A WORLD OF DATA: THE JOYS OF OCEANOGRAPHIC TIME-SERIES DATA

Cyndy Chandler

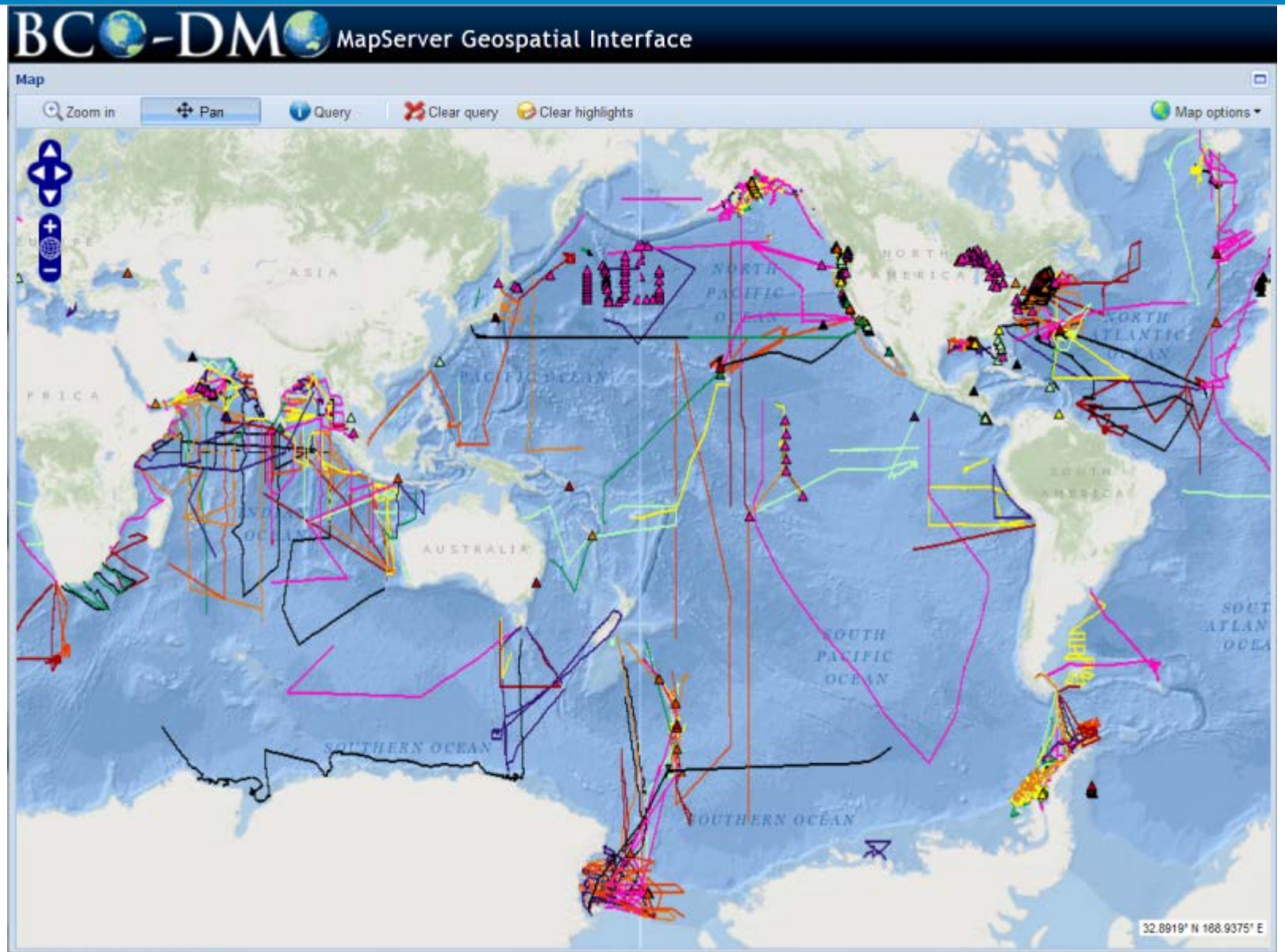
Woods Hole Oceanographic Institution

An International Time-series Methods Workshop

Bermuda Institute of Ocean Sciences

28-30 November 2012

THE WONDERFUL WORLD OF DATA



U.S. NSF-FUNDED RESEARCH DATA

BCO-DMO

Biological & Chemical Oceanography Data Management Office

The BCO-DMO mandate is to provide data management support throughout a research project for investigators funded by NSF OCE Biological and Chemical Oceanography Sections or NSF OPP ANT Organisms & Ecosystems Program, with the goal of improving access to NSF funded research data. BCO-DMO icludes staff from former U.S. JGOFS and U.S. GLOBEC programs.

