Data management – Archiving and Sharing

Dr. TVS Udaya Bhaskar

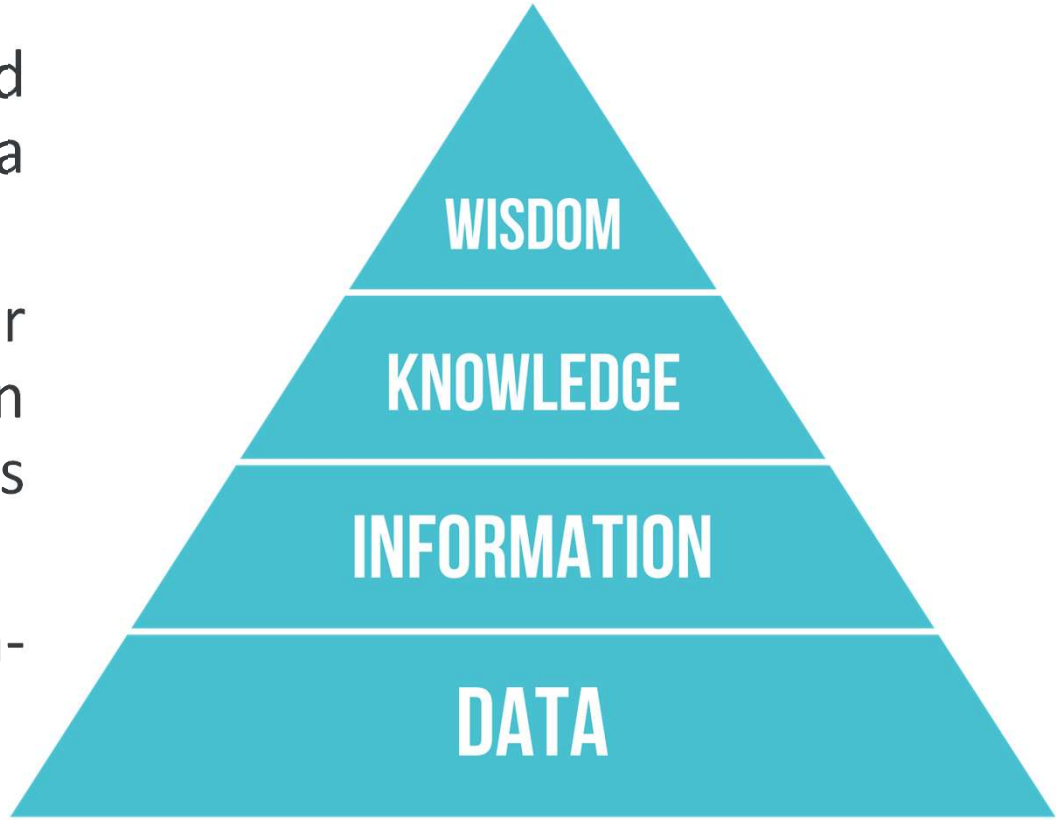
INCOIS, MoES, Govt of India, India

CLIVAR IORP/POGO Regional training workshop on observing the coastal and marginal seas in the western Indian Ocean

7 – 9, June 2022

Importance of data

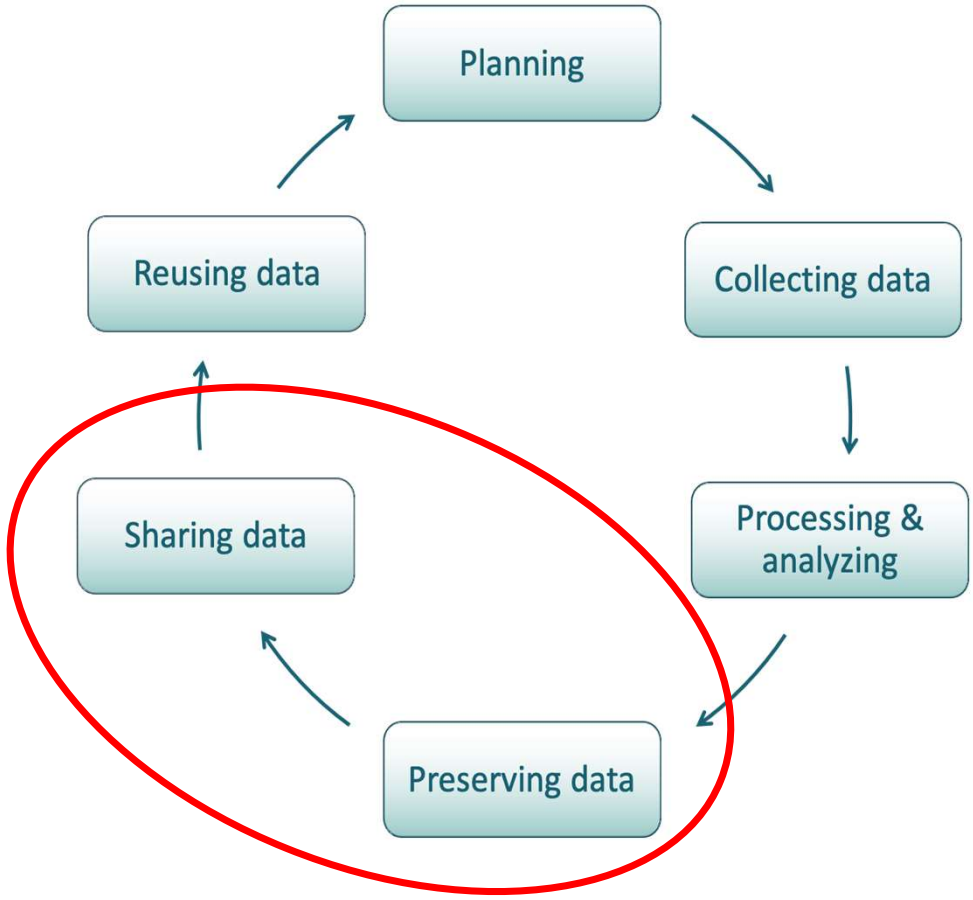
- Data is very valuable. It is collected for a reason, and collecting data requires financial support.
- Data is collected to gather information, this information brings knowledge and from this knowledge, we evolve to wisdom.
- This flow is also known as the Data-Information-Knowledge-Wisdom model, or DIKW pyramid.



