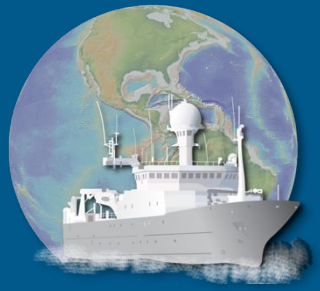


Rolling Deck to Repository (R2R) Updates and New Proposal

Columbia Climate School
Lamont-Doherty Earth Observatory

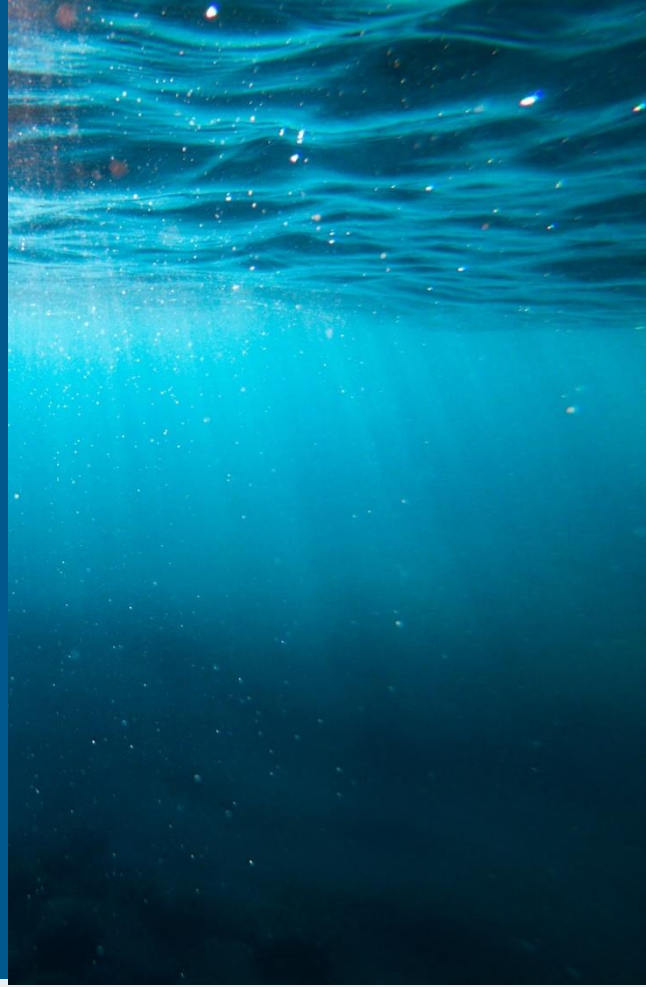
Woods Hole Oceanographic Institution
1930

Florida State University
1851

Scripps Institution of Oceanography

Overview

Karen Stocks



R2R at RVTEC

R2R Rolling Deck to Repository

Dru Clark

Gwynne Hayes

Frank Nitsche

Snehal Prabhu

Alex Strong

Rafael Uribe

George Dubinin

Becca Hudak

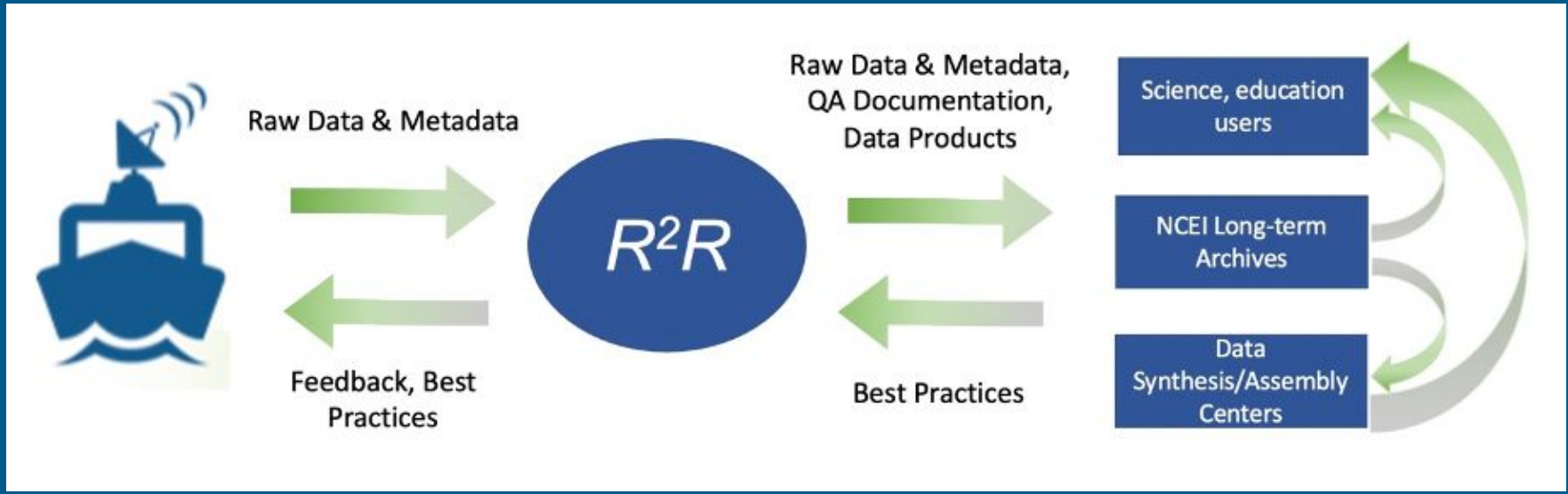
Shawn Smith

Karen Stocks

Please Come for a Vessel Check-up

Lunchtime Wednesday, Lamprey room

Data Flow


Raw Data & Metadata
Feedback, Best Practices
R2R
Raw Data & Metadata, QA Documentation, Data Products
Best Practices
Science, education users
NCEI Long-term Archives
Data Synthesis/Assembly Centers

Organize, Archive, and Disseminate Original Underway Data & Documents (~55,000 file sets)

