BCO-DMO
Biological and Chemical Oceanography Data Management Office

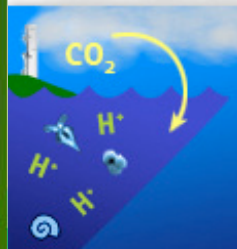
Better Practices for Shipboard Data Management

Cyndy Chandler
Biological and Chemical Oceanography Data Management Office

EPOCA
European Project on OCean Acidification

12 November 2009
Ocean Acidification Short Course
Woods Hole, MA USA

NSF



Ocean Acidification
Studying ocean acidification's effects on marine ecosystems and biogeochemistry

Discussion Topics

- Research Cruise
  - Allocation of sample (wire) time
  - Allocation of sample water
  - Cruise report
  - Data inventory
  - Cruise Sampling Event Log
- Data and Metadata Reporting
  - Data Quality (review)
  - Metadata and Standards
  - Data Centers and National Archives

What about experiments?

- presentation will be specific to cruise activities,
  but the concepts apply to lab experiments, perturbation
  or mesocosm experiments

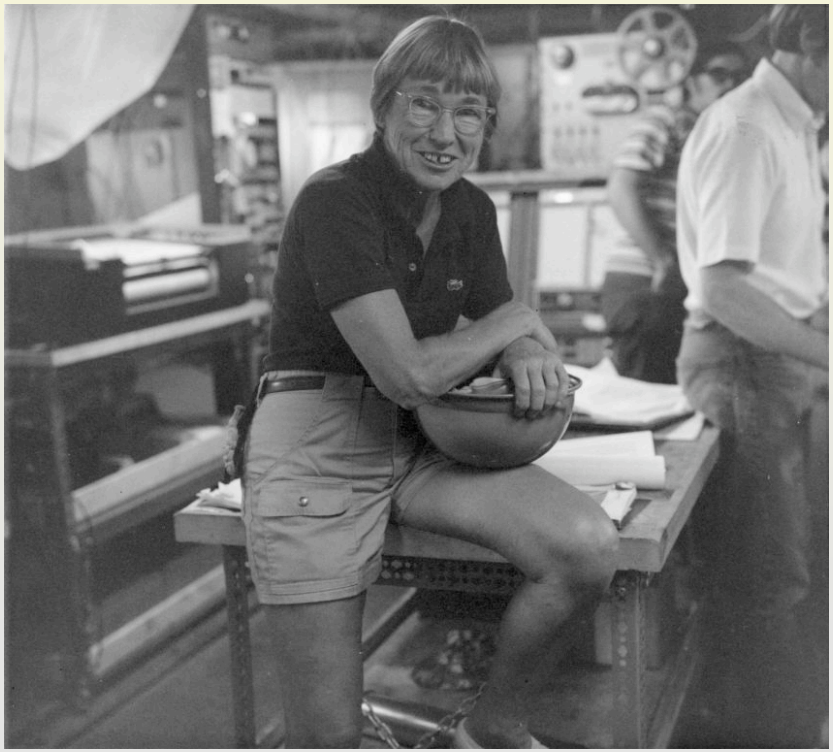
- European Project on OCean Acidification (EPOCA)
- "Guide for Best Practices on Ocean Acidification Research and Data Reporting“

  available on the EPOCA web site: http://www.epoca-project.eu/index.php/Home/Guide-to-OA-Research/

