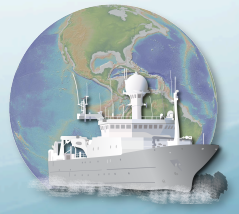
Introduction to Cruise Data Management and R2R

Karen Stocks
Geological Data Center, Scripps Inst. of Oceanography
kstocks@ucsd.edu

11 Feb 2016
UNOLS Chief Scientist Training Workshop, San Diego



I will send these slides around

No need to scribble

Some are primarily for later reference

Topics

- What is data management
- Cruise data management responsibilities: who does what
- Role of R2R in managing underway data
- Chief Scientist data responsibilities & good practices
  - Managing your data
  - Submitting to a data center

What is Data Management

- Having a plan from the beginning to ensure that
  - data and metadata are recorded accurately
  - are sufficient to support re-use
  - are preserved securely (backups)
  - and will be made accessible to others
- Management through full data 'life cycle', from proposal to preservation

Who has 1^0 Responsibility for Cruise Data Management

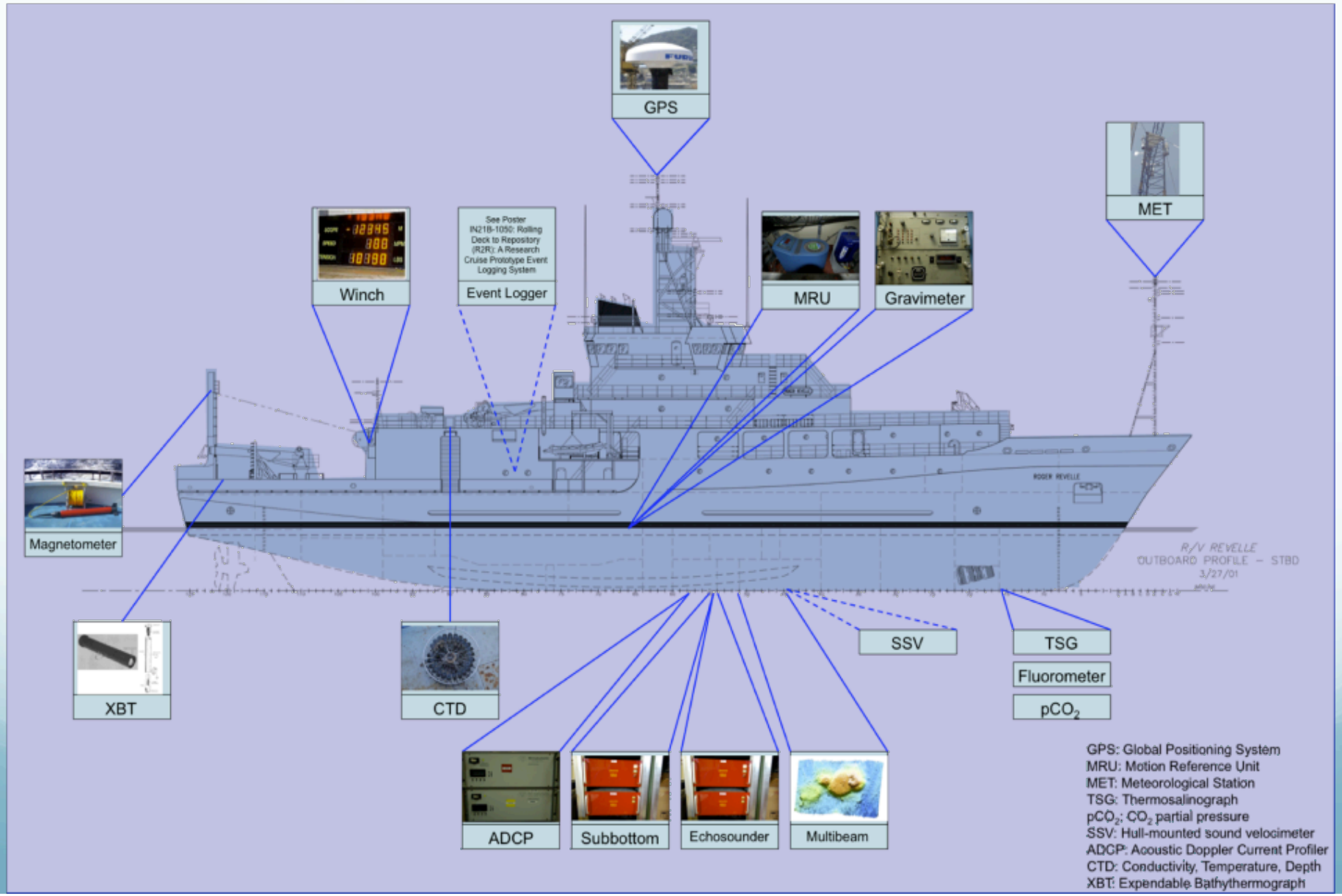
Routine underway data: from instruments permanently installed on the vessel, collected by the ship's technicians, unprocessed.

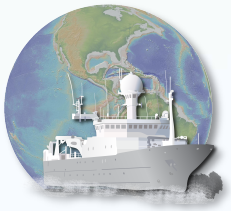
- Rolling Deck to Repository (R2R)

"**Science party data**": from instruments brought by science party, data collected by science party, most "station" data, products derived by science party.