BATS, CARIACO & HOT DATA

Time-series data from BCO-DMO:

- CARIACO data were added in 2005
- Recently added HOT and BATS Niskin data
- Other data types to be added (hopefully by end of February 2013)
  - Biogeochemistry
  - Primary Production
  - Phytoplankton
  - Zooplankton
  - Particulate Flux

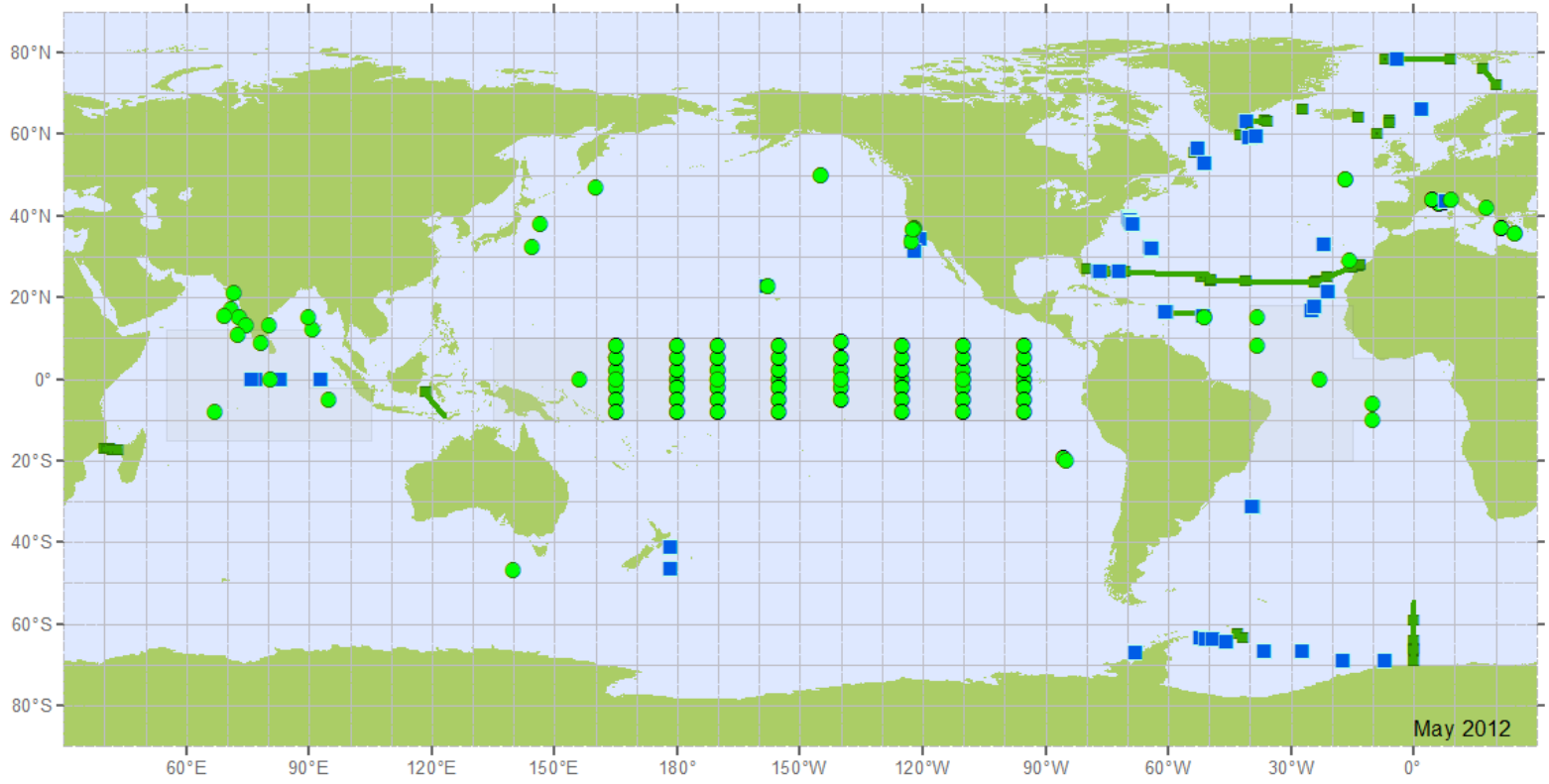
Terry McKee (WHOI)

