



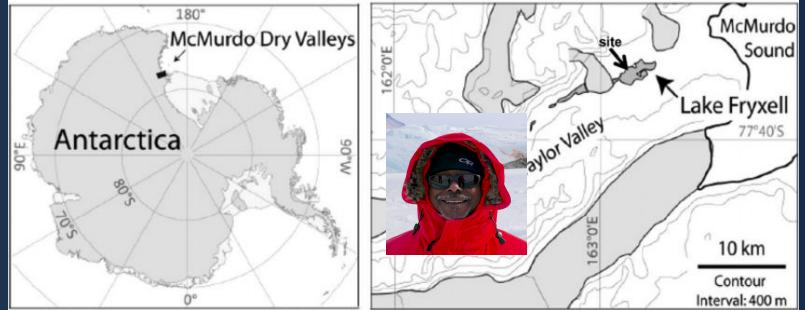
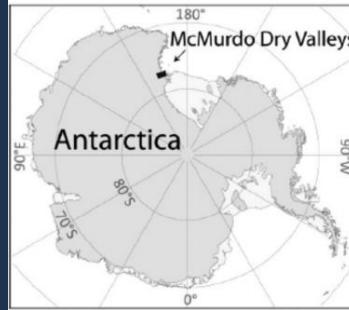
Data Schools

RDM & Open Science

*Presented by Steve Diggs
Atlanta SoRDS: 2022-10-13*

Adapted from original material by: S. "Venkat" Venkataraman

Who Am I?



Scientific Data: Hard to Find, Hard to Use



nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For Authors

Archive > Volume 533 > Issue 7604 > News Feature > Article

NATURE | NEWS FEATURE

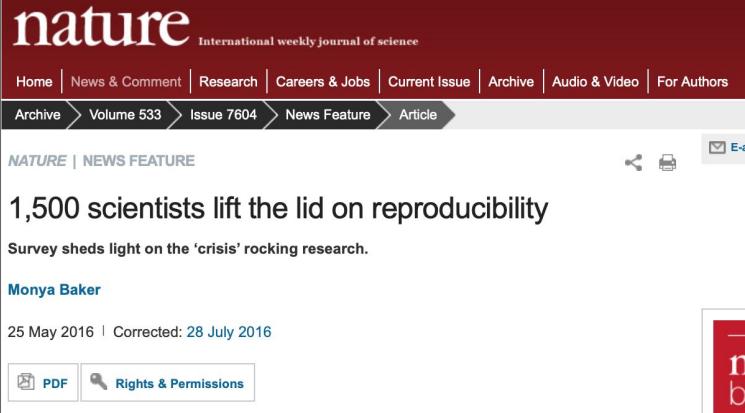
1,500 scientists lift the lid on reproducibility

Survey sheds light on the 'crisis' rocking research.

Monya Baker

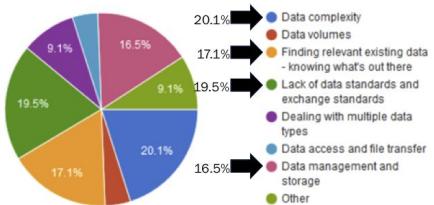
25 May 2016 | Corrected: 28 July 2016

PDF Rights & Permissions



Researcher Challenges with Data Use

The top four issues accounted for 73% of respondents



Data Management Skills Gap Analysis, April 7-2017
<http://bfe-inf.org/document/skills-gap-analysis>

BELMONT FORUM e-INFRASTRUCTURES DATA MANAGEMENT

nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For Authors

News & Comment > News > 2018 > September > Article

NATURE | NEWS

Scientists losing data at a rapid rate

Decline can mean 80% of data are unavailable after 20 years.

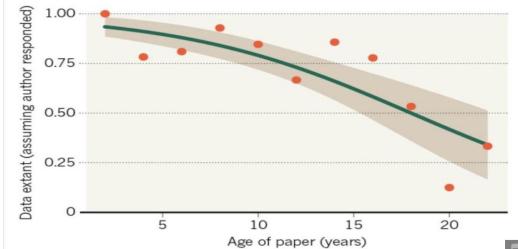
Elizabeth Gibney & Richard Van Noorden

19 December 2013

Rights & Permissions

MISSING DATA

As research articles age, the odds of their raw data being extant drop dramatically.



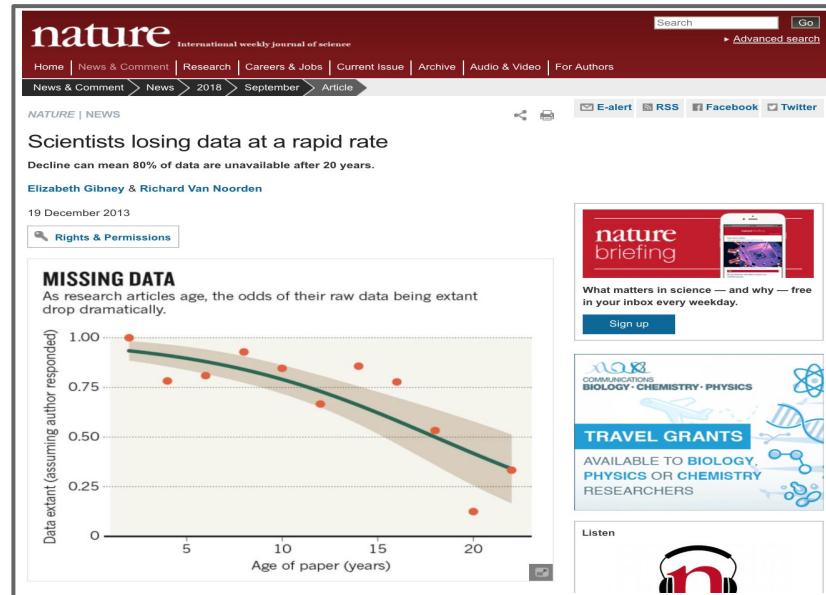
A scatter plot showing the relationship between the age of a paper (years) and the probability of its raw data being extant. The x-axis ranges from 0 to 20 years, and the y-axis ranges from 0 to 1.00. A green line shows a downward trend, indicating a exponential decay. Red dots represent individual data points, which generally follow the line but show significant variability.