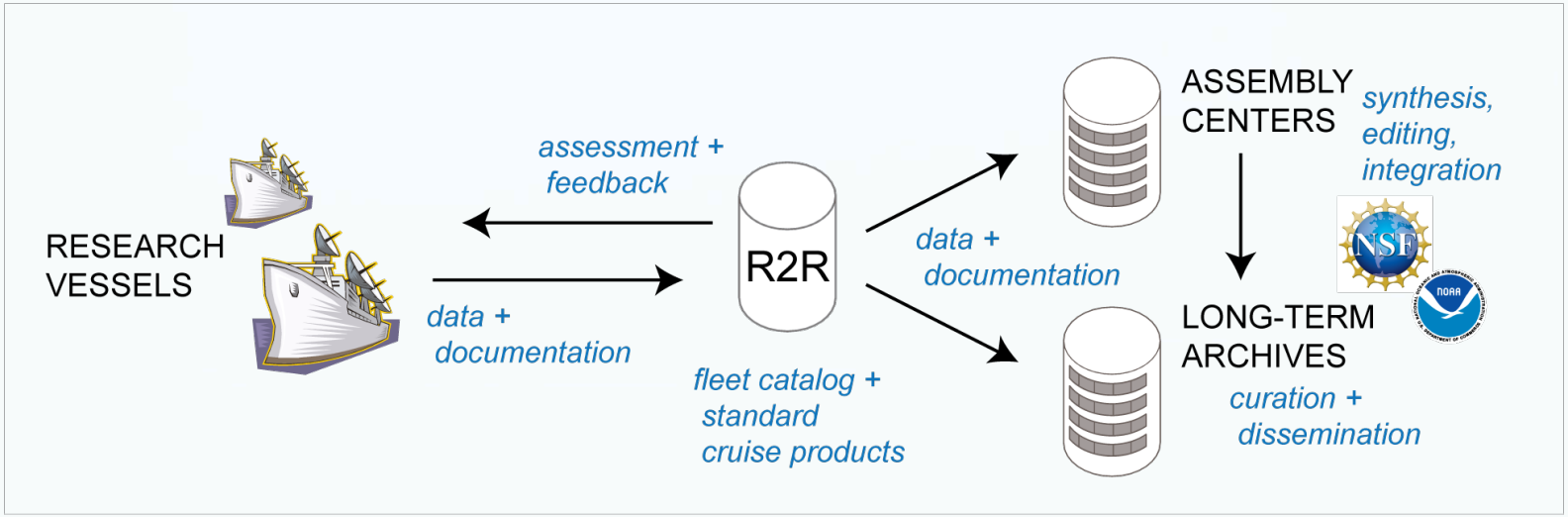
- Science Party

Common Underway Data

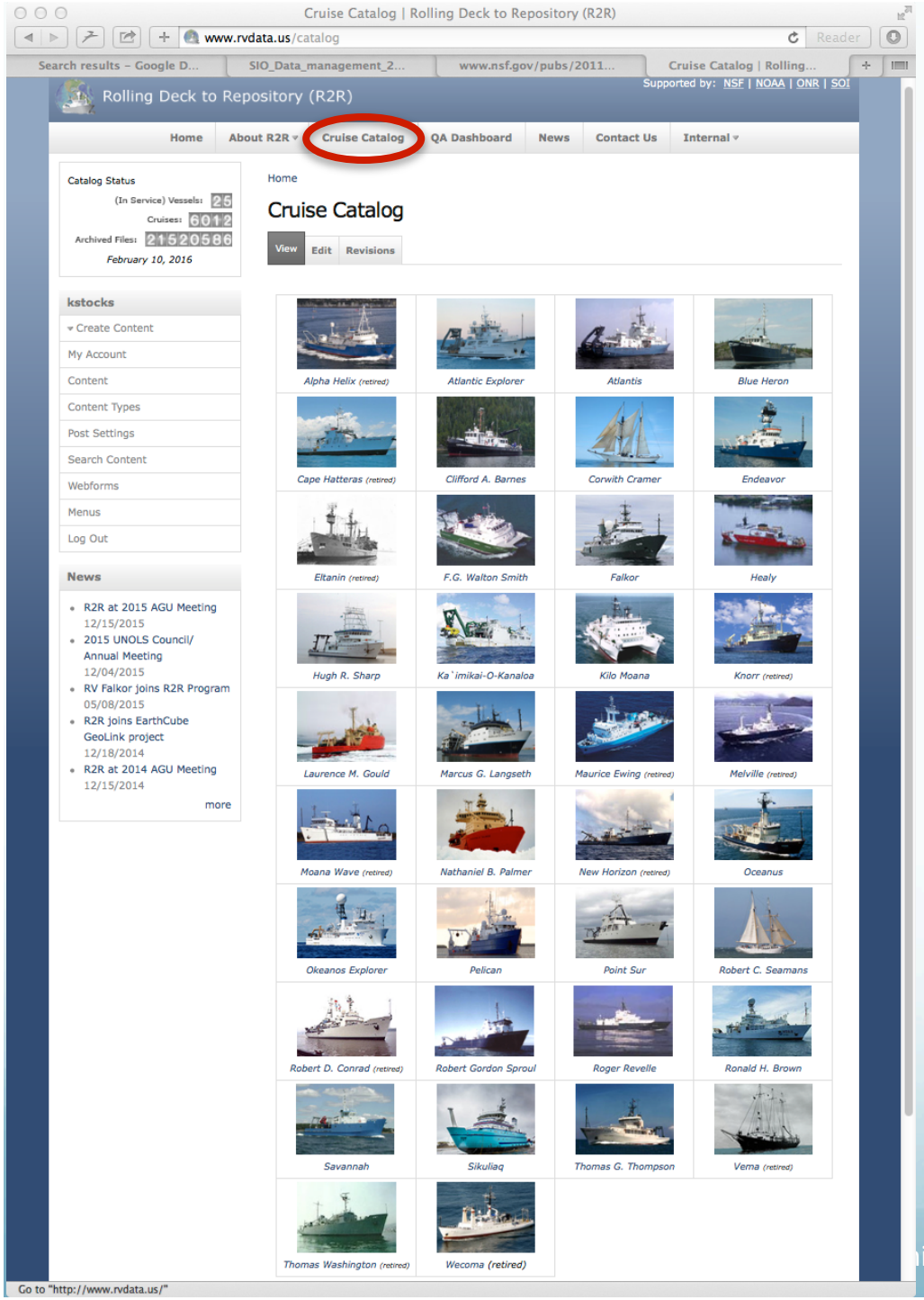




R2R Goals



- Migrate routine "underway" sensor data to long-term repositories
- Create catalog of cruises and standard products
  - Expedition details
  - Clean shiptrack navigation and geophysical data