OceanSITES

Taking the pulse of the global ocean

A worldwide system of deepwater reference stations providing:

The full depth of the ocean



OceanSITES Status Map - Operating Sites

OceanSITES Moorings and Observatories (139)

- OPERATING Real Time Data (92)
- OPERATING Delayed Mode Data (47)

Transport sites (16)

- OPERATING
- Transport Stations

OceanSITES

Taking the pulse of the global ocean

jcommops

JCOMM in-situ Observing Platform Support centre

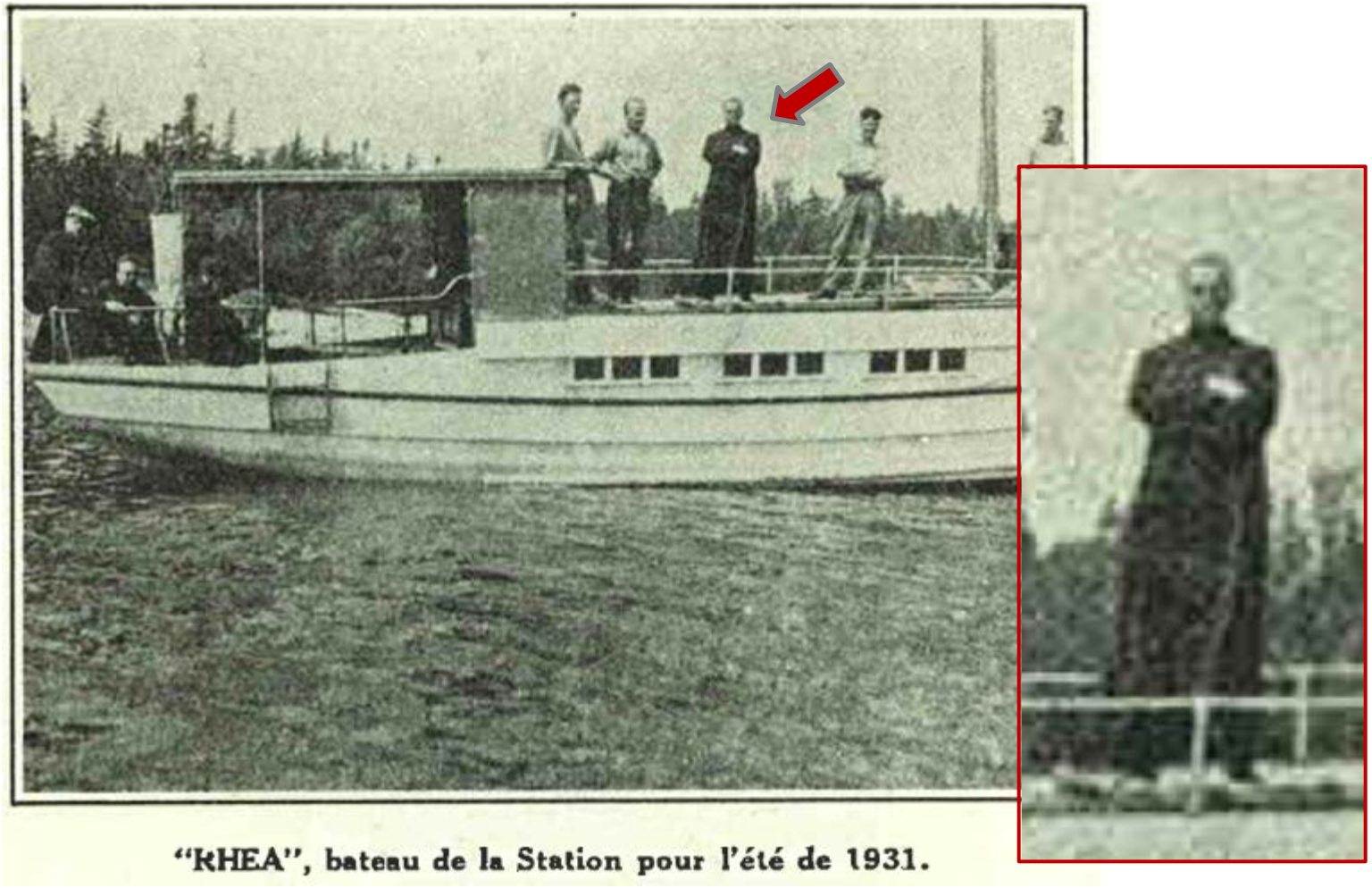
THE WONDERFUL WORLD OF DATA

" You can't play with the data without the metadata. Well, you can, but it's much less fun. " (Peter Wiebe, WHOI, 2009)

