) a unique identifier associated with the sample as an event

See Appendix C for examples of each header.

2.4 The data table

The data table includes detailed information for each ROI collected with an imaging-in-flow cytometer or submersible microscope instrument. Each row of the Level 1b data table corresponds to an individual ROI. The following column headers or 'fields' and associated units are the required standard field names for submitting imagery data to SeaBASS (at the end of this subchapter we will include suggestions to enable harvesting by OBIS). The following is a list of field names (i.e., measurement labels) that align with Darwin Core when possible. See Appendix B for the units for each data column and Appendix C for an example data file:

1. `associatedMedia`:
   a unique persistent identifier of the media associated with the occurrence. The field provides the unique imagery file name corresponding to the source of the ROI. Alternately, use this field to provide a URL pointing to a permanent landing page for the ROI image. In the latter case, instructions should be provided as comments in the header on how to construct the local file name based on the URL. If the local imagery file name cannot be constructed from the URL, then list both filenames. Use the pipe character '|' to separate the names, and do not use spaces.
2. `data_provider_category_automated`:
   (recommended but optional) A category used by the data provider to name the organism or particle for an automated classification, not necessarily a scientific name (e.g., pennate or detritus).
3. `scientificName_automated`:
   (recommended but optional) A scientific name from a recognized taxonomic reference database (e.g., WoRMS, AlgaeBase) at the lowest level that matches the data provider's category for an automated classification paired to a `scientificNameID`. Generally, the ROI corresponds to an occurrence assigned to a single taxonomic name.
4. `scientificNameID_automated`:
   An LSID from a recognized taxonomic reference database (e.g., WoRMS, AlgaeBase) at the lowest level that matches the data provider's category for an automated classification.
5. `data_provider_category_manual`:
   (recommended but optional) A category used by the data provider to name the organism or particle for a manual identification, not necessarily a scientific name.
6. `scientificName_manual`:
   (recommended but optional) A scientific name from a recognized taxonomic reference database (e.g., World Register of Marine Species, AlgaeBase) at the lowest level that matches the data provider's category, for a manual identification matched to `scientificNameID`. Generally, the ROI corresponds to an occurrence assigned to a single taxonomic name.
7. `scientificNameID_manual`:
   An LSID from a recognized taxonomic reference database (e.g., World Register of Marine Species, AlgaeBase) at the lowest level that matches the data provider's category for a manual identification.
8. `biovolume`:
   Biovolume for the target detected within the ROI determined by means specified in the biovolume calculation method or protocol document.

9. area_cross_section:
   Cross-sectional area of the target detected within the ROI determined by means specified in the image processing method or protocol document.
10. length_representation:
    Representation of length of the target detected within the ROI or largest mesh size for which the target could be retained, determined by means specified in the image processing method or protocol document.
11. width_representation:
    Representation of width of the target detected within the ROI or smallest mesh size through which the target could pass, determined by means specified in the image processing method or protocol document.
12. equivalent_spherical_diameter:
    (optional) Equivalent spherical diameter of the target detected within the ROI determined by means specified in the image processing method or protocol document.
13. area_based_diameter:
    (optional) Area-based diameter of the target detected within the ROI determined by means specified in the image processing method or protocol document.

The field name scientificName is included for human readability for the scientificNameID. scientificNameID is the lowest level, machine-readable classification in WoRMS that matches the data provider's category. For example, if the data provider's category is '*Katodinium* or *Torodinium*' then the scientificName and scientificNameID would be given for the higher-level classification, Dinophyceae and urn:lsid:marinespecies.org:taxname:19542, respectively, that classifies both organisms. If provided, genus and species names should be separated by an underscore, for example 'Karenia_brevis'. The fieldnames data_provider_category_automated, scientificName_automated, data_provider_category_manual, and scientificName_manual are recommended but optional for human readability. The only required field for the 'automated' and 'manual' identification is the scientificNameID. Note that the field names allow for the data submitter to include both automatic and manual classification on the same row for each ROI.

The fields of biovolume, area_cross_section, length_representation and width_representation, all representing a Darwin Core measurementType, are subjective to method and instrument type so the best equivalent representations of these units of measure should be reported using these fields. Refer to the instrument's analysis software documentation for these terms. Specify the terms in required metadata /length_representation_instrument_varname and /width_representation_instrument_varname and see SeaBASS's instructions to provide method of computing biovolume, area_cross_section, and representation of length and width of the target identified within each ROI. Optional fields include area_based_diameter and equivalent_spherical_diameter.

Supplementary lists of which taxonomic categories were assessed by manual and/or automatic classification methods are strongly recommended and are required as part of data submissions if not every ROI in a given datafile was classified. If every ROI was not classified, these lists are essential for the downstream creation of summary products involving the concentrations of phytoplankton taxa. When every ROI is classified, these lists are useful for determining absence.

These lists may be specific to a given water sample or datafile, e.g., if only diatoms are classified in a sample, or they may be comprehensive of every class in a classifier. Two lists are required for a given datafile: all LSIDs (defined in Subchapter 2.2) assessed for 'automated' and 'manual' (these terms are explained in more detail below.) Each of these files must be a simple plain-text file containing one `scientificNameID` per row. The file name must begin with `automated_assessed` or `manual_assessed`, in which automated refers to automatic classification while manual refers to manual annotation. To link a given SeaBASS datafile to its associated lists, include the header `associated_files` as its value, and provide the names of any associated files as a comma separated list (no spaces). For example: `/associated_files=Automated_assessed_id_D20180201T103729_IFCB102.txt,Manual_assessed_id_D20180201T103729_IFCB102.txt`. The same associated file names may be referenced in multiple data files, so it is only necessary to create additional files if different categories were assessed for different data files.