  Editors in chief:
  Ulf Riebesell, Victoria J. Fabry, Jean-Pierre Gattuso

Pre-cruise Planning

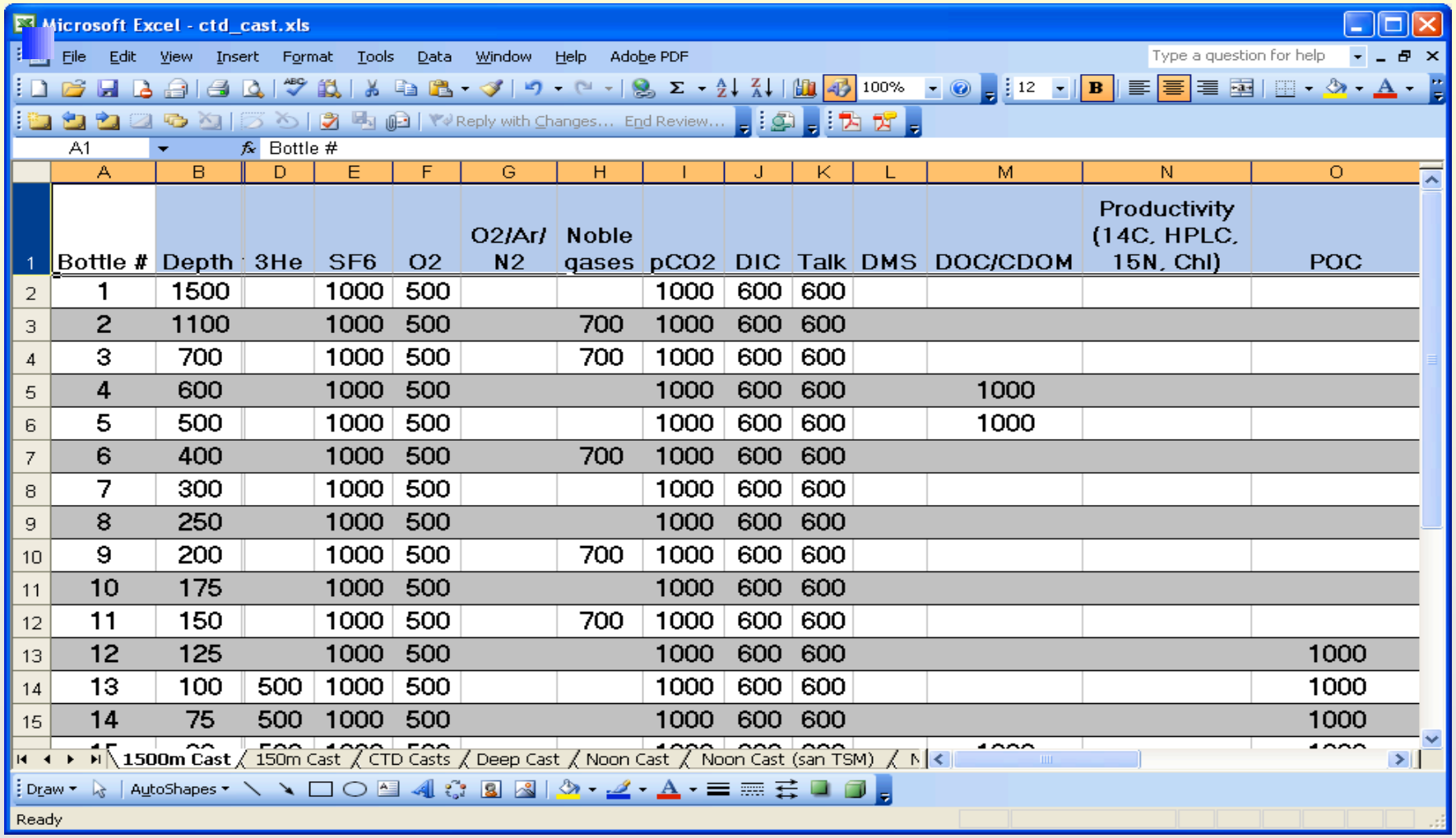
- Station plan
- Allocation of sample (wire) time
- Allocation of sample water



These arrangements should be made prior to the cruise and then reviewed at the first science briefing on board.

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

Cast Plan ~ Sampling Time and Water Allocation



Microsoft Excel - ctd_cast.xls

File Edit View Insert Format Tools Data Window Help Adobe PDF

A1 Bottle #

	A	B	D	E	F	G	H	I	J	K	L	M	N	O
1	Bottle #	Depth	3He	SF6	O2	O2/Ar/ N2	Noble gases	pCO2	DIC	Talk	DMS	DOC/CDOM	Productivity (14C, HPLC, 15N, Chl)	POC
2	1	1500		1000	500			1000	600	600				
3	2	1100		1000	500		700	1000	600	600				
4	3	700		1000	500		700	1000	600	600				
5	4	600		1000	500			1000	600	600		1000		
6	5	500		1000	500			1000	600	600		1000		
7	6	400		1000	500		700	1000	600	600				
8	7	300		1000	500			1000	600	600				
9	8	250		1000	500			1000	600	600				
10	9	200		1000	500		700	1000	600	600				
11	10	175		1000	500			1000	600	600				
12	11	150		1000	500		700	1000	600	600				
13	12	125		1000	500			1000	600	600				1000
14	13	100	500	1000	500			1000	600	600				1000
15	14	75	500	1000	500			1000	600	600				1000

