Which, where, and when?:

Sharing data products from automated *in situ* observations of plankton

GlobalHAB Symposium

23 August 2022

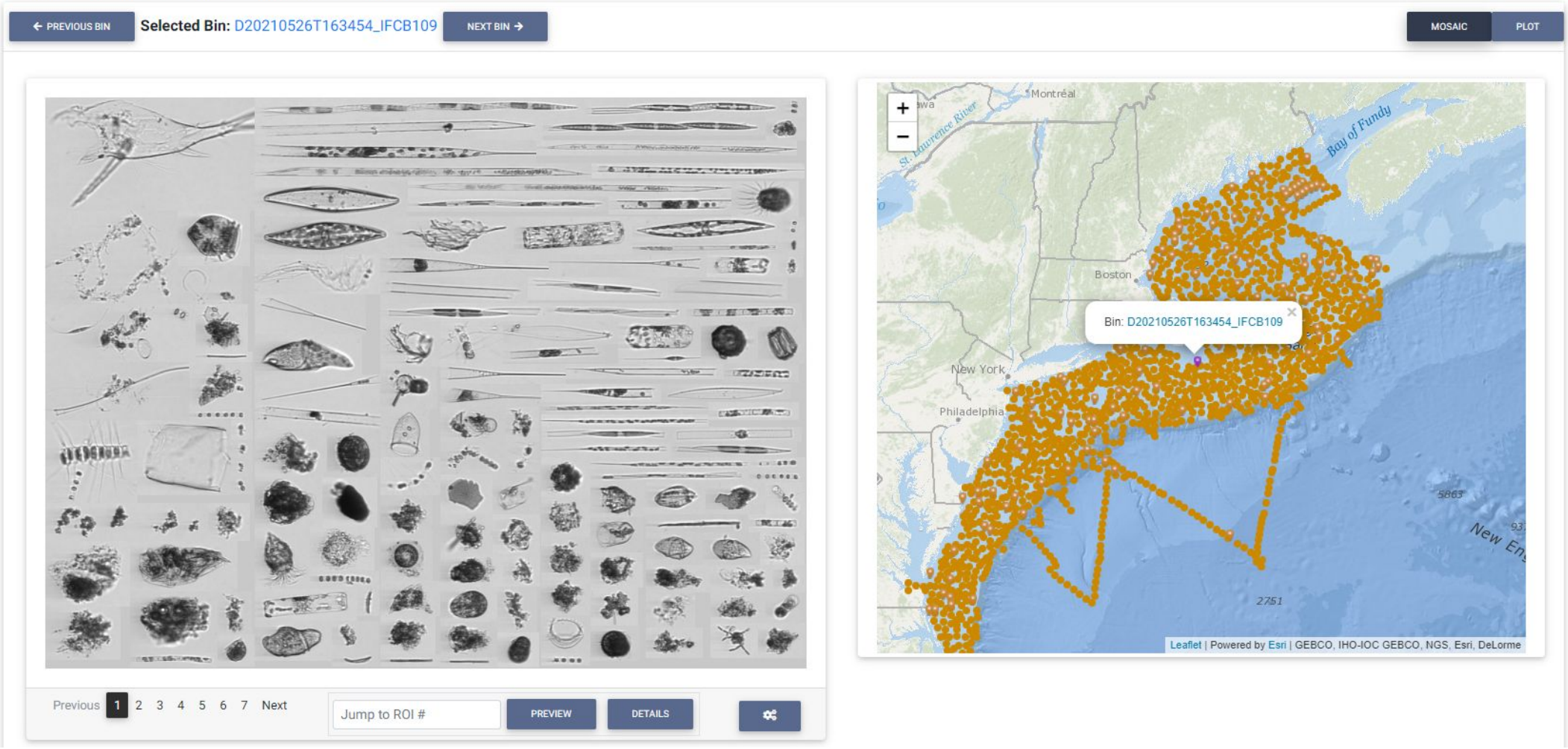
Stace Beaulieu (stace@whoi.edu)

WOODS HOLE OCEANOGRAPHIC INSTITUTION

https://tinyurl.com/sharing-HAB-data

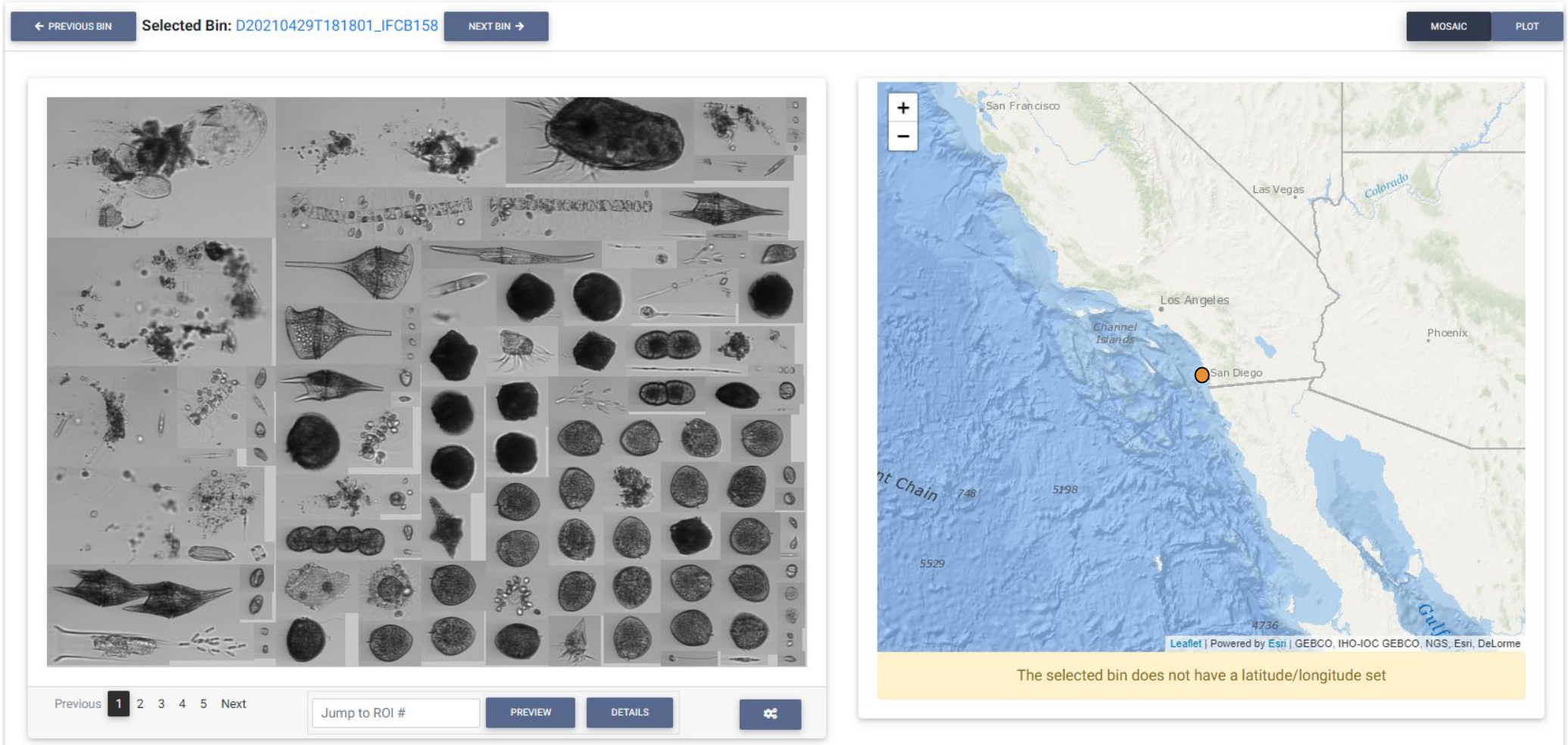
Which? Where?