2.4.1 Imperfect matches between data provider category and taxonomy

There are certain cases when the `data_provider_category` can be assigned to a level in a standard taxonomy, but that assignment does not reflect the same amount of detail. This includes annotation categories that are mixtures of taxonomic groups, such as cells that are too small or otherwise difficult to identify or multiple taxa that are morphologically indistinguishable given the image resolution available. One typical example for IFCB observations is a category such as 'miscellaneous nanoplankton' which may be used in the `data_provider_category` field with Eukaryota/86701 for the `scientificName` and taxon identifier with the following `scientificNameID`:

urn:lsid:algaebase.org:taxname:86701

There may also be instances when the morphological characteristics of the target in the ROI suggest a higher classification than the genus or species. For example, the user is confident that the particle in the ROI represents a pennate diatom but cannot classify it to a lower taxonomic level. In this instance, the user can use 'pennate' in the `data_provider_category` and the `scientificName` and taxon identifier for Bacillariophyceae with the following `scientificNameID`:

urn:lsid:marinespecies.org:taxname:148899

Another common example is when a ROI contains more than one identifiable target or a single target with some qualifying characteristic. In this case, the `scientificName` and `scientificNameID` can be assigned based on the dominant target in the ROI, with the `data_provider_category` providing the additional detail with appended strings (e.g., Thalassiosira_TAG_external_detritus; Guindardia_delicatula_internal_parasite).

2.4.2 Non-conforming ROIs

For those particles that are not identifiable (e.g., blurry images or image artifacts) or are non-living, the `scientificNameID` field must be filled, despite taxonomic reference databases not necessarily having a relevant entry and the images must be submitted for completeness of the data set. Non-living particles, such as detritus or fecal pellets, may be imaged by an imaging-in-flow cytometer or submersible microscope instrument. These particles may contribute to the oceanic carbon flux and are still important to quantify. Additionally, images that contain calibration beads or air bubbles require their own categories. For reporting such particles or beads, a machine-readable method is described below that allows the reporting of several common particles using a list (PTWG

namespace) hosted on SeaBASS. Additional terms will be added to it as needed, but the general technique also theoretically allows each laboratory to create 'data provider category' names that are clearly defined in the accompanying documentation.

2.4.3 Defining non-conforming ROIs

To facilitate identification of non-conforming ROIs, custom definitions not found in a taxonomic authority must be provided in an external document file. A Phytoplankton Taxonomy Working Group ("PTWG") custom namespace was created to define several standardized names for common terms that are not currently defined by WoRMS or Algae Base. As of December 2020, this includes: 'bad_image', 'bead', 'bubble', 'detritus', 'fecal_pellet', and 'other'. The term 'other' should only be used to describe a non-living particle. If the data provider is confident that the ROI is a living particle but cannot be identified to a specific taxonomic rank, then it should be classified to the rank of Eukaryota or Prokaryota (see Subchapter 2.4.1 for an example). These IDs are paired with definitions and are stored in a YAML [YAML Ain't Markup Language (YAML) Version 1.2".- YAML.org Retrieved 2019-05-28] formatted file in order to serve as a machine-readable configuration file for anyone working with the data files. To use the terms, combine the 'prefix' of the namespace (i.e., ptwg) with a given ID in the `scientificNameID` column, for example 'ptwg:bead' or 'ptwg:detritus'. The relevant `scientificName` and (if present) the recommended `data_provider_category` columns should be filled with the ID value (e.g., 'detritus'). If a submission uses the PTWG namespace, download the file and include it as part of the submission documents (see Appendix A).

2.4.4 Optional supplemental definitions of non-conforming ROIs

Optionally, non-conforming ROIs defined in the "PTWG" namespace may be supplemented by more specific higher-level definitions. For example, 'MYNICKNAME:opaque_detritus' could be used as a `scientificNameID` to enhance 'ptwg:detritus'. Custom terms must be defined in an external plain-text namespace file using YAML-format with each term containing 'id', 'definition', and 'associated_terms' (See Appendix A). The namespace requires a 'prefix', e.g., MYNICKNAME, which should be a short name without spaces that is unique to the laboratory or dataset. It is recommended to include the tag 'uri' containing a unique URL for the external namespace. The file must also be named uniquely without spaces, and it must be included in the documentation that is bundled with the data submission, which for SeaBASS includes listing the file name in the metadata header called `/documents=`. An example custom namespace is included in Appendix A.

2.4.5 Submitting data to OBIS

Although this document is primarily focused on data submission to SeaBASS (and to some extent BCO-DMO), it was important to include linkages to other important, global open-access data repositories, such as OBIS. To submit the above-formatted data to OBIS, several more steps are needed because some information must be added, renamed, or reformatted compared to how it is stored in SeaBASS files:

1. Add a column for `occurrenceID`: if only identifying a single taxon per ROI, the `occurrenceID` can be equivalent to (or remove prefix/suffix from) the unique entry in `associatedMedia`.
2. Add a column for `occurrenceStatus`, a statement that defines the presence or absence of a particular taxon in a sample.
3. Add or adapt columns for `eventDate`, `decimalLongitude`, and

`decimalLatitude` [Location information is also required by SeaBASS but uses different field names].

4. Add columns for `maximumDepthInMeters` and `minimumDepthInMeters` using depth from the header (the range can be your certainty in that depth).

5. Transform the table to long format with column `identifiedBy` to record automated or manual per row. The provider may consider specifying additional provenance in this column, for example by providing information about the auto-classification software or the name of the person for the manual annotation.

6. Add column `basisOfRecord` with the string `MachineObservation`. As of September 2020, OBIS developers are considering the use of Darwin Core terms `basisOfRecord` (with MachineObservation) and `identifiedBy` for data from plankton imaging systems.

7. In the long-format column for `scientificName`, replace underscores with spaces.

The above steps can yield an Occurrence Core submission to OBIS. In order to provide size information, an Event Core with Occurrence Extension and Extended Measurement or Fact Extension (three tables) should be generated. The submitter would provide an `eventID` unique to the sample (e.g., the bin number from IFCB Dashboard), and `measurementType`, `measurementValue`, and `measurementUnit` for `biovolume`, `area_cross_section`, `length_representation`, and `width_representation` need to be defined.