What matters in science — and why — free in your inbox every weekday.
Sign up

ALQI COMMUNICATIONS BIOLOGY · CHEMISTRY · PHYSICS

TRAVEL GRANTS AVAILABLE TO BIOLOGY, PHYSICS OR CHEMISTRY RESEARCHERS

Listen

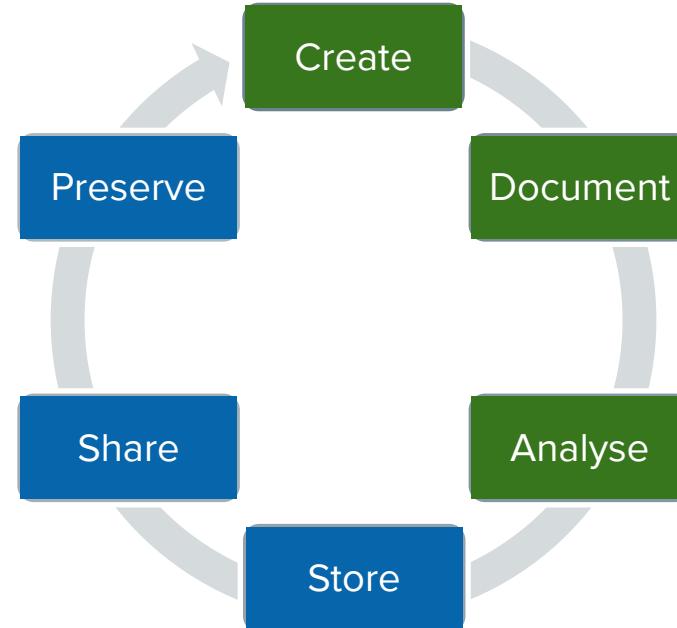


What is Research Data Management?



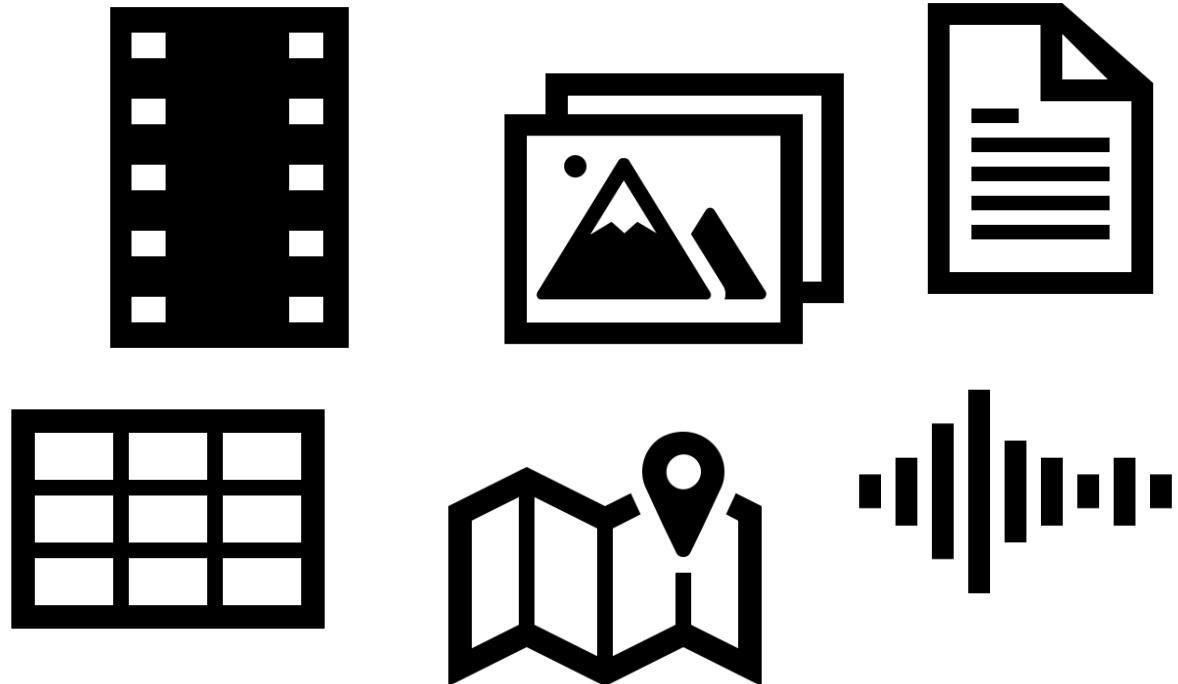
“the active management and appraisal of data (and related research outputs) over the lifecycle of scholarly and scientific interest”

Data management is part of good research practice.



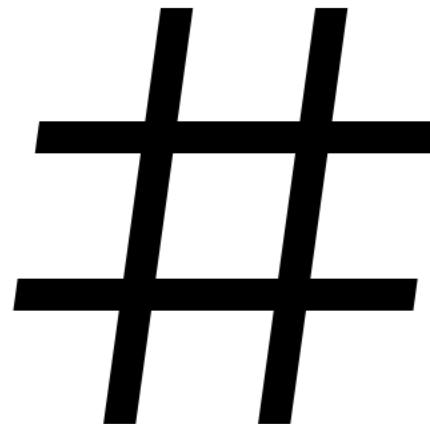
File formats

- Avoid **proprietary** and **lossy** formats
- Make sure **accessibility** ensured in long term preservation