https://ifcb-data.whoi.edu/timeline?dataset=NESLTER_broadscale&bin=D20210526T163454_IFCB109



When?

https://ifcb.caloos.org/timeline?dataset=del-mar-mooring&bin=D20210429T181801_IFCB158



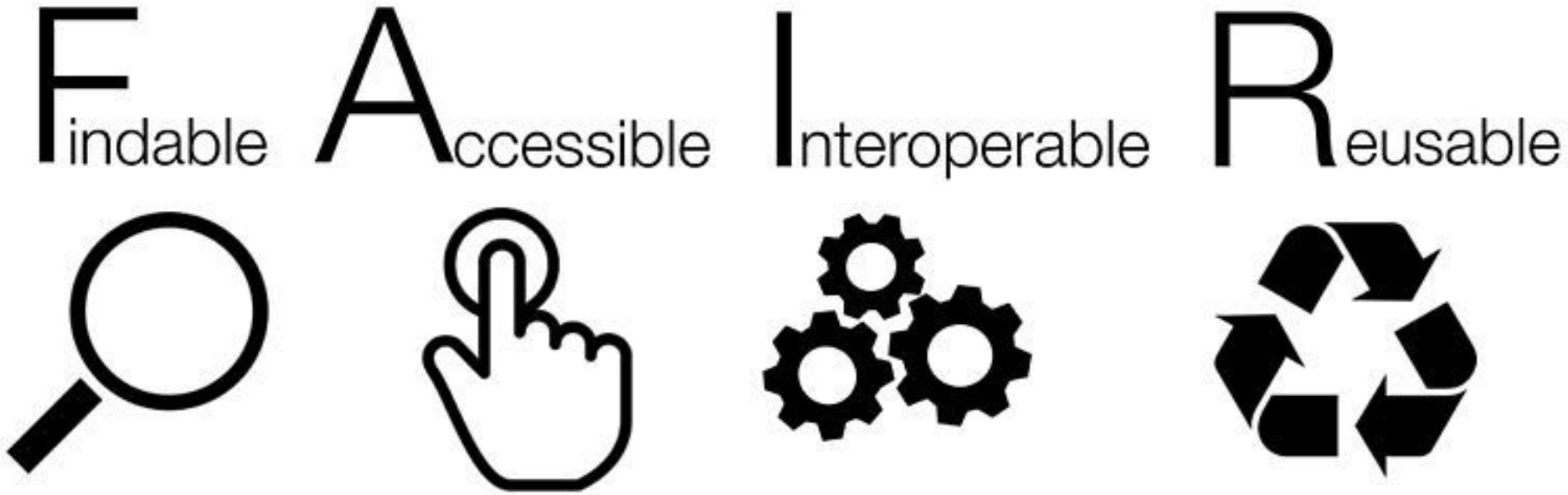
How?

Goal of this talk is to provide practical guidance how to share data products from automated *in situ* observations of plankton,

with a focus on standardizing data

https://tinyurl.com/sharing-HAB-data

Why?