UNDERWAY DATA SETS

DEVICE TYPE	MAKE-MODEL [LOCATION]	DOI	DATA
ADCP	Hawaii UHDAS	10.7284/147284	
Expendable Probe	Turo Quoll	10.7284/147278	
GNSS	Trimble BD982	10.7284/147280	
GNSS	Hemisphere V104S	10.7284/147294	
INS	Kongsberg Seapath 330+	10.7284/147285	
INS	iXSea iXSea [gps]	10.7284/147293	
Met Station [documentation]	SIO MET System	10.7284/147281	
Multibeam Sonar	Kongsberg EM712	10.7284/147282	
Multibeam Sonar	Kongsberg EM124	10.7284/147297	
Singlebeam Sonar	Knudsen 3260	10.7284/147283	
TSG	SeaBird SBE-45	10.7284/147279	

Assess Data Quality (Select Devices)

R2R QA Dashboard - Falkor Magnetometer

Rating	Fileset ID	Dataset	Cruise	Vessel		all_files_have_valid_checksum	all_files_in_valid_format	all_records_have_valid_time	minimal_metadata_exists	minimal_metadata_is_correct	minimum_layback_metadata_exists	percent_expected_data_files	percent_files_not_empty	percent_records_in_valid_igrf_range	percent_records_without_read_errors	percent_records_without_spikes
●	115750	Magnetometer *Geometrics G-882*	FK140418	Falkor	Docs & Plots	●	●	●	●	●	●	●	●	●	●	●
●	115753	Magnetometer *Geometrics G-882*	FK140613	Falkor	Docs & Plots	●	●	●	●	●	●	●	●	●	●	●
●	123609	Magnetometer *MarineMagnetics SeaSPY2*	FK151005	Falkor	Docs & Plots	●	●	●	●	●	●	●	●	●	●	●
▲	115749	Magnetometer *Geometrics G-882*	FK140307	Falkor	Docs & Plots	●	●	▲	●	●	●	●	●	●	●	●
▲	115752	Magnetometer *Geometrics G-882*	FK140502	Falkor	Docs & Plots	●	●	▲	●	●	▲	▲	●	●	●	●

Produce Quality-Controlled Data Products (Select Devices). ~14,000 products made

PROCESSED PRODUCTS

DATA TYPE	DEVICE TYPES(S)	PRODUCER	FORMAT	DATA
Bathymetry	singlebeam (Knudsen 3260)	Rolling Deck to Repository	r2rbathy_geocsv	
CTD	ctd (SeaBird SBE-911+)	Rolling Deck to Repository	seasoft-proc	
Expendableprobe	expendableprobe (Turo Quoll)	Rolling Deck to Repository	r2rxprobe_geocsv	
Navigation	ins (Kongsberg Seapath 330+)	Rolling Deck to Repository	r2rnav_geocsv	

Publish Cruise Catalog w/ Cruise DOIs

Sally Ride

Home / Search / Sally Ride / SR1918

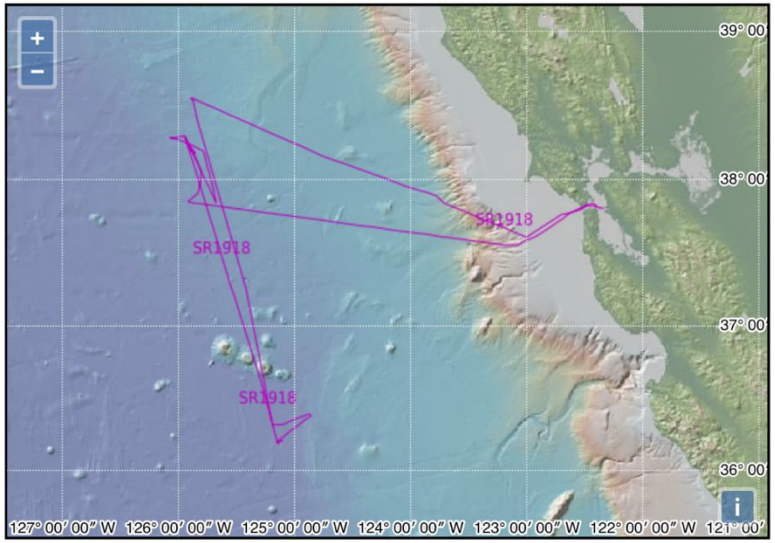
Operator: Scripps Institution of Oceanography