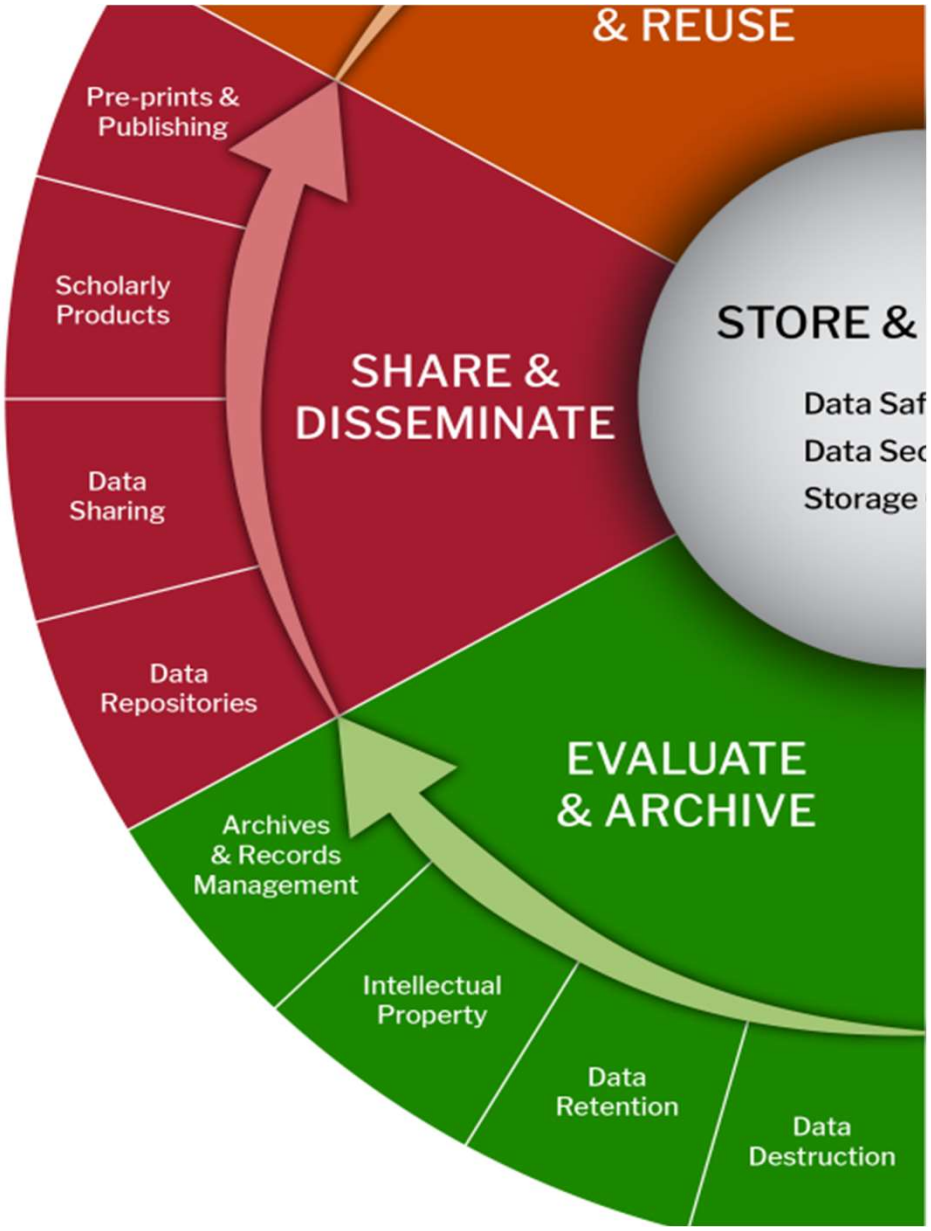
Credits: https://classroom.oceanteacher.org/course/view.php?id=739

Research Data Life Cycle (RDLC)

- The research data lifecycle is a key concept within Research Data Management (RDM).
- It describes the different stages research data go through before, during, and after a research project.
- I will deal with Sharing and Preserving/archiving data.



Credits: https://classroom.oceanteacher.org/course/view.php?id=739



Credits: https://classroom.oceanteacher.org/course/view.php?id=739

Data Preservation/Archival

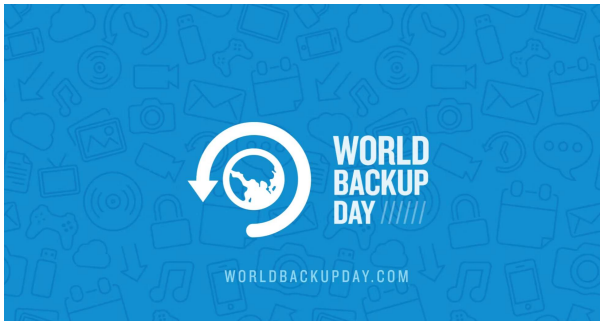
- Data preservation is defined as the act to conserve and maintain both the safety and the integrity of your data.
- The main goal of data preservation is to protect your data and information from being lost, or destroyed, and to contribute to the reuse of your data.
- The actual act of backing up your data is when you take the following steps:
  - you create a copy of your important data and information
  - you store this data and information in a secure and separate location
  - you recognize the back-up as a restoration method for your device

https://www.fosteropenscience.eu/content/cartoon-back

Credits: https://classroom.oceanteacher.org/course/view.php?id=739

Data back-up policy

- A data back-up policy will help to manage the expectations of users, and will provide specific guidance on the 'who, what, when, where and how' of the data back-up and restore process.
- Advantages of a data back-up policy:
  - It helps to clarify procedures, and responsibilities linked to data back-ups
  - It allows you to define (and enforce):
    - where backups are located
    - who can access backups and how they can be contacted
    - how often data should be backed up
    - what kind of backups are performed and
    - what hardware and software are recommended for performing backups