  - Event log
  - Operations report
- Assess data quality and provide feedback to R/V operators

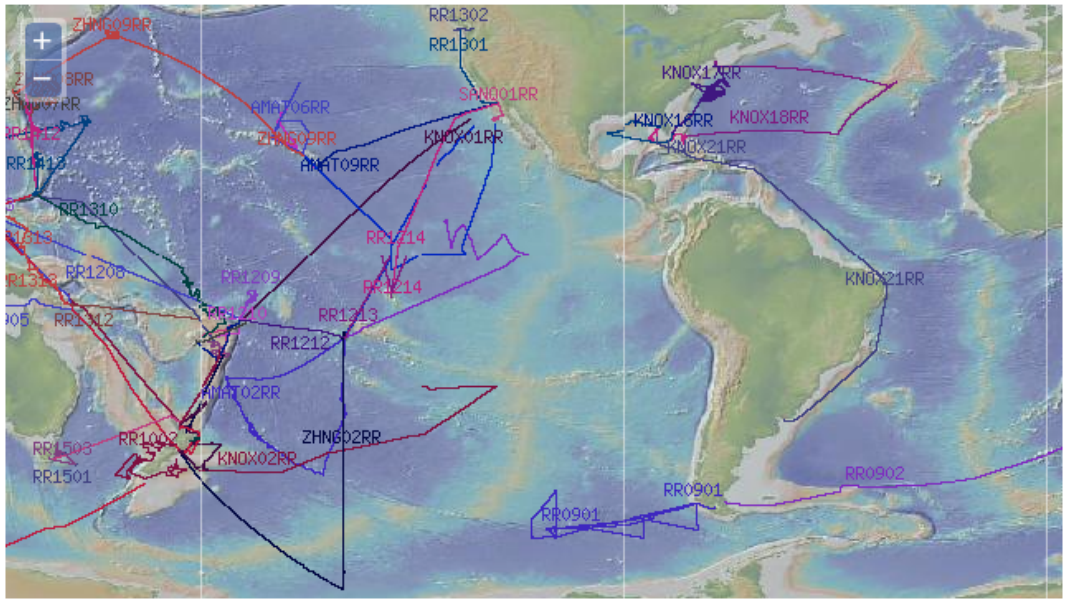


www.rvdata.us

R2R Cruise Catalog

R2R Team: Suzanne Carbotte, LDEO, Bob Arko, LDEO, Karen Stocks, SIO, Cyndy Chandler, WHOI, Shawn Smith, FSU