Cruise DOI: 10.7284/908624

Sally Ride

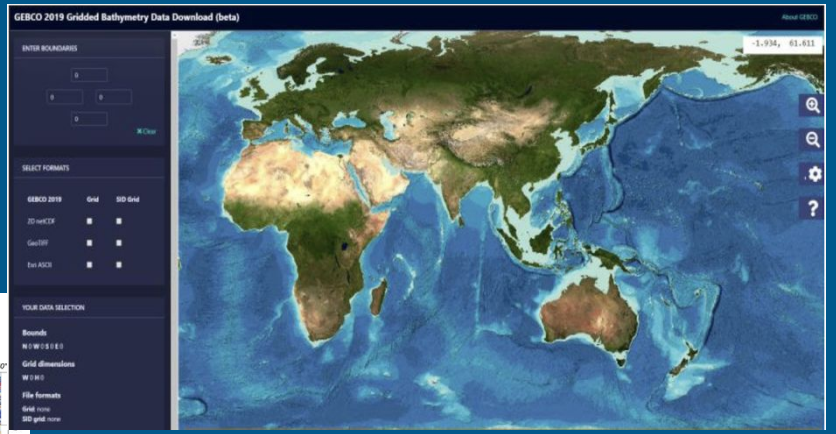
Cruise ID

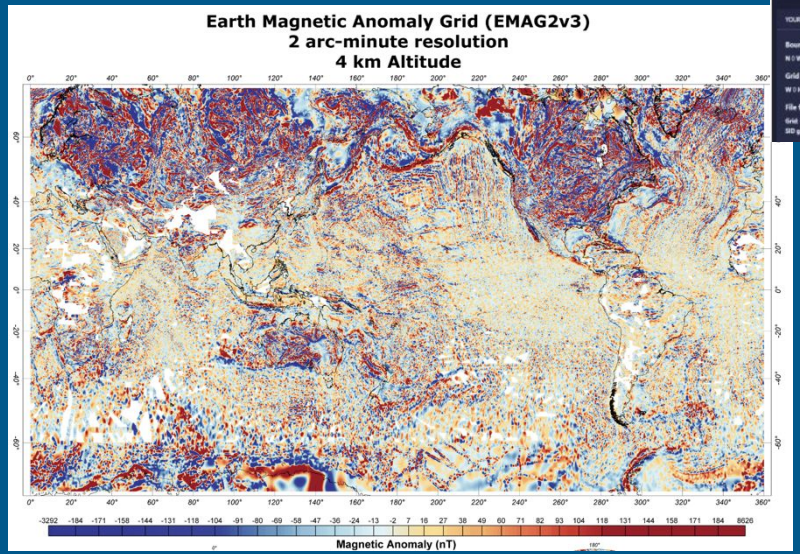
SR1918

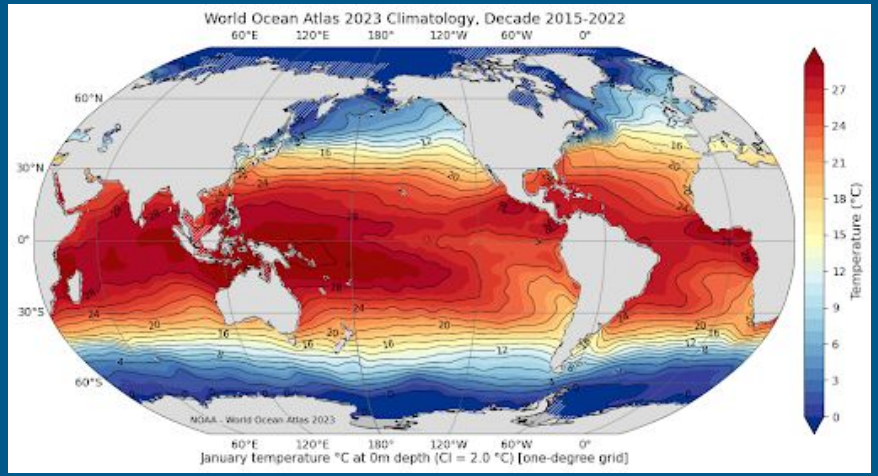


CRUISE ID	SUMMARY	START DATE	START PORT	END DATE	END PORT
SR1918	Project: Low-mode internal tide decay from the Mendocino Ridge Chief: Savage, Anna	2019-12-01	Alameda, California	2019-12-08	San Francisco, California

NCEI & Global Products







At-Sea Event Logging - Science and Technician Events (Becca and Snehal)

Return_to_Main_Page | R/V Atlantic Explorer | R/V Atlantis | R/V Endeavor | R/V F.G. Walton Smith | R/V Hugh R. Sharp | R/V Marcus G. Langseth | R/V Neil Armstrong | R/V R.G. Sproul | R/V Roger Revelle | R/V Sally Ride | R/V Savannah | R/V Sikuliaq | R/V Thompson | R/V NextGen

EN694-SE | EN695-SE | EN696-SE | en697-SE | en697b-SE | en697c-SE | en698-SE | EN699-SE | en700-SE | EN701-SE | EN702A-SE | en702b-SE | en703a-SE | en704-SE | en705-SE | EN706-SE | EN707-SE | EN708-SE | EN709-SE

"EN709", "R/V Endeavor", "", "", "", "", "", "", "", "", Page 1 of 10

ELOG

List | Find | Help

Summary | Threaded

-- Author -- | -- Instrument -- | -- Action --

192 Entries

Goto page 1, 2, 3 ... 8, 9, 10 Next All

Event	dateTimeUTC	GPS_Time	Instrument	Action	Transect	Station	Cast	Latitude	Longitude	Seafloor	Author	Comment
20230904.1303.001	20230904.1303	2023/09/04 13:03:49	Ship	startCruise	NaN	NaN	NaN	41.523698	-70.672452		lButler1	Lines off from WHOI dock
20230904.1303.002	20230904.1255	2023/09/04 12:55:14	Ship	other	NaN	NaN	NaN	41.523795	-70.672260	16 m	lButler1	Gangway removed from dock at WHOI and 01 deck to storage on board
20230904.1305.001	20230904.1305	2023/09/04 13:05:36	MeteorologicalSensor	other	NaN	NaN	NaN	41.523220	-70.672742	16 m	lButler1	met sensors operating and logging as usual
20230904.1305.002	20230904.1305	2023/09/04 13:05:59	Echosounder12	start	NaN	NaN	NaN	41.523095	-70.672790	16 m	lButler1	12 KHz started just after leaving the dock
20230904.1324.001	20230904.1454	2023/09/04 14:54:04	UHDAS for both ADCPs	start	NaN	NaN	NaN	41.352133	-70.922173	26 m	lButler1	UHDAS controls both OS75 and WH300 KHz ADCPs. Dan Torres modified the settings. Started south of Cuttyhunk.
20230904.1459.001	20230904.1230	2023/09/04 12:30:04	Other	other	NaN	NaN	NaN	41.523793	-70.672267		lButler1	Pre-cruise 10 minute orientaion
20230904.1500.001	20230904.1400	2023/09/04 14:00:44	Other	start	NaN	NaN	NaN	41.437463	-70.782648		lButler1	1.5 hour safety orientation video and talk by 2nd Mate
20230904.1951.001	20230904.1941	2023/09/04 19:41:49	Fluorometers on underway impeller	clean	NaN	NaN	NaN	41.779095	-70.487005		bClarke1	clean fluorometers before starting impeller pump flow through system and turning them on
20230904.2014.001	20230904.2002	2023/09/04 20:02:04	Other	start	NaN	NaN	NaN	41.786663	-70.442190		anOther	bag transfer
20230904.2014.002	20230904.2015	2023/09/04 20:15:19	Thermosalinographs on underway impeller	start	NaN	NaN	NaN	41.802885	-70.433955	24 m	lButler1	started sci sw impeller pump. SBE45 and SBE21 with fluorometers.
20230905.2233.001	20230905.1430	2023/09/05 14:30:39	Other	other	NaN	NaN	NaN	42.671317	-69.059072		lButler1	CTD training
20230907.1247.001	20230907.1248	2023/09/07 12:48:04	Thermosalinographs on underway impeller	other	NaN	NaN	NaN	43.339933	-61.930870		bClarke1	Turn off flow to SBE21 for maintenance. SBE21 salinity reading ~0.1PSU lower than value from SBE45. Stop SeaSave acq for SBE21
20230907.1251.001	20230907.1200	2023/09/07 12:00:49	Thermosalinographs on underway impeller	other	NaN	NaN	NaN	43.338192	-62.115668		lButler1	increase and decrease flow to possibly correct an offset in conductivity values between the SBE21 and SBE45
20230907.1351.001	20230907.1324	2023/09/07 13:23:59	Thermosalinographs on underway impeller	other	NaN	NaN	NaN	43.341535	-61.795250		bClarke1	Start Seasave acquisition file 07Sept2023a.hex after Triton-X and freshwater rinse of SBE21
20230907.1354.001	20230907.1328	2023/09/07	Thermosalinographs on	other	NaN	NaN	NaN	43.341657	-61.780355		bClarke1	SBE21 salinity reading ~.3PSU lower than SBE45. Stop SBE21 Seasave

