Foundational practices for Findable, Accessible, Interoperable, and Reusable (FAIR) data

Module 1
FAIR data training for Deep Ocean Early career Researchers (DOERs)
February 2023
https://tinyurl.com/DOERs-data-module1
Welcome to FAIR data training for DOERs!

New to iDOOS?
Implementing a Deep Ocean Observing Strategy project funded by the U.S. National Science Foundation

DOOS, more broadly:
an international, interconnected network of deep-ocean observing, mapping, exploration, and modeling programs endorsed as a UN Ocean Decade Programme
Introductions

Your instructor and helpers

Stace Beaulieu
Karen Stocks
Leslie Smith

iDOOS
SeaFAIRers

https://storymaps.arcgis.com/stories/6b70811732884162b35da102b5a23fab
iDOOS Principles of Collaboration [link]

iDOOS is committed to open science and to training a future generation of deep-sea leaders with the skills to create a FAIR and inclusive landscape. We seek to create a welcoming, respectful, inclusive, and professional environment that supports the open exchange of ideas and a diversity of perspectives.

iDOOS uses the Code of Conduct [here] established by Pangeo, a community of people working collaboratively to develop software and infrastructure to enable Big Data geoscience research.
How we will use Zoom

- Live transcription
- Chat
- Reaction icons:
  - Emojis
  - Yes, No, Slow down
  - Raise hand *
- Poll
- Breakout rooms

* Please try private chat with SeaFAIRer prior to Raise hand
What makes this course different?

It’s not just the content…
It’s how we will engage with the content, and how we will engage with each other

Learning by DOing
Shield-shaped model of interdisciplinary education

We are all data scientists!
Breadth of transferable skills

Bosque-Perez et al. (2016) https://doi.org/10.1093/biosci/biw042
Practical understanding of one or two other disciplines, sufficient to enable collaboration with researchers from those disciplines

Figure from Pennington et al. (2020) https://doi.org/10.1007/s11625-019-00735-3
Poll

Please rate your current level of satisfaction with your own data management practices.

How often do you code or program as part of your research?

I am confident with regard to how to make my own data Findable, Accessible, Interoperable, and Reusable (FAIR).
What will we learn in this course?

Foundational practices supporting the research data lifecycle, as well as practices specific to the ocean sciences

Infographic [link] (Stocks et al., 2022, doi:10.6084/m9.figshare.21766709.v1)

100+ yrs in < 8 hrs
What will we learn in this course?

When you complete the course, you may reach this level for some skills

Syllabus [link]

https://en.wikipedia.org/wiki/Bloom%27s_taxonomy
What will we learn in Module 1?

This module sets the foundation for effective data management, focusing on each individual’s practices, and introduces the concepts of Findable, Accessible, Interoperable, and Reusable (FAIR) data.

- **Apply** foundational data management practices to your own research data
- **Understand** concepts of Findable and Accessible (“F” and “A” in FAIR) data
- **Remember** benefits of making research data FAIR

[https://en.wikipedia.org/wiki/Bloom%27s_taxonomy](https://en.wikipedia.org/wiki/Bloom%27s_taxonomy)
What will we learn today? (1st session of Module 1)

Agenda ~15-min each:

• Introductions and Quick dip exercise
• Slides related to Wilson et al. (2017)
• Breakout rooms (cross-discipline): Deep dive exercise to discuss Wilson et al. (2017)
• Slides to lead into Homework
“Good enough practices”

Data management
Software
Project organization
(and 2nd session will touch on Keeping track of changes)

“Good enough practices” Data management

Save the raw data.

Ensure that raw data are backed up in more than one location.

Create the data you wish to see in the world.

Create analysis-friendly data.

Record all the steps used to process data.

Anticipate the need to use multiple tables, and use a unique identifier for every record.
[Module 3]

Submit data to a reputable DOI-issuing repository so that others can access and cite it.
[Module 4]

Wilson et al. (2017)
Let's highlight practices implemented in iDOOS

Implementing a Deep Ocean Observing Strategy (iDOOS) within the Global Ocean Observing System (GOOS)