https://www.go-fair.org/go-fair-initiative/

Image CC-BY Sangya Pundir

How? Ocean Best Practices Repository

https://www.oceanbestpractices.org/repository/

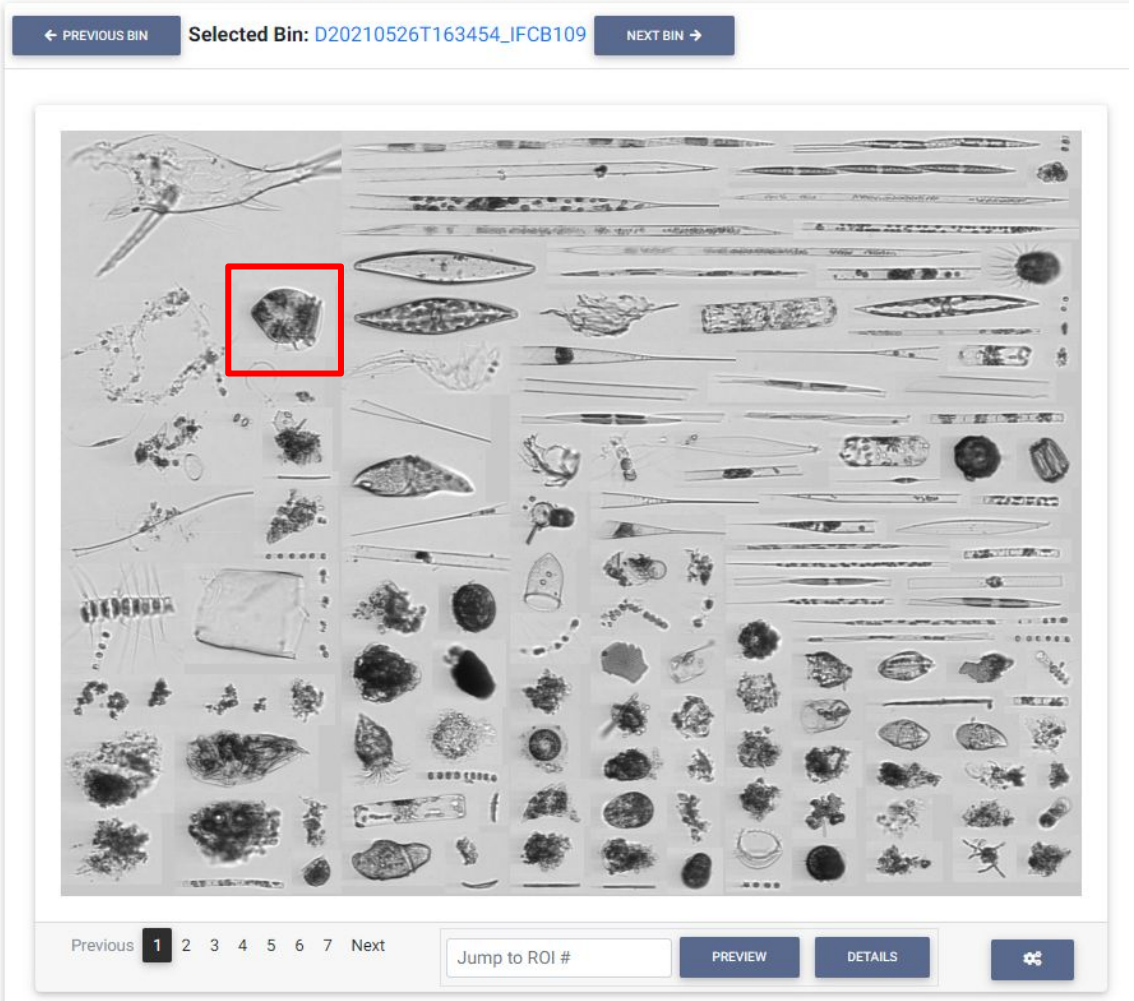
*

- Neeley, A., et al. (2021) Standards and practices for reporting plankton and other particle observations from images. Technical Manual. Ocean Carbon and Biogeochemistry Project Office.
- Martin-Cabrera, P., et al. (2022) Best practices and recommendations for plankton imaging data management: Ensuring effective data flow towards European data infrastructures. Version 1. JERICO-S3 project.
- Horton, T., et al. (2021) Recommendations for the Standardisation of Open Taxonomic Nomenclature for Image-Based Identifications. *Front. Mar. Sci.*

* discussion planned for Thursday

Example: Images from Imaging FlowCytobot (IFCB)

https://ifcb-data.whoi.edu/timeline?dataset=NESLTER_broadscale&bin=D20210526T163454_IFCB109



Let's look at a common workflow *

Autoclassifier to assign preliminary label per image

Manual annotation to verify auto-label or assign other label (with greater certainty)

Standardize to a taxonomic database (may prefer higher taxonomic rank for certainty)

Further standardize data/metadata to share with a community data repository

* note that Bengt showed 'Data flow and production of classifiers' in talk earlier today

Why standardize to a taxonomic database?

Image CC-BY CindyLouPhotos

Why standardize to a taxonomic database?

Table 2. Old and recent names of phytoplankton and zooplankton.

Old name	Recent name
Dactylococcopsis raphidioides	*Monoraphidium contortum*
Gomphosphaeria sp.	*Snowella* sp./*Woronichinia* sp./*Gomphosphaeria* sp.
Oscillatoria agardhii	*Planktothrix agardhii*
Oscillatoria limnetica	*Pseudanabaena limnetica*
Lyngbya limnetica	*Planktolyngbya limnetica*
Cyclotella laevissima	*Cyclotella meneghiniana*
Rhizosolenia minima	*Chaetoceros minimus*
Euglena spp.	*Euglena* spp. and/or *Eutreptiella* spp.
Richteriella botryoides	*Micractinium pusillum*
Tintinnopsis relicta	*Tintinnopsis fimbriata*
Brachionus pala	*Brachionus calyciflorus*
Triarthra longiseta	*Filinia longiseta*
Anuraea spp.	*Keratella* spp.

Finni, T., et al. (2001) *AMBIO: A J. of the Human Environment*

Why standardize to a taxonomic database?