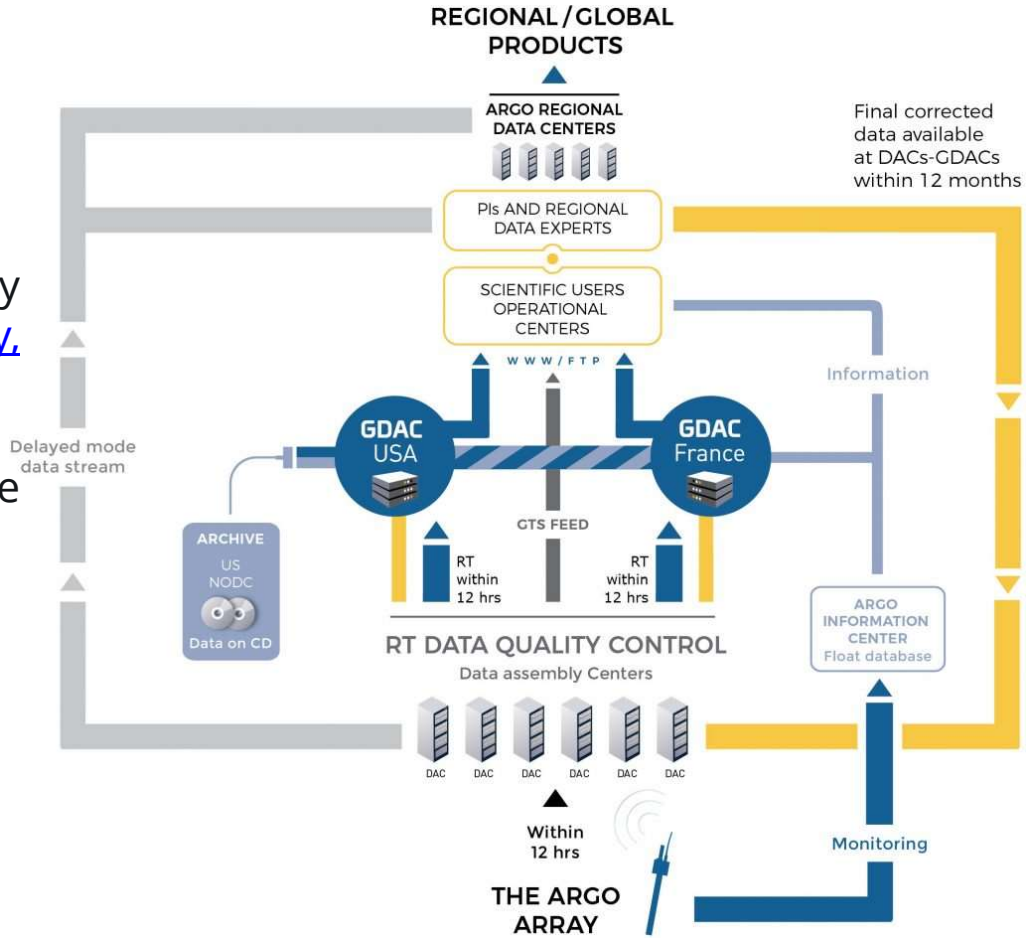
World Backup Day - March 31st

Credits: https://classroom.oceanteacher.org/course/view.php?id=739

Argo Float Data Preservation/Archival

Data are passed to Argo's two Global Data Assembly Centers (GDACs) in Brest, France and Monterey, California.

The GDACs synchronize their data holdings to ensure consistent data is available on both sites.