To complete a Darwin Core package for submission to OBIS, some metadata including a dataset title, license, description, contact, creators, and metadata providers must be specified. A 'dataset title' should include the following: "Taxonomic and size data for [what was imaged, e.g., phytoplankton and microzooplankton]" and "imaged with [instrument] [where] [when]". For 'description', the submitter must provide details on what, where, and when by referring to the required headers `experiment`, `cruise`, and `station` in the SeaBASS data file. Additionally, an identifier to the SeaBASS data package, such as the assigned Digital Object Identifier (DOI), must be included to ensure access to associated documentation (e.g., checklist and protocol). The contact would be the same as specified in the SeaBASS data file header; for creators and metadata providers, all contributors involved in the creation of the data and metadata must be listed.

3 Creating and submitting a file to SeaBASS

3.1 The file format

In Subchapter 2.3, we describe new metadata headers that were developed by this Working Group specifically for plankton and other particle data. In this section, we will describe additional, required metadata headers and instruct the data submitter on how a data file must be formatted specifically for submission to NASA's SeaBASS repository. The SeaBASS file format is a NASA Earth Science Data and Information Systems (ESDIS) approved standard. SeaBASS data files are flat, two-dimensional ASCII text files. They can be recognized by their ".sb" file extension and their internal structure. SeaBASS files begin with a series of self-describing metadata headers and are followed by a delimited data table, similar to a spreadsheet. In addition to the headers described in Subchapter 2.3, plankton and particle data files must always contain the following metadata headers: `start_date`, `start_time`, `end_date`, `end_time`, `north_latitude`, `south_latitude`, `east_longitude`, `west_longitude`; `measurement_depth` is also required unless it is provided as a field (data table column). Date-time information must be in UTC/GMT, and location information must be in units of decimal degrees using the World Geodetic System 1984 (WGS84). Latitude and longitude values must range between -90 to 90 and -180 to 180 degrees, respectively. If these values are identical for every data row then it is optional to include the equivalent information as columns in the data table (i.e. fields called: `date`, `time`, `lat`, `lon`, and `depth`). The date, time, and location headers must be included regardless. If the columns are included, then the headers are kept simply to summarize the relevant endpoints or extents found within the data table. Commas are reserved as the delimiter within the data table, so they are not otherwise allowed. Headers provide descriptive information (e.g., cruise name, date, missing values, fields and units, etc.) by means of machine-readable keyword-value pairs.

Many SeaBASS headers are required, as noted in the submission requirements. Certain rules apply to metadata headers (e.g., most header lines must begin with a forward slash "/"). Headers should not include any white space, so values containing multiple words are separated with underscores, not spaces. The only exception is for open-ended comment lines (beginning with "!"), which are allowed to contain spaces. All SeaBASS field names and units are standardized. A full explanation of the format, including the general rules and guidelines for creating files can be found on the SeaBASS website. An example of a Level 1b file type is provided with this document. See Appendix B for the full list of SeaBASS fields for imaging-in-flow cytometry data.

The data files are hosted in the SeaBASS archive and made publicly accessible via a variety of web-based search tools (Werdell et al. 2003). The SeaBASS website should be consulted for a complete list of format guidelines and submission instructions, but a few important points are listed here. Metadata headers containing appropriate experiment and cruise names - i.e. long-term project and deployment names, respectively - play an important role in cataloging data and should be selected thoughtfully. Submitters should consider if the submission is part of an existing project or time series and try to use consistent names. Otherwise, they should pick or suggest new unique names (ideally

25 characters or fewer). The experiment name is especially important for grouping data sets because it can link together multiple cruises and it becomes part of the assigned DOI. If users are preparing to submit a large amount of time series data, then it generally should be divided into several cruises to be manageably sized (e.g., by year).

File names must not contain spaces or special characters except for hyphens, underscores, and periods and must end in ".sb" suffix. File names must be unique within a submission, and ideally should be completely unique in SeaBASS. It is strongly recommended they are formed using descriptive patterns incorporating information or abbreviations of the measurement type, cruise name, date, depth or other information. For example: <EXPERIMENT>-<CRUISE>-<DATATYPE>_<YYYYMMDDHHMM_<R#>.sb, where R stands for the release number that is determined by the submitter. The Level 1b data file is a type of SeaBASS file with all the required components, headers, and fields specified above, with the data table containing 1 row per ROI. An example SeaBASS Level 1b file is provided in Appendix C and on the SeaBASS website.

3.2 Submission of images

Data submissions should include an organized directory containing the images and any relevant instrument metadata on which the Level 1b files are based. These should be provided even if a version of the annotated images is hosted at another repository such as EcoTaxa. SeaBASS will create a compressed tar file that is optionally available for these source files. If the submission is extremely large, then it should be split to create more reasonable sizes (e.g., by year if a long time series). The name or names of the highest-level directory must be provided in the metadata header called `/associated_files=`. If the images are hosted externally, then this localized directory may either contain the individual images, or alternately a more efficient data format (for example, sample-level endpoint files for IFCB data). If that header contains multiple values, then list them in a comma-separated format with no spaces. Check the SeaBASS plankton and particles page for further information and updates regarding image and metadata submissions.

3.3 Required documentation for data submission

Detailed documentation and a sample collection checklist should accompany a data submission. The following information should be included in this document:

1. Description of the instrument
2. Instrument calibration and maintenance
3. Sample collection method (e.g., by Niskin or flow through)
4. Instrument settings that affect types and sizes of particles imaged
5. Determining volume imaged per sample
6. Image processing method and version

7. Methods for automated and/or manual classification and taxonomic assignment
8. Summarization per sample (if applicable; applies to Level 2)
9. Additional data cleaning and quality assurance

3.3.1 Protocol documentation

Any data submission requires documentation that details information, such as software and hardware configurations, that determine how and what image data are collected and, therefore, how they will be interpreted in post processing. Detailed documentation of the software and hardware configurations will not only allow for duplication of such measurements but will also inform data users and interpretation. Moreover, information regarding sample collection technique, e.g., by Niskin or flow through, sampling depths, and any pre-filtering should be included in the documentation. See Appendix D for an example protocol document file.