So title of this talk becomes …

WORLD OF DATA:
THE JOYS OF WELL-MANAGED
OCEANOGRAPHIC TIME-SERIES DATA

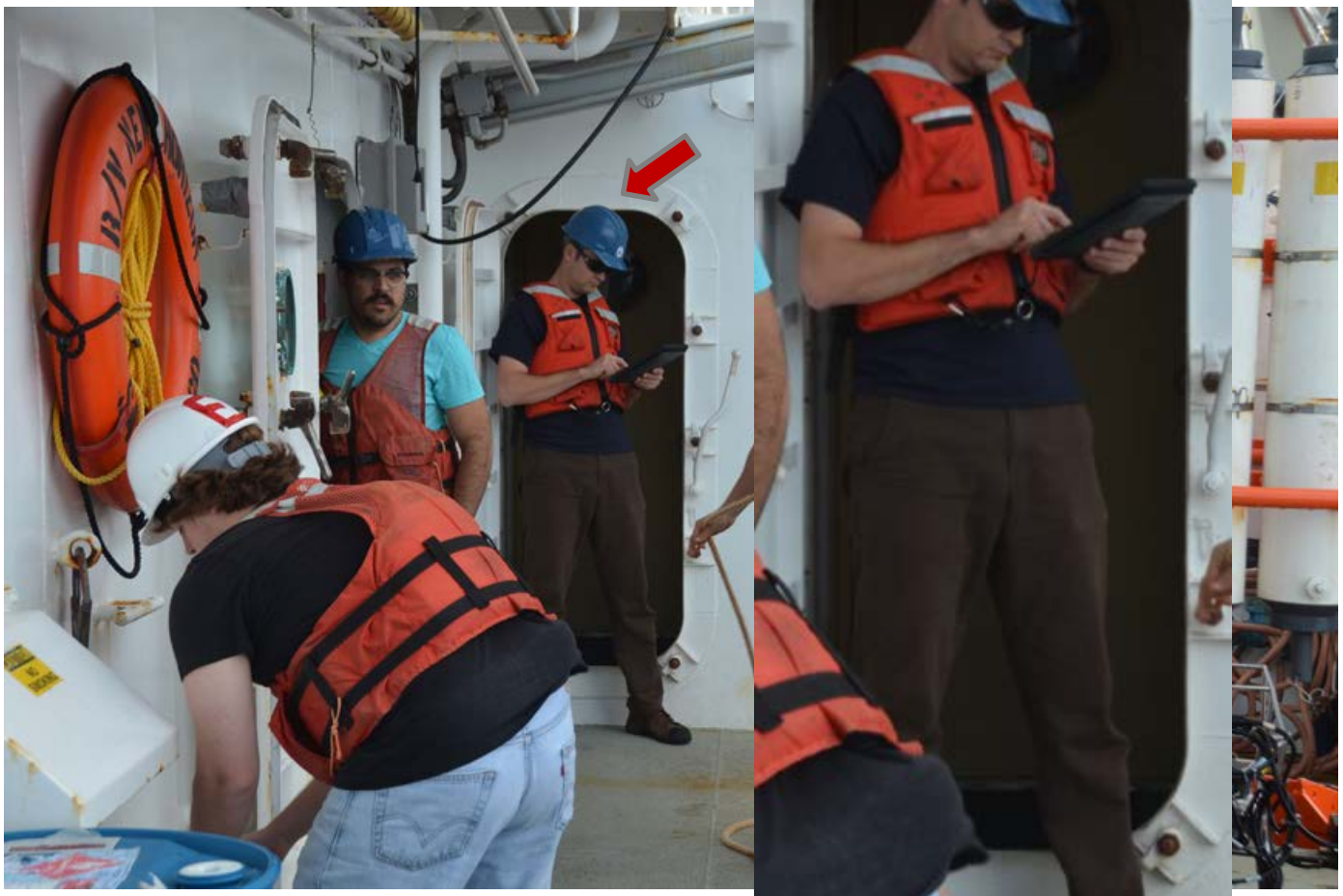
DATA MANAGEMENT ~ then



"RHEA", bateau de la Station pour l'été de 1931.