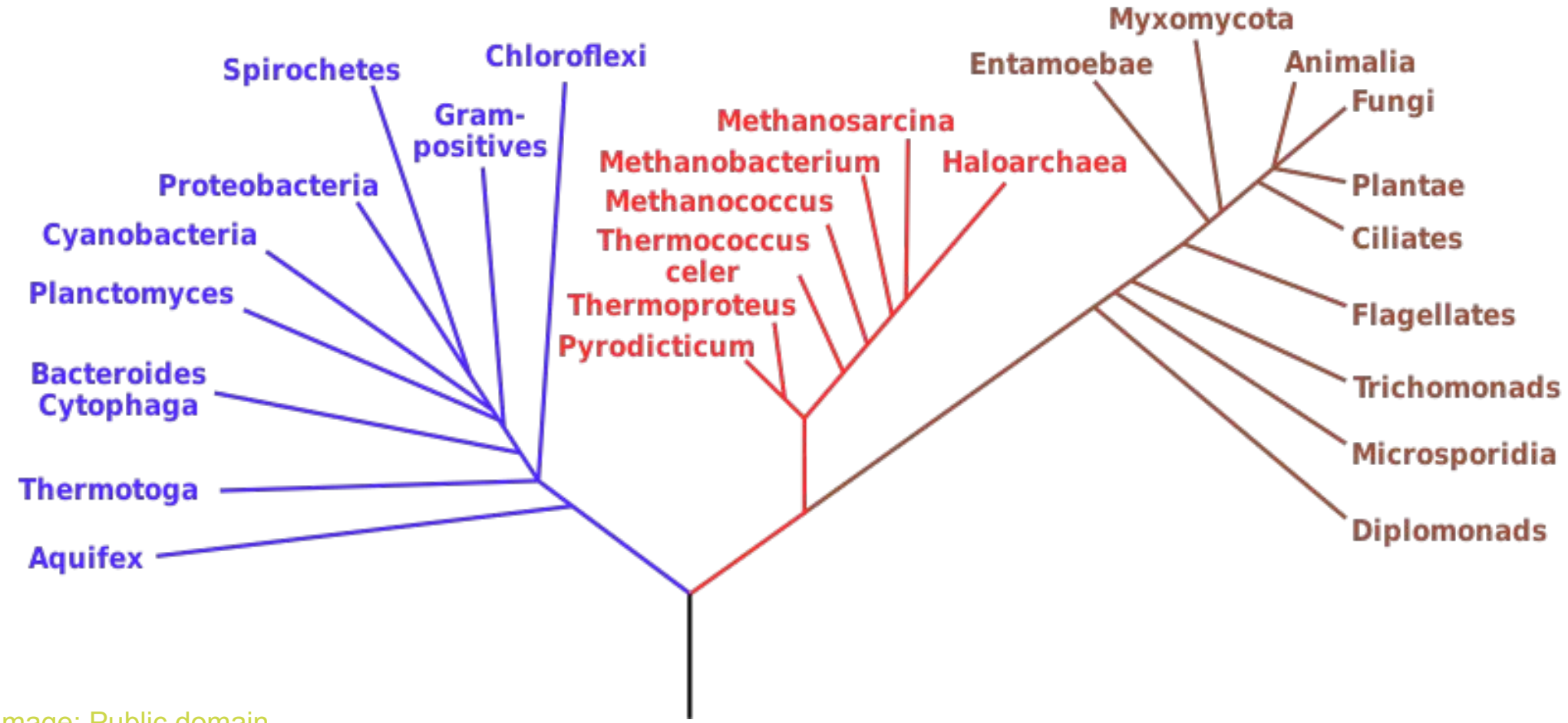


Image: Public domain

How? Standardize to a taxonomic database

- **Select a taxonomic database used by your community**
  - World Register of Marine Species (WoRMS) used by UNESCO's Harmful Algal Information System and Global HAB Status Report
  - IOC-UNESCO Taxonomic Reference List of Harmful Micro Algae http://www.marinespecies.org/hab/index.php
- **Community data repository may require or recommend a particular taxonomic database**
  - WoRMS is required by Ocean Biodiversity Information System (OBIS)
  - OBIS HAB node https://obis.org/node/33dec23c-af65-4fb1-a437-79543c562ef0

How? Standardize to a taxonomic database

- **Provide name from taxonomic database:**
  - e.g., you can use WoRMS Taxon Match Tool
- **Provide the paired identifier and/or kingdom:**
  - so that automated tools can ‘check’ against names that are in multiple branches of the tree of life
- **Retain your original label:**
  - often is a more descriptive name for a morphospecies

How? Standardize to a taxonomic database

- **Provide name from taxonomic database:**
  - scientificName
- **Provide the paired identifier and/or* kingdom**:**
  - scientificNameID , and/or*
  - kingdom**
- **Retain your original label:**
  - verbatimIdentification

* OBIS requires scientificNameID

** Global Biodiversity Information Facility (GBIF) recommends kingdom

How? Further standardize to share with repository

- **The specific terms on previous slide are from the Darwin Core standard,**

  used by Ocean Biodiversity Information System (OBIS) and the Global Biodiversity Information Facility (GBIF)

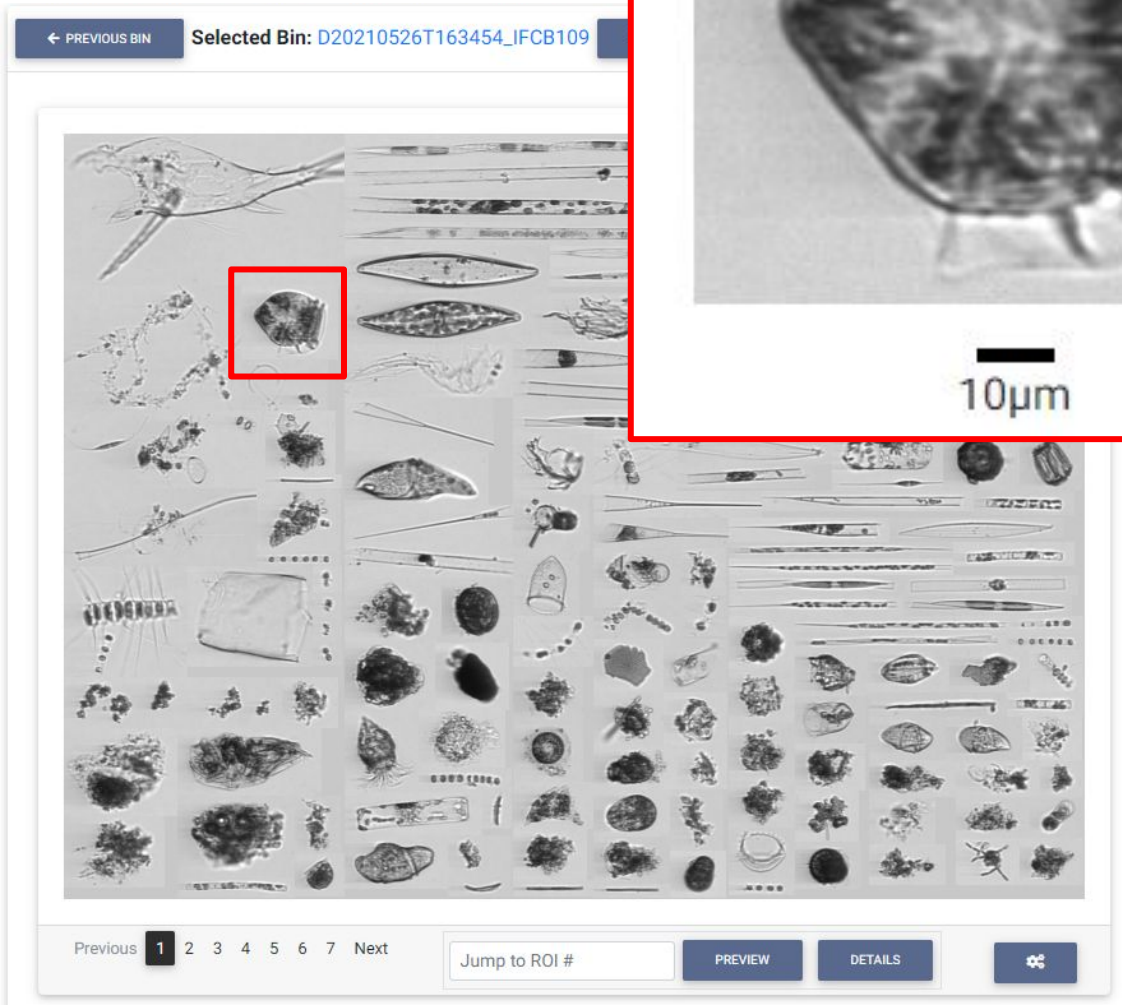
How? Further standardize to share with repository

