Small Working Group Phytoplankton Taxonomy

Data Standards and Practices for Taxon-Resolved Phytoplankton Observations

PIs: Heidi Sosik (WHOI), Aimee Neeley (NASA GSFC)
Christopher Proctor, Ivona Cetinić (NASA GSFC)

IFCB

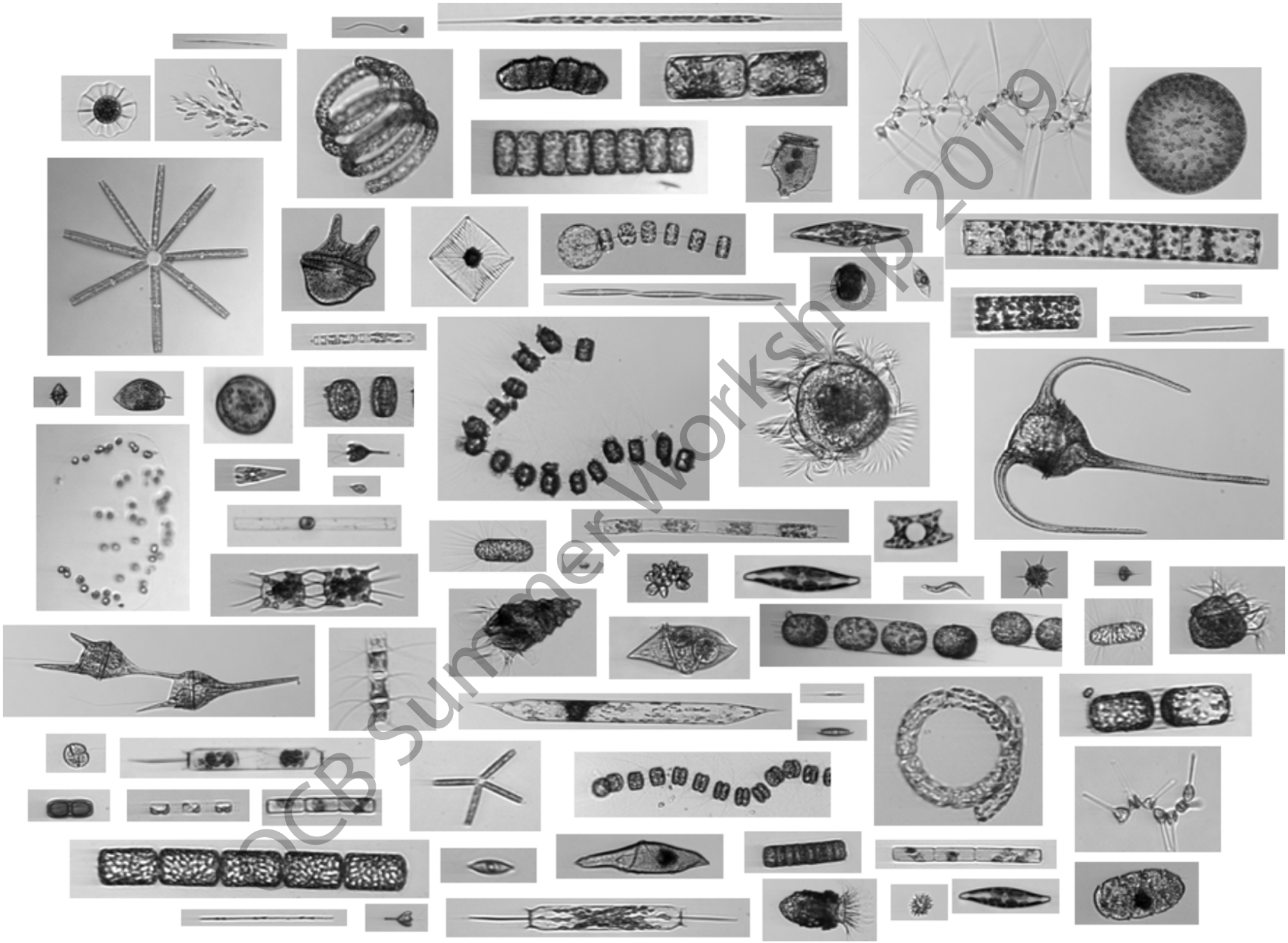
SPC

FlowCam

CytoSense