1500m Cast | 150m Cast | CTD Casts | Deep Cast | Noon Cast | Noon Cast (san TSM)

Draw AutoShapes

Ready

Keeping records (recording metadata)

- Log Sheets (formal way to record metadata)
  - station logs
  - sample logs
- Cruise report (cruise metadata)
- Data inventory (dataset metadata)
- Event log (device deployment metadata)

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 O2$_1$	Oxygens O2$_2$ O2$_3$	DIC/Alk	TOC/TON	Salts	Nuts	Bact	FRRF	POC/PON Vol=	HPLC Vol=	Chla	Flow Cyt	DOP	HP Hensell	15N/Φ	
1	0																	7	1
2	0	23.8		1		91	91		91		89 13					8	1	/14	2
3	0											73 1	67 -1	67 -1	67				3
4	20	23.4		2		92	92		92		88 12						2		4
5	20											74 2	68 -2	68 -2	68				5
6	40	21.2		3		93	93		93		87 11					7	3		6
7	40											75 3	69 -3	69 -3	69				7
8	50	20.9		4		94	94		94		86 10		70 -4	70 -4	70		4		8
9	60	20.5		5		95	95		95		85 9						5		9
10	60											76 4	71 -5	71 -5	71				10
11	70	20.2		6		96	96		96		84 8		72 -6	72 -6	72		6		11
12	80	20.0		7		97	97		97		83 7						7		12
13	80											77 5	73 -7	73 -7	73				13
14	90	19.9		8		98	98		98		82 6		74 -8	74 -8	74		8	/15	14
15	100	19.6		9		99	99		99		81 5						9		15
16	100											78 6	75 -9	75 -9	75				16
17	120	19.3		10		100	100		100		80 4								17
18	120											79 7	76 -10	76 -10	76				18
19	140	19.3		11		101	101		101		79-3								19
20	140											80 8	77 -11	77 -11	77				20
21	200	19.1		12		102	102		102		78-2	81 9							21
22	300	19.1		13		103	103		103			82* 10							22
23	500	18.6		14		104	104		104			83* 11							23
24	700	16.7		15		105	105		105		77-1	84* 12							24
												3L							

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

Better Practices for Shipboard Data

- **Data Management Best Practices Guide compiled by BCO-DMO based on experience from US GLOBEC and US JGOFS**

- a collection of better practice recommendations for management of data from research cruises

- available as a PDF download from: http://bco-dmo.org/resources

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

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 plethora of sampling devices deployed during a cruise
  - understand relative timing between events

a sampling event matrix

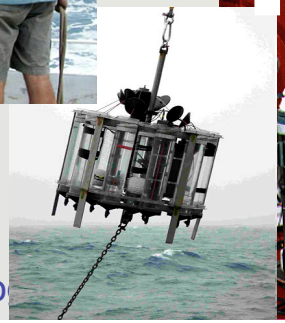
VERTIGO project KM0414 ALOHA cruise sampling event matrix

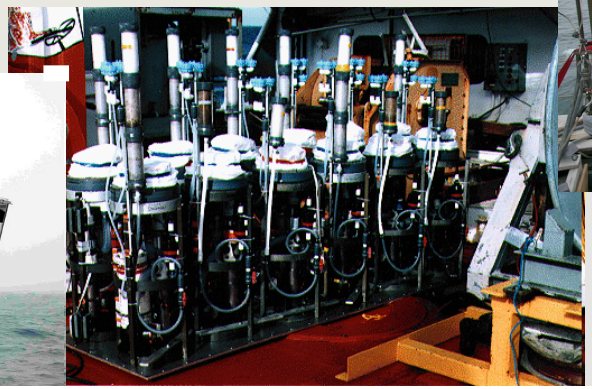
R/V Kilo Moana (University of Hawaii Marine Center)