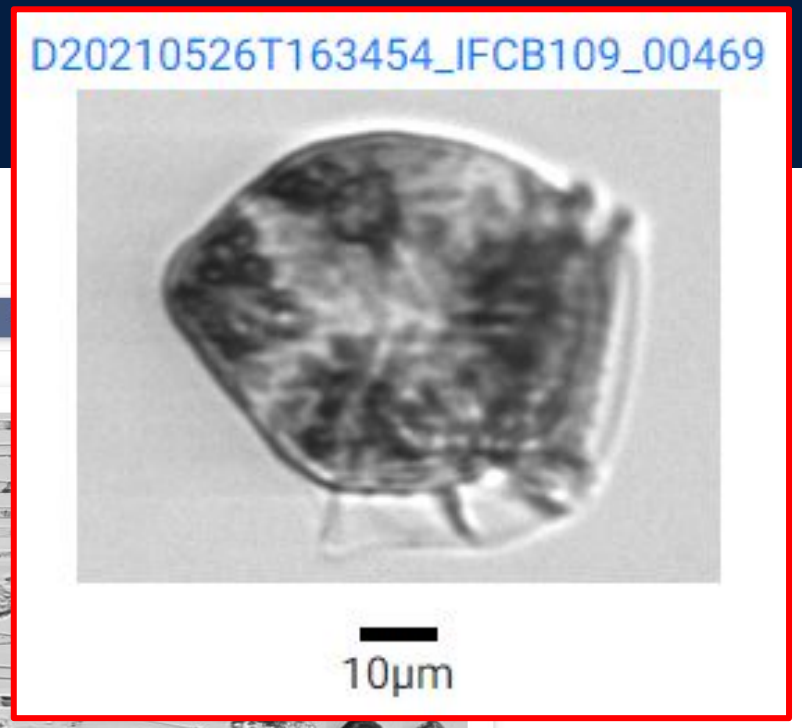
- **Report whether an automated identification has been verified:**
  - identificationVerificationStatus
    - PredictedByMachine, or
    - ValidatedByHuman

      Martin-Cabrera et al. (2022)
- **Explain the uncertainty of an image-based identification:**
  - identificationQualifier
    - select Open Nomenclature (ON) sign from flowchart

      Horton et al. (2021)

Example





Dinophysis_norvegica

⇩

(no manual annotation)

⇩

Dinophysis
urn:lsid:marinespecies.org:taxname:109462

⇩

- identificationVerificationStatus:
  - PredictedByMachine
- identificationQualifier:
  - stet.

But what about Where? When?

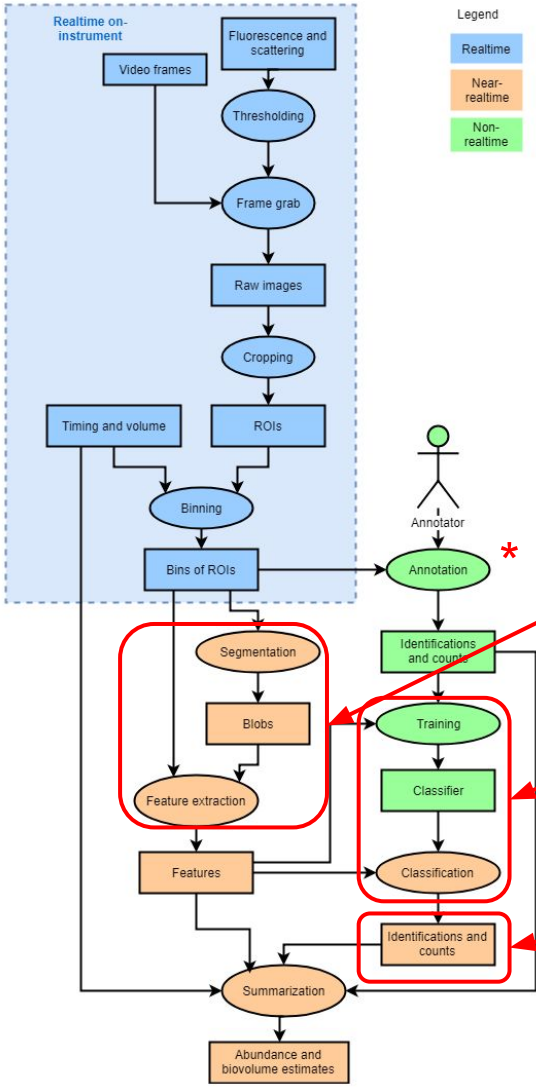
Main points since we have limited time today:

- Data for the Which, Where, and When tend to be generated and stored in different infrastructure, thus we often need a workflow to pull together these data
- Quick tutorials from recent OBIS-USA bio_mobilization_workshop for:
  - Where? See “Getting lat/lon to decimal degrees”
  - When?: See “Getting your dates in order”
- EurOBIS template highly recommends including depth*

* OBIS does not require minimumDepthInMeters or maximumDepthInMeters

When? Not just when sample was collected…

Imaging



"Level 2" data

[IFCB workflow chart]

Version Image processing software

Version Autoclassification software

Version Matching to taxonomic database

* see Bengt's talk Friday re: sharing of annotated images for classifier development