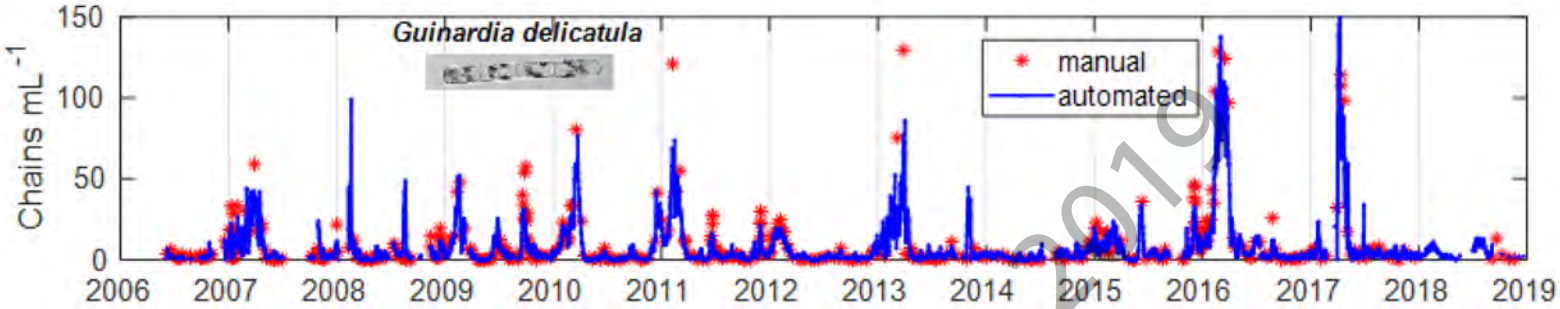
Jupiter Microscope

LISST-Holo

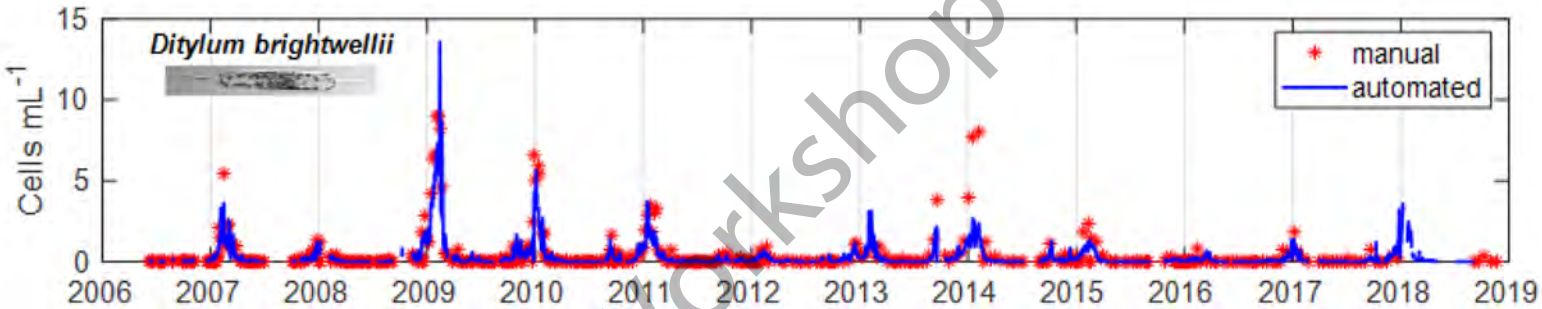


MVCO

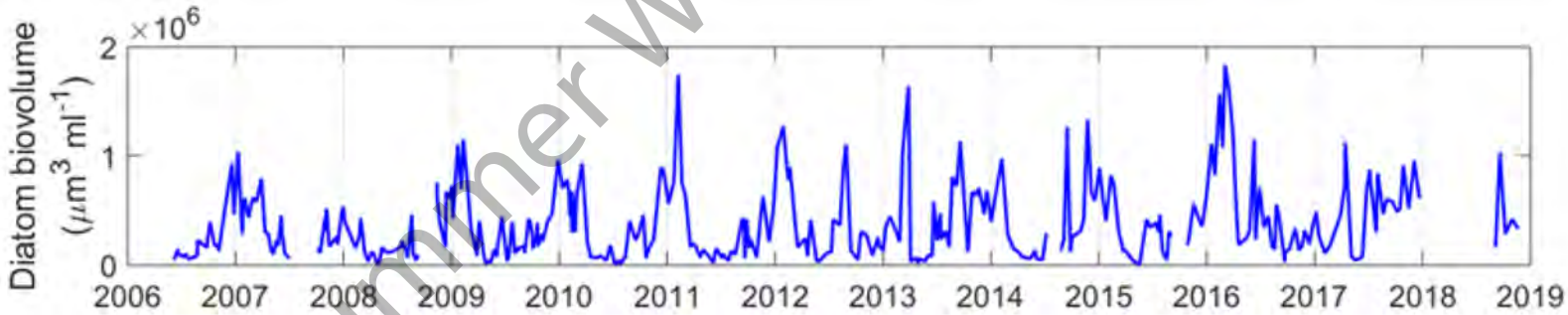
Species 1



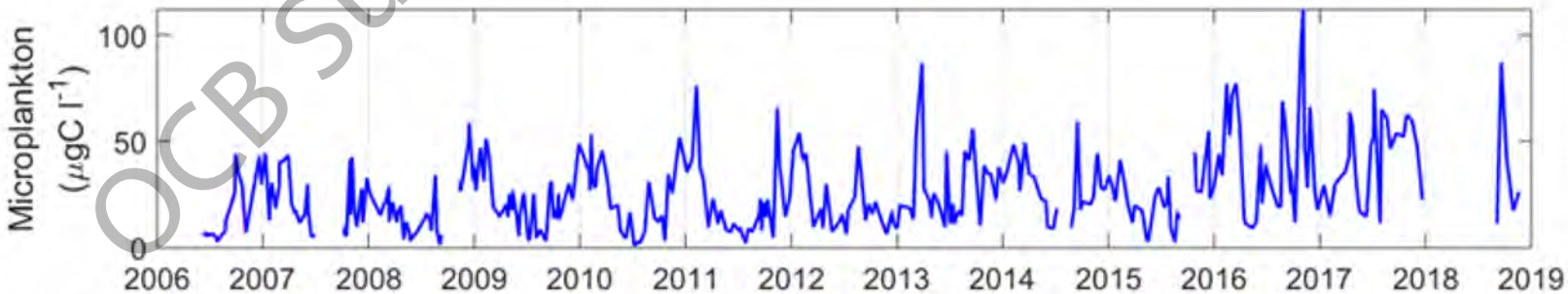
Species 2



All Diatoms



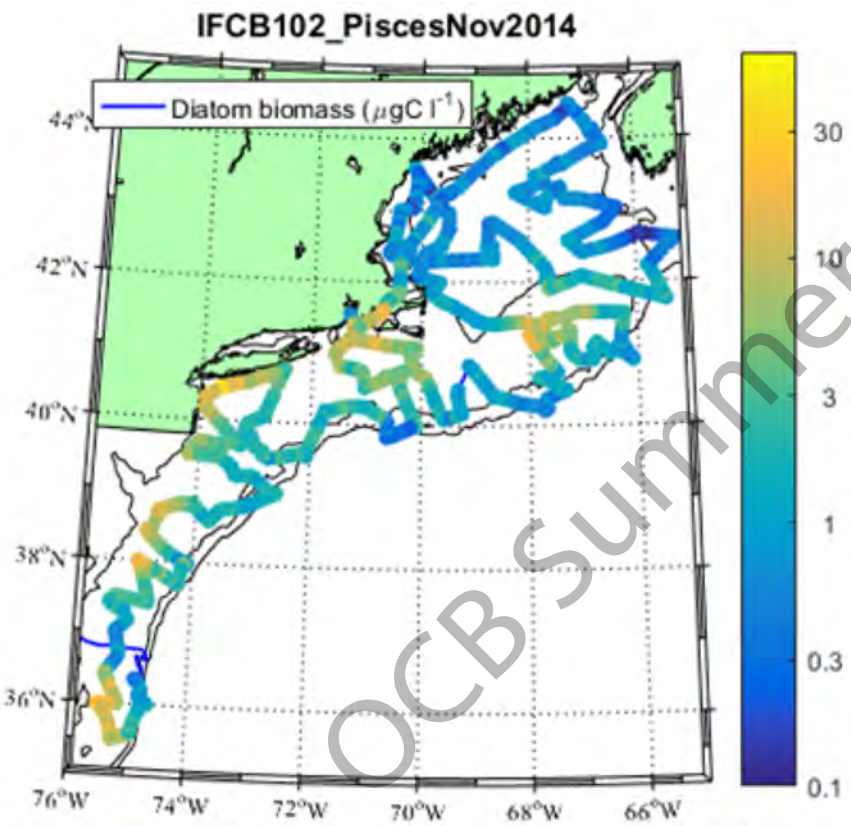
All Micro-Plankton

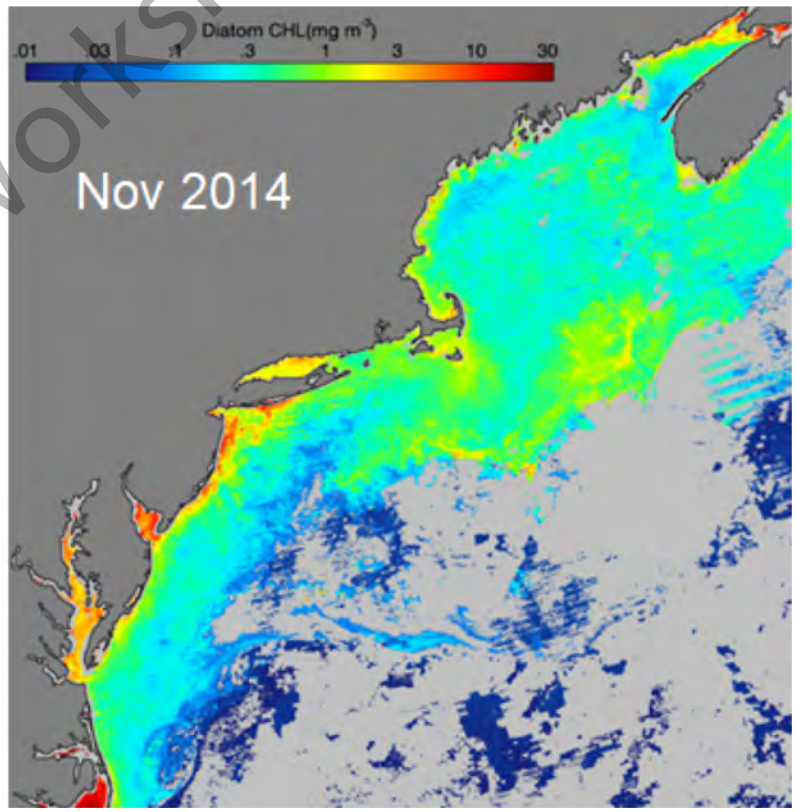


Diatom Biomass

MODIS PFT product

Courtesy of Kim Hyde





Objective: Develop a set of standards and best practices for phytoplankton taxonomy data to facilitate community-wide access to phytoplankton data products that support critical satellite algorithm development and validation