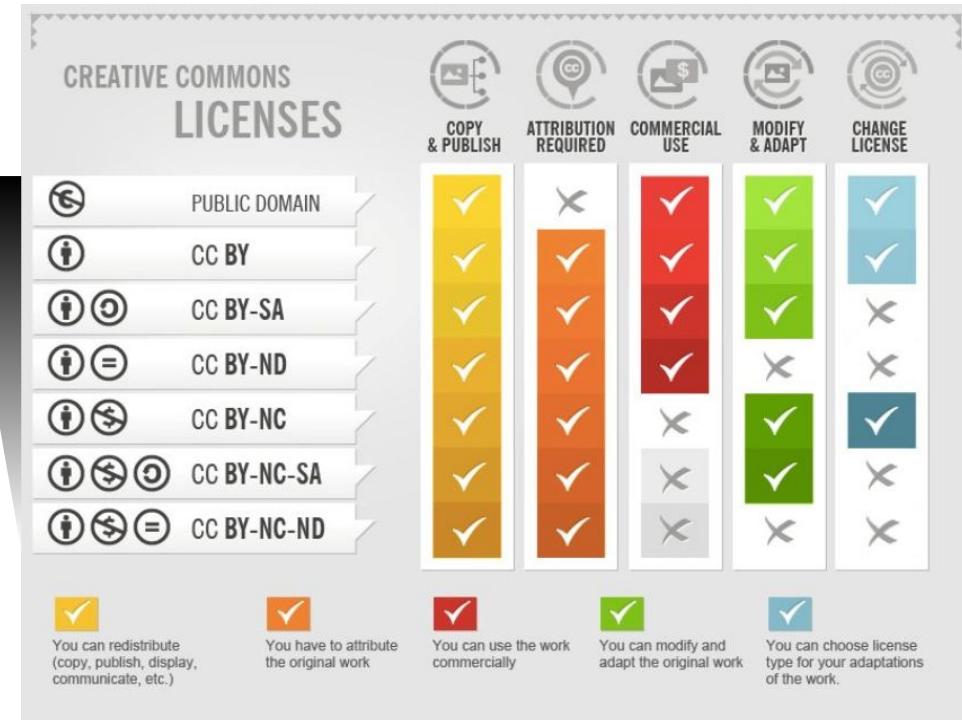
Metadata

- “data about data”
- Part of good documentation
- **Minimum information**
- Use common standards where possible to allow **interoperability**
- Reference:
<http://rd-alliance.github.io/metadata-directory/>; <https://rdamsc.dcc.ac.uk/>



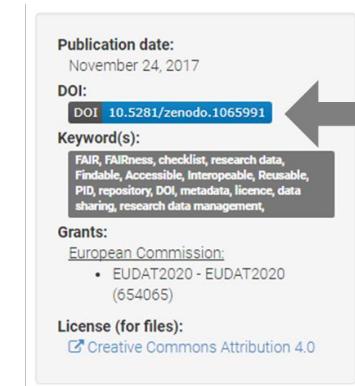
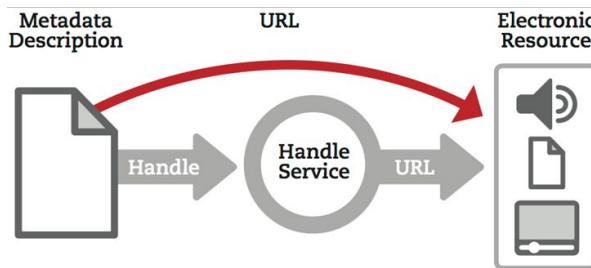
Licences

- Provide information to any potential data **reuser** of their rights
- Ensures clarity
- Creative commons (CC) licensing most used in research



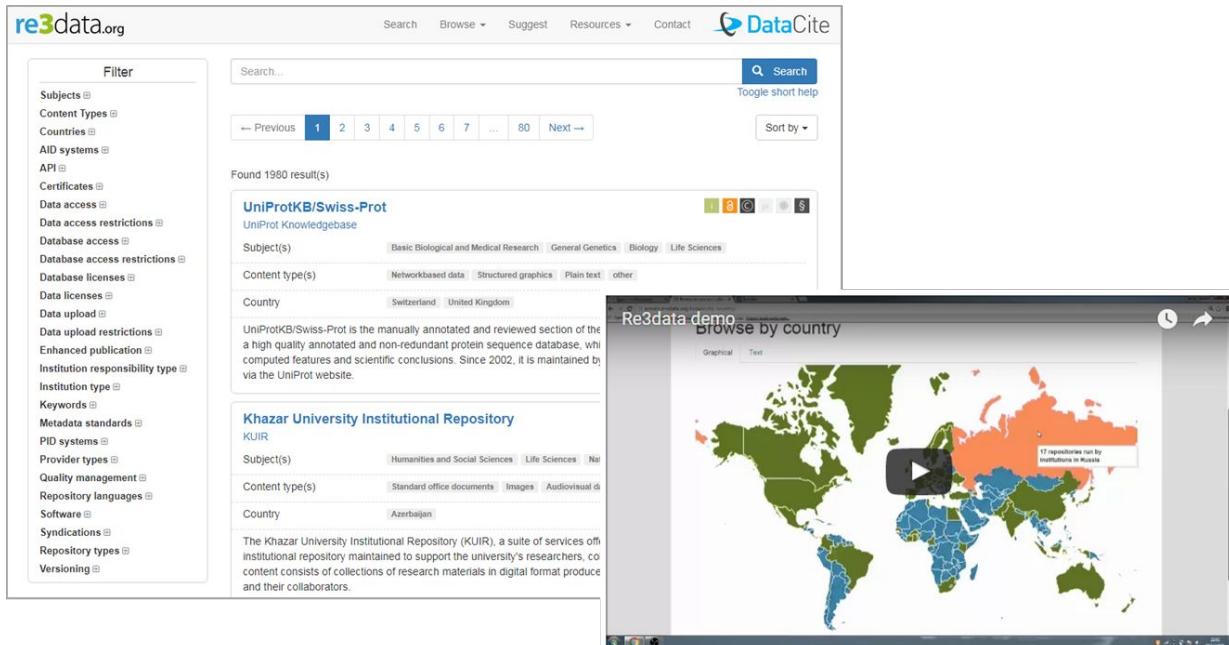
PIDs

- Unique, persistent identifiers can be used for different types of objects
- Ensures **disambiguation** and **findability**
- e.g. DOIs, ORCIDs, ISBNs