July 9th final summary of cruise activities

file VERTIGO final cruise activities.xls

Julian Day	172_173	173_174	174_175	175_176	176_177	177_178	178_179	179_180	180_181
ship plans- June 2004			particle collections	day 0	day 1	day 2	day 3	day 4	day 5
	20-Jun	21-Jun	22-Jun	23-Jun	24-Jun	25-Jun	26-Jun	27-Jun	28-Jun
hours	day 1	2	3	4	5	6	7	8	9
0		SS#1 CTD 2-18	SS#1 CTD 2-18	Trull trap in- 300	Survey CTD #24 & drifters	MOCNESS	MULVFS	NBST 300 out	MOCNESS
2		SS#1 CTD 2-18	SS#1 CTD 2-18	Launch Clap 150	Survey CTD #25 & drifters	SS	MULVFS	NBST 300 out	NBST 500 out
4		SS#1 CTD 2-18	recover 12hr NBST	STD bio cast- CTD #19 & Th CTD20	1000m CTD #26	CTD biocast #27 (shallow)	MULVFS	CTD #32	NBST 500 out
6	depart 0800	SS#1 CTD 2-18	SS#1 CTD 2-18	Launch Clap 300	Launch-optical trap	CTD biocast #28 (deep)	MULVFS	STD bio cast- CTD#33	NBST 500 out
8		SS#1 CTD 2-18	SS#1 CTD 2-18	Launch Clap 500	Launch-optical trap	MULVFS	NBST 150 out	Clap 300 out	Clap 500 out
10		recover 12 hr NBST	plankton net test- Silver/Tanner	Launch NBST 150	MOCNESS	MULVFS	NBST 150 out	CTD # 34-39	MOCNESS
12		SS#1 CTD 2-18	MOCNESS test	Launch NBST 300	MOCNESS	MULVFS	NBST 150 out	CTD # 34-39	MOCNESS
14		SS#1 CTD 2-18	MULVFS test	Launch NBST 500	Launch respirometer	MULVFS	Clap 150 out	CTD # 34-39	CTD # 40-44
16		deploy 12hr NBST test	MULVFS test	Survey CTD #21 & drifters	Go-Flo casts	Deep CTD 29- 3000m Ba/Th	Clap 150 out	CTD # 34-39	CTD # 40-44
18		SS#1 CTD 2-18	SS#1 CTD 2-18	Survey & drifters	Go-Flo casts	Deep CTD 29- 3000m Ba/Th	CTD #30 & 31	CTD # 34-39	CTD # 40-44
20	deploy 12hr NBST test	optical trap test	Launch 1 Siegel drifter	Survey CTD #22 & drifters	Launch C explorer	MULVFS	optical trap out	CTD # 34-39	CTD # 40-44
22	SS#1 CTD 2-18	SS#1 CTD 2-18	Trull trap in- 300	Survey CTD #23 & drifters	MOCNESS	MULVFS	bio cast-tow	MOCNESS	CTD # 40-44





Why is an event log important?

- the unique sampling event identifier helps to integrate observations from discrete data sets

- Example: CTD station 4 cast 2 is assigned event number 20080904.1342 … the Niskin bottle nutrient data and pigment data from that cast can be integrated using that event number

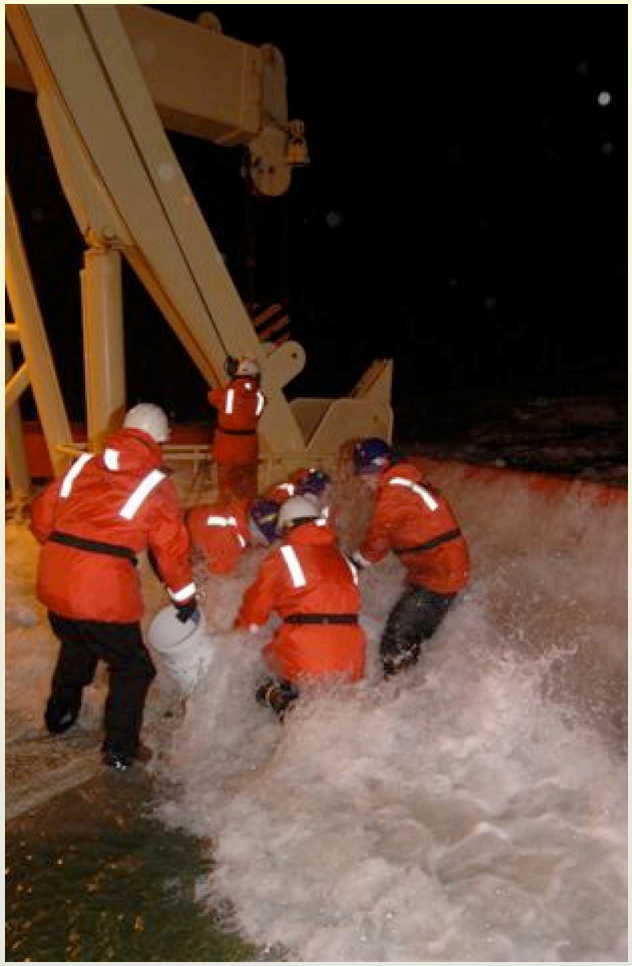
important event log fields . . .