Stace Beaulieu


Ivona Cetinić


Susan Craig


Emmanuel Devred


Joe Futrelle


Lee Karp-Boss

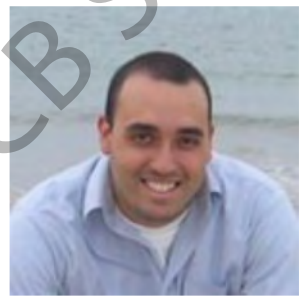

Aimee Neeley


Marc Picheral


Nicole Poulton


Collin Roesler


Chris Proctor


Adam Shepherd


Heidi Sosik

Approach

WG convened in spring 2017

4 in-person meetings

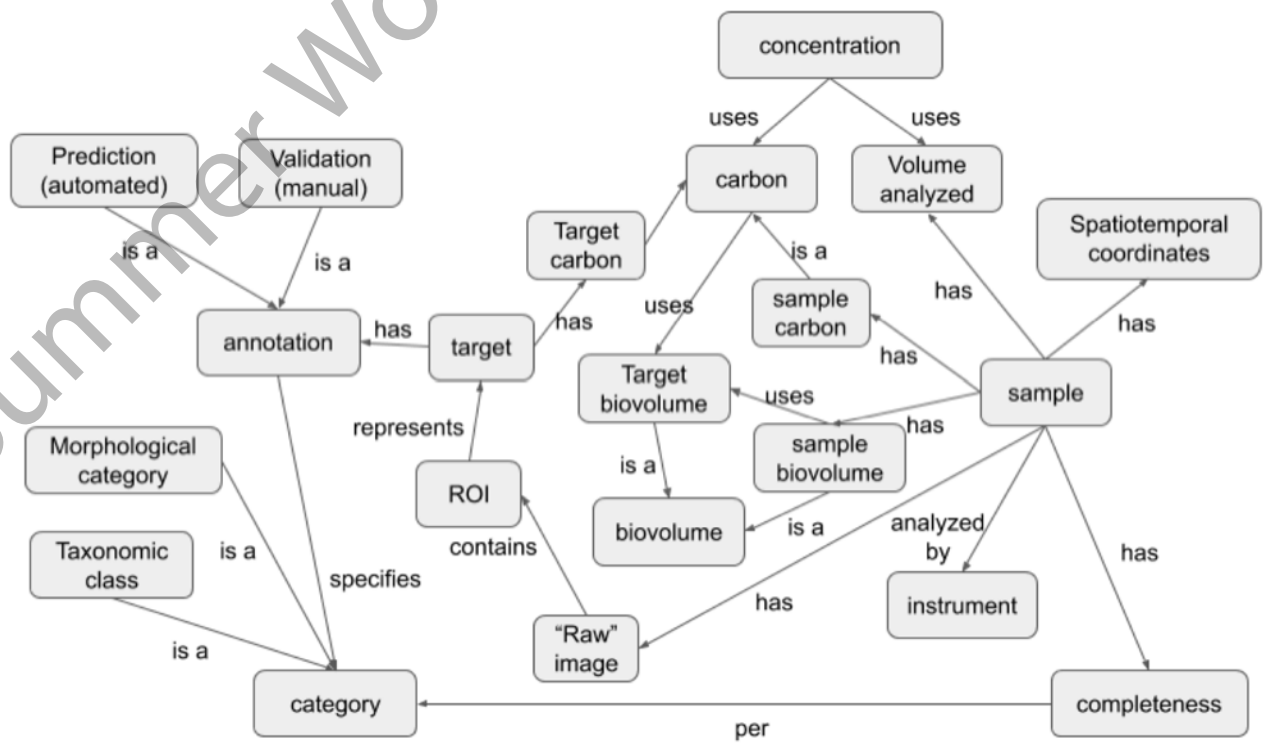
Series of web meetings

Shared documents

Formal use case based strategy

Iterative design and evaluation process

Example information model



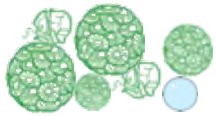
Developing a strategy for archiving taxonomic image data in SeaBASS and other archives

Currently: the archiving status of taxonomic image data has been in limbo… their complexity, size, and lack of standardization of these data present special challenges to data archives such as SeaBASS

Ongoing work with the OCB Phytoplankton Taxonomy Working Group has involved addressing challenges including prototyping how to:

- **Archive relevant info in a standardized file format**
- **Generate remote sensing products** (e.g., abundance, biovolume, carbon, or PSC estimates)
- **Preserve data and metadata** (enable reprocessing, track provenance)
- **Interact with existing systems** that are specifically designed to visualize and annotate these data (e.g., EcoTaxa and IFCB Dashboard)

Proposed strategy for archiving Taxonomic/Image data



What

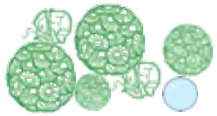
raw images and metadata:

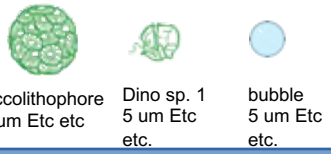
Bundles of instrument images and metadata

Notes

Preserve for reprocessing

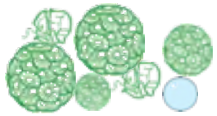
Proposed strategy for archiving Taxonomic/Image data





What	**raw images and metadata:** Bundles of instrument images and metadata	**processed images:** Classified and annotated
Notes	*Preserve for reprocessing*	*Specialized external hosts, e.g.* ***EcoTaxa*** *or* ***IFCB Dashboard***

Proposed strategy for archiving Taxonomic/Image data



Coccolithophore 15 um Etc etc.

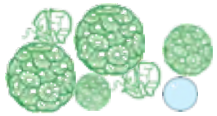
Dino sp. 1 5 um Etc etc.

bubble 5 um Etc etc.

Sample 1 (date, time, location, depth):
Coccolithophore, 15 um, http=abc, id = 12345
Dinoflagellate sp 1., 5 um, http=ab, id = 11198
Bubble, 5 um, http=ab912, special id = 5

What	**raw images and metadata:** Bundles of instrument images and metadata	**processed images:** Classified and annotated	**classified water sample data:** Info for every image in a water sample
Notes	*Preserve for reprocessing*	*Specialized external hosts, e.g.* ***EcoTaxa*** *or* ***IFCB Dashboard***	*SeaBASS File search*

Proposed strategy for archiving Taxonomic/Image data



Coccolithophore 15 um Etc etc.

Dino sp. 1 5 um Etc etc.

bubble 5 um Etc etc.