<table>
<thead>
<tr>
<th>Observing &amp; Exploration Networks</th>
<th>Data &amp; Modelling CI Networks</th>
<th>Management &amp; Policy Users</th>
</tr>
</thead>
<tbody>
<tr>
<td>Argo (Core, BGC, Deep)</td>
<td>CCHDO</td>
<td>AtlantOS</td>
</tr>
<tr>
<td>Challenger 150 (DOSI/SCOR)</td>
<td>CLIVAR GSOP</td>
<td>DOSI</td>
</tr>
<tr>
<td>EMSO</td>
<td>CLIVAR/OMDP &amp; CMIP</td>
<td>GEO-BluePlanet</td>
</tr>
<tr>
<td>GEO BON</td>
<td>ECCO</td>
<td>GOOS</td>
</tr>
<tr>
<td>GO-SHIP</td>
<td>EMODnet</td>
<td>Internat. CLIVAR</td>
</tr>
<tr>
<td>iAtlantic</td>
<td>ESIP/MDC</td>
<td>InterRidge</td>
</tr>
<tr>
<td>MBON</td>
<td>Esri</td>
<td>ISA</td>
</tr>
<tr>
<td>NDSF/UNOLS</td>
<td>FathomNet</td>
<td>IUCN</td>
</tr>
<tr>
<td>NOAA Ocean Exploration</td>
<td>ICODE/ODIS</td>
<td>POGO</td>
</tr>
<tr>
<td>OceanSITES</td>
<td>IRIS</td>
<td>UN Decade</td>
</tr>
<tr>
<td>OECI</td>
<td>ISA DeepData</td>
<td>UN Global Compact</td>
</tr>
<tr>
<td>ONC</td>
<td>Mercator Ocean</td>
<td>U.S. CLIVAR</td>
</tr>
<tr>
<td>OOI</td>
<td>OBIS</td>
<td>U.S. OCB</td>
</tr>
<tr>
<td>ReV Ocean</td>
<td>OBPS</td>
<td>U.S. Sanctuaries</td>
</tr>
<tr>
<td>Schmidt Ocean Institute</td>
<td>OceanPredict                &amp; Monuments</td>
<td></td>
</tr>
<tr>
<td>JTF SMART Cables</td>
<td>Seabed 2030</td>
<td></td>
</tr>
<tr>
<td>SOOS/SOCCOM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TPOS 2020</td>
<td></td>
<td></td>
</tr>
<tr>
<td>US-IOOS</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

https://www.deepoceanobserving.org/pages/about-doos
"Good enough practices" Data management

Save the raw data.

Ensure that raw data are backed up in more than one location.

Pulse: Yes (green check) or No (red x)

Do you need to save terabytes (or more) of data?

SeaFAIRer Michael F. Vardaro (UW/OOI-RCA)

<table>
<thead>
<tr>
<th>Name</th>
<th>Last modified</th>
<th>Size</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parent Directory</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A503CPPM/</td>
<td>2023-02-03 16:26</td>
<td></td>
<td>Coastal Pioneer MAB - At-Sea Test 3 Profiler Mooring</td>
</tr>
<tr>
<td>A503CP5H/</td>
<td>2023-02-03 16:23</td>
<td></td>
<td>Coastal Pioneer MAB - At-Sea Test 3 Surface Mooring</td>
</tr>
<tr>
<td>CE01ISSM/</td>
<td>2023-01-25 20:40</td>
<td></td>
<td>Coastal Endurance - OR Inshore Surface Mooring</td>
</tr>
<tr>
<td>CE01ISSP/</td>
<td>2022-10-26 19:03</td>
<td></td>
<td>Coastal Endurance - OR Inshore Surface Piercing Profiler Mooring</td>
</tr>
</tbody>
</table>
“Good enough practices” Data management

Create the data you wish to see in the world.
Create analysis-friendly data.


SeaFAIRer Carolina Berys
CCHDO (CLIVAR and Carbon Hydrographic Data Office)
“Good enough practices” Data management

Record all the steps used to process data.

Input data → Data processing
- Step 1
- Step 2
- Step n → Output products

Script(s)
Place a brief explanatory comment at the start of every program.
Give functions and variables meaningful names.
Make dependencies and requirements explicit.

Always search for well-maintained software libraries that do what you need. [Module 3]
Test libraries before relying on them. [Module 3]
Submit code to a reputable DOI-issuing repository. [Module 4]

(highlighting 3 of 10 practices)

Wilson et al. (2017)
“Good enough practices” Software

Place a brief explanatory comment at the start of every program. Give functions and variables meaningful names. Make dependencies and requirements explicit.

Pulse: Yes or No
Do you need high-performance computing to process/analyze your data?