Cruise Catalog: Roger Revelle

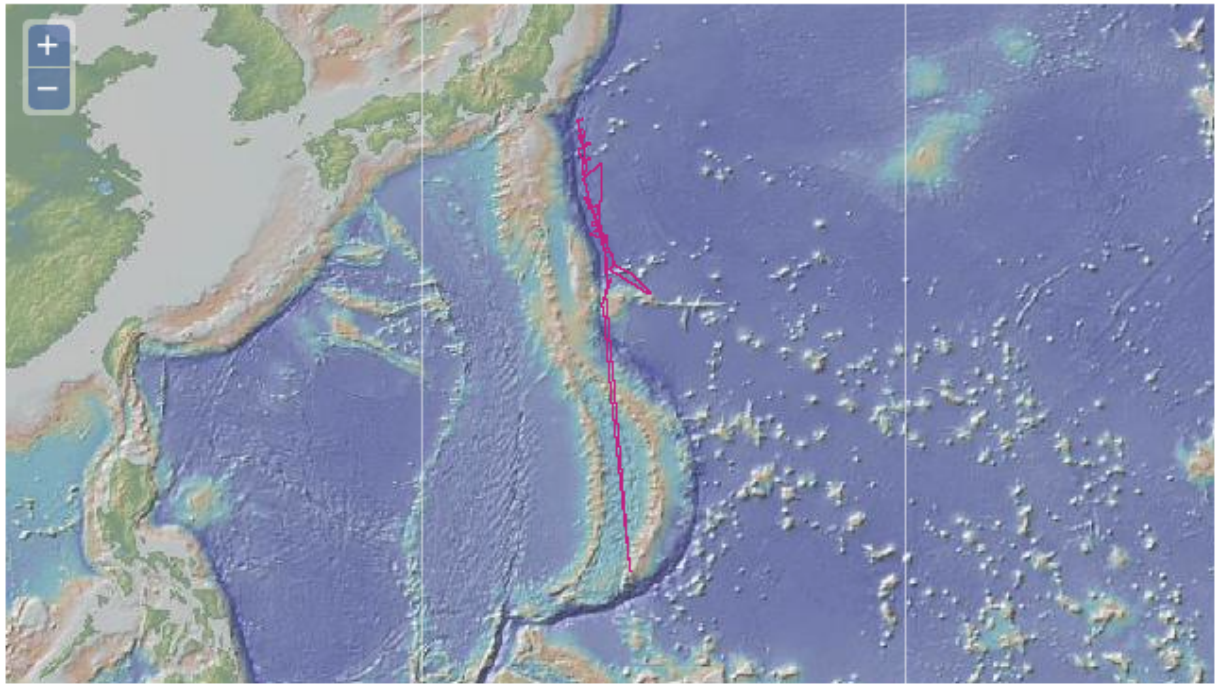


Operator: Scripps Institution of Oceanography

Cruise ID *(Click for Details)*	Start Date	Start Port	End Date	End Port
	Summary			
RR1509	2015-06-20	Napier, New Zealand	2015-06-30	Napier, New Zealand
	Project: Hikurangi Ocean Bottom Investigation of Tremor and Slow Slip (HOBITSS) (Info) *Chief:* Wallace, Laura (UTIG)			
RR1508	2015-05-19	Auckland, New Zealand	2015-06-17	Napier, New Zealand
	Project: Subduction Thrust Investigation of New Zealand using Geothermics and Seismics (STINGS) (Info) *Chief:* Harris, Robert (OSU)			
RR1507	2015-04-21	Auckland, New Zealand	2015-05-14	Auckland, New Zealand
	Project: Collaborative Research: Geochemical effects on the functional microbial community dynamics of hydrothermal deposits along the Eastern Lau Spreading Center			

(261 cruises)

Cruise Catalog: RR1412



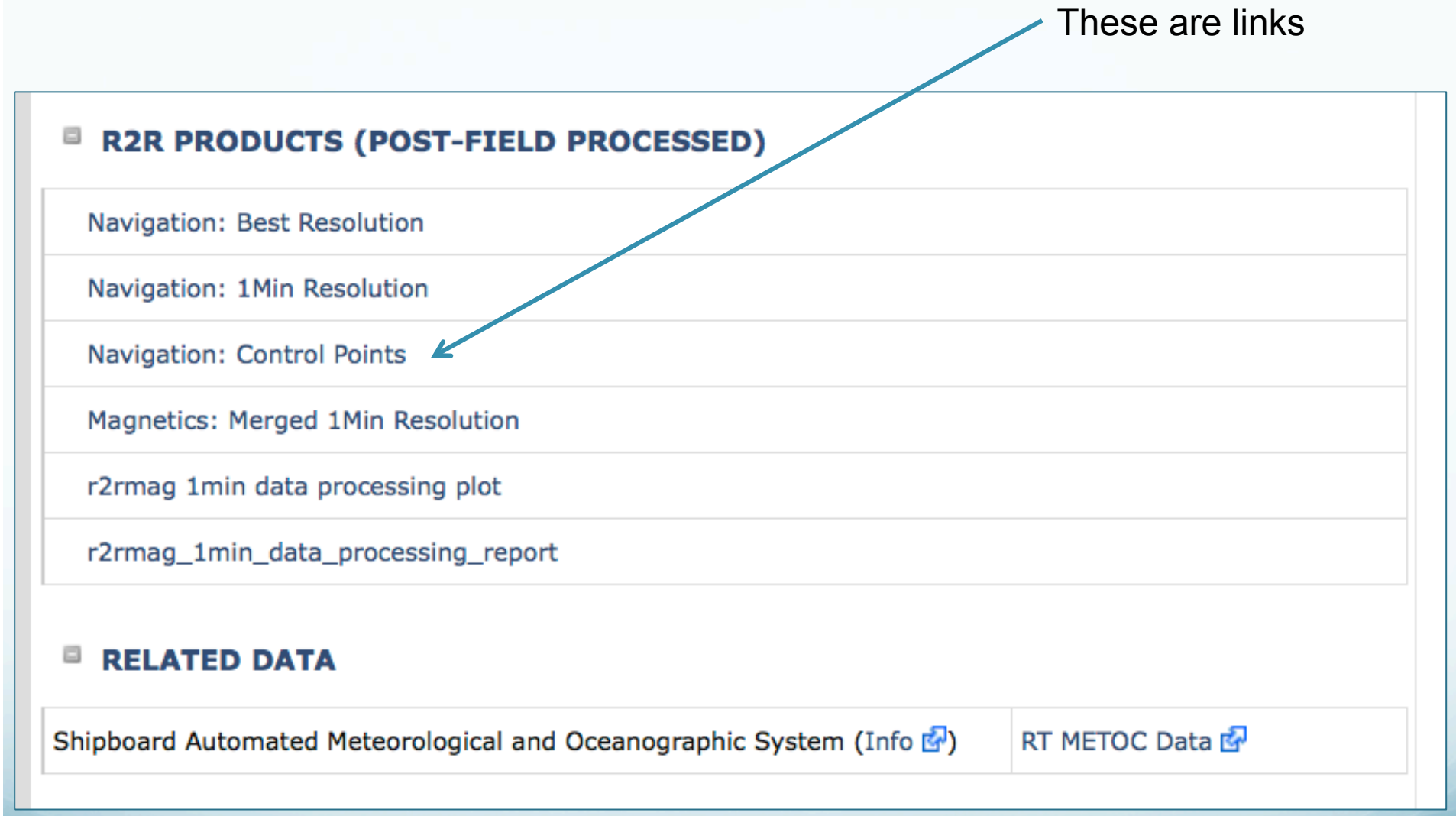
Operator: Scripps Institution of Oceanography
Vessel: Roger Revelle