Vessel RHEA and the "metadata monks", summer of 1931

DATA MANAGEMENT



Logging the CTD/Niskin cast event during NH1208

DATA LIFE CYCLE

Good data management practices account for all aspects of the data cycle, featuring a stewardship philosophy in all phases from “proposal to preservation”.

DATA MANAGEMENT GOALS

- Support the research needs of program
- Protect and preserve the data

- Provide access to the data for colleagues
- Enable accurate use and re-use of data

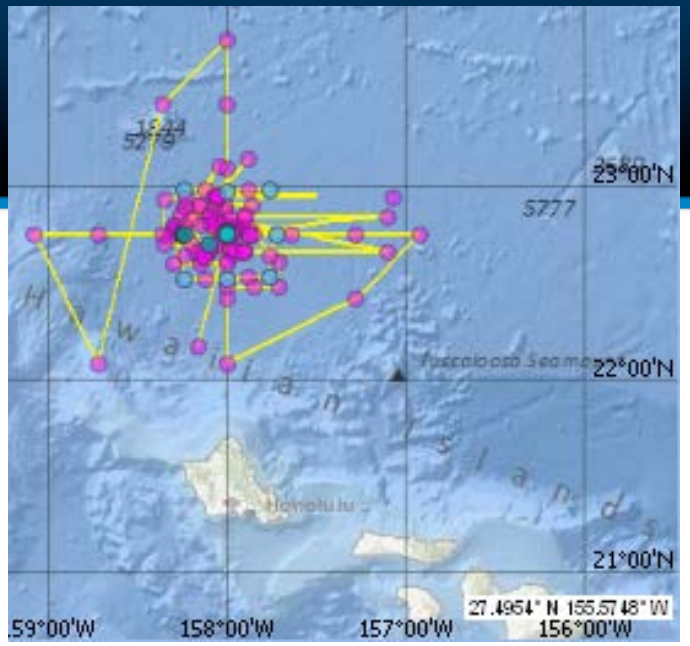
PROPOSAL

- Be aware of data policies associated with funding, including recent changes
- Understand the expectations
  - funding agency
  - research program
  - national

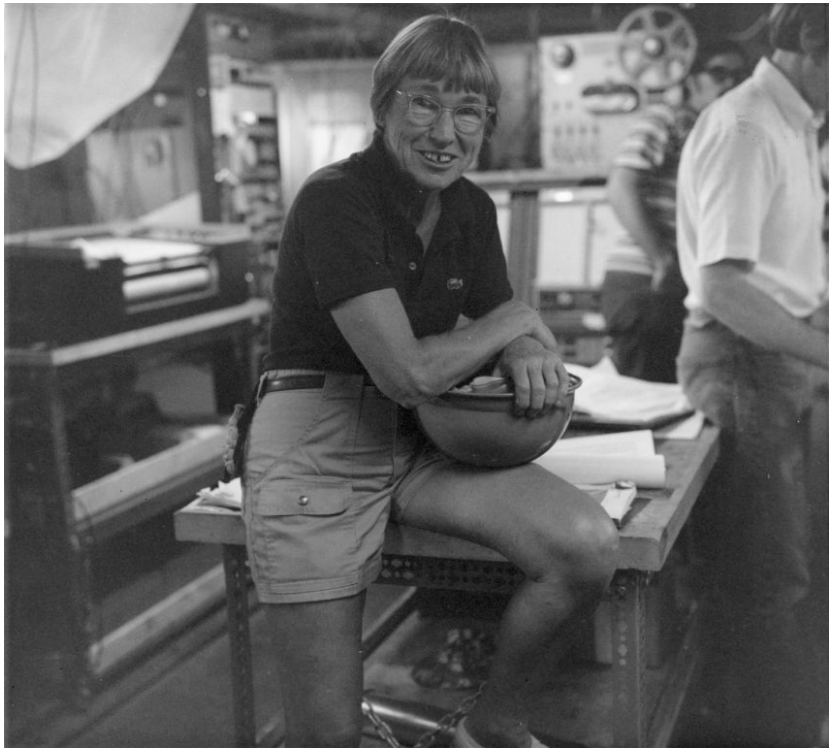
DATA ACQUISITION

- Document protocols & procedures
- Event log of sampling activities
- Controlled term lists
  - instruments, actions, people names
  - names of measurements

DATA ACQUISITION



- Station plan
- Allocation of sample (wire) time
- Allocation of sample water



Arrangements should be made prior to the start of sampling and then reviewed periodically. Set the expectations, and understand the resources needed to meet them.