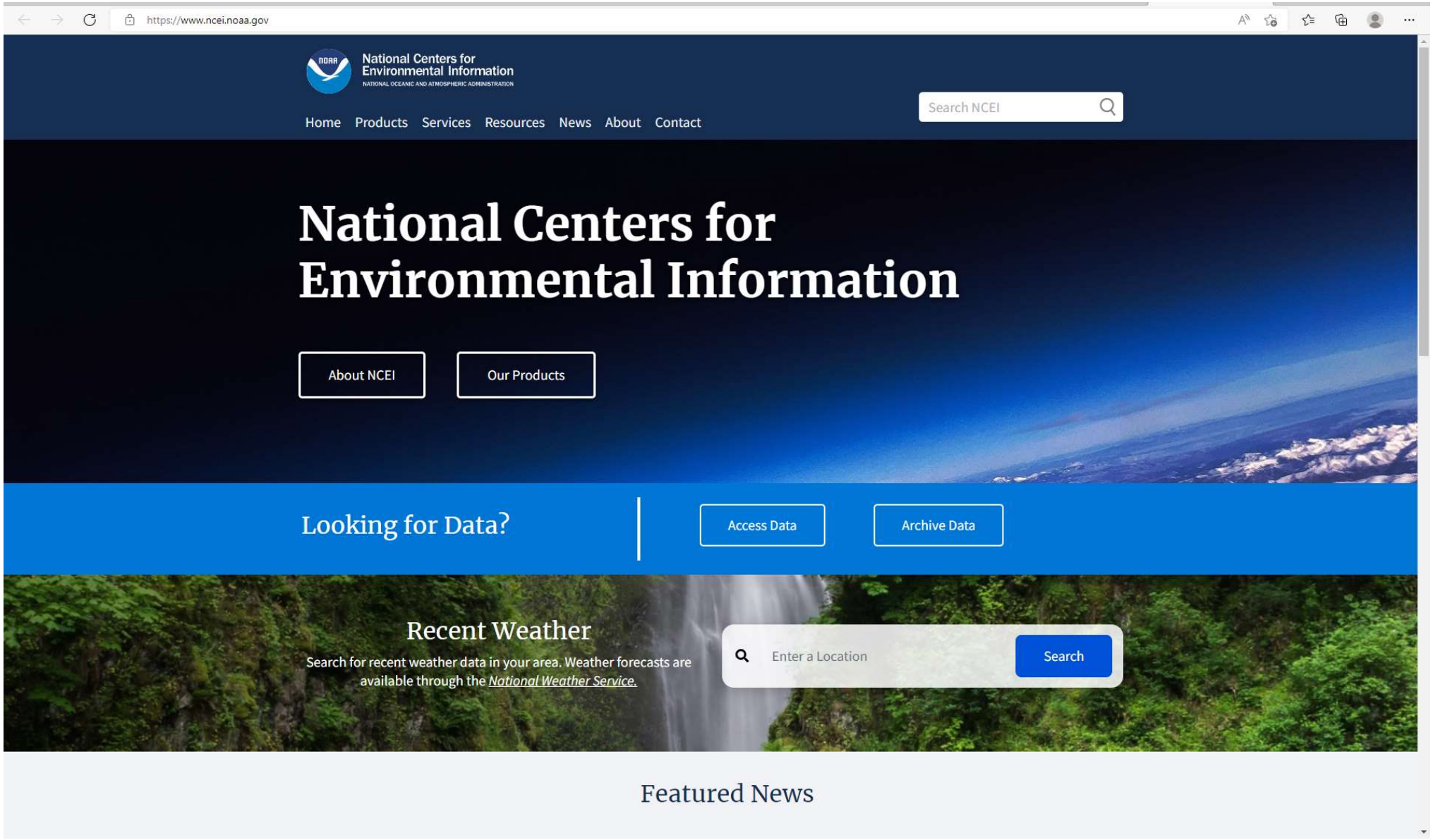
Data repositories

- A data repository - sometimes also referred to as a data library or data archive - is a storage space for researchers to deposit the data set(s) associated with their research, for long-term archiving and preservation.
- Usually, a repository will also provide a permanent link (permalink strategy) for online citation and instant access so that researchers may offer a direct link to their data and ancillary files in a later published article or conference paper. This is usually provided through a Digital Object Identifier (DOI), which allows later linking of data and possibilities for interoperability and mashing up of data archives

Credits: https://classroom.oceanteacher.org/course/view.php?id=739

- Data repositories should meet all of the following requirements:
  - Ensure long-term persistence and preservation of datasets (minimum of 5 years after publication).
  - Be supported by a research community or research institution.
  - Provide deposited datasets with stable and persistent identifiers.
  - Allow access to data without unnecessary restrictions.
  - Provide clear terms of data use and data access on each dataset landing page.
  - Facilitate anonymous reviewer access for embargoed data.

Credits: https://classroom.oceanteacher.org/course/view.php?id=739


https://www.ncei.noaa.gov
National Centers for Environmental Information
NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION
Home
Products
Services
Resources
News
About
Contact
Search NCEI
National Centers for Environmental Information
About NCEI
Our Products
Looking for Data?
Access Data
Archive Data
Recent Weather
Search for recent weather data in your area. Weather forecasts are available through the National Weather Service.
Enter a Location
Search
Featured News

https://www.ncei.noaa.gov/products/global-argo-data-repository

National Centers for Environmental Information
NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION

Search NCEI

Home Products Services Resources News About Contact

Home / Products / Global Argo Data Repository

Global Argo Data Repository

The Argo Ocean Profiling Network is a global ocean observing system developed to address the lack of data coverage in parts of the world ocean, as well as the need for regular capture intervals to enable both short and long-term climate predictions. NCEI operates and manages the Global Argo Data Repository (GADR), which provides long term archive services to store and preserve data. NCEI also implements reanalysis updates and corrections provided by the U.S. Global Ocean Data Assimilation Experiment (GODAE) and French Institute for Research and Exploration of the Sea (IFREMER) Global Data Assembly Centers (GDACs), which provide access to real-time and near real-time data and perform initial quality control measures.

Data Access | Data Assembly Centers Inventory | File Naming Conventions

Data Repositories and Assembly Centers

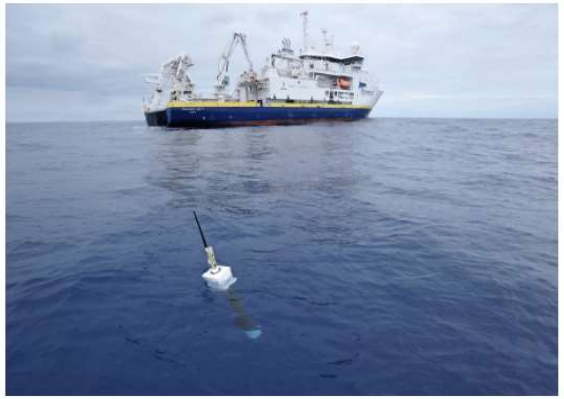
Quality controlled Argo data from both the Global Argo Data Repository (GADR) and Global Data Assembly Centers (GDACs), available in Mono-Profile and Multi-Profile NetCDF formats.

Global Argo Data Repository (GADR)

HTTP | FTP | THREDDS

Global Data Assembly Centers (GDAC)