Repositories

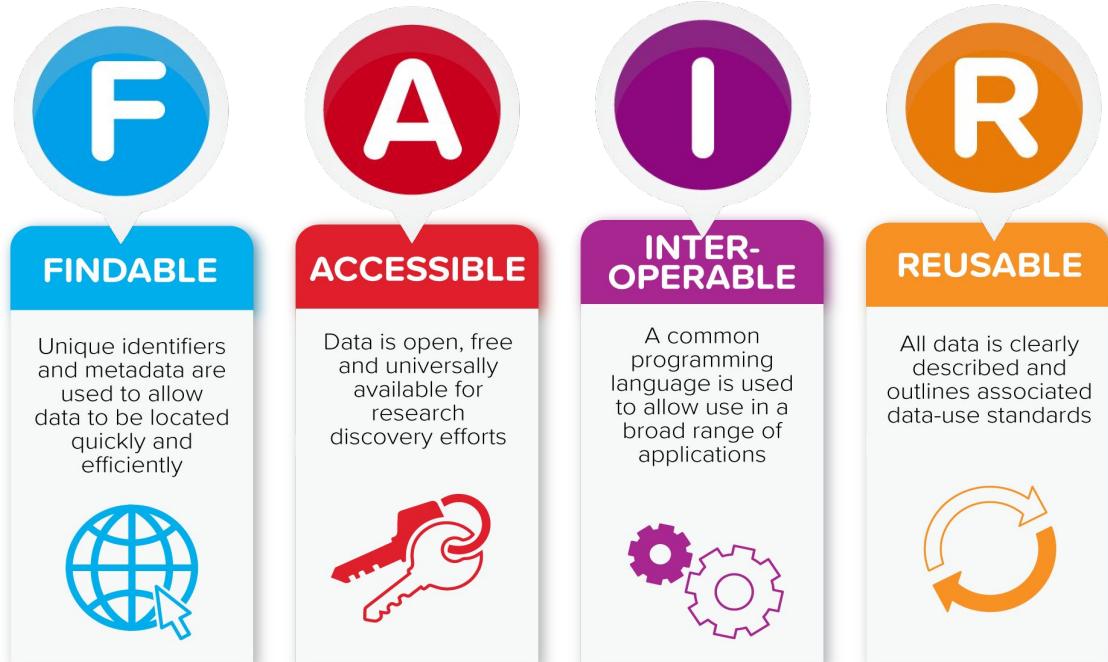
- If possible, choose a suitable repository at the very start of a project
- Use re3data.org to search for repos
- Many data problems can be resolved by selecting and using the right data repository



The image shows two screenshots of the re3data.org platform. The left screenshot displays the search results for 'UniProtKB/Swiss-Prot'. It includes a sidebar with various filters like Subjects, Content Types, Countries, and API. The main results page shows basic information about the dataset, including its subject (Basic Biological and Medical Research), content type (Networkbased data, Structured graphics, Plain text, other), and countries it spans (Switzerland, United Kingdom). The right screenshot shows a world map titled 'Browse by country' where different countries are color-coded according to the number of repositories they host. A callout box highlights that 17 repositories are run by institutions in Russia.

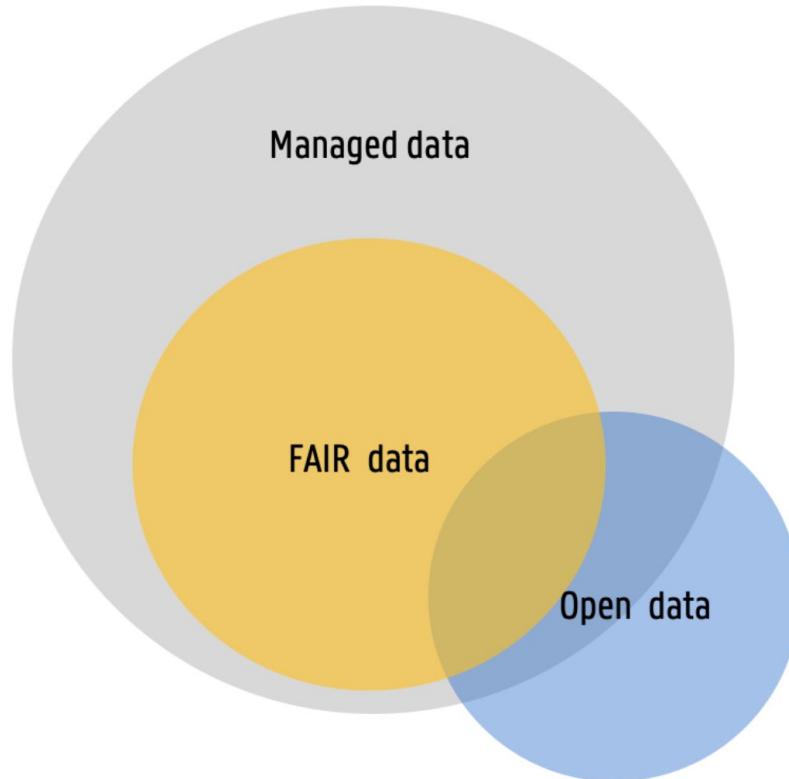
FAIR

- Findable, Accessible, Interoperable & Reusable
- Part of a growing movement to increase the value of data by ensuring their long term preservation
- A distillation of good RDM practices



Open Science

- Where possible, any publicly funded research outputs should be made publicly available
- FAIR ≠ Open
- “As open as possible, as closed as necessary”



Data Management Plans (DMPs)

- A document written before the start of a project
- A “living document”
- A way of neatly tying together all the info discussed previously
- Can also be helpful to the **current project** as well as **future users**
- Try the DCC’s [DMPonline](#) tool to get you started



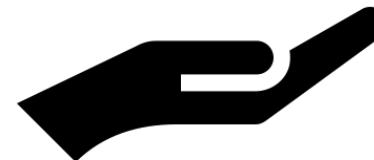
A Quick Review:

Proper Data Management can ...