Have a plan, write it down, communicate it early and often.

SAMPLING PLAN

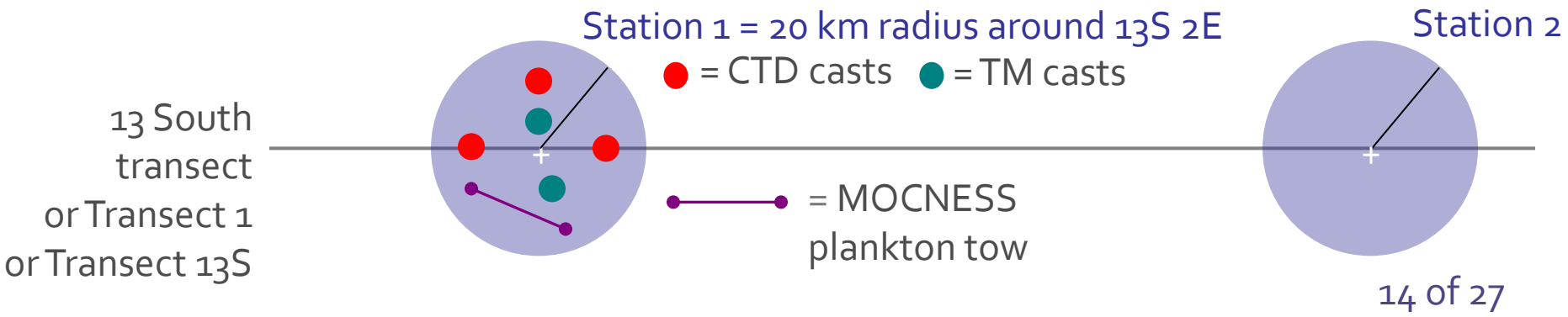
determine the sampling plan in advance, consider using these concepts, and explain the plan to everyone involved

transect – a series of sampling locations often in a line

site – a fixed sampling location (e.g. time-series or benthic)

station – a geographic sampling location, e.g. anything within a 20 km radius of a surface (latitude/longitude) position

cast or tow – deployment of a sampling device



Log Sheet per Sampling Device

CTD/Rosette cast

Cruise: EDDIES | Leg: 1 | Cast: 8 | Type: A1 (0-700m)

Date: 16 Jun 04 | Time: 0253 | Lat: 33 41.809 | Long: 63 9.912 | Samplers: Nathan, Sarah, Grace, Dennis, Tom

Date: 16 Jun 04 | Time: 0331 | Lat: 33 41.84 | Long: 63 9.74

N #	Depth	Niskin Temp	Helium	Oxygens		DIC/Alk	TOC/TON	Salts	Nuts	Bact	FRRF	POC/PON	HPLC	Chla	Flow Cyt	DoP	HP Headwall	$^{15}N/\phi$
				$O2_1$	$O2_2$ $O2_3$							Vol=	Vol=					
1	0																	7
2	0	23.8		1		91	91		91		89 13					6	1	/14
3	0											73 1	67 -1	67 -1	67			
4	20	23.4		2		92	92		92		88 12						2	
5	20											74 2	68 -2	68 -2	68			
6	40	21.2		3		93	93		93		87 11					7	3	
7	40											75 3	69 -3	69 -3	69			
8	50	20.9		4		94	94		94		86 10		70 -4	70 -4	70		4	
9	60	20.5		5		95	95		95		85 9						5	
10	60											76 4	71 -5	71 -5	71			
11	70	20.2		6		96	96		96		84 8		72 -6	72 -6	72		6	
12	80	20.0		7		97	97		97		83 7						7	
13	80											77 5	73 -7	73 -7	73			
14	90	19.9		8		98	98		98		82 6		74 -8	74 -8	74		8	/15
15	100	19.6		9		99	99		99		81 5						9	
16	100											78 6	75 -9	75 -9	75			
17	120	19.3		10		100	100		100		80 4							
18	120											79 7	76 -10	76 -10	76			
19	140	19.3		11		101	101		101		79-3							
20	140											80 8	77 -11	77 -11	77			
21	200	19.1		12		102	102		102		78-2	81 9						
22	300	19.1		13		103	103		103			82* 10						
23	500	18.6		14		104	104		104			83* 11						
24	700	16.7		15		105	105		105		77-1	84* 12						
												3L						

What is a ‘Cruise Sampling Event Log?