HTTP | FTP | THREDDS



Basins Data

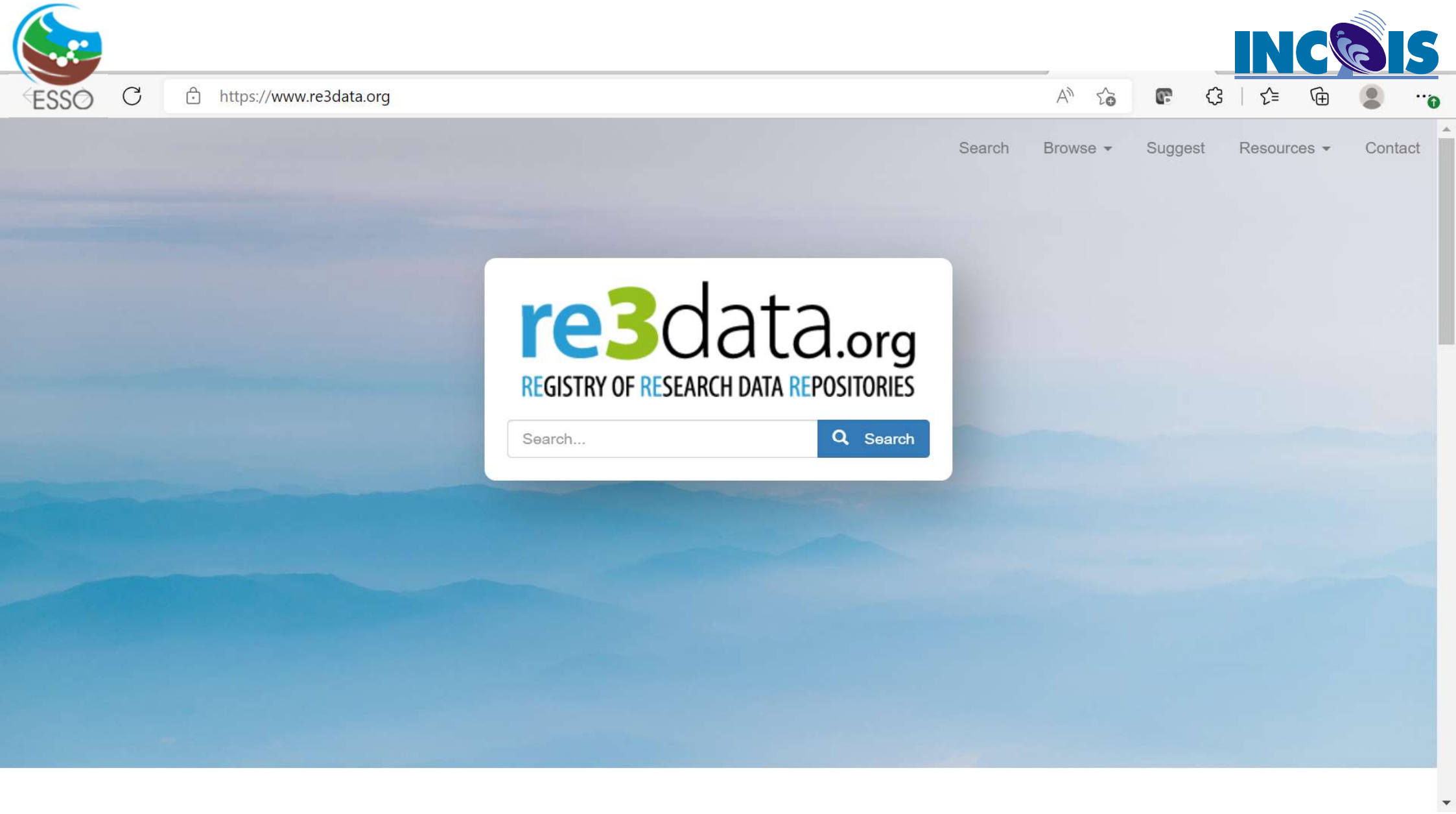
Registry of Research Data Repositories

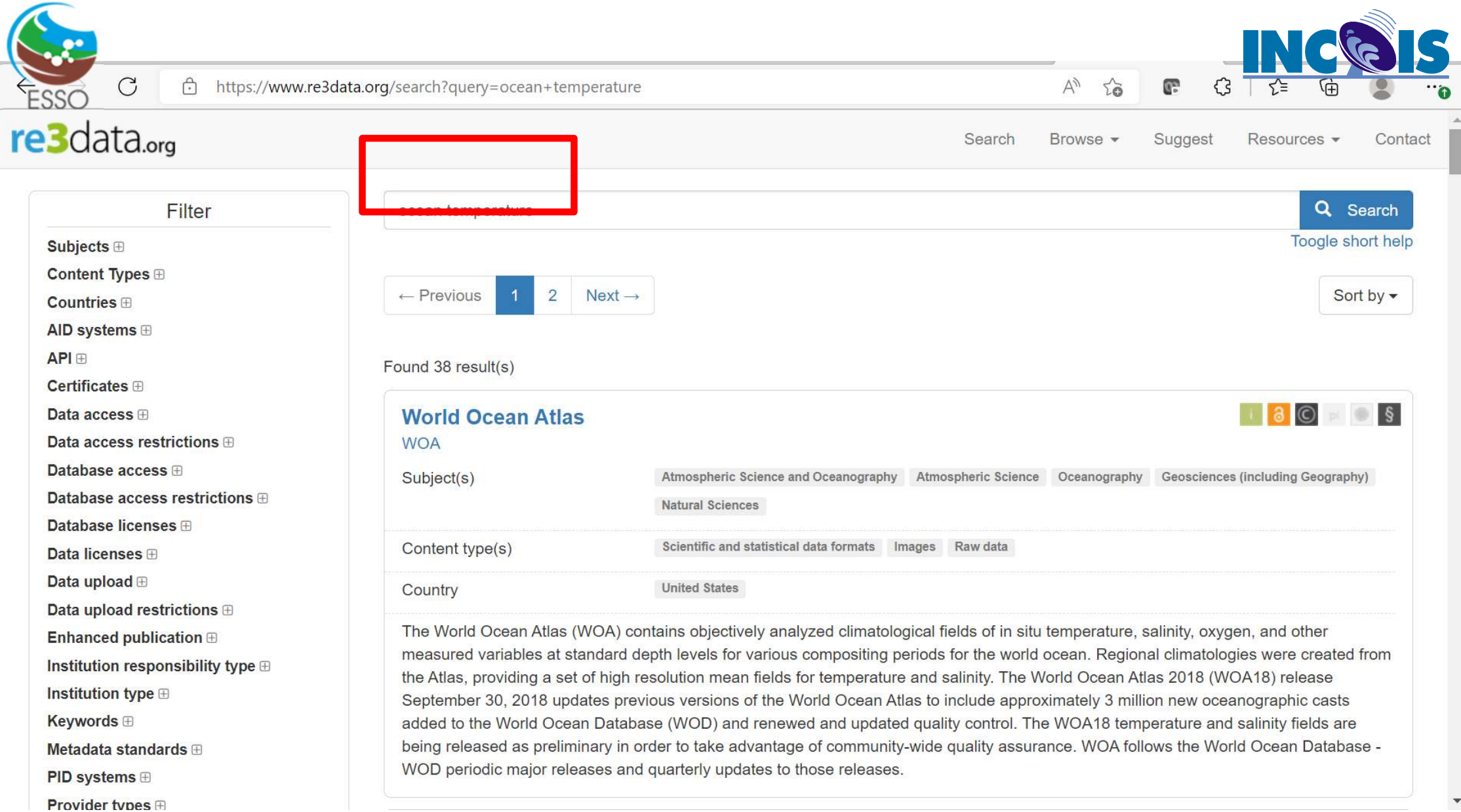
- A vast amount of repositories already exist on a global scale. So how can you choose the most suitable one? You could either select a discipline specific or a generic repository. But even then, your choices are still vast…

re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

- In 2012, a global registry of research data repositories was launched, covering research data repositories from different academic disciplines. (https://www.re3data.org/)