Example: Python code to get Argo profiles stored on Argovis
“Good enough practices” Project organization

Project directory (folder)
|-- data
   |-- birds_count_table.csv
   |-- metadata.xml
|-- doc
   |-- manuscript.md
   |-- changelog.txt
|-- LICENSE
|-- README
|-- results
   |-- summarized_results.csv
|-- src
   |-- sightings_analysis.py

Modified from
Box 3. Project layout

Wilson et al. (2017)
**“Good enough practices” Project organization**

Former SeaFAIRer Diana LaScala-Gruenewald

[https://github.com/iobis/dataset-edna](https://github.com/iobis/dataset-edna)

```text

```
```

Biologists: Also see “FAIR in (biological) practice” [Exercise 4: Typical folder organizations](#)
“Good enough practices”
https://doi.org/10.1371/journal.pcbi.1005510

Box 1. Summary of practices

Data Management *

* We can start the discussion with Data management, but please feel free to discuss Software or Project organization if that is of more interest to you.
Follow-up to breakout rooms

Feel free to share in the Zoom Chat which practice you chose to discuss in the breakout room

“Good enough practices” [https://doi.org/10.1371/journal.pcbi.1005510](https://doi.org/10.1371/journal.pcbi.1005510)
Relating “Good Enough Practices” to FAIR

It’s not just data, but also the metadata...

It’s not just code, but the environment in which the code is run...

And then it’s how you organize all of that...

And then it’s how you share all of that...
Learning by DOing: Finding and Accessing “F” and “A” in FAIR data

Can we Find and Access the data and/or software that support the results in:

https://doi.org/10.1002/ecy.3418
Open Research

Dates and locations of colonization sampler deployments and recoveries (Mullineaux 2020b) are available from the Woods Hole Open Access Server (WHOAS) repository at https://doi.org/10.26008/1912/bco-dmo.733210.3. Counts of colonists collected from colonization surfaces are provided in Data S1 (CountsPerSpeciesPerSandwich_EPR.csv) and are a subset of Mullineaux (2020a) from the Woods Hole Open Access Server (WHOAS) repository at https://doi.org/10.26008/1912/bco-dmo.733173.2. Modality assignments for each functional trait are provided in Data S1: ModalitiesPerTraitPerSpecies_EPR.csv and in Dykman et al. (2021) on the Woods Hole Open Access Server (WHOAS) repository at http://doi.org/10.26008/1912/bco-dmo.844993.1. Scripts used for this analysis (Dykman 2021) are provided on Zenodo: http://doi.org/10.5281/zenodo.4625160.

https://doi.org/10.26008/1912/bco-dmo.733173.2
http://doi.org/10.5281/zenodo.4625160
Homework

Choose a journal article relevant to your research and determine if you can Find and Access the data and/or software that support the results.

Use this form to report if you were able to Find and Access:

https://tinyurl.com/DOERs-data-homework1
Feedback “minute card” for 1\(^{st}\) session

Please fill out the form below to let us know what you liked about today and what you didn't. These surveys are anonymous, and we'll use your answers to help improve the course for the next session:

https://tinyurl.com/DOERs-data-feedback
Foundational practices for Findable, Accessible, Interoperable, and Reusable (FAIR) data

Module 1
FAIR data training for Deep Ocean Early career Researchers (DOERs)
February 2023
https://tinyurl.com/DOERs-data-module1
Welcome to 2nd session of Module 1

Your instructor and helpers

- Stace Beaulieu
- Karen Stocks
- Leslie Smith

You!

https://storymaps.arcgis.com/stories/6b70811732884162b35da102b5a23fab
iDOOS Principles of Collaboration [link]

iDOOS is committed to open science and to training a future generation of deep-sea leaders with the skills to create a FAIR and inclusive landscape. We seek to create a welcoming, respectful, inclusive, and professional environment that supports the open exchange of ideas and a diversity of perspectives.

iDOOS uses the Code of Conduct [here] established by Pangeo, a community of people working collaboratively to develop software and infrastructure to enable Big Data geoscience research.
How we will use Zoom

Live transcription *(note: we are not recording)*

Chat

Reaction icons:

- Emojis
- Yes, No, Slow down
- Raise hand *

*[no Poll today]*

Breakout rooms

* Please try private chat with SeaFAIRer prior to Raise hand
Recap 1st session Poll results

Your current level of satisfaction with your own data management practices:

How often you code or program as part of your research:

Your confidence with regard to how to make your own data FAIR:
Recap 1\textsuperscript{st} session

Focused on “Good Enough Practices” article, in particular the Data Management practices (Wilson et al. 2017)

Minute Card Feedback, and feedback through Zoom chat and Slack:
• Added next slide for “Good enough practices” Project organization

Included a slide for Results from Homework (at end of Google slides, not included in PDF)
“Good enough practices” Project organization

Example in previous slides is aligned to Project organization for a publication