What types of information are important to record in the event log?

- the unique sampling event identifier
- Example: YYYYMMDD.hhmm
  YYYYMMDD + time/2400
  SCYYMMDD.hhmm (SC=ship code)
- station and cast number
- date and time (UTC)
- position (latitude and longitude)

other important fields in the event log . . .

- instrument [package] type
  CTD, TM, drifter net, penguin
- name of person
  responsible for sampling event
- activity descriptor
  e.g. deployment, recovery
  start, max depth, end, abort



additional possible fields . . .

- description of activity and/or comments (free text)
  e.g. first cast after retermination
- local time (important for biology cruise)
- timezone
- cruise notebook or subsample log page
- position relative to a feature (eddy center or treatment patch)

shipboard sampling event log

generated automatically using some algorithm

controlled vocabulary

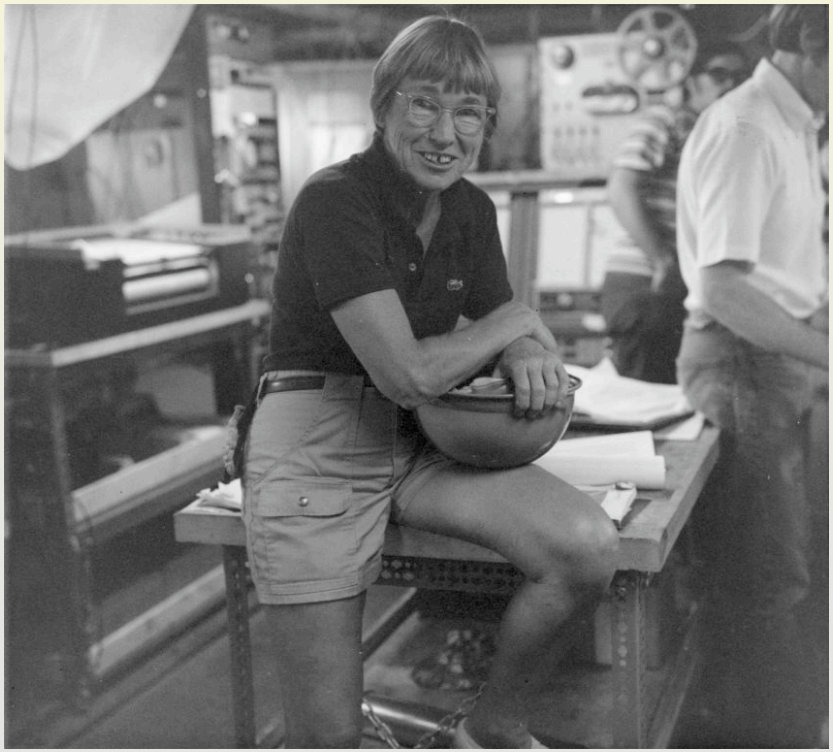
event	date	time	time_L	sta	lon	lat	ev_type	person	activity
0212208	20020121	2208	1108	TEST	-175.220	-53.572	CTD001	nd	CTD001
0230442	20020123	0442	1742	0	-171.480	-55.398	CTD002	Wang	CTD002
0231556	20020123	1556	0456	0	-171.583	-55.407	ZooTow	Landry	ZooplankTow
0232351	20020123	2351	1351	1	-171.521	-55.334	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:



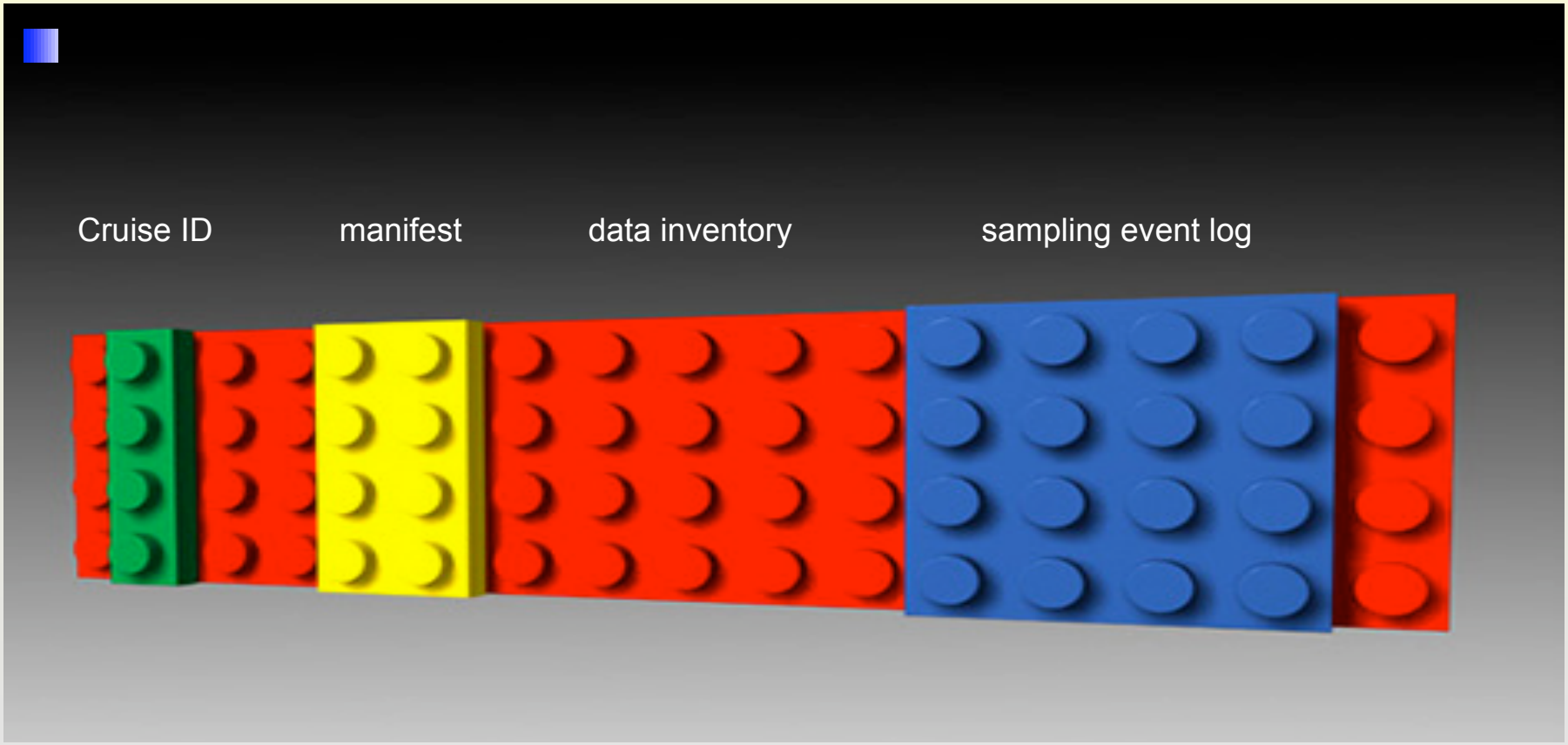
- date and time
  shipboard network and UTC
  (*not your wristwatch*)

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

Final Event Log

- should be an electronic file in plain text (TSV or CSV)
- many researchers record events on paper logs in the main lab, and then enter the records into Excel
- if the original event log entries were made on paper log sheets, scan the originals and convert to PDF
- some research vessels support event logging applications, and NSF is funding the development of an event logger for use on UNOLS vessels (R2R project, rvdata.us)

Constructing the cruise report . . .