Credits: https://classroom.oceanteacher.org/course/view.php?id=739


https://www.re3data.org
Search
Browse
Suggest
Resources
Contact
re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES
Search...
Search


https://www.re3data.org/search?query=ocean+temperature
re3data.org
Search
Browse
Suggest
Resources
Contact
Filter
Subjects
Content Types
Countries
AID systems
API
Certificates
Data access
Data access restrictions
Database access
Database access restrictions
Database licenses
Data licenses
Data upload
Data upload restrictions
Enhanced publication
Institution responsibility type
Institution type
Keywords
Metadata standards
PID systems
Provider types
ocean temperature
Search
Toogle short help
← Previous
1
2
Next →
Sort by
Found 38 result(s)
World Ocean Atlas
WOA
Subject(s)
Atmospheric Science and Oceanography
Atmospheric Science
Oceanography
Geosciences (including Geography)
Natural Sciences
Content type(s)
Scientific and statistical data formats
Images
Raw data
Country
United States
The World Ocean Atlas (WOA) contains objectively analyzed climatological fields of in situ temperature, salinity, oxygen, and other measured variables at standard depth levels for various compositing periods for the world ocean. Regional climatologies were created from the Atlas, providing a set of high resolution mean fields for temperature and salinity. The World Ocean Atlas 2018 (WOA18) release September 30, 2018 updates previous versions of the World Ocean Atlas to include approximately 3 million new oceanographic casts added to the World Ocean Database (WOD) and renewed and updated quality control. The WOA18 temperature and salinity fields are being released as preliminary in order to take advantage of community-wide quality assurance. WOA follows the World Ocean Database - WOD periodic major releases and quarterly updates to those releases.

https://www.seanoe.org

SEANOE Sea scientific open data publication

About
Publish your data

ODATIS SeaDataNet

Search everywhere | Search document

988 Result(s)

Order by newest | 1 2 3 ... 50

Reset filters

PUBLICATION YEAR

- 2022 (90)
- 2021 (173)
- 2020 (163)
- 2019 (113)
- 2018 (153)
- 2017 (125)
- 2016 (73)
- 2015 (56)

DISCIPLINE

- Administration and dimensions (42)
- Atmosphere (30)
- Biological oceanography (211)
- Chemical oceanography (138)
- Cross-discipline (63)

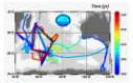
GEOGRAPHICAL AREA

- Adriatic Sea (22)
- Aegean Sea (4)
- Arabian Sea (8)
- Arctic Ocean (49)
- Atlantic Ocean (639)
- Balearic Sea (1)
- Baltic Sea (4)
- Barents Sea (2)

LICENCE CC

- CC-BY (471)
- CC-BY-NC (249)
- CC-BY-NC-ND (95)
- CC-BY-NC-SA (91)
- CC-BY-ND (20)
- CC-BY-SA (17)
- CC0 (43)
- LGPLv3 (2)

OISO cruises: S-ADCP data



Metzl Nicolas, Lo Monaco Claire, Kartavtseff Annie

Measurements of the currents were collected with the Ship-mounted Acoustic Doppler Current Profilers (S-ADCP) in the South Indian and Southern oceans during cruises carried out since 2000 in the framework of the OISO program (Océan Indien Service d'Observations). The OISO program initiated in 1998 (Metzl and Lo Monaco, 1998, https://doi.org/10.18142/228), collects measurements of CO2 and associated parameters in both surface and water column along the repeated lines of R.V. Marion-Dufresne in [...]

2022. Dataset — Open access

Perceptions of managers about Brazilian PAs

Borges Rebecca, Eyzaguirre Indira

Brazil is one of the mega biodiverse countries on the planet. As a strategy to safeguard its biodiversity, the country has heavily relied on protected areas (PAs). However, gaps in research and management of PAs, in Brazil and in other countries, still exist, so studies that include Rapid Assessment and Prioritization of Protected Areas Management (RAPPAM) are essential to evaluate the role and performance of PAs. The aim of this data colection was, among others, to assess the effectiveness of [...]

2022-05-01. Dataset — Open access

Wave tank testing of a multi-purpose platform with aquaculture, wind turbine and wave energy converters

Ohana Jeremy, Horel Boris, Merrien Arnaud, Arnal Vincent, Bonnefoy Felicien, Giulio Brizzi, Bouscasse Benjamin

This dataset was collected during the experimental model tests carried out in the frame of the Blue Growth Farm project (EU–H2020 project 774426). The project aims at the development and demonstration of an automated, multifunctional platform for open sea farm installations of the Blue Growth Industry. This modular and environmentally friendly platform gathers aquaculture systems, a wind turbine and a set of wave energy converters. Experiments are performed to prepare the outdoor tests of a 1/15 [...]

2022-05-03. Dataset — Open access

Mean length and weight at-age of anchovy and sardine estimated during the PELGAS survey in the Bay of Biscay in springtime

Doray Mathieu, Duhamel Erwan, Boiron-Leroy Anne, Marchand Laetitia, Bled--Defruit Geoffrey, Petitgas Pierre

The Pélagiques Gascogne (PELGAS, Doray et al., 2000) integrated survey aims at assessing the biomass of small pelagic fish and monitoring and studying the dynamics and diversity of the Bay of Biscay pelagic ecosystem in springtime. PELGAS has been conducted within the EU Common Fisheries Policy Data Collection Framework and Ifremer's Fisheries Information System. Details on survey protocols and data processing methodologies can be found in Doray et al., (2018a, 2021). This dataset comprises the [...]