Example for IFCB published in community repository

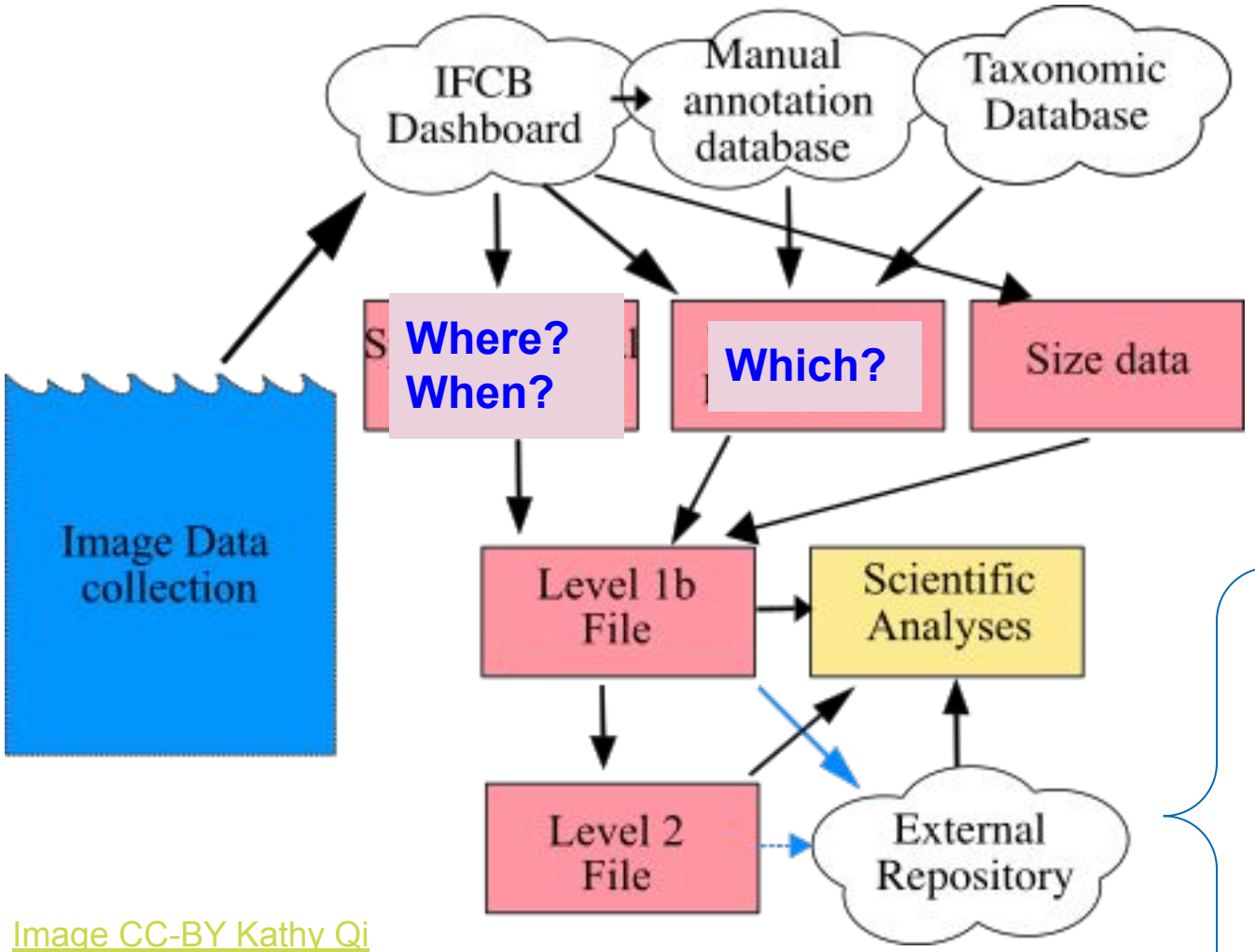


Image CC-BY Kathy Qi

Sosik, H., et al. (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 ver 1. Environmental Data Initiative.

Example for IFCB to OBIS

IFCB_OBIS_with_manual_annotation.ipynb

- Jupyter Notebook with code in Python
- prototype for testing purposes
- output tables for Darwin Core Archive package*

Out[15]:

occurrenceID	verbatimIdentification	identifiedBy	scientificNameID	scientificName	kingdom	eventID
D20180406T033616_IFCB115_00003_Bacillariophyceae	Bacillariophyceae	hsosik	urn:lsid:marinespecies.org:taxname:148899	Bacillariophyceae	Chromista	D20180406T033616_IFCB115
D20180406T033616_IFCB115_00013_ciliate	ciliate	hsosik	urn:lsid:marinespecies.org:taxname:1348	Spirotrichea	Chromista	D20180406T033616_IFCB115
D20180406T033616_IFCB115_00016_Euglena	Euglena	hsosik	urn:lsid:marinespecies.org:taxname:8012	Euglena	Protozoa	D20180406T033616_IFCB115
D20180406T033616_IFCB115_00019_Chaetoceros	Chaetoceros	kkenitz	urn:lsid:marinespecies.org:taxname:148985	Chaetoceros	Chromista	D20180406T033616_IFCB115

* include Event and Occurrence (not yet Extended Measurement or Fact)

In summary: Sharing data products from automated *in situ* observations of plankton

- **Standardizing will help make your data FAIR (Interoperable and Reusable)**
- **Consider using the Darwin Core standard used by OBIS and GBIF**
  - Refer to best practices [links in previous slides]
  - Refer to OBIS and GBIF manuals
  - Tutorials from OBIS workshops
  - Join a community of practice, such as Standardizing Marine Bio Data (SMBD) Working Group

Acknowledgements

Funding from NOAA

- At GlobalHAB Symposium: Talk with Raphe and Kasia (see their presentation on Thursday) and Mike to learn more about PCMHAB project

Funding from NSF

Thanks to Standardizing Marine Bio Data (SMBD) Working Group!

Thanks to Karina and Trevor @ Axiom!

Thanks to Sosik lab and collaboration with NOAA NEFSC EcoMon!

Relevant to other communities with automated *in situ* observations of plankton

- Boss, E., et al. (2022) Recommendations for Plankton Measurements on OceanSITES Moorings With Relevance to Other Observing Sites. *Front. Mar. Sci.*, https://doi.org/10.3389/fmars.2022.929436
- Clayton, S., et al. (2022) Bio-GO-SHIP: The Time Is Right to Establish Global Repeat Sections of Ocean Biology. *Front. Mar. Sci.*, https://doi.org/10.3389/fmars.2021.767443

EuroGOOS Biological Observations Working Group

https://eurogoos.eu/biological-observations-working-group/

Key Objectives include:

> “The standardisation and SOPs for Imaging and evaluating how these data may contribute the most effectively to EOVs (Essential Ocean Variable) and EBVs (Essential Biodiversity Variables).”

xx

IOOS Bio Data Mobilization Workshop

Marine Data Mobilization Workshop for Biology and Ecosystem Essential Ocean Variables (Bio-Eco EOV)