Image processing and analysis occurs after image collection and before image classification and annotation. Image processing software is used for image thresholding, segmentation, and feature extraction for particle size, shape and texture. Some types of imaging-in-flow cytometers and submersible microscope instruments have their own software and associated terminology for image processing, and additional image processing software from third parties or from one's own lab group may be used for feature extraction, with inevitable updates to such software. The documentation should also include a description of image processing software and any computations (including equations and references therein) of feature-based products, such as biovolume. Inclusion of links to code repositories is strongly encouraged.

The classification of plankton and other particles requires software, whether done automatically using a machine-learning classifier with script to interpret class scores or manually recording into an annotation database. A description of processing methods for automated classifications, including the method (e.g., feature-based random forest, or deep-learning convolutional neural network), metrics used to select a "winning" class score, and the version of software is required. If appropriate, a link to a code repository should be provided. For manual annotations, it is assumed that only annotations with high confidence will be submitted to a repository. It is recommended to report if annotations were selected from a "higher power" annotator or report only those annotations with verifications. In the event the provider cannot confidently generate any classification for a given ROI, fill values should be used (i.e., use SeaBASS's numeric 'missing' value). Additional quality assurance steps are recommended before submission, such as the confirmation that the values for `biovolume`, `area_cross_section`, `length_representation`, and `width_representation` in the Level 1b data table are within expected ranges. The protocol file naming convention must be specific to the data submission, for example: protocol_plankton_and_particles_<CRUISE>.txt

3.3.2 The checklist

The submission checklist, essentially an abbreviated version of the protocol document, is designed

to standardize and preserve critical methods and analysis details that are needed for intercomparison, reprocessing, and to assist in evaluating the data for satellite validation or inclusion in algorithm development datasets. The checklist will also provide guidance as to which fields and headers to include essential information that must be added to the comments section of the data file, how to arrange data matrices, and determine the critical documentation that must be included with the data submission. If multiple formats are offered for download (e.g., rich text and plain text), choose one and fill out the necessary sections. Rename the file in a relevant way to make it unique (e.g., add the cruise name to the end of the file name), and add it to the other documents and calibration files that are part of the submission. The checklist naming convention must be specific to the data submission, for example: checklist_plankton_and_particles_<CRUISE>.txt. An example of a checklist is provided in Appendix E.

4 Appendix A: The YAML File

```
# namespace_MYNAMESPACENICKNAME.yaml # replace MYNAMESPACENICKNAME
with a short name of your choosing
# This namespace supplements identifications of non-conforming ROIs
defined in the ptwg namespace
- prefix: ptwg
  description: Ocean Carbon and Biogeochemistry Phytoplankton Taxonomy
Working Group
  uri: "https://seabass.gsfc.nasa.gov/ptwg_namespace_v1/"
- prefix: MYNAMESPACENICKNAME
  description: custom descriptions my lab uses to provide extra infor-
mation about non-conforming ROI identifications
  uri: "https://example.org/my_namespace/" # choose a unique URL that
you know nobody else will use
  terms:
  - id: transparent_detritus # this is a custom term that maps to one
or more IDs in other namespaces
    definition: transparent unidentified marine debris
    associated_terms: # these map to one or more terms in the ptwg
namespace
    - id: "ptwg:detritus"
  - id: opaque_detritus
    definition: opaque unidentified marine debris
    associated_terms:
    - id: "ptwg:detritus"

# In this example, two hypothetical identifiers were created to sup-
plement non-conforming ROIs defined in the PTWG namespace.
# For example, you can use the following identifier in the SeaBASS
scientificNameID data values:
#
# MYNAMESPACENICKNAME:transparent_detritus
#
# and we know that the full URI of the term is
#
# https://example.org/my_namespace/transparent_detritus
#
# We also know that it maps to ptwg id "detritus"
#
# ptwg:detritus (using the prefix) or
# https://seabass.gsfc.nasa.gov/ptwg_namespace_v1/ (using the full
URI)
```

5 Appendix B: SeaBASS Fieldnames

SeaBASS Fields	Units	Data type	Optional/ Required	Can it be used as a header?
associatedMedia	none	String	Required	No
data_provider_category_ automated	none	String	Optional	No
scientificName_ automated	none	String	Optional	No
scientificNameID_ automated	none	String	Required	No
data_provider_category_ manual	none	String	Optional	No
scientificName_manual	none	String	Optional	No
scientificNameID_manual	none	String	Required	No
biovolume	um^3	Floating point	Required	No
area_cross_section	um^3	Floating point	Required	No
length_representation	um	Floating point	Required	No
width_representation	um	Floating point	Required	Yes
volume_sampled_ml	um	Floating point	Required as field or header	Yes
volume_imaged_ml	um	Floating point	Required as field or header	Yes

SeaBASS Fields	Units	Data type	Optional/ Required	Can it be used as a header?
pixel_per_um	um	Floating point	Required as header or field	Yes
equivalent_spherical_ diameter	um	Floating point	Optional	No
area_based_diameter	none	Floating point	Optional	No
associated_files	none	String	Optional	Yes
associated_file_types	none	String	Optional	Yes
lat	degrees	Floating point	Optional	Yes, different header name
lon	degrees	Floating point	Optional	Yes, different header name
date	yyyymmdd	Date [UTC]	Optional	Yes, different header name
time	hh:mm:ss	Time [UTC]	Optional	Yes, different header name
depth	m	Floating point	Optional	Yes, different header name