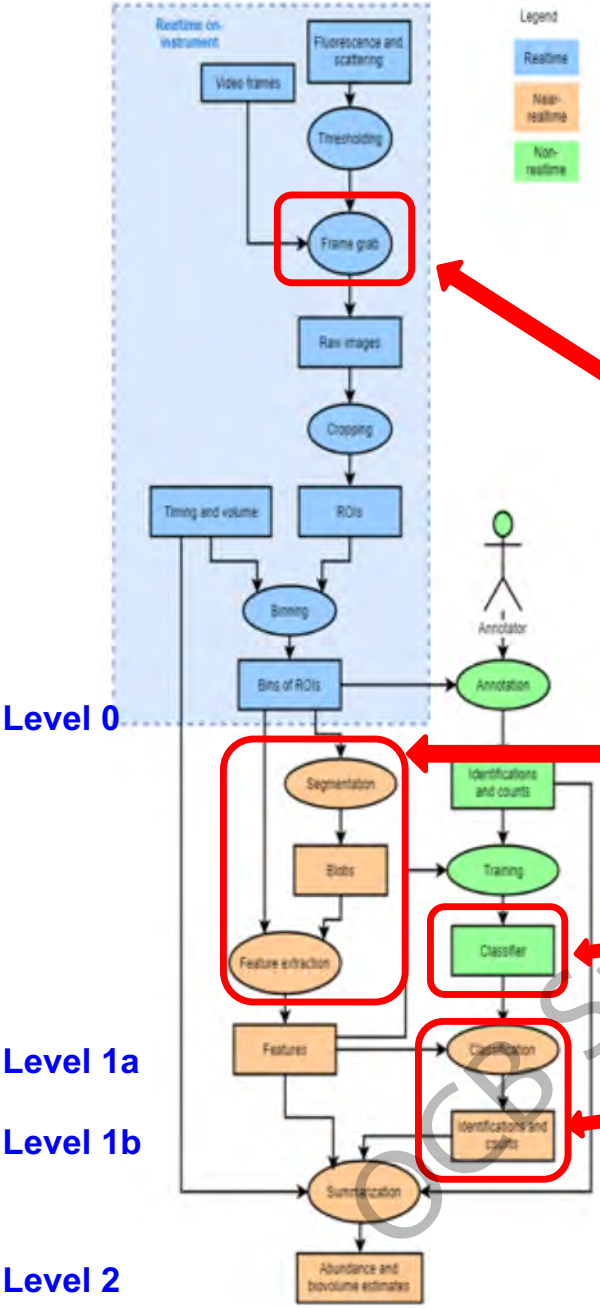
Sample 1 (date, time, location, depth):
Coccolithophore, 15 um, http=abc, id = 12345
Dinoflagellate sp 1., 5 um, http=ab, id = 11198
Bubble, 5 um, http=ab912, special id = 5

Sample 1: date,loc,dino_biovol = 50 m^3 L-1
Sample 2: date,loc,dino_biovol = 12 m^3 L-1
Sample 3: date,loc,dino_biovol = 12 m^3 L-1
Sample 4: date,loc,dino_biovol = 12 m^3 L-1
Sample 5: date,loc,dino_biovol = 3 m^3 L-1

What	**raw images and metadata:** Bundles of instrument images and metadata	**processed images:** Classified and annotated	**classified water sample data:** Info for every image in a water sample	**summary products:** e.g., abundance, biovolume, or carbon for phytoplankton groups
Notes	*Preserve for reprocessing*	*Specialized external hosts, e.g.* ***EcoTaxa*** *or* ***IFCB Dashboard***	*SeaBASS File search*	*SeaBASS File search* *Made /w standardized code*

****Create a prototype for contributing IFCB data and other imagery data to SeaBASS****

IFCB workflow overview chart



What is important for provenance of IFCB data?

Instrument settings that affect types and sizes of particles imaged

Image processing method and version

Classification algorithm and version

Interpretation for automated classification

Best practices revolve around

- **How to specify your taxa so that others can understand your taxon groups.**

  i.e. align your taxon groups with taxonomic authorities (e.g., WoRMS)

- **How to provide sufficient metadata so that others can reuse your data.**

  i.e., enable creation of summary products based on taxa and size classes

- **How to structure and format the data and metadata for interoperability and reuse.**

  i.e., select certain formats/file types to facilitate downstream workflows

All of the above facilitate downstream workflows to create standardized files and summary products

SeaBASS file format

ASCII text organized into two sections:

1. Metadata Headers

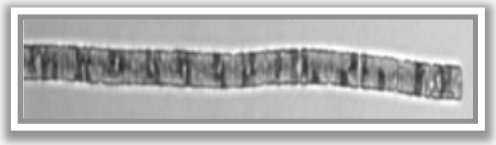
- Self-describing keyword-value pairs

```
/begin_header
/keyword=value
! This is a comment
/end_header
```

2. Data Matrix

- Describes all identified images within a water sample

Each row describes a classified image



Metadata for L1B files (classified images in a water sample)

Structured metadata headers specify:

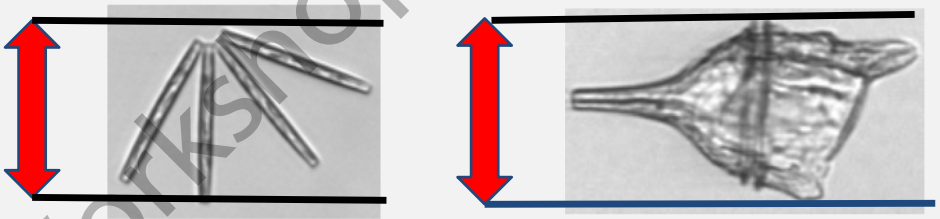
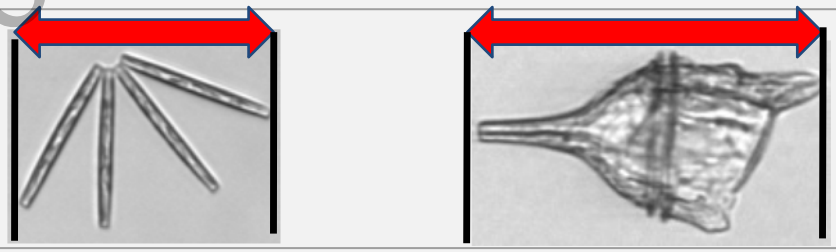
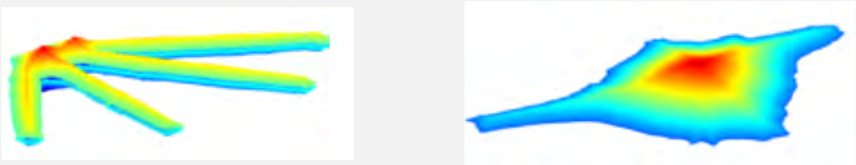
- spatial/temporal info
- volume_sampled
- volume_imaged
- pixel_per_µm
- instrument name/manufacturer
- which groups assessed were completely assessed
- external namespace file

```
/begin_header
/identifier_product_doi=10.5067/SeaBA…
/investigators=John_Smith,Mary_Johnson
/affiliations=State_University
/contact=jsmith@state.edu
/experiment=EXAMPLE_A
/cruise=cal0101
/station=93
/data_file_name=pigments_cal0101.dat
/documents=cal0101_readme.txt
/calibration_files=turner_cals_12.txt
/data_type=pigment
/start_date=20010314
/end_date=20010314
/start_time=16:01:30[GMT]
/end_time=16:30:45[GMT]
/north_latitude=42.135[DEG]
/south_latitude=42.055[DEG]
/east_longitude=-72.375[DEG]
/west_longitude=-72.420[DEG]
! Comments
! Comment lines can include extra info
/missing=-9999
/below_detection_limit=-8888
/delimiter=tab
/fields=time,depth,CHL,CHL_SD,PHAEO,Tpg
/units=hh:mm:ss,m,mg/m^3,mg/m^3,mg/m^3
/end_header
```

Non-structured metadata headers specifies:

- Trigger type (comment)
- Biovolume calculation method

Size information to derive PSCs or carbon abundance

Field name	Units	Description
feret_diameter_min	µm	
feret_diameter_max	µm	
biovolume	μm^3	
area_cross_section	μm^2	

Each identification will specify a “Namespace”

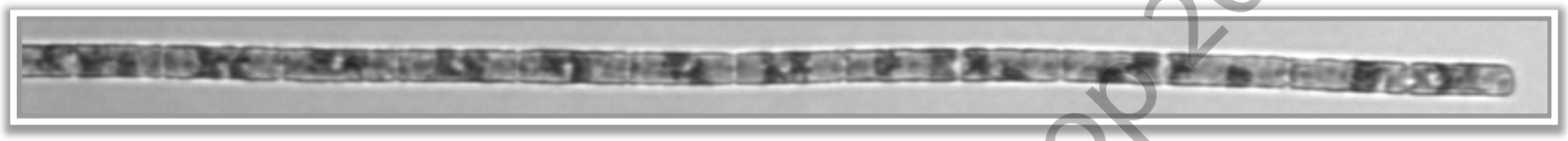
1. Contributors may need to use a different primary standardized reference library than WoRMS/AphiaID (e.g., UniEuk)
2. Contributors can create and use customized namespaces. They must be defined in a linked YAML file. The OCB PTWG is providing a basic custom library to supplement AphiaIDs with non-living IDs (e.g., detritus)

namespace_automated	identification_automated	namespace_manual	identification_manual
aphiaid	149109	aphiaid	149109
aphiaid	#	aphiaid	-9999
ptwg	‘bubble’	ptwg	‘bubble’
aphiaid	148985	aphiaid	148912

Each sample will have columns for 2 types of identification:

- Manual (i.e., human validated)
- Automated (i.e., machine/algorithm)

Use reference libraries (e.g., WoRMS AphiaID) to indicate classification/identification



AphiaID also links parent taxons

AphiaID: 149112
rank: **Species**
scientific name: Guinardia delicatula

↑

AphiaID: 149119
rank: **Genus**
scientific name: Guindardia

↑

AphiaID: 149068
rank: **Family**
scientific name: Rhizosoleniaceae

↑

order, superorder, class, etc...

In Summary…

Our Objective: Develop a set of standards and best practices for phytoplankton taxonomy data to facilitate community-wide access to phytoplankton data products that support critical satellite algorithm development and validation.

- Imaging technologies are becoming part of standard data collection for large scale oceanographic field campaigns (e.g., EXPORTS and NAAMES) and time series studies (e.g., LTER).
- We initially focused on IFCB-derived products. File formats for other instrument and data platforms (e.g., FlowCAM and EcoTAXA) are in progress.
- Similar approaches could be applied for other kinds of biological imaging, such as the many zooplankton imaging systems (e.g., UVP).
- We want your input! We expect this to be an iterative process.