Cruise DOI: 10.7284/904113

Cruise ID	Start Date	Start Port	End Date	End Port
	Details			
RR1412	2014-10-29	Apra, Guam	2014-11-23	Apra, Guam
	Project: Geochemical variation beneath the Izu-Bonin arc and implications for the generation of arc magmas **SCIENCE PARTY** **FILE MANIFEST** **UNDERWAY DATA SETS (ORIG FIELD DATA)** **R2R PRODUCTS (POST-FIELD PROCESSED)** **RELATED DATA**			

UNDERWAY DATA SETS (ORIG FIELD DATA)

Device Type	Make-Model [Location]	Files	Archive Status	
adcp	Hawaii UHDAS	List	*(submitted to NCEI)*	
expendableprobe	Sippican MK21	List	NCEI	Download
gnss	Ashtech ADU2	List	R2R	Download
gnss	Furuno GP-150	List	R2R	Download
gyrocompass	iXSea Phins	List	R2R	Download
magnetometer	Marine Magnetics SeaSPY	List	R2R	Download
metstation	SIO MET-System	List	R2R	Download
multibeam	Kongsberg EM122	List	NCEI	Download
singlebeam	Knudsen 320B/R	List	NCEI	Download
winch	Markey DUSH-5	List	R2R	Download
winch	Markey DUTW-9-11	List	R2R	Download


These are links
R2R PRODUCTS (POST-FIELD PROCESSED)
Navigation: Best Resolution
Navigation: 1Min Resolution
Navigation: Control Points
Magnetics: Merged 1Min Resolution
r2rmag 1min data processing plot
r2rmag_1min_data_processing_report
RELATED DATA
Shipboard Automated Meteorological and Oceanographic System (Info)
RT METOC Data

Rolling Deck to Repository (R2R)

Supported by: NSF | NOAA | ONR

Home | About R2R | Cruise Catalog | QA Dashboard | News | Contact Us | Internal

R2R Quality Assessment Dashboard

This dashboard provides information about Quality Assessment (QA) tests performed by R2R. QA tests do *not* assess the scientific utility of data, but are intended to identify *suspicious data* which may indicate sensor problems. QA is performed on data files as originally delivered from vessels. This dashboard shows the output of the initial implementation of R2R QA protocols, and includes details of QA tests, thresholds used and results. We welcome your input; please contact us HERE with comments or suggestions.

SEARCH BY:

Vessel: Melville Cruise: MV1204

Device Type: All Go->

All Filesets for MV1204

Cruise Title: *Transit*
Vessel: Melville
Total Filesets: 8

Click column headings to sort

Rating	Device	Cruise	Vessel			
	GNSS *Ashtech ADU2*	MV1204	Melville	QA Summary	QA Configuration	File Manifest
	Multibeam Sonar *Kongsberg EM122*	MV1204	Melville	QA Summary	QA Configuration	File Manifest
	Magnetometer *Marine Magnetics SeaSPY*	MV1204	Melville	QA Summary	QA Configuration	File Manifest
	GNSS *Furuno GP-150*	MV1204	Melville	QA Summary	QA Configuration	File Manifest
	Acquisition Sys *Hawaii UHDAS*	MV1204	Melville	QA Summary	QA Configuration	File Manifest
	INS *CodaOctopus F-185*	MV1204	Melville	QA Summary	QA Configuration	File Manifest

What you need to know about R2R
as a chief scientist

- You are not responsible for submitting routine underway data to a repository.
- Post-cruise you will receive an email asking you to release data – you can decline (and we'll ask again periodically.) R2R does not release data publicly without permission.
- After you release, you can point people to R2R if they ask you for underway data. You can use the standard products & QC dashboard.

Pop Quiz

Question: You are given a "Science" directory on the ship. What do you put on it?

A. Scientific data relating to the cruise.

B. A list of everyone's name, email and address so they can stay in contact after the cruise.

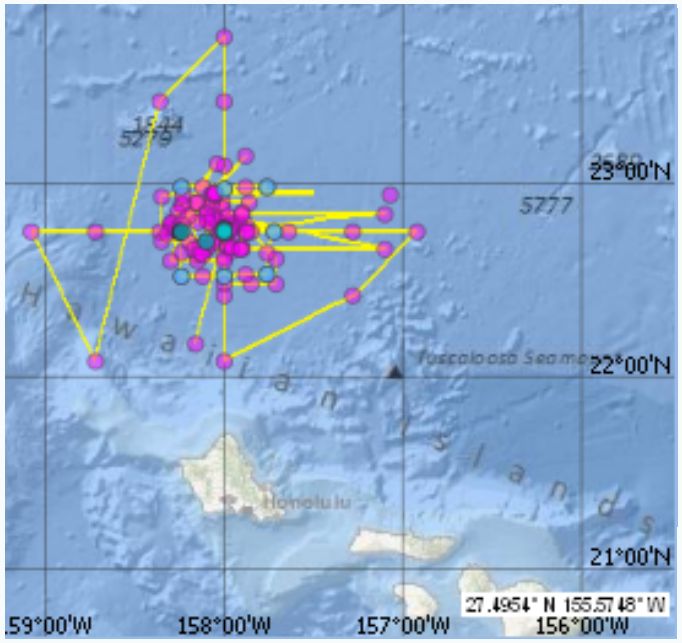
C. Movies and iTunes lists to share

D. Pics of the science party having fun at the last port.

Science Party Data

Chief Scientist Responsibilities and Best Practices

Just as a Chief Scientist needs to develop a sampling plan



S/he needs to develop a data management plan.