SAMOS - Near Real Time MET and TSG Evaluation (Shawn)

History

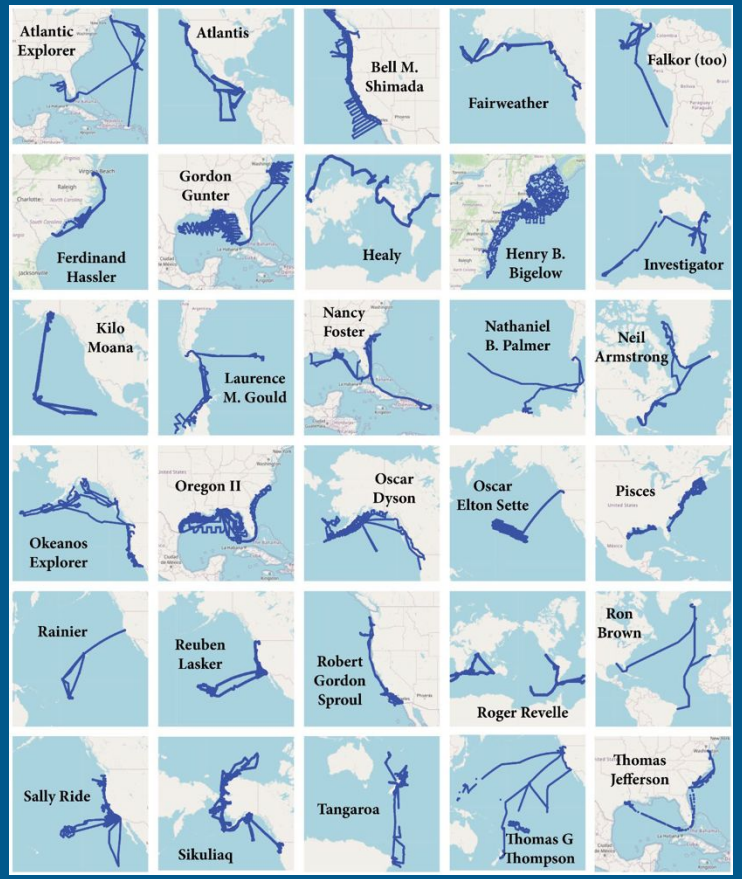
- Providing high-quality underway meteorological and oceanographic data from research vessels (RV) to the scientific community since 2005

Users

- Satellite algorithm developers
- Air-sea exchange (flux) researchers
- Ocean and atmosphere modelers
- Operational forecasters
- Geoinformatics, Ocean Best Practices

Status

- 30 vessels active in 2023
  - 2024: *Nuyina* added, *LM Gould* retired
- 9.23 million 1-minute data records processed in 2023
- Global coverage, concentrated around North America.



Coming Next

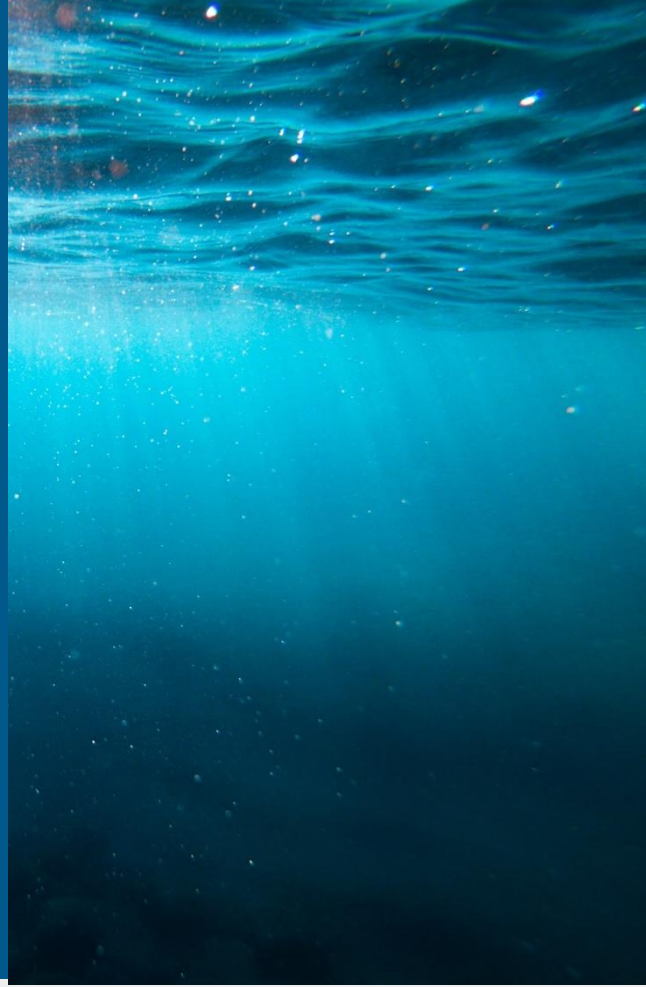
New proposal pending

- R2R and FAIR Principles
- Event Log Harvesting

Recommended Practices

R2R and FAIR Principles

Shawn Smith



Summary of Principles

- The Findability (F) principles focus on making data and digital resources discoverable.
  - To achieve this, data should be assigned globally unique and persistent identifiers (such as DOIs or URIs), cataloged, and described using rich metadata for search and discovery.