Extra Slides

Each sample will have columns for 2 types of identification:

- **Manual (i.e., human validated)**
- **Automated (i.e., machine/algorithm)**

identification_automated	identification_manual
149109	149109
149109	-9999
148985	148912

1. Automated & manual identification matched
2. Manual not attempted
3. Automated & manual identification results differed

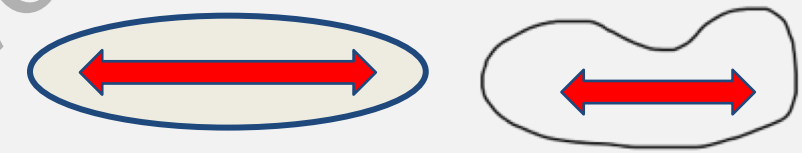
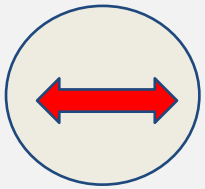
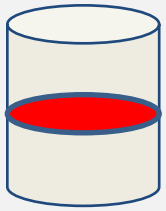
SeaBASS data submission example

How to conclude?
Discuss other element of submission?
Discuss PTWG next steps? (e.g., prototype FlowCam and other data measurements to ensure that our solutions are fully flexible/universal?)
TM?

A full dataset (e.g., one cruise) for contribution to SeaBASS would include:

- Level 1B files (.sb)
- Namespace table
- Protocol document file
- Project metadata are included in SeaBASS headers; additional details may be provided in linked external documentation files
- Raw/instrument data files

Size information to derive PSCs or carbon abundance

Field name	Units	Description
feret_diameter_min	µm	
feret_diameter_max	µm	
biovolume	µm^3	 $\frac{\pi}{6} \times W^3$  $\frac{\pi}{4} \times L \times W^2$
area_cross_section	µm^2	 $A = \pi r^2$

SeaBASS data submission example

A full dataset (e.g., one cruise) for contribution to SeaBASS would include:

- **Level 1B files (.sb)**
- Namespace table
- Protocol document file
- Project metadata are included in SeaBASS headers; additional details may be provided in linked external documentation files
- Raw/instrument data files

YAML Example

A markup language file

```
- prefix: ptwg
  description: Ocean Carbon and Biogeochemistry Phytoplankton Taxonomy Working Group
  uri: "http://ocbptwg.org/ns#"
  terms: # terms in the namespace
  - id: detritus # an example of a term with no taxonomic id
    definition: unidentified marine debris
  - id: bead
    definition: plastic calibration target
  - etc.
```

One of the reasons / use cases these data are useful to NASA and other users (e.g., modelers) are calculated products:

Abundances, biovolumes, carbon, or size class estimates

Herein, we consider how the data can be used and distributed by NASA or other databases. So, we might reference SeaBASS, but we're aiming for generalizable/adaptable solutions and protocols

We've developed a prototype for archiving IFCB data in SeaBASS. It is extendable to other measurement types, but more work needs to be done (e.g., FlowCam, Flow Cytometry).

Since looking at a text file that looks like a giant spreadsheet isn't very visually appealing for a large-group presentation, we'll break apart our best practices and focus on them and highlight particular components of the prototype

Intro best practices
Focus on the taxa
Focus on the sizes
Above 2 lead into 2nd prototype data table format
Focus on the provenance
Leads into 2nd prototype metadata formats

OCB 2019 Image Data

Introduce the concept of data being submitted to repositories (e.g., SeaBASS & BCODMO)

General standards to ensure interoperability, sharing, etc (briefly)

Here are all the challenges that we get into when we try to create files for imagery data XYZ…
Keep processing information
Deal with many instrument types and formats

We need solutions for imagery data which is even more complicated than data types we typically archive

Developing best practices and a format to store the data

Is a slide needed for PACE and validation data? I'm not sure how much we need to set the stage vs. delve into technical details

SeaBASS data types

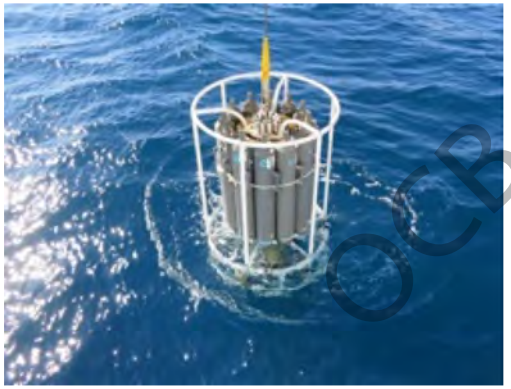
Data archived in SeaBASS are collected from ships, moorings, autonomous buoys and other platforms. Measurements come from a variety of instruments, such as profilers, hand-held sensors, and laboratory analyzers.

Diverse data types include:

- apparent and inherent optical properties
- phytoplankton pigments
- carbon stocks
- hydrography
- other biogeochemical & atmospheric measurements

- not much phytoplankton imagery yet…

1



2

3

4

Images provided by Javier Concha[1,2,3] and Chris Proctor[4]