Data Management - Before

Before the Cruise

- Develop a data collection plan: what information will be collected, when and by whom? How will files be named, stored, and copied? What metadata standards will be used?
- Create data and metadata templates, and establish controlled vocabularies and critical identifiers.
- Coordinate with other science efforts on the ship and shipboard staff to ensure consistent data naming and collection practices.
- Determine and order hardware and software. Ensure compatibility among components, have redundancy in critical systems. Don't forget the peripherals: cables. batteries. converters. connectors.

Biological Sampling in the Deep Sea, ch 16

Data Management - During

On the Cruise

- Collect all data and metadata.
- Use an event logger to record all science party sampling.
- Create regular backups on multiple systems.
- Don't leave the ship without all required information.

Biological Sampling in the Deep Sea, ch 16

Data Management - After

After the Cruise

- Quality assess and quality control all data.
- Track the individuals responsible for different components of the data during analysis (e.g. taxonomists identifying particular taxa), and how data will be re-integrated.
- In the short term, copy data onto maintained servers, such as institutional systems, when available.
- In the longer term, submit data to a recognized repository; register data with appropriate registries and catalogs.

Biological Sampling in the Deep Sea, ch 16

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)

Develop Consistent Identification scheme for all important concepts

- Project code
- Cruise ID -> R2R, ship tech
- Vessel -> ICES codes
- Sampling event: Cast / Tow / Dredge etc.
- Site / Station / Transect
- Instrument / Sensor
- Protocol identifier (sampling or processing method)

Why Consistent Identification?

From a computer perspective, these samples are *not* the same:

- 2A,
- 2-A
- II-A
- 2:A are not the same.

Define the concepts, define the identification scheme, document it, share it.

Details…

- Use information-rich filenames. “Data.xls” is bad, but “TAN0104_228_20010419_Sled.xls” to indicate data from epibenthic sled sample 228 taken on voyage 0104 of the RV Tangaroa, collected 19 April 2001, is good. Consider versioning and subsampling.
- Never record times without indicating the time zone – doesn’t have to be in each record, but must be in the metadata.
- Use decimal degrees for latitude and longitude, with a preceding minus sign (“–“) to indicate west or south.
- Express depth as positive meters below sea surface, unless you have reason not to (and document if so.)

Pop Quiz: Which is the right place to develop your data plan?

Removed:
Picture of happy people in an office

Data Sharing

Why bother?

Your President Expects You to Share (PARR)

February 2013, OSTP memo

The Office of Science and Technology Policy (OSTP) hereby directs each Federal agency with over $100 million in annual conduct of research and development expenditures to develop a plan to support increased public access to the results of research funded by the Federal Government.

Your Funding Source Requires Data Sharing

- NSF, NOAA, DOE, DOD, NIH all have policies relating to data sharing.
- NSF OCE: PIs are required to submit, at no more than incremental cost and within a reasonable time frame (but *no later than two (2) years after the data are collected)*, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF/OCE grants to the appropriate Data Center (See appendices below or consult with the cognizant NSF Program Officer).

http://www.nsf.gov/pubs/2011/nsf11060/nsf11060.pdf

Your Journal Requires You to Share

e.g. Science:

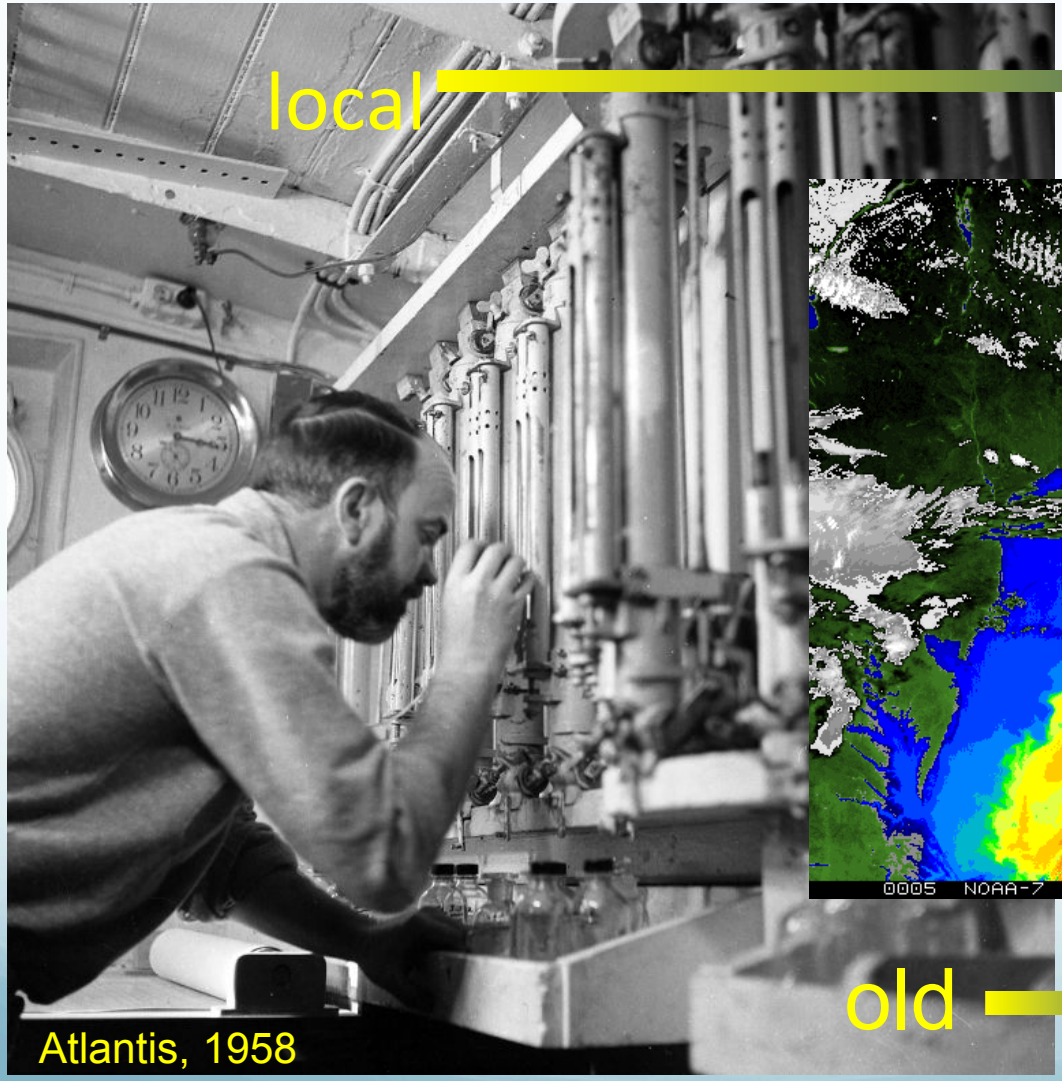
- *Before publication, large data sets...must be deposited in an approved database and an accession number provided for inclusion in the published paper*