- The Accessibility (A) principles emphasize that data and digital resources should be accessible with persistent metadata through an open, free, standardized communication protocol that allows for authentication and authorization procedure when appropriate.

Reproduced from Peng, G., R. R. Downs, H. K. Ramapriyan, Y. Wei, B. Ramachandran, M. Parsons, Z. Liu, and NASA O'FAIR WG, 2024: A Practical Guide for Open, Free & FAIR NASA Earth Science Data Products. Document ID: NASA-OFAIR-ESDSWG-DOC-0002. Version: v00r05-20240614. CC0 1.0 + Attribution. https://doi.org/10.5067/DOC/ESCO/ESDSWG-0002V1

Summary of Principles

- The Interoperability (I) principles aim to facilitate seamless integration and exchange of data and digital resources across different systems, tools, services, and research domains.
  - To achieve this, it suggests that data and digital resources should be structured using standardized and widely accepted data models, formats, and vocabularies, including references to other relevant data and metadata, for example, a reference to utilized metadata standards.

- The Reusability or R principles promote the creation of data and digital resources that can be readily understood, interpreted, and (re)used.
  - This involves adhering to domain-relevant standards and providing comprehensive data documentation, including clear data usage license and detailed provenance.

Reproduced from Peng, G., R. R. Downs, H. K. Ramapriyan, Y. Wei, B. Ramachandran, M. Parsons, Z. Liu, and NASA O'FAIR WG, 2024: A Practical Guide for Open, Free & FAIR NASA Earth Science Data Products. Document ID: NASA-OFAIR-ESDSWG-DOC-0002. Version: v00r05-20240614. CC0 1.0 + Attribution. https://doi.org/10.5067/DOC/ESCO/ESDSWG-0002V1

R2R FAIR Status

- Findability
  - R2R data packages (bags) include
    - Standard, machine readable metadata (DataCite formats)
    - Persistent identifiers (DataCite DOI)
    - Linked to well-curated R2R cruise catalog
    - Reciprocal links to other data repositories
  - Cruise metadata also curated with DOIs
  - R2R website supports schema.org metadata for discovery
- Accessibility
  - Datasets are linked via URLs that resolve either to NCEI or R2R landing pages
  - APIs supported by R2R are listed here https://service.rvdata.us

R2R FAIR Status

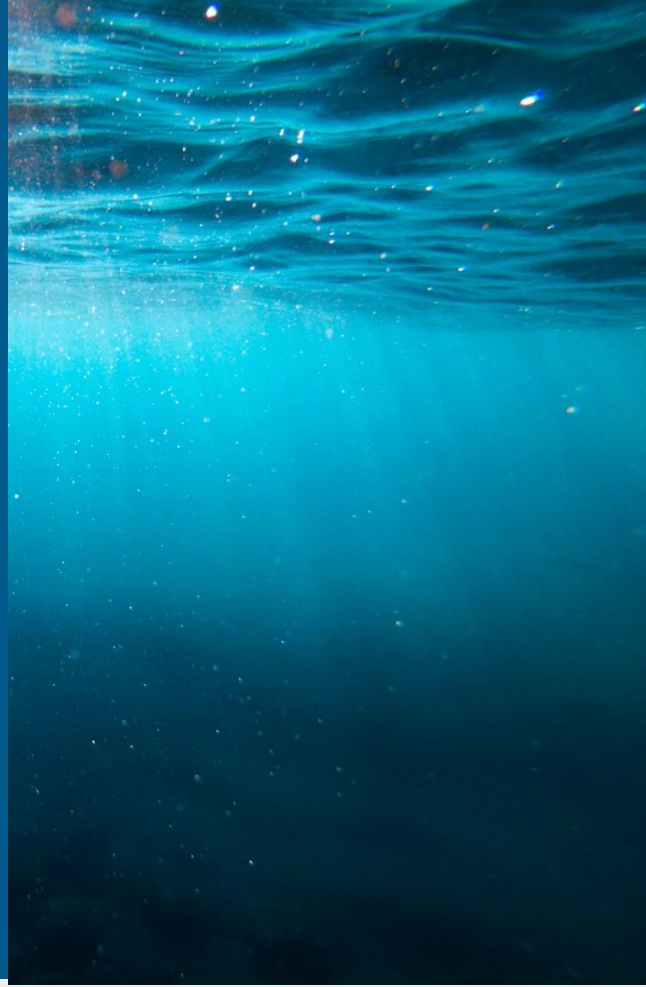
- Interoperability
  - R2R has done a reasonably good job of using controlled vocabs for the metadata it generates (https://www.rvdata.us/about/technical-details/vocabularies) but **inside** the files things are often uncontrolled.
  - Only some data sets/device types have defined data formats
  - Multiple data acquisition system capture original device data in a variety of formats, sampling rates, etc.
- Reusability
  - R2R uses Creative Commons CC0 license
    - Is this applicable/transferable when NCEI serves data submitted by R2R?
  - Comprehensive dataset documentation is lacking or hard to identify for some datasets
  - Domain-relevant standards may not exist.
  - Data provenance not always clear between datasets and observing systems/devices

Proposed FAIR Activities

- R2R plans to move the ARF towards FAIR underway data collection, documentation, and distribution practices by establishing a FAIR data working group
- Composition of WG
  - R2R team members
  - Marine techs
  - Data managers and archivists
  - FAIR experts
- Charge to WG
  - Review existing FAIR implementation plans and procedures in the geosciences
  - Assess the FAIR readiness for ARF underway data using existing assessment tools
  - Develop device-specific plans moving the ARF towards FAIR at the time of data collection
  - Develop guidelines and protocols to align ARF underway data with documented principles.