6 Appendix C: Single Sample Data File Example

Note: The optional fields of `equivalent_spherical_diameter` and `area_based_diameter` are not included in this example.

```
/begin_header
/identifier_product_doi=10.5067/SeaBASS/NESLTER/DATA001
/investigators=Heidi_Sosik
/affiliations=Woods_Hole_Oceanographic_Institution
/contact=hsosik@whoi.edu
/experiment=NESLTER
/cruise=NESLTER_transect
/station=-9999
/data_file_name=NESLTER-NESLTER_transect_plankton_and_
particles_201802011037_R1.sb
/eventID=D20180201T103729_IFCB102
/data_type=flow_thru
/data_status=final
/documents=protocol_plankton_and_particles_NESLTER_IFCB102.
txt,checklist_plankton_and_particles_NESLTER_IFCB102.txt,namespace_
ptwg_nonconforming_roi_v1.txt
/calibration_files=NESLTER_IFCB102_calibration.txt
/start_date=20180201
/end_date=20180201
/start_time=10:37:29[GMT]
/end_time=10:37:29[GMT]
/north_latitude=41.3250[DEG]
/south_latitude=41.3250[DEG]
/east_longitude=-70.5650[DEG]
/west_longitude=-70.5650[DEG]
/water_depth=50
/measurement_depth=5
/instrument_model=Imaging_FlowCytobot_IFCB102
/instrument_manufacturer=McLANE_Research_Laboratories_Inc
/volume_sampled_ml=5
/volume_imaged_ml=3.99
/pixel_per_um=2.77
/associatedMedia_source=http://ifcb-data.whoi.edu/NESLTER_transect/
D20180201T103729_IFCB102.html
/length_representation_instrument_varname=maxFeretDiameter
/width_representation_instrument_varname=minFeretDiameter
!
```

```
! Please include with the submission two files with the list of the
automated_assessed_id and manual_assessed_id.
! For each one, create a text file with the scientificNameID_automated
or scientificName_manual, respectively.
! Please include the following headers:
/associated_files=automated_assessed_id_D20180201T103729_IFCB102.
txt,manual_assessed_id_D20180201T103729_IFCB102.txt, raw_images.tar.gz
/associated_files_type=metadata
!
! Comments
! This data file is an example and data are not necessarily real
! To construct each image filename from associatedMedia: extract the
string after the last / and replace .html with .png
!
/missing=-9999
/delimiter=comma
/fields=associatedMedia,data_provider_category_
automated,scientificName_automated,scientificNameID_automated,data_
provider_category_manual,scientificName_manual,scientificNameID_
manual,biovolume,area_cross_section,length_representation,width_
representation
/units=none,none,none,none,none,none,none,um^3,um^2,um,um
/end_header
http://ifcb-data.whoi.edu/NESLTER_transect/D20180201T103729_
IFCB102_00002.html,other,-9999,ptwg:other,Katodinium_
or_Torodinium,Dinophyceae,urn:lsid:marinespecies.
org:taxname:19542,647.186,149.748,23.731,9.614
http://ifcb-data.whoi.edu/NESLTER_transect/D20180201T103729_
IFCB102_00003.html,detritus,-9999,ptwg:detritus,detritus,-9999,ptwg:de
tritus,9091.825,1006.138,76.624,35.301
http://ifcb-data.whoi.edu/NESLTER_transect/D20180201T103729_
IFCB102_00004.html,detritus,-9999,ptwg:detritus,misc
ellaneous_nanoplankton,Eukaryota,urn:lsid:algaebase.
org:taxname:86701,26.397,14.857,7.975,2.888
http://ifcb-data.whoi.edu/NESLTER_transect/D20180201T103729_
IFCB102_00008.html,Rhizosolenia,Rhizosolenia,urn:lsid:marinespecies.
org:taxname:149069,Rhizosolenia,Rhizosolenia,urn:lsid:marinespecies.
org:taxname:149069,10144.506,1994.292,332.908,8.57
http://ifcb-data.whoi.edu/NESLTER_transect/
D20180201T103729_IFCB102_00009.html,Thalassiosira_TAG_
external_detritus,Thalassiosira,urn:lsid:marinespecies.
org:taxname:148912,Cylindrotheca,Cylindrotheca,urn:lsid:marinespecies.
org:taxname:149003,12.746,15.118,19.137,3.61
http://ifcb-data.whoi.edu/NESLTER_transect/D20180201T103729_
IFCB102_00116.html,bead,-9999,ptwg:bead,bead,-9999,ptwg:be
ad,1160.738,229.77,29.966,13.238
```

```
http://ifcb-data.whoi.edu/NESLTER_transect/
D20180201T103729_IFCB102_02121.html,bubble,-9999,ptwg:b
ubble,pennate,Bacillariophyceae,urn:lsid:marinespecies.
org:taxname:148899,8061.75,536.042,33.126,24.188
http://ifcb-data.whoi.edu/NESLTER_transect/D20180204T103729_
IFCB102_00283.html,fecal_pellet,-9999,ptwg:fecal_pellet,fecal_pellet,-
9999,ptwg:fecal_pellet,39902.14,2429.199,122.485,25.271
```

7 Appendix D: Example Protocol Document

Document author and contact info: Heidi Sosik (WHOI), hsosik@whoi.edu

Description of the instrument: Imaging of phytoplankton and other particles with Imaging FlowCytobot (IFCB; McLane Research Laboratories, Inc, Falmouth, MA). The IFCB is an imaging-in-flow cytometer. As such, it measures not only individual particle fluorescence and light scattering, but also captures a high resolution (~1 µm) image of each cell or chain in the size range ~5-150 µm width. Controlled flow and illumination conditions ensure a very high rate of images containing in focus, single targets aligned in the flow such that the largest cross-section is imaged. Images can be collected at up to ~15 Hz, depending on particle concentrations encountered. Images have a resolution of 2.77 pixels per micrometer.