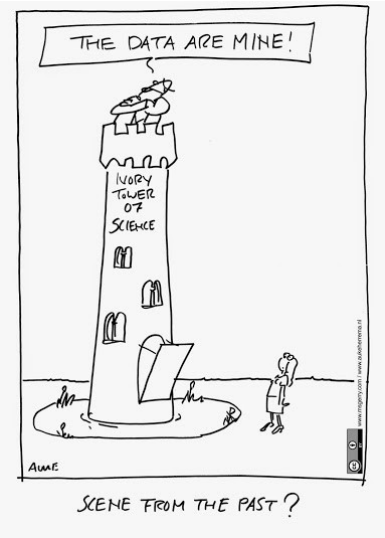
Data Archiving Summary

- Making sure that data remain accessible for future use can be accomplished by storing them in a data repository.
- Data repositories allow data to become part of the public domain.
- The Registry of Research Data Repositories (Re3Data) was developed to help people find their way through data repositories to either search for data or to find the best suitable repository to store their own data.
- References:
  - https://dataoneorg.github.io/Education/bestpractices/create-and-document.html
  - https://us.norton.com/internetsecurity-how-to-the-importance-of-data-back-up.html
  - https://en.wikipedia.org/wiki/Data_preservation
  - Cindy Parr, Heather Henkel, DataONE (Aug 30, 2011) "Best Practice: Create and document a data backup policy". Accessed through the Data Management Skillbuilding Hub at https://dataoneorg.github.io/Education/bestpractices/create-and-document
  - https://dataoneorg.github.io/Education/bp_step/preserve/
  - https://dataoneorg.github.io/Education/lessons/06_protect/06_protect.pdf
  - https://digitalguardian.com/blog/what-data-repository
  - https://www.infotoday.com/cilmag/apr16/Uzwyshyn--Research-Data-Repositories.shtml
  - https://guides.library.oregonstate.edu/research-data-services/data-management-archive-preserve
  - https://www.nature.com/sdata/policies/repositories

Credits: https://classroom.oceanteacher.org/course/view.php?id=739

Sharing Data

- This phase of the Research Data Life Cycle - data sharing - is entered mostly after the actual research has been done.
- Data were collected, analyses were done and first results are available.
- At that point, the research data can be published for re-use, either by making it available in a repository, provided with metadata and a license, or published as a data paper, or - most preferably - a combination of both.



Credits: https://classroom.oceanteacher.org/course/view.php?id=739

Data Sharing …..

- For effective sharing of data one need to:
  - Define a proper data license
    - Discuss and provide the best license for the data
  - Create and populate Digital Object Identifiers for datasets
  - Publish Data papers

Credits: https://classroom.oceanteacher.org/course/view.php?id=739

(i) Data Licenses

- A data licence is a legal instrument that specifies a standard set of terms and conditions regarding sharing and re-use of research data.
- Data which is shared with a licence becomes Open Data.
- A licence tells a user exactly what they can and cannot do with the data, through unambiguous conditions.
- A general rule of thumb when it comes to assigning a license is “the simpler, the better”.
- There are a lot of different licences available, but when you are looking at maximal reuse of your data, try to adopt a licence standard which is already widely (globally) in use, such as the Creative Commons licenses.

Credits: https://classroom.oceanteacher.org/course/view.php?id=739

The following scheme provides a clear overview of these conditions.

Icons	Licence	Guidelines
	BY	**Attribution**: You let others can copy, distribute, perform and remix your work if they credit your name as specified by you.
	SA	**Share Alike**: You let others copy, distribute, display, perform, and modify your work, as long as they distribute any modified work on the same terms. If they want to distribute modified works under other terms, they must get your permission first.
	NC	**Non-commercial**: You let others copy, distribute, display, perform, and (unless you have chosen NoDerivatives) modify and use your work for any purpose other than commercially unless they get your permission first.
	ND	**No Derivatives**: You let others copy, distribute, display and perform only original copies of your work. If they want to modify your work, they must get your permission first.

License Icon	Attribution	Licence Elements
CC BY	Attribution (CC BY)	Allows others distribute, remix, tweak, and build upon your work, even commercially, as long as they credit you for the original creation. This is the most accommodating of licenses offered and recommended for maximum dissemination.
CC BY SA	Attribution-ShareAlike (CC BY-SA)	Allows others remix, tweak, and build upon your work even for commercial purposes, as long as they credit you and licence their new creations under the identical terms. This is the licence used by Wikipedia.
CC BY NC	Attribution-NonCommercial (CC BY-NC)	Allows others remix, tweak, and build upon your work non-commercially, and although their new works must also acknowledge you and be non-commercial, they don't have to license their derivative works on the same terms.
CC BY NC SA	Attribution-NonCommercial-ShareAlike (CC BY-NC-SA)	Allows others remix, tweak, and build upon your work non-commercially, as long as they credit you and licence their new creations under the identical terms.
CC BY ND	Attribution-NoDerivs (CC BY-ND)	Allows others redistribute, even commercially, as long as it is passed along unchanged and in whole, with credit to you.
CC BY NC ND	Attribution-NonCommercial-NoDerivs (CC BY-NC-ND)	Allows others to download your works and share them with others as long as they credit you, but they can't change or use commercially. This licence is the most restrictive CC licence.

Credits: https://classroom.oceanteacher.org/course/view.php?id=739

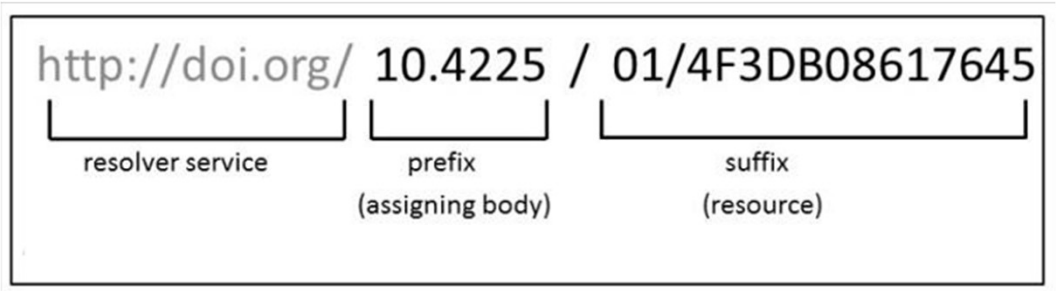
(ii) Digital Object Identifiers (DOIs) for datasets

- Digital Object Identifiers (DOIs) have been widely used by publishers of peer reviewed journals for over more than 10 years to uniquely identify a specific article.
- Associating DOIs to scientific publications has not only increased the traceability of the cited literature, but also simplified the maintenance of citation indexes which serve today to assign academic credit to scientists for their work.