Event Log Harvesting

Becca Hudak



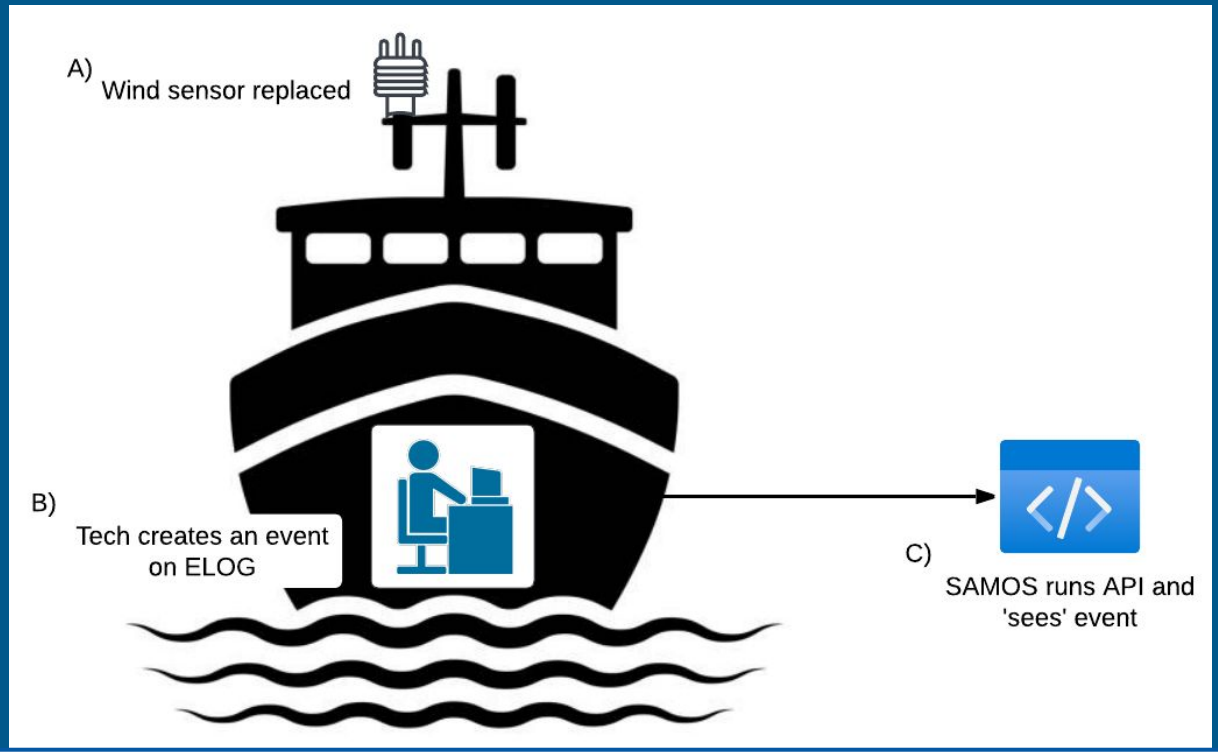
Machine to Machine Harvesting of the R2R Eventlog

- The environmental sensor data streams collected during a cruise are of high value for later reuse, particularly as they contribute to building global syntheses and historical time series observations of ocean properties.

- M2M access to Eventlog data may allow SAMOS to

  (a) maintain up-to-date metadata (e.g., sensor swaps, cal changes) and

  (b) improve data flagging/QC by noting elog documented sensor failures or sensor impacting events (e.g., icing, roosting birds, etc).



qq_082223-SE

qq_082223-SE, Page 1 of 1

ELOG

New | Edit | Delete | Duplicate | Find | Help

Summary | Threaded

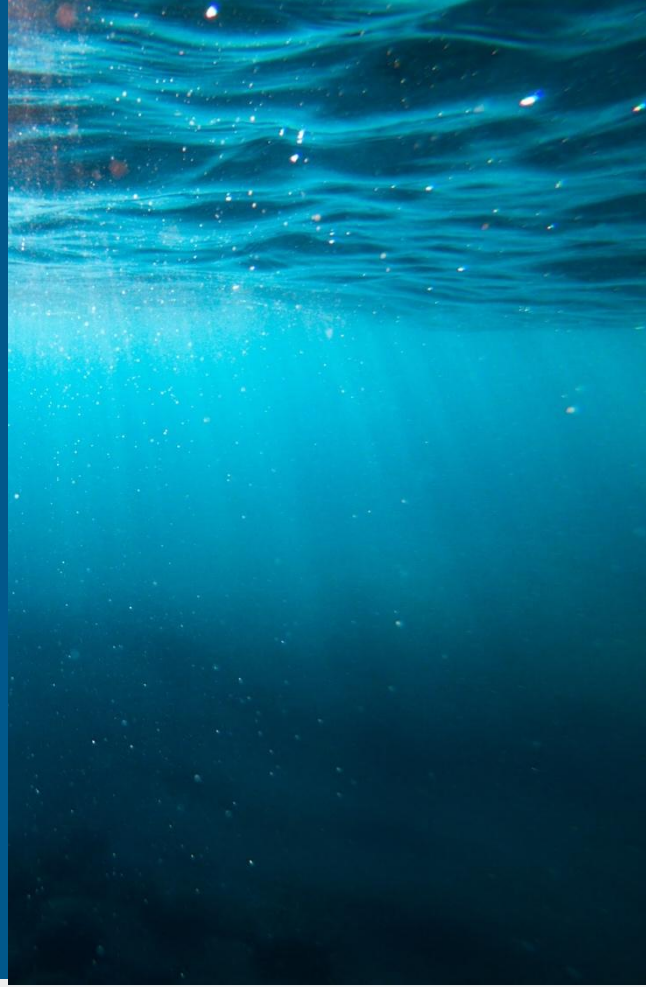
-- Author -- | -- Instrument -- | -- Action -- | 3 Entries

Event	dateTimeUTC	GPS_Time ▲	Instrument	Action	Transect	Station	Cast	Latitude	Longitude	Seafloor	Author	Comment
20230831.1348.001	**20230831.0900**	**2023/08/31 08:59:58**	**Vaisala WXT520**	**service**	**NaN**	**NaN**	**NaN**	**41.524553**	**-70.671048**		**rHudak1**	**Sensor 111348 swapped out for sensor 111874; calibration files attached**
20230831.1347.001	20230831.1347	2023/08/31 13:48:00	Vaisala WXT520	start	NaN	NaN	NaN	41.524502	-70.671041		rHudak1	Sensor started at dock, cruise starts tomorrow.
20230912.1348.001	20230901.0948	2023/09/01 09:48:21	Vaisala WXT520	stop	NaN	NaN	NaN	41.524507	-70.670988		rHudak1	Met sensor collection stopped due to carriage return issue; will restart when issue is resolved

ELOG V2.9.0-2411

Recommended Data Practices

George Dubinin, Gwynne Hayes



Timely Data Submission

- Our goal is to receive new cruise data distributions (distros) quarterly basis or more frequently.
- These data are important to science
- Timely data submission expedites distro ingestion and gives R2R more time to verify and validate network-transmitted data.