Your Professional Society Expects You to Share

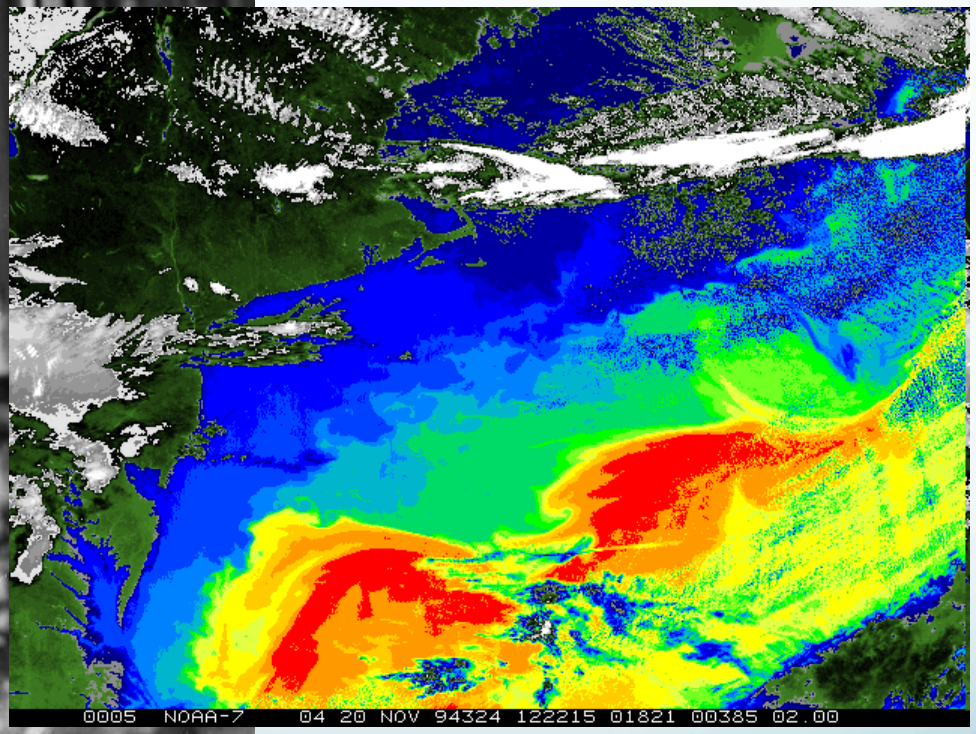
E.g. AGU draft data position statement:

- *Earth and space sciences data bases are a world heritage that should be made available to the scientific community and public as soon as possible (in some cases in real-time), should be organized and preserved in useable format, and should be conserved long-term for future use*

Data Sharing Facilitates Scientific Progress



to global



old → new

Sharing Your Data Will Probably Benefit You

Costello (2009): Sharing data can:

- boost scientists' recognition
- generate invitations to meetings
- present consulting and collaboration opportunities, and
- increase citation rates because their productivity will be more visible.

Costello, M.J. (2009). Motivating online publication of data. *BioScience*, 59: 418-427. doi:10.1525/bio.2009.59.5.9

Continuous Plankton Recorder Example

Stevens et al. (2006) found a positive correlation between their funding and the amount of data they were sharing over time

"We believe that the approach at SAHFOS to increase data accessibility and provide new visualisation tools has enhanced awareness of the data and led to the financial security of the organisation"

Stevens D., Richardson A.J., Reid P.C. (2006). Continuous Plankton Recorder database: evolution, current uses and future directions. *Marine Ecology Progress Series*, 316:247-255

You will most likely be responsible for submitting your data to a data center.

In an appropriate format and with metadata sufficient for discovery and use…by a stranger.

This is not a trivial task.

How to Share Your Data

Pop Quiz

Question: The way for me to share my data is?

A. If someone asks me for it, I'll send them a copy.
B. I'll post it on my lab website.
C. I'll submit it to the appropriate data center to ensure long-term archiving and access

What is the “appropriate” data center?