Former SeaFAIRer Diana LaScala-Gruenewald

https://github.com/iobis/dataset-edna

At the end of the hour today, I can show you some examples from my own work to address questions posted in DOERs Slack:

“How to create or organize our Project folder for Multidisciplinary project …? Do we need to create project folder for each individual separately or for entire project data…”
What will we learn in 2nd session of Module 1?

Agenda ~15-min each:

• Recap and Quick dip Results from Homework

• Slides related to Wilkinson et al. (2016)

• Breakout rooms: Deep dive exercise to discuss “Thing 6: Data organization and management”

• Concluding slides and resources to learn more
System Output

“Relies on the adoption and use of FAIR principles”
What is FAIR Data?

In order to properly understand FAIR, you should keep four things in mind:

1) Both humans and machines are intended as digesters of data.
2) The FAIR principles apply to both data and metadata.
3) The principles are not necessarily about open data.
4) The FAIR principles are not rules or standards.

https://www.howtofair.dk/what-is-fair/
https://doi.org/10.1038/sdata.2016.18
Findable – the “F” in FAIR data

F1. (meta)data are assigned a globally unique and persistent identifier

F2. data are described with rich metadata

F4. (meta)data are registered or indexed in a searchable resource

(highlighting 3 of 4 principles) Wilkinson et al. (2016)

“To understand and address global deep-sea challenges, we must find and leverage data from across national and international data facilities and programs” - Karen Stocks, our lead iDOOS SeaFAIRer [link]
Accessible – the “A” in FAIR data

A1. (meta)data are retrievable by their identifier using a standardized communications protocol

A2. metadata are accessible, even when the data are no longer available [or not online]

Wilkinson et al. (2016)
Results from Homework: “F” and “A”

F1. (meta)data are assigned a globally unique and persistent identifier

For the comment about the AGU journal:

“AGU requires that the underlying data … be available … Additionally, authors should make available software that has a significant impact on the research. This entails:
1. Depositing the data and software in a community accepted, trusted repository, as appropriate, and preferably with a DOI”

A1. (meta)data are retrievable by their identifier using a standardized communications protocol

For my demo article in Ecology:

“As a condition for publication in ESA journals, all underlying data and novel statistical code pertinent to the results … must be made available in a permanent, publicly accessible data archive or repository …”
Interoperable – the “I” in FAIR data

11. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

12. (meta)data use vocabularies that follow FAIR principles

(highlighting 2 of 3 principles)

Wilkinson et al. (2016)

SeaFAIRer Carolina Berys
https://cchdo.ucsd.edu/

CF netCDF format data now Available
2021-03-13
CCHDO has begun producing CF compliant netCDF files to improve interoperability and machine readability. These will be offered in addition to the existing formats, and will roll out over time as data are converted.

https://cfconventions.org/
Reusable – the “R” in FAIR data

R1. meta(data) are richly described…

R1.1. (meta)data are released with a clear and accessible data usage license

R1.2. (meta)data are associated with detailed provenance

R1.3. (meta)data meet domain-relevant community standards

Wilkinson et al. (2016)

Former SeaFAIRer Diana LaScala-Gruenewald

https://github.com/iobis/dataset-edna
Benefits of making research data FAIR

"...can offer different benefits to... researchers, particularly in terms of opportunities to manage time and storage costs in a more efficient way, and improve collaboration across scientific communities" [Modules 1 & 2]

"...enable easier data integration within and across disciplines, supporting worldwide, multi- and interdisciplinary research endeavours addressing global challenges such as climate change, ... or the realisation of the sustainable development goals" [Modules 3 & 4]

Engelhardt et al. (2022)

The Horror Story links to a YouTube video *. Rather than a Horror Story, let’s talk solutions to challenges...

* Data Sharing and Management Snafu in 3 Short Acts
CC-BY Karen Hanson, Alisa Surkis and Karen Yacobucci
https://youtu.be/N2zK3sA1r-4

Image courtesy SeaFAIRer Dawn Wright Esri
“Good enough practices” Keeping track of changes

Back up (almost) everything created by a human being as soon as it is created.

Keep changes small.


https://doi.org/10.1371/journal.pcbi.1005510
TOP 10 FAIR DATA & SOFTWARE THINGS: Oceanography

Thing 6: Data Organization and Management

Activity 2:
Identifying vulnerabilities
- **Scenario 1:** Your entire office/lab building burns down overnight. No one is harmed, because no one was there, but all electronics in the building perish beyond hope of repair. The next morning, can you access any of your data?