Avoiding Distro Updates

When we send data to an archive, like NCEI, they are difficult to modify

- It's much harder to update/change a distro than to add new files.
- If data already has been sent to NCEI this makes updates complicated. We prefer:
  - to serve data as-is and fill in gaps with documentation rather than manual changes
  - to receive the distros complete

Separating Oversize Filesets

- We would like to receive massive filesets (+1TB, eg SBP-29 & EK-80) as separate tarballs from the regular underway data.
- Please preserve the directory structure such that, if merged into the original distro, it would appear unmodified

Include cruise.md5deep Files

- Ensure that your distro contains a .md5deep file or an md5_summary.txt file
- Check that it has a plausible number of lines (not suspiciously truncated)
- Avoid skipping the generation of the md5 hashes
    - Files will often have a checksum resembling “*********” - the integrity of these files cannot be verified.

Safe and Effective File Naming

- It’s helpful to have date strings in filenames. We use these strings to verify what date corresponds to which file.
  - Avoid Julian Day date/time as it interferes with our automated date checking processes.
- Don’t put special characters in filenames.
  - *Seapath 330 User's Guide.pdf* <- this shouldn’t have an apostrophe!!
  - *Knudsen comparison with SBP39_top line of Oceanus Z_W-bound_same line, same time.jpg* <- this shouldn’t have commas or spaces

Standardizing the Directory Structure

By consolidating device formats you make it easier for us to maintain our mapping to ensure that we are picking up all useful data.

- Acquisition systems are a means of standardizing the directory structure for serial devices.
- Many of these perform other bookkeeping like enforcing valid filenames.
- Consistency helps us identify missing data.
- A standardized directory structure helps scientists who work across multiple vessels.

Network Transfers (Globus)

Our preferred network transfer method is Globus

- Globus performs checks on all transmitted data to ensure file integrity remains uncompromised.
- Pushing data to R2R's endpoint does not require a subscription.
  - If you are pushing a particularly large distro, please let us know first via info@rvdata.us beforehand!
  - We will help you with setting up Globus!
- Globus also ensures security of the transfer.

Data-Specific Asks

- Double-check for empty files i.e. not zero-byte but missing required data.
- Geophysical data: Include gravity ties and magnetometer layback in the distro to ensure R2R has current information for QA/DP.
  - Include documentation for devices in a top-level docs directory
- Confirm water column sonar data contains navigation.
  - Lack of navigation in wcd data creates downstream issues for NCEI archival.

Regularly Maintain Device Format Documentation

- R2R serves device data format documentation to the end user
  - The end user directly benefits from detailed and current device documentation
- R2R is looking to overhaul our system to work with these documents to introduce automation and checks on the data

```
Data Format Description:
The file contains GGA, VTG, and ZDA strings, alternating line-by-line,
preceeded by decimal unix system timestamps.  Definitions follow.

Definition of Unix time:
    Decimal seconds since 00:00:00 UTC - January 1, 1970 (minus leap seconds)



Definition of GGA:
GGA - essential fix data which provide 3D location and quality data.

Example Record: $GPGGA,123519,4807.038,N,01131.000,E,1,08,*47

where:
     GGA          Global Positioning System Fix Data
     123519       Fix taken at 12:35:19 UTC
     4807.038,N   Latitude 48 degrees 07.038 minutes North
     01131.000,E  Longitude 11 degrees 31.000 minutes East
     1            Fix quality: 0 = fix not available or invalid
                               1 = GPS Standard Positioning Service (SPS) mode, fix valid
                               2 = differential GPS (DGPS), SPS mode, fix valid
                               3 = GPS Precise Point Positioning (PPS) mode, fix valid
     08           Number of satellites being tracked
     *47          Checksum, always begins with *
```

Format Documentation Example

- High level description of the type of data
- Listing of the NMEA tags found in the data