- a chronological record of all scientific sampling events that happened during a cruise, wherein each sampling event is assigned a unique identifier

Why is an event log important?

- event logs with unique sampling event identifiers help to …
  - integrate observations from the many sampling devices deployed during a cruise
  - understand relative timing between events

ELOG is a freely-available open source digital logbook system

hhttps://midas.psi.ch/elog/

R2R event log: *http://www.rvdata.us/about/eventlog*

shipboard sampling event log

generated automatically using some algorithm

controlled vocabulary

event	date	time	time_L	sta	lon	lat	ev_type	person	activity
0212208	20020121	2208	1108	TEST	-175.220	-53.572	CTD001	nd	CTD001
0230442	20020123	0442	1742	0	-171.480	-55.398	CTD002	Wang	CTD002
0231556	20020123	1556	0456	0	-171.583	-55.407	ZooTow	Landry	ZooplankTow
0232351	20020123	2351	1351	1	-171.521	-55.334	CTD003	nd	CTD003
0240153	20020124	0153	1453	1	-171.490	-55.329	TM001	Wang	TM001
0240356	20020124	0356	1656	1	-171.336	-55.314	CTD004	Bailey	CTD004
0240745	20020124	0745	2045	1	-171.408	-55.335	Pump_Cast	Andrews	PumpCast01
0241133	20020124	1133	0033	1	-171.405	-55.324	TM002	Wang	TM002
0241319	20020124	1319	0219	1	-171.384	-55.333	CTD005	Timothy	CTD005
0241435	20020124	1435	0335	1	-171.385	-56.333	HPT	Tanner	HandPlankTow
0241520	20020124	1520	0420	1	-171.383	-55.337	TM003	Landry	TM003

date, time and position from shipboard system

Event Log Data Sources

- arrangements were made, agreed upon and reviewed at the first science briefing on board . . .

- and everyone agreed on the common data source for:

- date and time
  shipboard network and UTC
  (*not your wristwatch*)

- position information
  decimal degrees lat/lon
  (agree on required precision)

Cruise Report

- basic cruise metadata
  - Cruise ID - a way to identify the cruise
    - KN195-08 (ship, voyage and leg)
    - KM0908 (ship, 2 digit year and sequential voyage for year)
  - dates and ports
- personnel manifest
  - list of everyone on board and contact information
  - their role during the cruise
- data inventory
  - list of who is expecting to collect what data during cruise
- event log
  - list of every device deployment

Data Inventory (list of expected measurements)

Instrument	Measurement	PI_name	co-PI_name
TMR	Bottle O2	Casciotti	Frame;Sieracki
TMR	Nitrate isotopes	Casciotti	nd
TMR	Uptake Expts-Fe Cd Zn Hg Ni	Cox	Saito
CTD	Productivities; selected stations	DiTullio	nd
CTD	Pigments	DiTullio	nd
CTD	Uptake Expts-carbon C14	Ditullio	Riseman
ON_DECK_PUMP	Incubation Expts-Iron;DMSP effects	DiTullio	nd
TMR	N2O	Frame	Casciotti
TMR	Methyl Mercury	Hammerschmidt	nd
CTD	nifH gene expression	Hilton	Zehr;Webb
TMR	FeL	Lam	Buck
MCLANE	Fe-Metal Particulates	Lam	nd
MCLANE	POC	Lam	nd
nd	Aerosol metals	Lamborg	nd
nd	Sediment trap fluxes including metals	Lamborg	nd
TMR	Total Dissolved Mercury	Lamborg	nd
TMR	DOC	Morris	Carlson
CTD	Heterotrophic bacterial counts-act	Morris	nd
CTD	Proteomics	Morris	Rocap
CTD	Pro and Syn phylogeny-ecotype	Rocap	Webb
ON_DECK_PUMP	Incubation Expts-Phosphate	Rocap	nd
LAB	Sampling Event Log	Saito	nd

ANALYSIS & SYNTHESIS

- Backups to prevent data loss
- Document protocols & procedures
  - including quality control
  - use of standards and reference materials
- Cite the procedures document for your time-series
- Sufficient information to support accurate re-use of data

Data Quality

- It’s important to understand that reporting on the ‘quality’ of a dataset (a set of measurements) is not a statement about it’s value to the community.