Instrument calibration and maintenance: Main calibration issues are (1) ensuring sample volume is properly quantified (a function design criteria set during manufacture; user verification is good practice, but experience suggests this does not need to be repeated unless there are hardware changes in the instrument); and (2) determination of image scaling (micrometers per pixel; user determined with particles of interest).

Instrument settings that affect types and sizes of particles imaged: Images in this dataset were triggered by chlorophyll fluorescence, thus mainly representing phytoplankton but include herbivorous microzooplankton. IFCB trigger thresholds were set to image as wide a size range as possible, i.e., 5 to 150 micrometers, with quantitative observations in a narrower range. Images have resolution 2.77 pixels per micrometer.

Sample collection method: The cruise was on R/V Endeavor and identified as EN608, with the following Digital Object Identifier (DOI) https://doi.org/10.7284/908133. The IFCB was operated with chlorophyll fluorescence and scattering triggers enabled and it was configured to automatically sample 5 ml in 20-minutes intervals from the uncontaminated seawater flow (diaphragm pump source, pre-debubbler to ensure minimal damage to cells). IFCB was also used to analyze discrete samples from Niskin bottles (some with chlorophyll fluorescence triggering only). The samples were pre-filtered with a 150 µm nitex mesh to prevent system clogs.

Determining volume imaged per sample: The IFCB draws in 5 ml per sample but does not image the entire volume. The volume imaged per sample is provided in the online IFCB dashboard under Basic Info as Volume Analyzed.

Image processing method and version: Full resolution images are stored, though only the portion of the camera field that contains the target of interest (real time segmentation is done during acquisition). Images were processed with software for segmentation and feature extraction to determine size parameters per ROI (IFCB Features Version 4). Results from image processing are provided in the IFCB dashboard per sample as a features_v4.csv file. We selected a subset of 4 features to provide per

ROI (Area, Biovolume, maxFeretDiameter, minFeretDiameter). The biovolume calculation method is described by Moberg and Sosik (2012) https://doi.org/10.4319/lom.2012.10.278. Image processing yielded no features for only 7 of the 144,281 ROIs in this dataset. For direct access to images, replace the .html extension in the Level 1b associatedMedia with .jpg or .png. Processing code and wiki-based documentation is available at: https://github.com/WHOIGit/ifcb_classifier.

Methods for automated and/or manual classification and taxonomic assignment: Identifications to morphological categories were done manually using annotation software with a database that also records the annotators and the number of times an annotation has been verified. We queried the database to export manual annotations for the geographic subset of IFCB102 samples and then further divided into subsets that include only those samples for which every ROI in the sample was verified by a high-power annotator (to increase certainty in the manual identifications). In this version of this data product, each ROI corresponds to an occurrence of a single taxon (in future versions we may account for categories or tags for a small number of ROIs that represent multiple taxa). Most of the morphological category names could be resolved to accepted taxonomic names and machine-readable identifiers in the World Register of Marine Species (WoRMS). The level of taxonomic identification varies, but some distinctive taxa can be identified to species level. Some morphological categories could only be matched to WoRMS at a higher taxonomic level, for example mix_elongated is a morphological category of diatoms. Some morphological categories could not be matched to WoRMS but were matched to Eukaryota in AlgaeBase. We also resolved taxonomic names and identifiers for the full list of categories in the annotation database, to be able to indicate absence for those categories not observed in samples from this cruise (i.e., occurrenceStatus absent in a future version of this dataset). Several categories are not organisms thus taxonomic names are NotApplicable (e.g., bubble, detritus). Code and classification lookup tables are available in GitHub: https://github.com/klqi/EDI-NES-LTER-2019/tree/master/namespace_validation.

Summarization per sample: Summarization per sample is possible because we are only providing data for samples for which every ROI had a manual annotation. Code for the summarization from the Level 1b to the Level 2 data table is available in GitHub: https://github.com/klqi/EDI-NES-LTER-2019/tree/master/taxon_abundance. Concentrations per taxon per sample may be calculated by dividing the abundance or biovolume by the volume imaged. Abundance does not correspond necessarily to cell counts because chain- or colony-forming organisms may be imaged as a single ROI. Note concentrations will be underestimates for taxa that do not always trigger fluorescence.

Additional data cleaning and quality assurance: Additional data cleaning and metadata template assembly were performed with code available on GitHub: https://github.com/WHOIGit/nes-lter-ifcb-transect-winter-2018. We renamed or added attributes to enable harvesting of the Level 1b data table as an occurrence table for the Ocean Biodiversity Information System (OBIS, e.g., occurrenceID, eventDate, decimalLongitude, decimalLatitude, occurrenceStatus, basisOfRecord). We assured that the geographic and temporal coverage and values for attributes were within expected ranges.

Key method references:

Olson, R. J., and H. M. Sosik. 2007. A submersible imaging-in-flow instrument to analyze nano- and microplankton: Imaging FlowCytobot. *Limnology and Oceanography: Methods* 5: 195-203.

Sosik, H. M., and R. J. Olson. 2007. Automated taxonomic classification of phytoplankton sampled with imaging-in-flow cytometry. *Limnology and Oceanography: Methods* 5: 204-216.