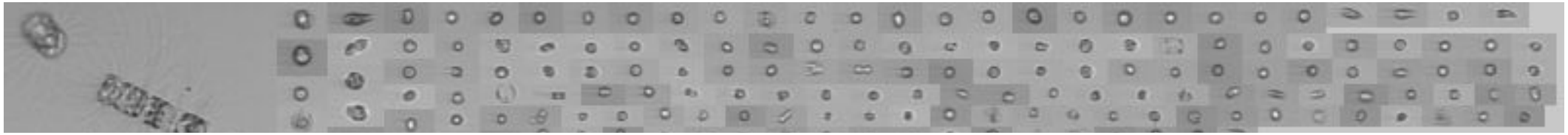
March 14-15, 2022 (virtual)

https://github.com/ioos/bio_mobilization_workshop

slide from previous version presentation

My goal for this workshop:

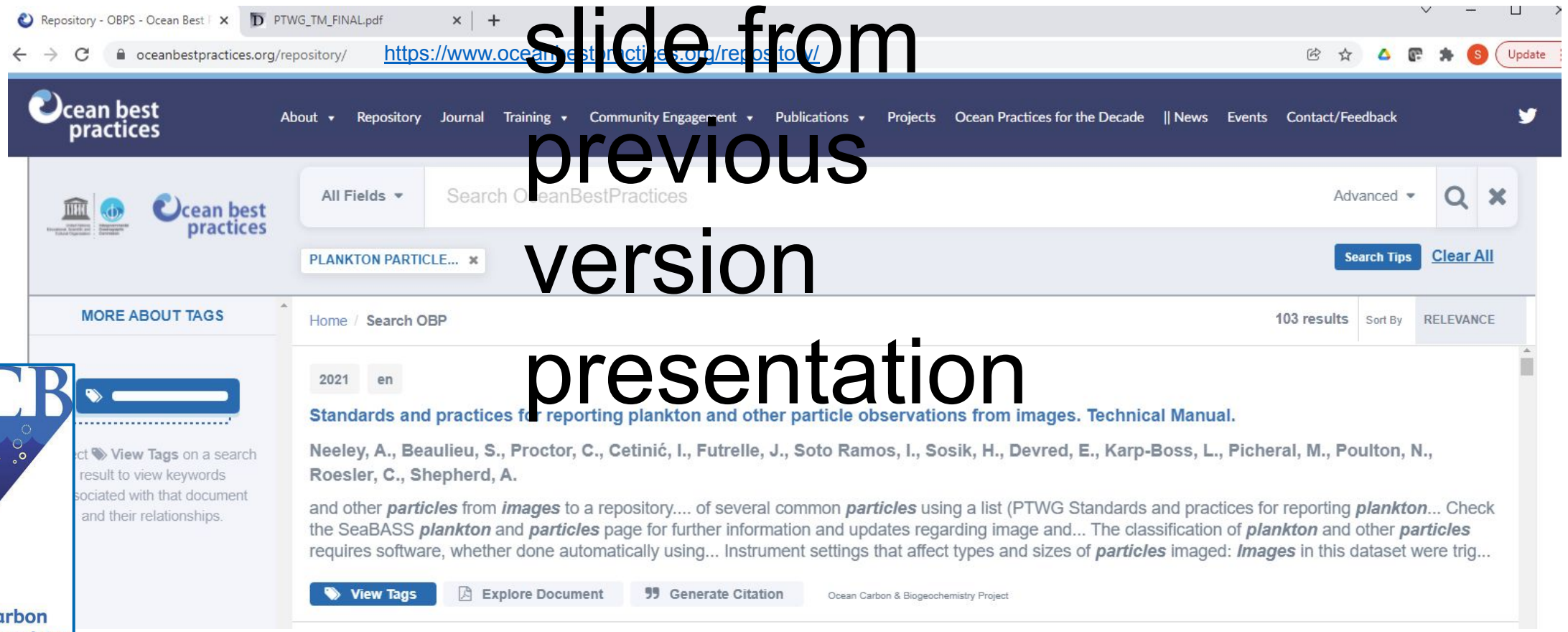
Create demo notebook showing how to transform from OCB PTWG data model + metadata into tables for Darwin Core Archive Event Core with OBIS-ENV-DATA approach



https://sccoos-ifcbdb.srv.axds.co/timeline?dataset=CA-IFCB-158

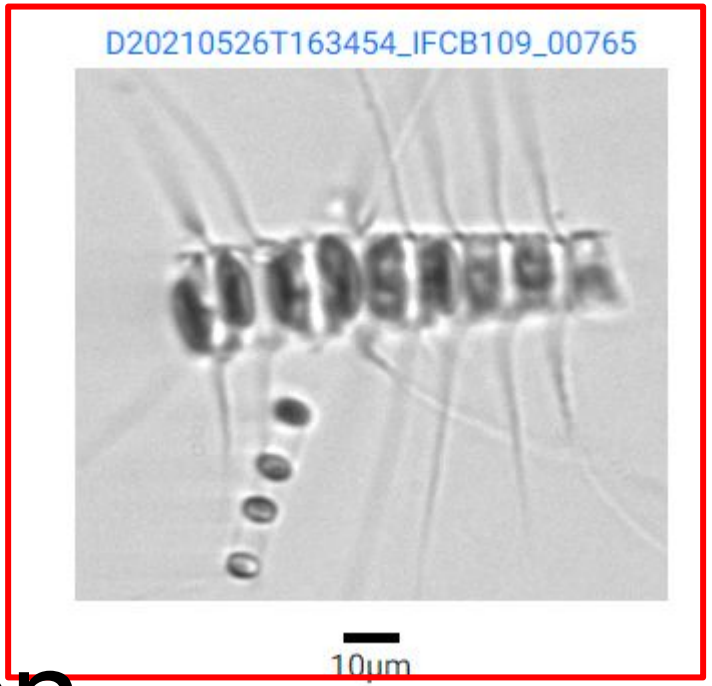
Data model to accommodate taxa and sizes

Developed by [Ocean Carbon & Biogeochemistry Phytoplankton Taxonomy Working Group]



slide from previous version presentation

Data model to accommodate taxa and sizes



slide from previous version presentation

Highest autoclass score:	Dinophysis_norvegica	Chaetoceros_didymus
Max/Min Feret Diameter:	63 um / 51 um	82 um / 58 um
Major/Minor Axis:	60 um / 50 um	84 um / 39 um
Equiv. Diameter:	54 um	49 um

Data model to accommodate taxa and sizes

Neeley et al. (2021)