- Just because a dataset might be ranked lower on some scale used to assess quality, does not mean it is of less value to the community.

- A dataset of known quality is more valuable than one that lacks the quality assessment metadata.

Data Quality (of your data)

Questions asked during sampling and analysis (related to accuracy and precision):

- How “good” do I need the measurement to be? (QA)
- How well did I make the measurement? (QC)

Data Quality Metadata:

- report the questions above and answers with the data
- much of this information still fits in the methods section of the peer-reviewed publication – but the problem is that all the data no longer fit in that same publication
- important to report the data quality information with the published dataset (reported as metadata)

FINAL DATA SETS

- CONTRIBUTE DATA TO A REPOSITORY
  - Consider national and program-specific policies and expectations
  - Building a community resource
  - Data and documentation sufficient to support data use and re-use
- USE & REUSE OF DATA
  - Role as data producer and consumer
  - Data integration from multiple sources

FINAL DATA SETS

- PUBLICATION
  - Publish the data as a “citeable reference” (may be in addition to time-series data server)
    - Digital Object Identifier (DOI) from library
    - Earth Systems Science Data (ESSD) journal http://earth-system-science-data.net
  - Cite the data in the manuscript
- PRESERVATION
  - Contribute data and documentation to a permanent archive
  - National or World Data Centers

RESOURCES

Data Management Resources

- IOC/IODE Ocean Teacher
  http://classroom.oceanteacher.org/
- BCO-DMO Resources
  http://bco-dmo.org/resources/
- GO-SHIP Repeat Hydrography Manual
  http://www.go-ship.org/HydroMan.html

Data

- Rolling Deck to Repository (R2R) (US Academic Fleet)
  http://www.rvdata.us/catalog/
- SeaDataNet (Pan-European)
  http://www.seadatanet.org/
- BCO-DMO, National and World Data Centers

THANK YOU

Questions?

Woods Hole, Massachusetts, USA

DISCUSSION SUMMARY

Notes added 1 December 2012

1. ability to embargo data

If data must be embargoed (access limited for some period of time), then the recommendation is to make the metadata about the data publicly available with information on the estimated time when the data would be available for public access. The lead PIs from several sites suggested that data from "regular" time-series site cruises be made available as soon as possible, but that data from process study type cruises done near the time-series site locations, would be shared at a later data. Data required for student thesis work are often exempt from expected data reporting dates, and released only after the student thesis has been published.

2. cruise report

Someone asked about suggestions for tips on generating the cruise report. The recommendation is that the Chief Scientist is responsible for filing a cruise report after a cruise, but for time-series cruises, it is highly recommended to start with a cruise report template, and fill in the sections for each cruise. Chief Scientists should recommend that each cruise participant contribute a section on their work during the cruise. Several PIs commented that hydrographic data are released in preliminary form (not yet quality controlled) very soon after a cruise.

3. The participants were asked to share the expected times described in their data policies. There was a broad range of expectations: near real-time reporting of preliminary data; data made available quarterly or annually as quality control is completed; data not publicly available due to local logistical limitations (lack of infrastructure); data contributed to the National Data Center within five years or data made available only per written request.

4. The participants were asked to send information (URLs) to the workshop coordinators to document where their data are being made available.

DISCUSSION SUPPORT SLIDES

Additional slides (not shown during talk)

- Proposed quality flag system
- Metadata guidelines

QF FLAG PROPOSAL: PRIMARY FLAG LEVEL

Code	Definition	Test criteria
0	Good	Passed all applied documented QC tests
1	Quality not evaluated	
4	Questionable/suspect (i.e. inconclusive)	Failed non-critical documented metric or subjective QC tests
8	Bad	Failed critical documented metric QC tests
9	Missing data	

This scheme has been proposed as an IODE standard:
URL: https://sites.google.com/site/gebichwiki/data-qa-qc

METADATA

WHO

- list of investigators and their affiliations
- funding sources

WHAT

- list of measurements and observations and units
- names and descriptions

METADATA

WHEN

- date/time in UTC
- and timezone and/or local time

WHERE

- latitude, longitude, depth/height

METADATA GUIDELINES

WHY

- describe science plan and research focus

HOW

- methods
- procedures and protocols
  - time-stamped, versioned
  - living document that must be kept up to date