Sosik, H. M., J. Futrelle, E. F. Brownlee, E. Peacock, T. Crockford, and R. J. Olson. 2016. hsosik/ifcb-analysis: IFCB-Analysis software system, initial formal release at v2 feature stage [Data set]. Zenodo. http://doi.org/10.5281/zenodo.153978

Sosik, H.M., E. Peacock, and M. Santos. (2020). Abundance and biovolume of taxonomically-resolved phytoplankton and microzooplankton imaged continuously underway with an Imaging FlowCytobot along the NES-LTER Transect in winter 2018 Ver1. Environmental Data Initiative. https://doi.org/10.6073/pasta/74775c4af51c237f2a20e4a8c011bc53

Peacock, E.E., E. T. Crockford, and H.M. Sosik. 2018. IFCB at sea user guide. https://docs.google.com/document/d/14IfQBriV2AZs1akefM8JYirSAApnVFbDG2XQ74klIOI/

8 Appendix E: Example Checklist

CHECKLIST FOR SeaBASS SUBMISSION: Imaging-in-flow cytometry
V20200729
Please fill out the Collection, Measurement, and Analysis methods sections. Answer below each number. When finished, rename this file to be specific for your data, e.g., "checklist_plankton_and_particles_MyCruiseName.txt"

Experiment Name: _____
Cruise Name: _____
Bundled images submitted? _____
Assessed ID list(s) for automated and/or manual classification submitted and referenced in '/associated_files' metadata headers? _____

- SAMPLE COLLECTION METHODS -
---

1. How were the water samples collected? (Niskin bottle, bucket etc.)
2. Standard depths of sample collection (surface, chl max etc.)
3. Was the sample prefiltered? If so, type of filter (e.g., nitex, pore size)
4. How was the sample introduced to the instrument (pipetted, drawn from a larger vessel, syringe-fed)?

- SAMPLE MEASUREMENT METHODS -
---

1) List the instrument make, model and accessories (if applicable):
2) List instrument calibration and maintenance performed (including date):
3) Measurement mode (autoimage, trigger fluorescence only, trigger including scatter):
4) Objective (magnification):
5) Flow cell type (catalog number, size/depth):
6) Sampling Flow rate:
7) Image collection speed (Hz, fps):
8) Method of focus (e.g., Culture, beads):
9) Size range of particles imaged:

- DATA ANALYSIS METHODS -
---

1) Classifier used (including date of most recent update):
2) Taxonomic authority used:
3) Were all ROIs annotated?
4) Are Lists of all Life Science Identifiers assessed for 'automated' and for 'manual' included in your submission?

9 References

Behrenfeld, M. Climate-mediated dance of the plankton. *Nature Climate Change* 4, 880–887 (2014). https://doi.org/10.1038/nclimate2349

Boss, E., A. M. Waite, J. Uitz, S. G. Acinas, H. M. Sosik, K. Fennel, I. Berman-Frank, M. Cornejo, S. Thomalla, and H. Yamazaki. 2020. Recommendations for plankton measurements on the GO-SHIP program with relevance to other sea-going expeditions. SCOR Working Group 154 GO-SHIP Report.

Boyd, P.W., Claustre, H., Levy, M. et al. Multi-faceted particle pumps drive carbon sequestration in the ocean. *Nature* 568, 327–335 (2019). https://doi.org/10.1038/s41586-019-1098-2

Gobler, C.J. (2020). Climate Change and Harmful Algal Blooms. Insights and perspective, *Harmful Algae*, 91: 101731. https://doi.org/10.1016/j.hal.2019.101731.

IOCCG (2014). Phytoplankton Functional Types from Space. Sathyendranath, S. (ed.), Reports of the International Ocean-Colour Coordinating Group, No. 15, IOCCG, Dartmouth, Canada. http://dx.doi.org/10.25607/OBP-106.

IOCCG (2020). Synergy between Ocean Colour and Biogeochemical/Ecosystem Models. Dutkiewicz, S. (ed.), IOCCG Report Series, No. 19, International Ocean Colour Coordinating Group, Dartmouth, Canada. http://dx.doi.org/10.25607/OBP-711.

Kissling WD, Ahumada JA, Bowser A, Fernandez M, Fernández N, García EA, Guralnick RP, Isaac NJ, Kelling S, Los W, McRae L. Building essential biodiversity variables (EBV s) of species distribution and abundance at a global scale. *Biological Reviews*. 2018 Feb; 93(1):600-25.

Le Quere, C., Harrison, S.P., Colin Prentice, I., Buitenhuis, E.T., Aumont, O., Bopp, L., Claustre, H., Cotrim Da Cunha, L., Geider, R., Giraud, X. and Klaas, C., 2005. Ecosystem dynamics based on plankton functional types for global ocean biogeochemistry models. *Global Change Biology*, 11(11), pp.2016-2040.

Menden-Deuer, S., Morison, F., Montalbano, A. L., Franzè, G., Strock, J., Rubin, E., ... & Marrec, P. (2020). Multi-Instrument Assessment of Phytoplankton Abundance and Cell Sizes in Mono-Specific Laboratory Cultures and Whole Plankton Community Composition in the North Atlantic. *Frontiers in Marine Science*, 7, 254. https://doi.org/10.3389/fmars.2020.00254.

Mouw, C.B., Barnett, A., McKinley, G.A., Gloege, L. and Pilcher, D., 2016. Phytoplankton size impact on export flux in the global ocean. *Global Biogeochemical Cycles*, 30(10), pp.1542-1562.

Muller-Karger, F.E., Miloslavich, P., Bax, N.J., Simmons, S., Costello, M.J., Sousa Pinto, I., Canonico, G., Turner, W., Gill, M., Montes, E. and Best, B.D., 2018. Advancing marine biological observations and data requirements of the complementary essential ocean variables (EOVs) and essential biodiversity variables (EBVs) frameworks. *Frontiers in Marine Science*, 5, p.211.

Olson, R. J., and H. M. Sosik. 2007. A submersible imaging-in-flow instrument to analyze nano- and microplankton: Imaging FlowCytobot. *Limnology and Oceanography: Methods* 5: 195-203.

Peacock, E.E., E. T. Crockford, and H.M. Sosik. 2018. IFCB at sea user guide. https://docs.google.com/document/d/14IfQBriV2AZs1akefM8JYirSAApnVFbDG2XQ74klIOI/

Sosik, H.M. and R.J. Olson., 2007. Automated taxonomic classification of phytoplankton sampled with imaging-in-flow cytometry. *Limnology and Oceanography: Methods*. 5: 204-216.