"Standards and practices for reporting plankton and other particle observations from images"

slide from previous version presentation

Image identifier	Type					Size					
	Automated				Manual	8	9	10	11	12	13
	class label used by lab group	taxonomic name/ID pair from reference database									
D20210526T163454_IFCB109_00469	Dinophysis_norvegica	(match to WoRMS)						63	51		54
D20210526T163454_IFCB109_00765	Chaetoceros_didymus	(match to WoRMS)						82	58		49

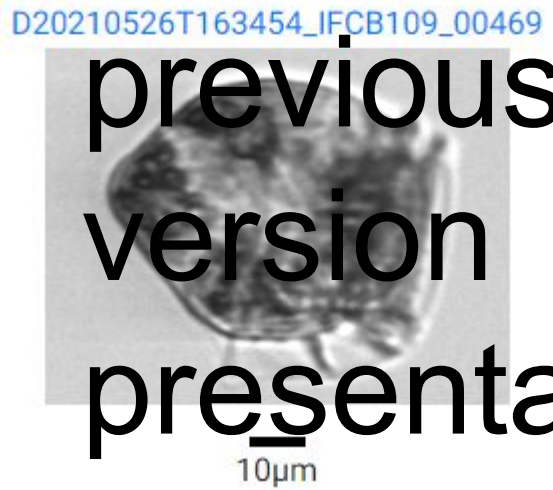
“Level 1b” data table

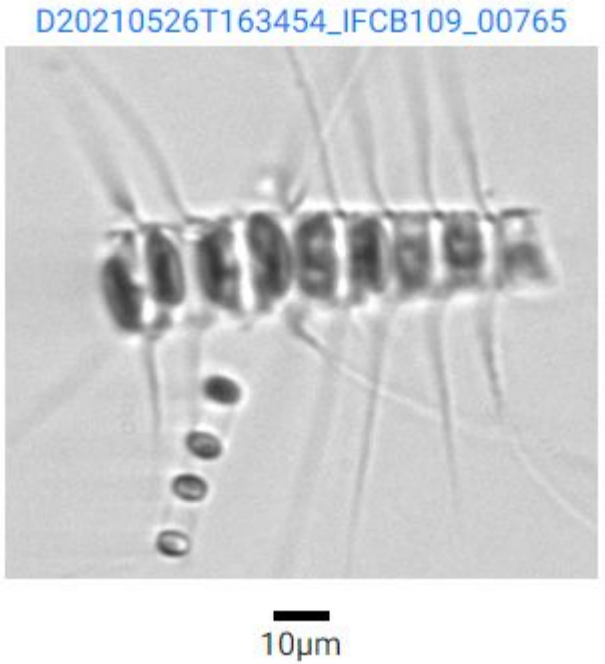
8. biovolume
9. area_cross_section
10. length_representation
11. width_representation
12. equivalent_spherical_diameter
13. area_based_diameter

Not (yet) accommodated by data model

- Includes size metrics, but other morphological metrics may be of interest (e.g., shape)
- Does not explicitly include uncertainty

slide from previous version presentation





Highest autoclass score:	Dinophysis_norvegica	Chaetoceros_didymus

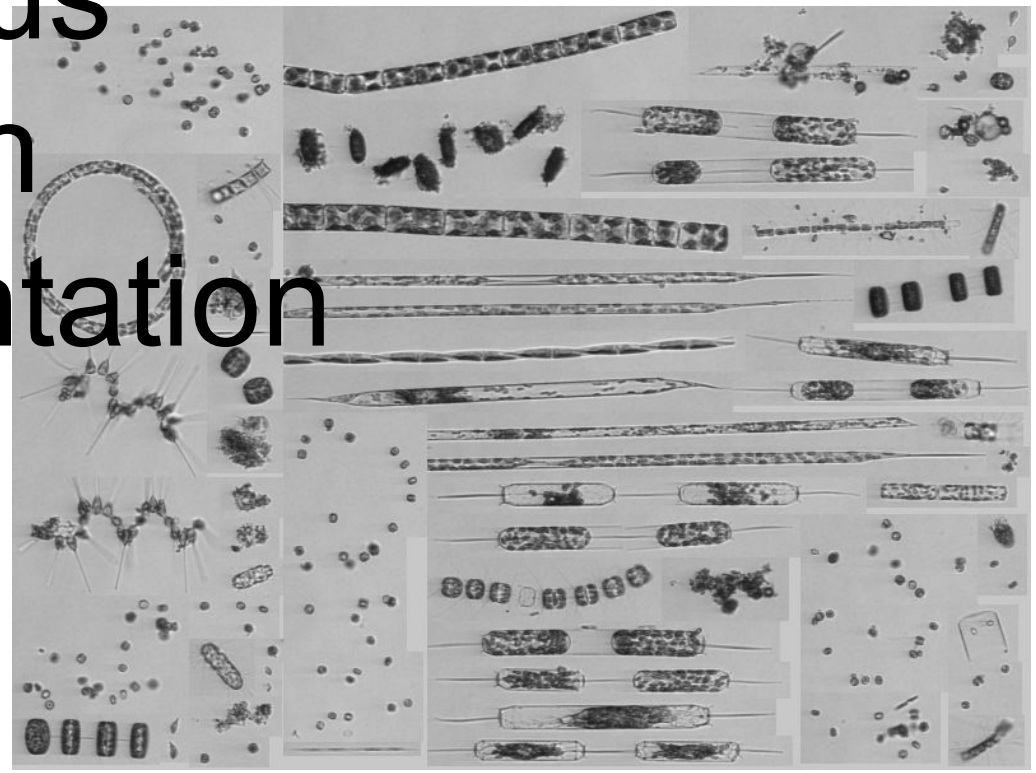
Data package published in EDI

First data published with this data model to Environmental Data Initiative (EDI) repository:

Sosik et al. (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”

slide from previous version presentation

- Additional columns added to better standardize to Darwin Core as step towards sharing with Ocean Biodiversity Information System (OBIS)



Data submitted to BCO-DMO

More (and more!) data products from different types of plankton imaging systems, including IFCB, ZooSCAN, ISIIS, and more

BCO-DMO participating in discussions to match size metrics to controlled vocabularies

- *Example:* Matching between terms used in EcoTaxa, Darwin Core, and BODC vocabularies

slide from previous version presentation

Dataset Term	Dataset term Definition	Match Type	Term URI
object_esd	Object Equivalent Spherical Diameter	skos:broadMatch?	https://vocab.nerc.ac.uk/collection/S06/current/S0600260/

BCO-DMO
Biological & Chemical Oceanography Data Management Office