- Increase **value**
- Increase **reproducibility**
- Increase **provenance**
- Increase **integrity**
- Increase **accountability**
- Reduce **risks**
- Reduce **costs**
- Reduce **fraud**

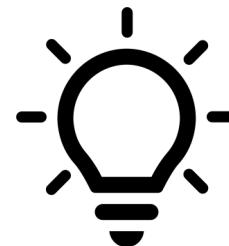
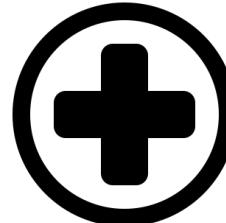
Building institutional support

- There is a growing trend in institutions looking to provide the necessary tools for researchers
- Increase **ownership**
- Decrease use of third party (commercial) solutions
- Try the research infrastructure self evaluation ([RISE](#)) tool yourself to see how your institution fares



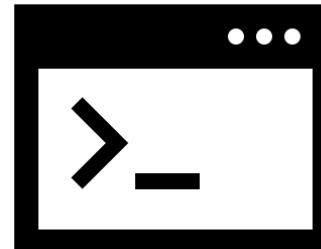
Sensitive data

- Increasing interest in how to align these data with wider research data
- New, specific rules need to be developed
- Doesn't only mean clinical data – geospatial, IPR, etc
- Some examples: [Reproducible Health Data Services](#), [Raising FAIRness in health data and health research performing organisations \(HRPOs\)](#)



Software

- Growing movement to apply FAIR to software and code
- Still treated as a “data” object
- [CURE-FAIR](#)
- See also Lamprecht, Anna-Lena et al. ‘Towards FAIR Principles for Research Software’. Data Science, vol. 3, no. 1, pp. 37-59, 2020. [DOI: 10.3233/DS-190026](https://doi.org/10.3233/DS-190026)
- See also [RSEs](#)



Online tools available to you

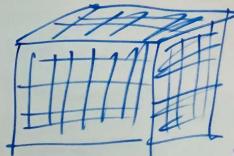
- [FOSTER](#) – especially their [handbook](#) – very useful!
- [FAIR Cookbook](#) – great resource for implementing and teaching FAIR!
- *How to be FAIR with your Data - A teaching and training handbook for higher education institutions* – expected to be published by Dec 2021 – check [FAIRsFAIR](#) for announcement!
- [OpenAIRE](#) – especially resources under the “support” category
- [EOSC](#) – also the huge number of tools that you can use in your research
- [CESSDA](#) – aimed at social sciences
- MOOCs – e.g. [Coursera](#), [Futurelearn](#), [Open Science MOOC](#)
- [DCC – guides](#) and other resources
- Join [RDA](#)!
- ...and many, many more!

Data Stewardship

- Increasingly seen as important
- Dedicated FTE roles in many institutions internationally; some part-time while being researchers at the same time
- Role is to act as main contact for researchers looking for assistance in RDM workflow at their institution
- Many efforts underway to “professionalize” role

NO FAIR

It took years to create,
It should take years to use!



You are a
big researcher.
Why waste time
on metadata?

Speed is of the
essence!

- Use fortran binary or
- Use CSV - no headers
- Document in Lotus Notes
- Resulting code has smaller
footprint!



| Force users to come to you!



(Advice on writing for a scientist in another field)

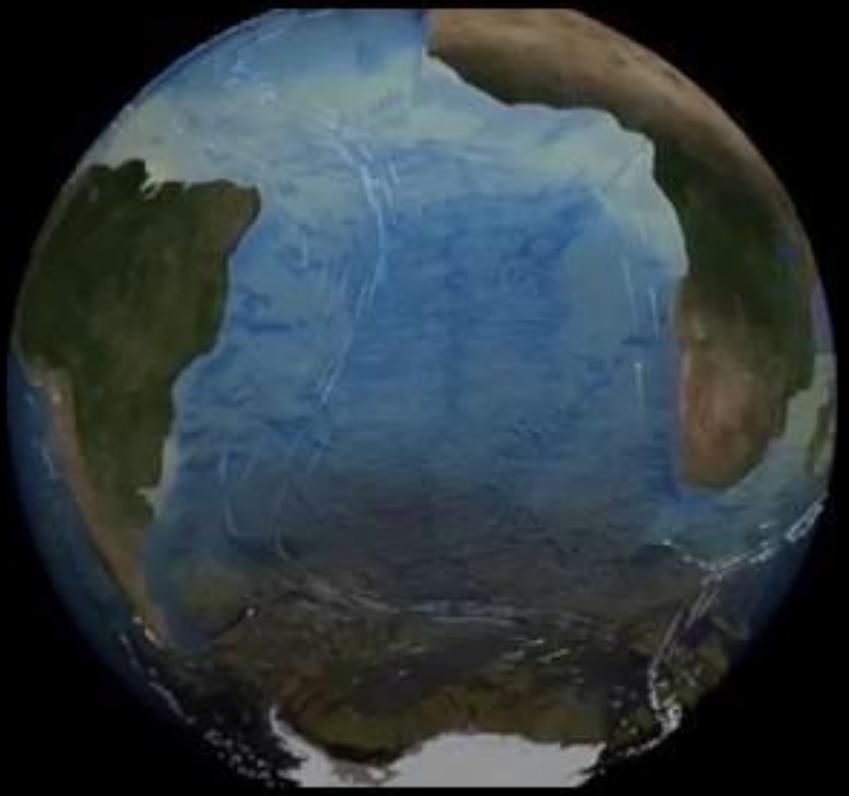
Don't underestimate your readers' intelligence, but
don't overestimate their knowledge of a particular field.