Sosik, H. M., J. Futrelle, E. F. Brownlee, E. Peacock, T. Crockford, and R. J. Olson. 2016. hsosik/ifcb-analysis: IFCB-Analysis software system, initial formal release at v2 feature stage [Data set]. Zenodo. http://doi.org/10.5281/zenodo.153978

Sosik, H.M., E. Peacock, and M. Santos. (2020). Abundance and biovolume of taxonomically-resolved phytoplankton and microzooplankton imaged continuously underway with an Imaging FlowCytobot along the NES-LTER Transect in winter 2018 Ver1. Environmental Data Initiative. https://doi.org/10.6073/pasta/74775c4af51c237f2a20e4a8c011bc53

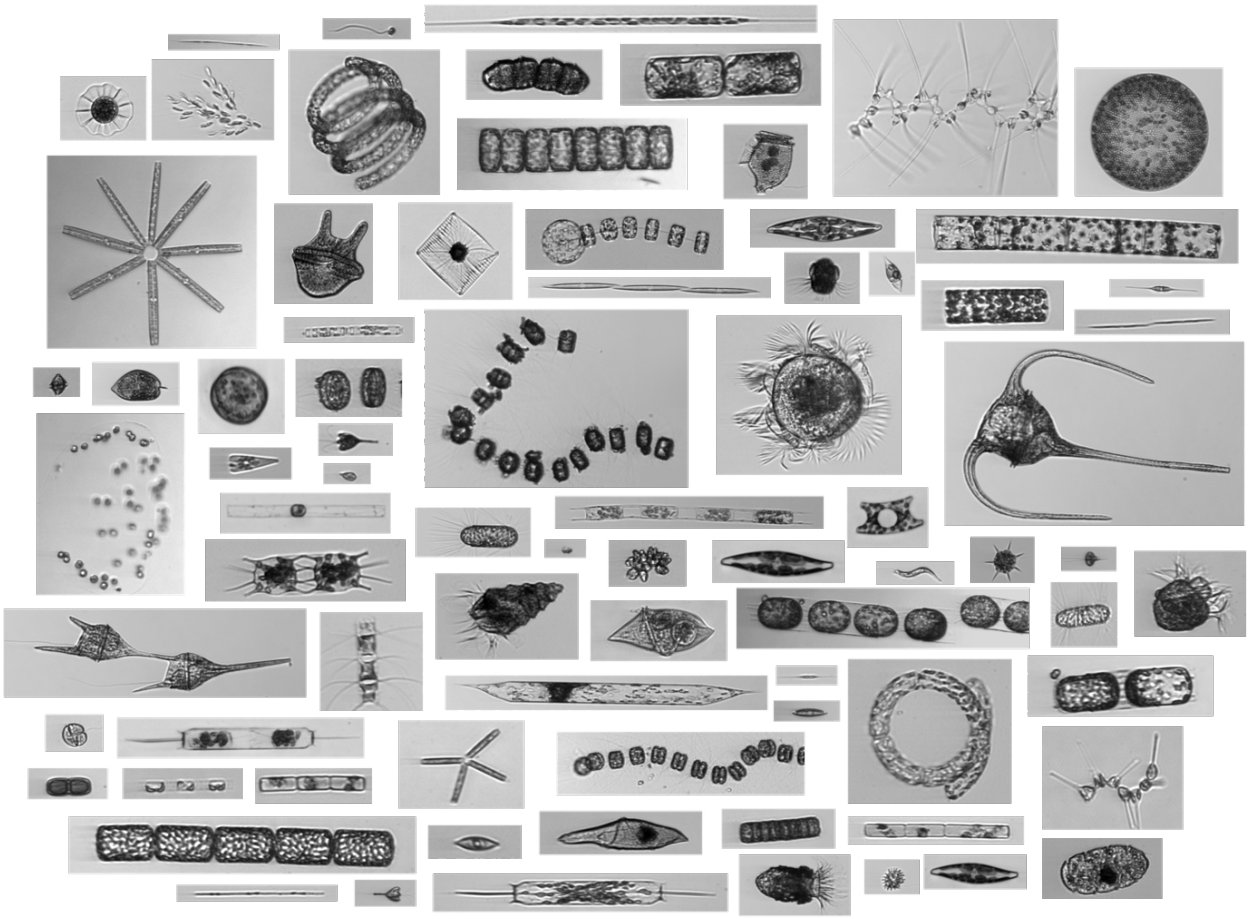
Walcutt, N.L., Knörlein, B., Cetinić, I., Ljubesic, Z., Bosak, S., Sgouros, T., Montalbano, A.L., Neeley, A., Menden-Deuer, S. and Omand, M.M. (2020), Assessment of holographic microscopy for quantifying marine particle size and concentration. *Limnology and Oceanography: Methods*, https://doi.org/10.1002/lom3.10379.

Werdell, P.J., Bailey, S.W., Fargion, G.S., Pietras, C., Knobelspiesse, K.D., Feldman G.C., and McClain, C.R., 2003: Unique data repository facilitates ocean color satellite validation. *EOS Transactions. AGU*, 84(38), pp.377.

Werdell, P. J., Behrenfeld, M. J., Bontempi, P. S., Boss, E., Cairns, B., et al. (2019). The Plankton, Aerosol, Cloud, Ocean Ecosystem Mission: Status, Science, Advances. *Bulletin of the American Meteorological Society*, 100(9), 1775-1794. https://doi.org/10.1175/BAMS-D-18-0056.1.

Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018. https://doi.org/10.1038/sdata.2016.18

For more information visit:
us-ocb.org/phytoplankton-taxonomy-working-group/



Ocean Carbon & Biogeochemistry Program
Woods Hole Oceanographic Institution
266 Woods Hole Road MS#25, Woods Hole, MA 02543
hbenway@whoi.edu | 508.289.2838
www.us-ocb.org | Twitter @us_ocb

OCB acknowledges support from these US agencies:

This report was developed with federal support of NSF (OCE-1558412) and NASA (NNX17AB17G). Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsoring agencies.