Discussion Topics

- Research Cruise
  - Allocation of sample (wire) time
  - Allocation of sample water
  - Cruise report
  - Data inventory
  - Cruise Sampling Event Log
- Data and Metadata Reporting
  - Data Quality (review)
  - Metadata and Standards
  - Data Centers and National Archives

Data Quality

- Quality Assessment
  It is important to have a system in place to assure outside users that the analytical results produced are of proven and known quality. (Andrew Dickson, 2009)

- to make an assessment of quality we must also keep in mind the time [the data were] collected, the methodology – the capability of the time (A.K. Sinha, Virginia Tech, 2006)

- Data Quality involves
  quality assurance – done prior to measurement
  quality control – done after measurement

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? (QA)
- How good did I get 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 document the data quality assessment with the published dataset (reported as metadata)

Data Quality (of colleague's data)

Thank you to Andrew Dickson (previous lecture)

In order to assess 'fitness for purpose' ← important

- it helps to know why the measurement was made
- ones ability to ascertain 'good enough' is related to the uncertainty associated with the measurement
- uncertainty relative to your needs as defined by the research topic

Metadata

- Needed to automate the process of data discovery
  - like using a library catalog to locate a resource
- Needed to determine the fitness of a data set for use
  - particularly regarding quality (“fitness for purpose”)
- Needed to facilitate accurate data interpretation
  - e.g. units of measurement, data format

- Metadata records are expensive to generate
  - and may require additional expertise to define
- But the benefits are substantial
  - metadata make it possible to find data sets, and use them effectively
  - they allow the benefits of investments in data to be realized

Metadata

- in the US, 1995 was the year that state government agencies started devoting resources to metadata capture

- motivated by:
  - 1995 Paperwork Reduction Act (104th US Congress)
  - Internet and HTML = World Wide Web
  - desire to automate public access to government documents following the requirement that agencies establish ‘locator services’ for federal information

- Dublin Core Metadata Initiative (2001)

Work in Progress . . .

There is no cookbook of instructions – or at least the book isn't finished – and establishing best practices will continue to be iterative.

- research themes are becoming more complex
- cost of doing research will continue to increase
- *in situ* data can not be collected 'again'
  The water samples collected in March 2009 from 22° 45'N, 158° 00'W can not be collected again.