*When writing about science, don't simplify the science;
simplify the writing*

Ocean Observations

101



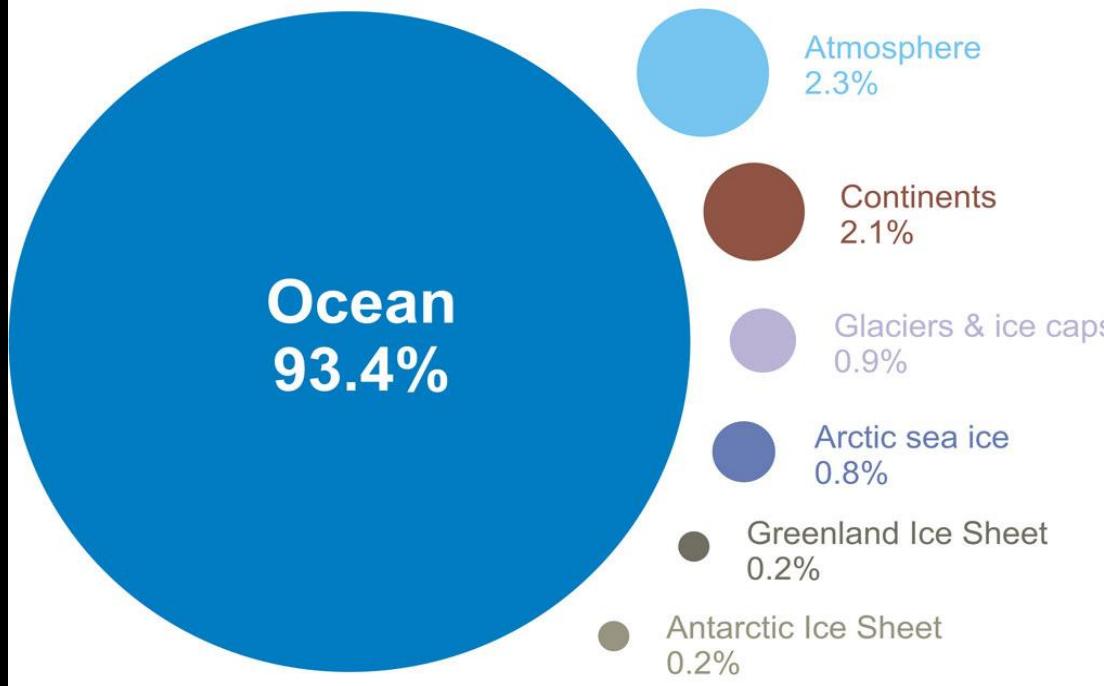


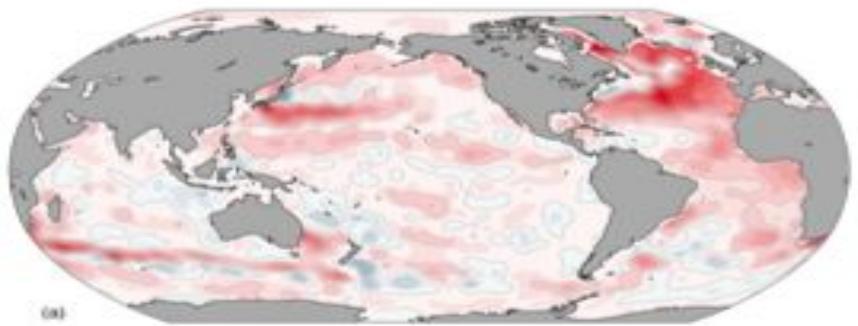
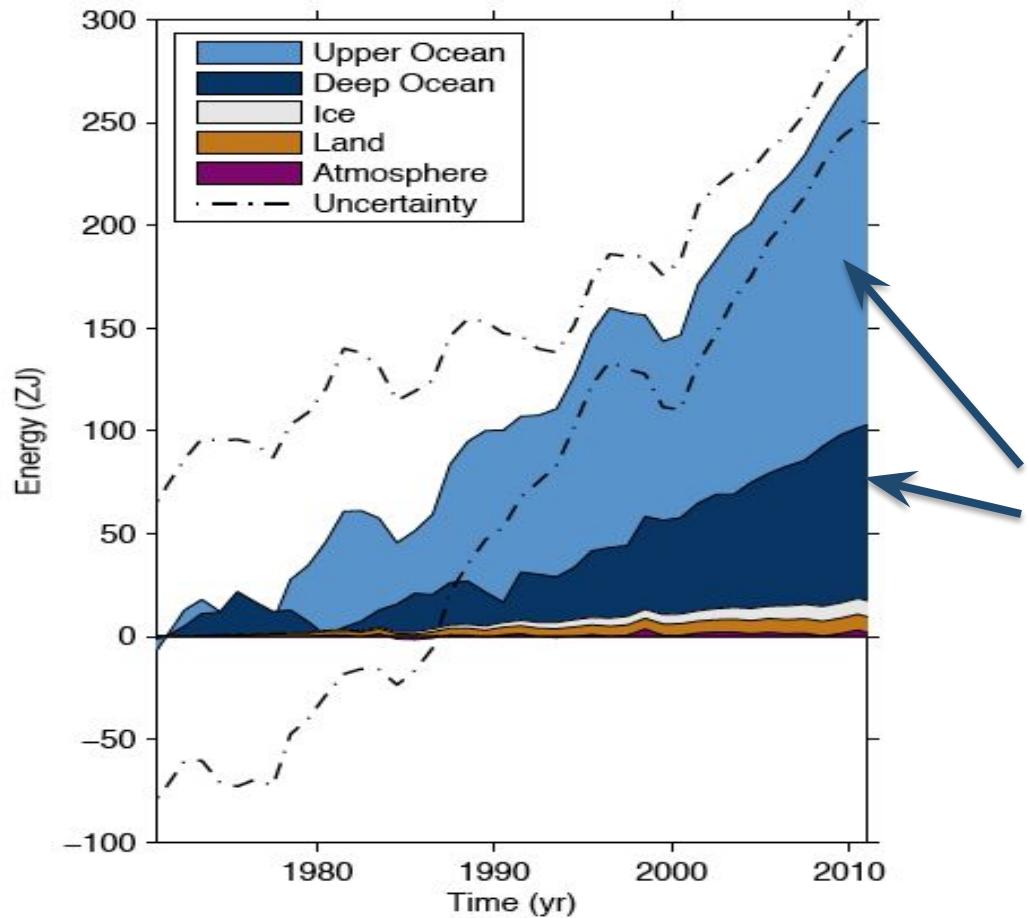
**1.332 x 10²¹ Liters of Water
~352 Quintillion Gallons**



Oceans are the “flywheel of climate”

Where is global warming going?