Credits: https://classroom.oceanteacher.org/course/view.php?id=739

Description of a DOI

- A DOI Name (DOI) can be assigned to any object that is a form of intellectual property. DOI should be interpreted as 'digital identifier of an object' rather than 'identifier of a digital object'.
- A DOI consists of a unique, case-insensitive, alphanumeric character sequence that is divided into parts separated by a forward slash.



Resolver service https://doi.org	This ensures the DOI resolves to an online metadata record about the dataset or collection
Prefix	Denotes a unique naming authority and is assigned by a DOI Registration Agency. All DOIs start with the number 10, followed by a period '.'. The numbers following the period identify the registrant or assigning body..
Suffix	The DOI suffix consists of a character string of any length chosen by the registrant. Each suffix is unique to the prefix element that precedes it.

Credits: https://classroom.oceanteacher.org/course/view.php?id=739

DOI continued …..

- Assigning DOIs to your dataset can facilitate these processes for the following reasons:
  - The DOI system can serve to identify and locate a specific version of a dataset
  - Datasets to which a DOI is assigned should stay persistently accessible.
  - After assigning a DOI to a dataset, the data should no longer be changed. This will assure that claims based on data can always be checked.
  - Minimum metadata - including the names of the data author(s) – needs be associated with the DOI during registration. Having the data authors associated with the DOI will assure that they can receive the appropriate credit and recognition for their work
  - DataCite will store the associated metadata and disseminate it to other initiatives like the Data Citation Index (Thomson Reuters) and ORCID. The Date Citation Index will link the indexed datasets to publications citing them. ORCID allows researchers to claim authorship of their datasets, which will help resolve confusion caused by researches with a similar name.
  - By assigning a DOI with the dataset, the data publisher attests that the dataset answers to certain quality standards and contains all the necessary metadata to make the dataset usable by other scientists. This will encourage other scientists to use it for their analyses and then cite your dataset in their publications.

Credits: https://classroom.oceanteacher.org/course/view.php?id=739

(iii) Data Papers

- A data paper is a peer reviewed document describing a dataset, published in a peer reviewed journal.
- It takes effort to prepare, curate and describe data. Data papers provide recognition for this effort by means of a scholarly article.
- Unlike a conventional research article, the primary purpose of a data paper is to describe data and the circumstances of their collection, rather than to report hypotheses and conclusions.

Credits: https://classroom.oceanteacher.org/course/view.php?id=739

Data Papers continued ….

- Several tools have been developed over the years to assist researchers in writing a data paper. Depending on requirements and personal choice, the following tools can be used:
  - GBIF Integrated Publishing Toolkit (IPT): https://www.gbif.org/ipt
  - NephilaPaper: https://ferramentas.sibbr.gov.br/nephila/ (Português)
  - Arpha Writing Tool: https://arpha.pensoft.net/tips/Start-a-manuscript

Credits: https://classroom.oceanteacher.org/course/view.php?id=739

BMC Bioinformatics

BMC Bioinformatics. 2011; 12(Suppl 15): S2.
Published online 2011 Dec 15. doi: 10.1186/1471-2105-12-S15-S2

PMCID: PMC3287445
PMID: 22373175

The data paper: a mechanism to incentivize data publishing in biodiversity science

Vishwas Chavan[✉#1,2] and Lyubomir Penev[#1,2]

▸ Author information ▸ Article notes ▸ Copyright and License information Disclaimer

Abstract

Go to: ▸

Background

Free and open access to primary biodiversity data is essential for informed decision-making to achieve conservation of biodiversity and sustainable development. However, primary biodiversity data are neither easily accessible nor discoverable. Among several impediments, one is a lack of incentives to data publishers for publishing of their data resources. One such mechanism currently lacking is recognition through conventional scholarly publication of enriched metadata, which should ensure rapid discovery of 'fit-for-use' biodiversity data resources.

Discussion

We review the state of the art of data discovery options and the mechanisms in place for incentivizing data publishers efforts towards easy, efficient and enhanced publishing, dissemination,

Data Sharing Summary

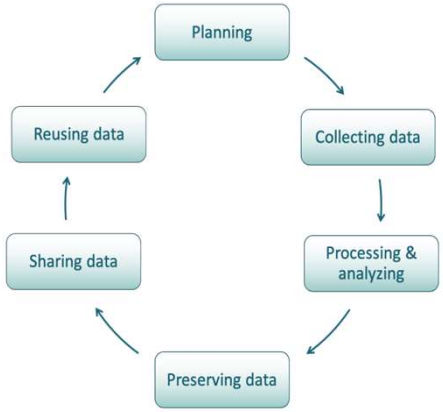
- Sharing starts taking place mostly after the actual research was completed.
- It is advised to assign a license to research data.
- Create and assign a Digital Object Identifier (DOI) to the data to get more visibility and citations.
- Check for the possibility of data papers.
- References:

  - https://www.library.yorku.ca/web/open/overview/data-licensing/
  - https://www.cessda.eu/Training/Training-Resources/Library/Data-Management-Expert-Guide/6.-Archive-Publish/Publishing-with-CESSDA-archives/Licensing-your-data
  - Creative Commons
  - European Data Portal Licensing Assistant.
  - https://www.doi.org/
  - Piwowar HA, Day RS, Fridsma DB (2007) Sharing Detailed Research Data Is Associated with Increased Citation Rate. PLoS ONE 2(3): e308. https://doi.org/10.1371/journal.pone.0000308

Credits: https://classroom.oceanteacher.org/course/view.php?id=739

Summary and Conclusion

- Data is very valuable. It is collected for a reason.
- There are 6 Phases in a Research Data Life Cycle (RDLC)
  - Phase 1: Planning
  - Phase 2: Collecting data
  - Phase 3: Processing and analysing data
  - Phase 4: Preserving data
  - Phase 5: Sharing data
  - Phase 6: Reusing data



- Data preservation/archival is defined as the act to conserve and maintain both the safety and the integrity of your data.
  - you create a copy of your important data and information
  - you store this data and information in a secure and separate location
  - you recognize the back-up as a restoration method for your device
- Data sharing - is entered mostly after the actual research has been done.
  - Define a proper data license
  - Discuss and provide the best license for the data
  - Create and populate Digital Object Identifiers for datasets
  - Publish Data papers