- **Scenario 2:** The cloud server you use (everything from Google Drive to GitHub) crashes. Can you still access your most up to date data?

Discussion 1:
- From either of the two scenarios, can your data survive a disaster? What are some of the things that you think you are doing incorrectly to prevent data loss?

Discussion 2:
- Think about a time when you had or potentially had a data disaster - how could the disaster have been avoided? What, if anything, have you changed about your data storage and workflow as a result?
TOP 10 FAIR DATA & SOFTWARE THINGS: Oceanography

Thing 6: Data Organization and Management

Alternative choice for discussion, relating back to 1st session

Activity 1:
Considerations for basic data organization and management

Group Discussion 1:
- Is your data file structure something that a new lab member could easily learn, or are datasets organized in a more haphazard fashion?
- Do you have any documentation associated describing how to navigate your data structures?

Group Discussion 2:
- Talk about where/how you are currently storing data you are working with. Would another lab member be able to access all your data if needed?
Follow-up to breakout rooms

Feel free to share in the Zoom Chat which discussion you chose in the breakout room

**Discussion 1:**
- From either of the two scenarios, can your data survive a disaster? What are some of the things that you think you are doing incorrectly to prevent data loss?

**Discussion 2:**
- Think about a time when you had or potentially had a data disaster - how could the disaster have been avoided? What, if anything, have you changed about your data storage and workflow as a result?

**Group Discussion 1:**
- Is your data file structure something that a new lab member could easily learn, or are datasets organized in a more haphazard fashion?
- Do you have any documentation associated describing how to navigate your data structures?

**Group Discussion 2:**
- Talk about where/how you are currently storing data you are working with. Would another lab member be able to access all your data if needed?

Thing 6: Data Organization and Management [https://librarycarpentry.org/Top-10-FAIR/2019/01/18/oceanography/](https://librarycarpentry.org/Top-10-FAIR/2019/01/18/oceanography/)

51
Importance of Data Management

From SeaFAIRer Sarah O’Connor @ NOAA

Prevent data loss

• Natural disaster
• Server hardware/software failure
• Format obsolescence
• Malicious attack by human or automated agents
• Human error
• Loss of institutional commitment

From International Coral Reef Symposium (ICRS) Data Management Workshop
Importance of Data Management

From SeaFAIRer Sarah O’Connor @ NOAA

Overwhelming amounts of data
  • Refer to Pulse reactions in 1st session

Better accessibility and reuse
  • “A” and “R” in FAIR data
Importance of Data Management

From SeaFAIRer Sarah O’Connor @ NOAA

More often required than not

Many funding sources (agencies/NGOs) now require:

• A documented plan for organization [Module 2]
• Metadata [Module 3]
• Long-term storage of the research data

Open Access [Module 2]
"Data management is a continuous process"

Sharing at the last minute

Planning for sharing

CC-BY Tomasz Zielinski and Andrés Romanowski
Selected resources to learn more

Data management

• ESIP Data Management Training Clearinghouse (search includes multiple languages https://www.dmtc-devel.org/search)


• “A Graduate Student's Road Map to Data Management Training” Roberts-Pierel et al. (2021), https://doi.org/10.6084/m9.figshare.14384456.v1

Software

• Software Carpentry lessons https://software-carpentry.org/lessons/

• Barker, M., Chue Hong, N.P., Katz, D.S. et al. (2022) Introducing the FAIR Principles for research software. *Sci Data* 9, 622, https://doi.org/10.1038/s41597-022-01710-x

Project organization

• CyVerse Foundational Open Science Skills 2023: Introduction to Project Management
I can:

Apply foundational data management practices to my own research data

Understand concepts of Findable and Accessible (“F” and “A” in FAIR) data

Remember benefits of making research data FAIR
Acknowledgements

Development of these materials was supported by the AccelNet-Implementation: Implementing a Deep Ocean Observing Strategy (iDOOS) project funded by the National Science Foundation (#2114717).

This training program is based on a workshop series developed by data professionals at Woods Hole Oceanographic Institution (Beaulieu et al. 2020, https://doi.org/10.1575/1912/26103)
How to reuse these materials

The authors (Stace E. Beaulieu, ORCID 0000-0002-2609-5453; iDOOS SeaFAIRers) will publish these materials with a CC BY Attribution 4.0 International license.

For questions about this course, contact lead instructor Stace Beaulieu (stace@whoi.edu). For questions about the DOERs program, the iDOOS project, or DOOS, contact Leslie Smith (leslie.smith@youroceanconsulting.com).