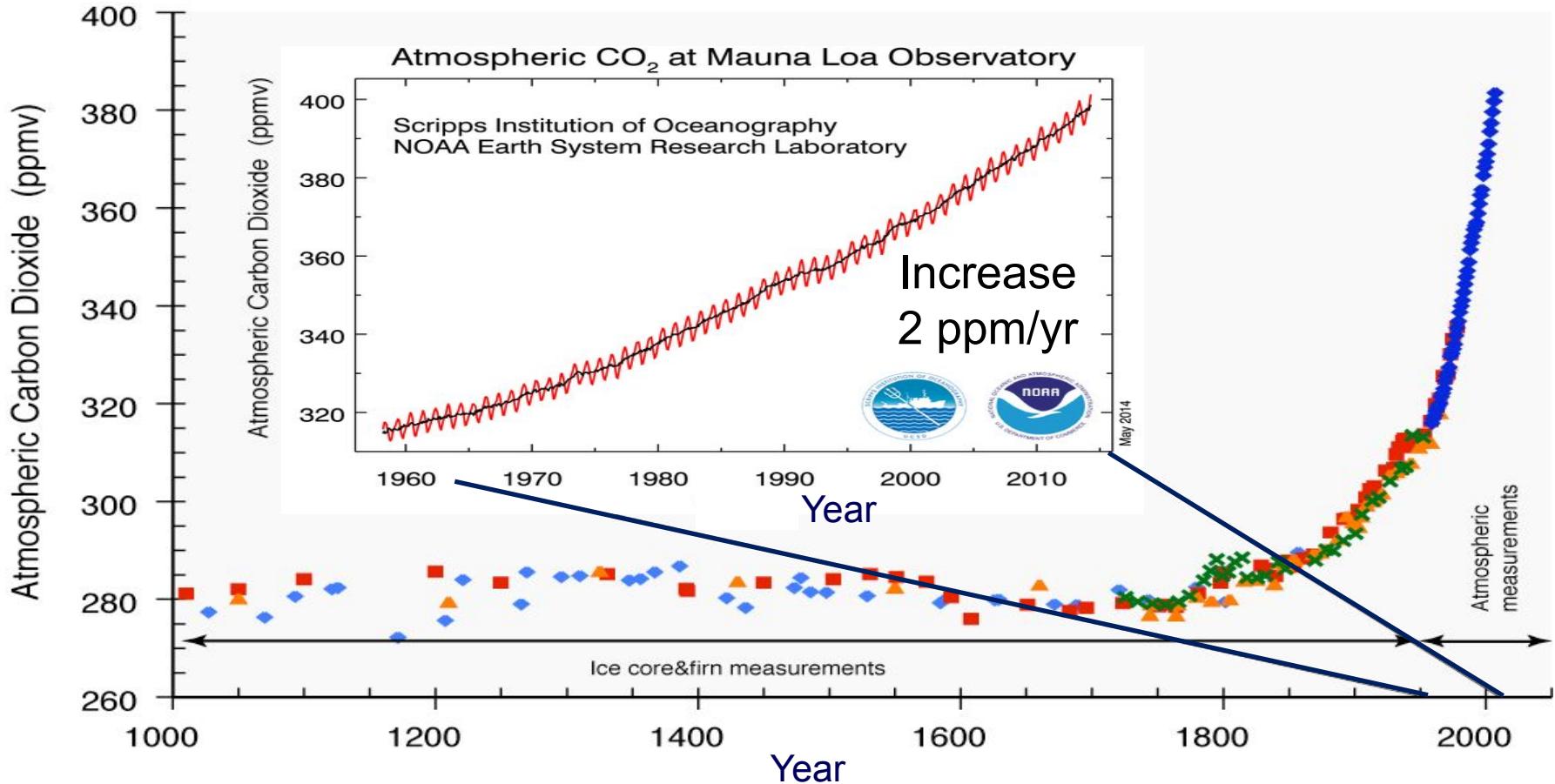




The global energy excess is mostly in the ocean

If the excess heat in the system that is now in the ocean were in the atmosphere, surface air would be 100°C warmer.

Atmospheric CO₂ was steady for at least 1,000 years pre-industrial revolution



Adapted from Sarmiento and Gruber 2002 using Trends online data



Ocean Acidification

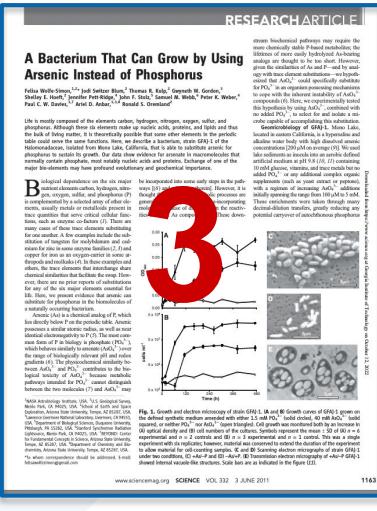
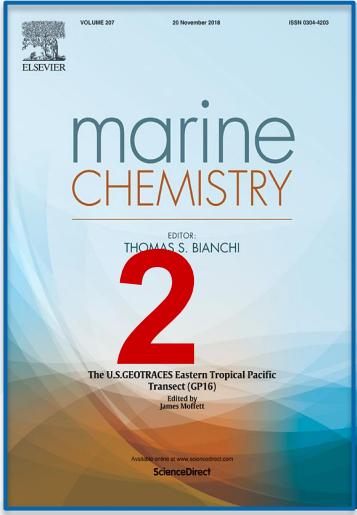
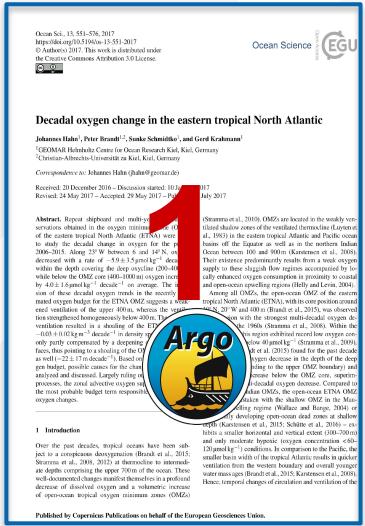
- Approximately 28% of the CO₂ generated by human activities since the mid-1700s has been absorbed by the oceans.
- Ocean acidity has increased 30% since the start of the industrial age.
- Ocean acidity is projected to increase 100-150% percent by 2100.
- Current rate of acidification is nearly 10x faster than any period over the past 50 million years.