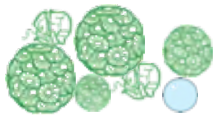
http://seabass.gsfc.nasa.gov

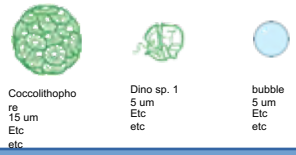
Custom Namespaces provided as YAML files

```yaml
# In the following example I'm just making up values, there's no attempt to use
# correct ones. For example for custom namespaces we can use whatever URI we want to
# and the ones I'm including are just placeholders.

# The basic structure is a list of namespaces. Each namespace minimally has
# a prefix (for reference) and a full URI. In addition for custom namespaces
# terms can be defined. Each term has a local ID and a definition and optionally
# a link to associated taxonomic IDs

- prefix: worms # short prefix to refer to this namespace
  description: World Register of Marine Species
  uri: "urn:lsid:marinespecies.org:taxname:" # full URI of namespace
- prefix: ptwg
  description: Ocean Carbon and Biogeochemistry Phytoplankton Taxonomy Working Group
  uri: "http://ocbptwg.org/ns#"
  terms: # terms in the namespace
  - id: detritus # an example of a term with no taxonomic id
    definition: unidentified marine debris
  - id: bead
    definition: plastic calibration target
  - etc.
- prefix: sosik
  description: Blah blah blah
  uri: "tag:hsosik@whoi.edu,2019:ns:"
  terms:
  - id: guinardia_parasite # this is a term with taxonomic id(s)
    definition: Guinardia delicatula interacting with a parasite
    associated_terms: # can map to multiple terms
    - id: "worms:149112"
      label: Guinardia delicatula
    - id: "sosik:guinardia"
      label: Guinardia

# In this example, we can use the following identifier:
#
# sosik:guinardia_parasite
#
# and we know that the full URI of the term is
#
# tag:hsosik@whoi.edu,2019:ns:guinardia_parasite
#
# We also know that it maps to worms id 149112
# and from our description of worms we know that that's got the id
#
# worms:149112 (using the prefix) or
# urn:lsid:marinespecies.org:taxname:149112 (using the full URI)
#
# In the seabass record we can now simply use the namespace prefix
# followed by the id, e.g.,
# - ptwg:detritus
# - sosik:guinardia_parasite
# - worms:149112
#
# This is following best practices in LoD for namespace-scoped ids.
# It assumes that you can construct URIs by appending a term ID
# onto the end of a base URI which is common / best practice
```

Proposed strategy for archiving Taxonomic/Image data

Columns indicate data format and storage location





Sample 1 (date, time, location, depth):
Coccolithophore, 15 um, http=abc, id = 12345
Dinoflagellate sp 1., 5 um, http=ab, id = 11198
Bubble, 5 um, http=ab912, special id = 5

Sample 1: date,loc,dino_biovol = 50 m^3 L-1
Sample 2: date,loc,dino_biovol = 12 m^3 L-1
Sample 3: date,loc,dino_biovol = 12 m^3 L-1
Sample 4: date,loc,dino_biovol = 12 m^3 L-1
Sample 5: date,loc,dino_biovol = 3 m^3 L-1

	1.	2.	3.	4.
What	**Raw images and metadata:** Bundles of instrument data grouped by cruise/deployment	**Processed images:** classified and annotated If hosted online, each image has unique permalink	**SeaBASS files, 1 per water sample:** 1 data row per ROI (Region Of Interest; i.e., image)	**SeaBASS file:** Rows and columns describe entire water samples summarized into **remote sensing products**
Where	*SeaBASS (optional access)*	*External to NASA: Images displayed via e.g.* ***EcoTaxa*** *or* ***IFCB Dashboard***	*SeaBASS File search*	*SeaBASS File search* *Made by SeaBASS staff /w standardized code*
Extra Details	*Essential to preserve so data can be reclassified to update #2 onward. e.g., due to advances in machine learning or human effort*	*Processing and display systems are hosted by external organizations. Hypothetically, nodes could be run at NASA, but with issues of security, cost, and/or redundancy*	*SeaBASS columns describe each ROI, e.g.: biovolume, min & max feret diameter, ID # (e.g. WoRMS ID), "permalinks" point to #2*	Satellite products: Abundance (cell L^-1) Biovolume (um^3 L^-1) Carbon (ug L^-1) Calculated by taxa or size e.g., diatom_carbon, total_phyto_abundance

Preserve raw data – outline how SeaBASS metadata header includes core information, but had to be supplemented for the specific needs of these data types

/water_depth=15.0
/measurement_depth=4
/instrument=PLACEHOLDER for GCMD instrument best matched to IFCB; note we also considered adding /instrument_model= and /instrument_manufacturer=
! Instrument, Model Number: Imaging FlowCytobot, IFCB010
! Instrument settings that affect types and sizes of particles imaged: e.g., images triggered by autofluorescense (and/or scattering) and tuned for particles on order of 10 to 100 um; FlowCAM can be operated in either fluorescence-triggered mode or auto trigger (auto image) mode
/volume_sampled=5[ml]
/volume_imaged=2.09[ml]
/pixel_per_um=3.4
! Image processing method and version: for IFCB we can refer to a code library with documentation (https://github.com/hsosik/ifcb-analysis/tree/master/feature_extraction); for FlowCAM proprietary VisualSpreadsheet software
! Biovolume calculation method: we had discussed citing an article e.g. "distance map algorithm, originally developed for IFCB", or based on equivalent spherical diameter using the area provided in the data table; note from Aimee her software can compute 4 different ways, some discussion needed among PIs for which to use for FlowCAM
! Automated classification method and version: should refer to the machine learning approach, and a code library (e.g., https://github.com/hsosik/ifcb-analysis/tree/master/classification)
! Interpretation for automated classification: e.g.,"top score wins", "wins above adhoc threshold" would be fine for Random Forest, but CNN outputs weights not probabilities so thresholding does not make sense
/contributor_namespace_file_name=Sosik_lab_namespace_v1.csv
! We are considering a yaml file that could be inserted into the header to represent the contributor namespace table
! IDs for all categories assessed per namespace for identification_manual; namespace prefix given here (repeat if also providing for identification_automated); we'd want this as structured metadata once SeaBASS can ingest at "/"
! ptwg:bead,detritus,bubble
! sosik: ciliate_mix, Chaetoceros, Corethron, Guinardia, mix
!/url_source=https://ifcb-data.whoi.edu/mvco/D20170505T153648_IFCB010.html