Metadata ~ the goal

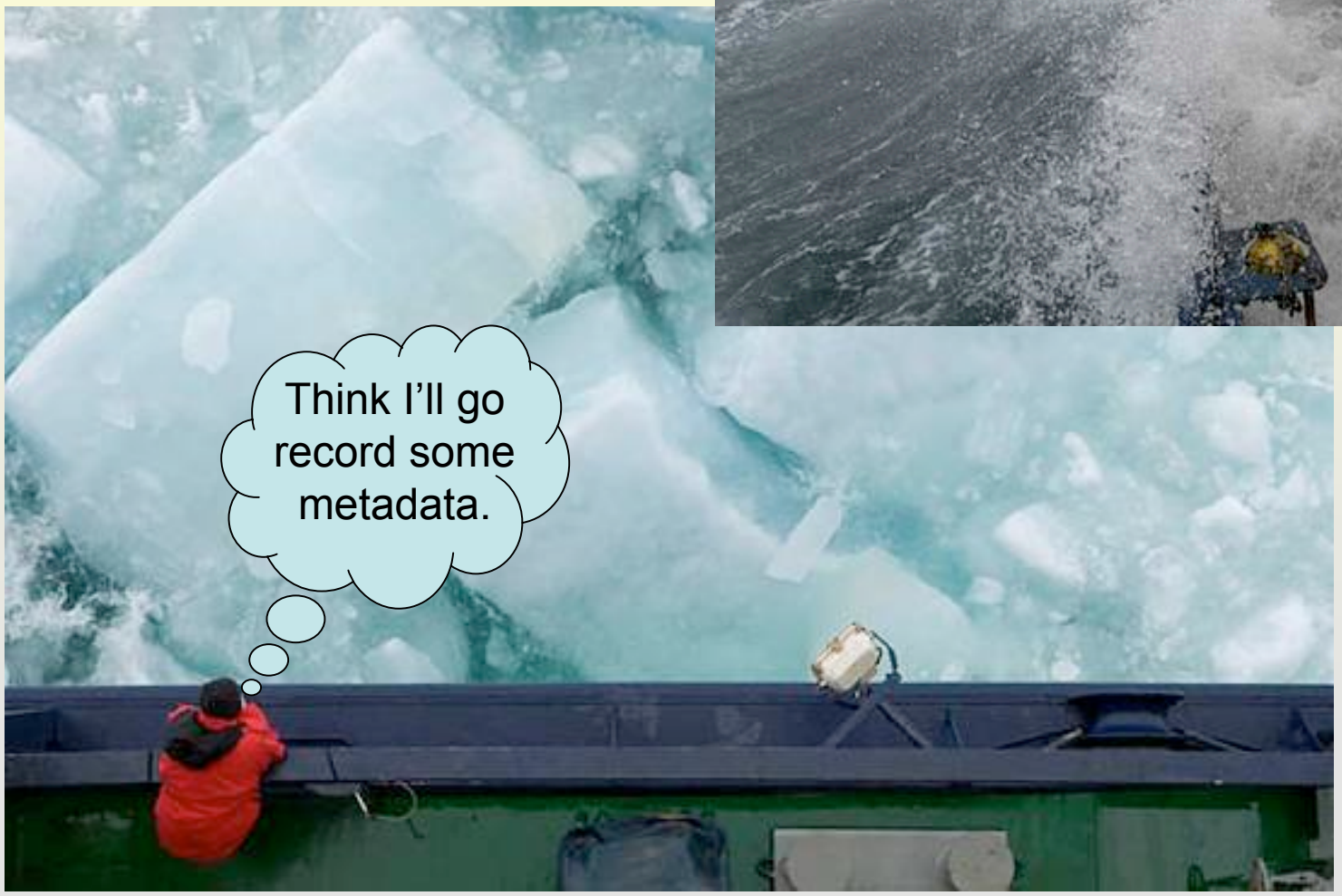
- document the quality assurance and control measures applied to the measurements during sampling and analysis
  - what protocols were followed (include reference)
  - were replicates done, include results of control chart
  - were inter-comparisons done (analytical techniques, different labs?)
  - were reference materials used (which ones)
  - what was done to account for T and P dependencies
  - were data adjusted based on results of quality control procedures
- objective: to report sufficient metadata to support
  - determination: are these data 'fit for purpose'
  - accurate re-use of the data
- metadata reporting is especially important when the protocols are still being developed (e.g. OA sampling and analytical techniques)
- remember local v global (in space and time); resultant data will be used and re-used by colleagues

- European Project on OCean Acidification (EPOCA)
- "Guide for Best Practices on Ocean Acidification Research and Data Reporting“

  available on the EPOCA web site: http://www.epoca-project.eu/index.php/Home/Guide-to-OA-Research/

  Editors in chief:
  Ulf Riebesell, Victoria J. Fabry, Jean-Pierre Gattuso

recording metadata at sea
. . . is problematic



… this is the office !

Metadata matter

- ocean acidification research is and will continue to be . . .
  - expensive (research cruises are resource intensive)
    - fuel costs
    - equipment allocation
    - people time (highly trained people at sea)
  - collaborative
    - team projects are more complicated than individual research
  - important – answers are needed to enable science-based decision support for legislative policies
- means the metadata matter more

Data management partners: BCO-DMO

- metadata forms (http://bcodmo.org/resources)
  - Program
  - Project
  - Deployment (e.g. cruise)
  - Dataset metadata contributed with the data
- data contributed in any format (often as Excel files)
- researchers work in partnership with BCO-DMO staff members to manage data through all phases of a project

Standards

community adopted standards for . . .

- sampling and analytical protocols
  - assessment of quality assurance and control
- metadata content standards
  - FGDC or ISO may be required by some Data Centers

Use of standards can facilitate data integration.

At the moment, most of the effort relating to standards is being handled by data centers.

Data Centers and National Archives

- BCO-DMO
  - Biological and Chemical Oceanography Data Management Office
  - for researchers funded by US NSF OCE
- CDIAC (Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory)
- NODC
  - National Oceanographic Data Center
  - permanent data archive for US researchers funded by NOAA or NSF
- analogous *ODC or WDC in other nations


Questions?
conclusion part 2 of 2
thank you

end of day 1
part 2 of
total 3 part data management section

Part 1: Monday, ODV Introduction, Reiner Schlitzer

Part 2: Thursday, Data Management: Introduction
and Shipboard Data, Cyndy Chandler

Part 3: Friday, Contributing Data to Data Centers, Alex Kozyr