Workshop Time

Problem Description: Teamwork!

Your interdisciplinary science team will quickly review one of these peer-reviewed papers

https://bit.ly/2022_SoRDS-ATL_papers



Introduction

Over the past decades, tropical oceans have been subjected to a tropiccean deoxygenation (Brandt et al., 2015; Tournaya et al., 2008, 2012) at thermocline to intermediate depths comprising the upper 700 m of the ocean. These deoxygenation changes manifest themselves in a profound decrease of dissolved oxygen and a volumetric increase of oxygen-poor tropical oxygen minimum zones (OMZs) and only moderate hypoxic (oxygen concentration < 60 $\mu\text{mol L}^{-1}$) conditions. In comparison to the Pacific, the smaller basin with the tropical Atlantic results in quicker ventilation from the western boundary and overall younger water masses (Brandt et al., 2015; Karstensen et al., 2008). Hence, temporal changes of circulation and ventilation of the

Published by Copernicus Publications on behalf of the European Geosciences Union.

Problem Description

- Your group has decided to try to replicate the results from the research described in the paper
- Your assignment is as follows:
 1. Scan the paper, look for clues as to where the data used in this research might reside
 2. Find and try to download those data
 3. **Be mindful throughout of what makes the data easy/hard to:**
 - Locate
 - Download
 - READ/UNDERSTAND (possibly with other data)

Example

ANTARCTIC OCEANOGRAPHY IN A CHANGING WORLD >> SIDEBAR

Sixty-Four Days of Hydrography and Storms: RVIB *Nathaniel B. Palmer*'s 2011 SO4P Cruise

BY JAMES H. SWIFT AND ALEJANDRO H. ORSI

THE OCEANIC MAGAZINE OF THE OCEANOGRAPHIC SOCIETY
Oceanography

DOI:10.5675/1525-3043.1010
Published online in Oceanography in March 2012. © 2012 by the American Meteorological Society.
http://dx.doi.org/10.5675/1525-3043.1010
The copyright holder for this article is The Oceanographic Society, which has granted permission to The Oceanography Society to license this article in perpetuity. Authorization to copy items for internal or personal use under the Copyrights Law is granted by The Oceanography Society for libraries and other users registered with the Copyright Clearance Center (CCC), 27 Congress Street, Salem, MA 01970, for the flat fee of \$10.00 per article. Send payment to CCC.

Downloaded from https://academic.oup.com/ocean/article/25/1/101/261043 by guest on 09 April 2018

carbon dioxide, and other related parameters.

In brilliant Antarctic weather and rarely seen open waters in McMurdo Sound, we set out on the icebreaking research vessel *Nathaniel B. Palmer* from the ice pier at the US Antarctic

systematic section's circulation country's ty and ocean ate nthal Ocean erarching anges hwater,

In addition, we aimed to close off key CLIVAR meridional transects to the Antarctic shelf break, including completion of transects along 150°W and 170°W.

With nominal spacing of 30 nm, each station consisted of a full-depth deployment of a 36-place rosette/CTD equipped with dual temperature/conductivity channels, pressure and dissolved oxygen instruments, a reference thermometer, a transmissometer, a fluorometer, an altimeter, and an acoustic Doppler current profiler (ADCP). Water samples were collected for measurements of salinity, dissolved oxygen, nutrients, chlorofluorocarbons, dissolved inorganic and organic carbon, total alkalinity, pH, colored dissolved organic matter

ADCP, surface temperature/salinity/ $p\text{CO}_2$, and other seawater properties, meteorology, solar radiation, and aerosols/precipitation.

US data and accompanying documentation are publicly available at the CLIVAR and Carbon Hydrographic Data Office (via <http://ushydro.ucsd.edu>) and the Carbon Dioxide Information Analysis Center (<http://cdiac.ornl.gov>).

As we neared Cape Adare to start the first station, winds rose well past 30 knots and continued to roughen the seas during the day. This weather was a taste of the future, because storms frequently interrupted our work (e.g., 105 hours were lost in the first two weeks of the cruise alone), but the new data were fascinating from the start.

ADCP, surface temperature/salinity/ $p\text{CO}_2$, and other seawater properties, meteorology, solar radiation, and aerosols/precipitation. US data and accompanying documentation are publicly available at the CLIVAR and Carbon Hydrographic Data Office (via <http://ushydro.ucsd.edu>) and the Carbon Dioxide Information Analysis Center (<http://cdiac.ornl.gov>).

Your Workspace

Use the provided template 1-2 slides on what you found, and be prepared to discuss your findings in about 12 minutes. You can decide on a presenter any time between now and when you present.

<https://bit.ly/SoRDS-ATL-slides>



Template: Your Team's Research Data Assessment Report

- Our Team consisted of ...
 - Name 1 Name N
- Persistent Identifier for the data
 - PID
- We found the data we were looking for here:
 - Name of repository / data center and URL
- The data was discoverable/accessible
 - Were you able to read the data with ODV or other software?
 - Could not find the data (why?)
- Overall FAIRness assessment (1=*awful* to 10=*great*)

What did you find?



Hypotheses come and go but data remain. Theories desert us, while data defend us. They are our true resources, our real estate, and our best pedigree.

A handwritten signature in cursive script, appearing to read "S. Ramón y Cajal".

Santiago Ramón y Cajal
Nobel Prize in Medicine (1906)