- Short description for each NMEA tag about what its data describes

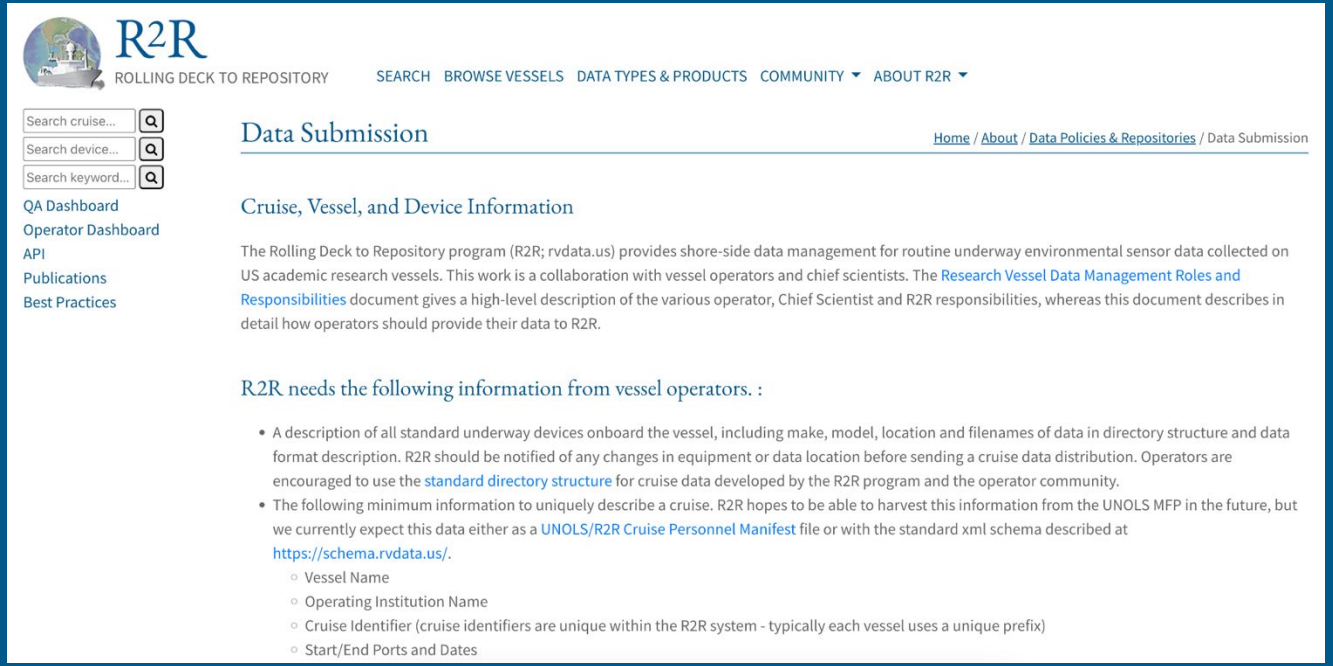
- Breakdown of what each column value in an example data line corresponds to

```
Data Format Description:
The file contains GGA, VTG, and ZDA strings, alternating line-by-line,
preceeded by decimal unix system timestamps.  Definitions follow.

Definition of Unix time:
    Decimal seconds since 00:00:00 UTC - January 1, 1970 (minus leap seconds)



Definition of GGA:
GGA - essential fix data which provide 3D location and quality data.

Example Record: $GPGGA,123519,4807.038,N,01131.000,E,1,08,*47

where:
     GGA          Global Positioning System Fix Data
     123519       Fix taken at 12:35:19 UTC
     4807.038,N   Latitude 48 degrees 07.038 minutes North
     01131.000,E  Longitude 11 degrees 31.000 minutes East
     1            Fix quality: 0 = fix not available or invalid
                               1 = GPS Standard Positioning Service (SPS) mode, fix valid
                               2 = differential GPS (DGPS), SPS mode, fix valid
                               3 = GPS Precise Point Positioning (PPS) mode, fix valid
     08           Number of satellites being tracked
     *47          Checksum, always begins with *



Definition of ZDA:
ZDA - Date and Time

Example Record: $GPZDA,201530.18,04,07,2002,01,00*60

where:
     201530.18  hours, minutes, and seconds expressed as hhmmss.ss
                 20       2-digit hour [24 hour clock]
                 15       2-digit minute
                 30.18    decimal seconds
     04         2-digit day,
     07         2-digit month
     2002       4-digit year
     01         2-digit local timezone hours: -13 to 13
     00         2-digit local timezone minutes: 0 to 59
     *60        Checksum, begins with *
```

Documentation Request List

- Md5deep checksum file for entire distro
- Magnetometer layback (usually part of the cruise-level metadata doc)
- Gravity ties (usually a PDF in the same directory as magnetometer layback)
- Format documents for all devices (in the format previously shown)

R2R Data Submission

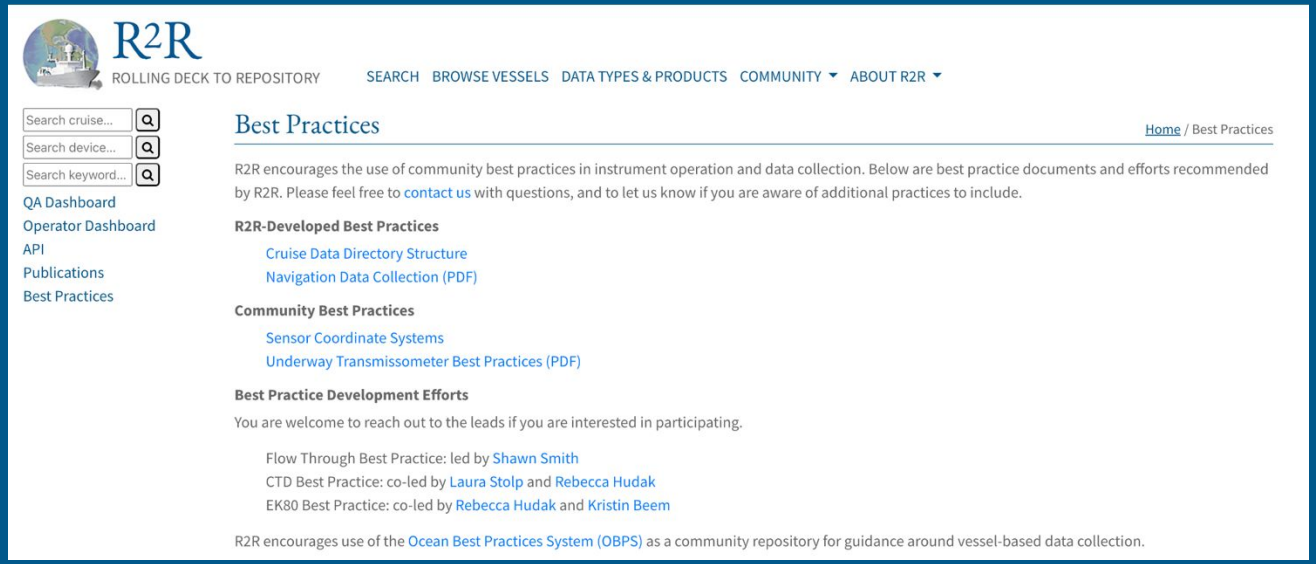


For more information refer to the Data Submission page of rvdata.us

https://www.rvdata.us/about/data-policies-and-repositories/data-submission

R2R Best Practices

For more information refer to the Best Practices page of rvdata.us



https://www.rvdata.us/best_practices

Contact Us: info@rvdata.us

Dru Clark, SIO

Karen Stocks, SIO

George Dubinin, SIO

Snehal Prabhu, WHOI

Rebecca Hudak, WHOI

Rafael Uribe, LDEO

Frank Nitsche, LDEO

Gwynne Hayes, LDEO

Alex Strong, LDEO

Shawn Smith, FSU

Questions for the Community:

- What is your biggest challenge when working with R2R?
- Is there any extra guidance that R2R can provide based on our requests? Other ways we could help?
- Any other feedback?

Acknowledgements

Lamont-Doherty Earth Observatory
Columbia University | Earth Institute

Woods Hole Oceanographic Institution 1930

Florida State University 1851

Scripps Institution of Oceanography

NSF

Office of Naval Research ONR Science & Technology

NOAA National Oceanic and Atmospheric Administration U.S. Department of Commerce

Schmidt Ocean Institute

University National Oceanographic Laboratory System UNOLS