It depends on:

- Who funded you
- What kind of data you have

NSF-OCE Named Repositories

- BCO-DMO: The Biological and Chemical Oceanography Data Management Office. www.bco-dmo.org.
  - For projects funded by the Biological and Chemical Oceanography Sections and the Division of Polar Programs Antarctic Organisms & Ecosystems Program
- IEDA: Integrated Earth Data Applications. www.iedadata.org
  - Solid earth data (geology, geochemistry, and geophysics)
- CCHDO: CLIVAR data
- US NCEI: National Centers for Environmental Information (formerly NODC, NGDC, NCDC)
- Thematic data repositories: OBIS, GenBank, CDIAC

Note: policy also covers physical samples

- Know your Policy
- Know your Data Center
- *Before* you sail, find out what is required for data submission

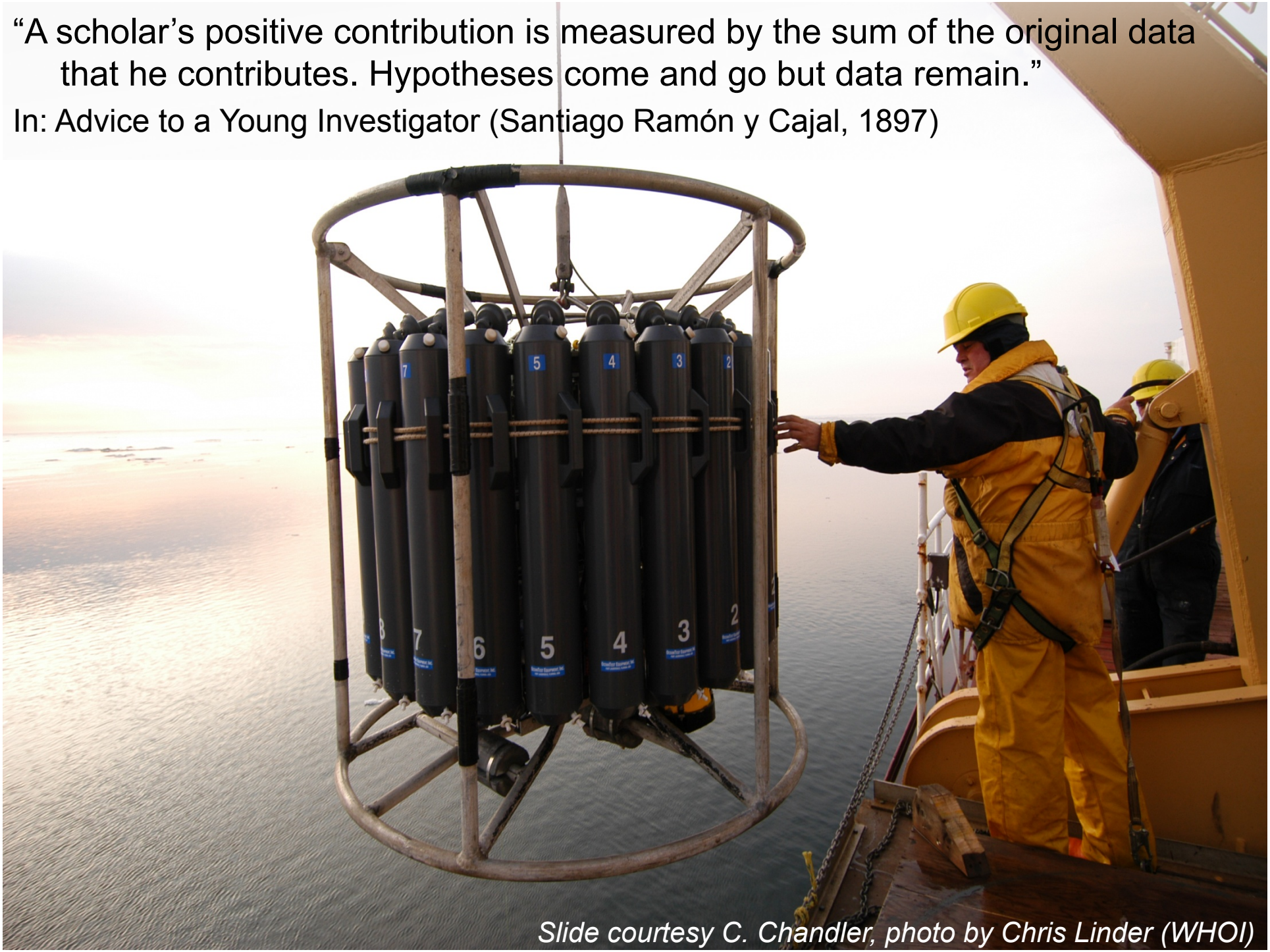
Resources

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

  http://bco-dmo.org/resources
- GO-SHIP Repeat Hydrography Manual
  - sampling and analytical protocols for water column measurements http://www.go-ship.org/HydroMan.html
- Stocks et al. 2016. Information Management Strategies for Deep-Sea Biology. In Costello & Consalvey (eds) Biological Sampling in the Deep Sea. Wiley.

“A scholar’s positive contribution is measured by the sum of the original data that he contributes. Hypotheses come and go but data remain.”

In: Advice to a Young Investigator (Santiago Ramón y Cajal, 1897)



Slide courtesy C. Chandler, photo by Chris Linder (WHOI)

Questions?

Karen Stocks: kstocks@ucsd.edu

info@rvdata.us

R2R Team: Suzanne Carbotte, LDEO, Bob Arko, LDEO, Karen Stocks, SIO, Cyndy Chandler, WHOI, Shawn Smith, FSU