Data Management Guide for Researchers

National Environmental Science Program
Marine and Coastal Hub

Eric Lawrey[1], Emma Flukes[2]

1 Australian Institute of Marine Science
2 University of Tasmania

Creative Commons licence

Cataloguing data

This publication (and any material sourced from it) should be attributed as: Lawrey, E., Flukes, E. (2023). Data Management Guide for Researchers. Report to the National Environmental Science Program, Marine and Coastal Hub. Australian Institute of Marine Science and University of Tasmania

This publication is available on the NESP Marine and Coastal Hub website: https://www.nespmarinecoastal.edu.au

Disclaimer

The NESP Marine and Coastal Hub advises that this publication comprises general advice on managing research data. The reader is advised and needs to be aware that such advice was developed based on historic projects undertaken in the NESP program and that future projects may have different requirements. The hub has exercised due care and skill in preparing and compiling the information in this publication. Notwithstanding, the hub, its employees and advisers disclaim all liability, including liability for negligence and for any loss, damage, injury, expense, or cost incurred by any person as a result of accessing, using, or relying on any of the information or data in this publication to the maximum extent permitted by law.

Acknowledgements

This work was undertaken for the Marine and Coastal Hub, a collaborative partnership supported through funding from the Australian Government’s National Environmental Science Program (NESP).

Document control

Version	Date of issue	Author	Reason for change
0.1	29/04/2023	Eric Lawrey	First draft
0.2	05/05/2023	Emma Flukes	Minor amendments
0.3	14/06/2023	Duane Fraser	Review of indigenous data management section.
0.4	24/07/2023	Emma Flukes	Incorporating Peter Walsh’s changes (minor)
1.0	07/09/2023	Eric Lawrey	Incorporate review changes

Contents

Practical guidance for research data management

NESP emphasizes open and enduring access to environmental and climate research data, fostering informed decision-making and long-term public benefits.

This document is intended as a starting guide on data management for researchers working on National Environmental Science Program (NESP) Marine and Coastal (MaC) Hub funded projects. It covers the Hub expectations on data management and general guidance on how research data should be managed and published, in line with the [Australian Code for the Responsible Conduct of Research](). This document is an extension to Marine and Coastal Hub Data Management Strategy ([https://www.nespmarinecoastal.edu.au/data-management/]()).

The National Environmental Science Program (NESP) is a long-term research funding program by the Australian Government. This program funds environment and climate research that:

- Provides evidence for the design, delivery, and on-ground outcomes for environmental programs.
- Helps decision-makers, including Indigenous communities, build resilience.
- Supports positive environmental, social, and economic outcomes.

The NESP round 2 research is organised into [four hubs](), each focused on a particular theme of research: Marine and Coastal (present Hub), Climate Systems, Resilient Landscapes, Sustainable Communities and Waste.

The Marine and Coastal hub focuses on research that informs the management of Australia's marine and coastal environments, including estuaries, coast, reefs, shelf and deep-water. Within the NESP MaC Hub, projects are divided into northern and southern node projects. The northern node projects are administered by the Reef and Rainforest Research Centre while the southern node projects are administered by University of Tasmania (Figure 1).

NESP is publicly funded research so its outputs should be publicly available by default. This ensures that all those who can benefit from the research have ready access to it. In many cases the core information delivery mechanisms for projects are through papers, reports, presentations, and workshops. Data gathered and generated by NESP projects are also key project outputs. Publishing data enables future projects can build off existing data collections and analysis, shortening the time to future discoveries; encourages collaboration; and enables reproducibility.

The NESP MaC hub provides two Data Wranglers to assist research projects with all issues associated with their data management and data publication.

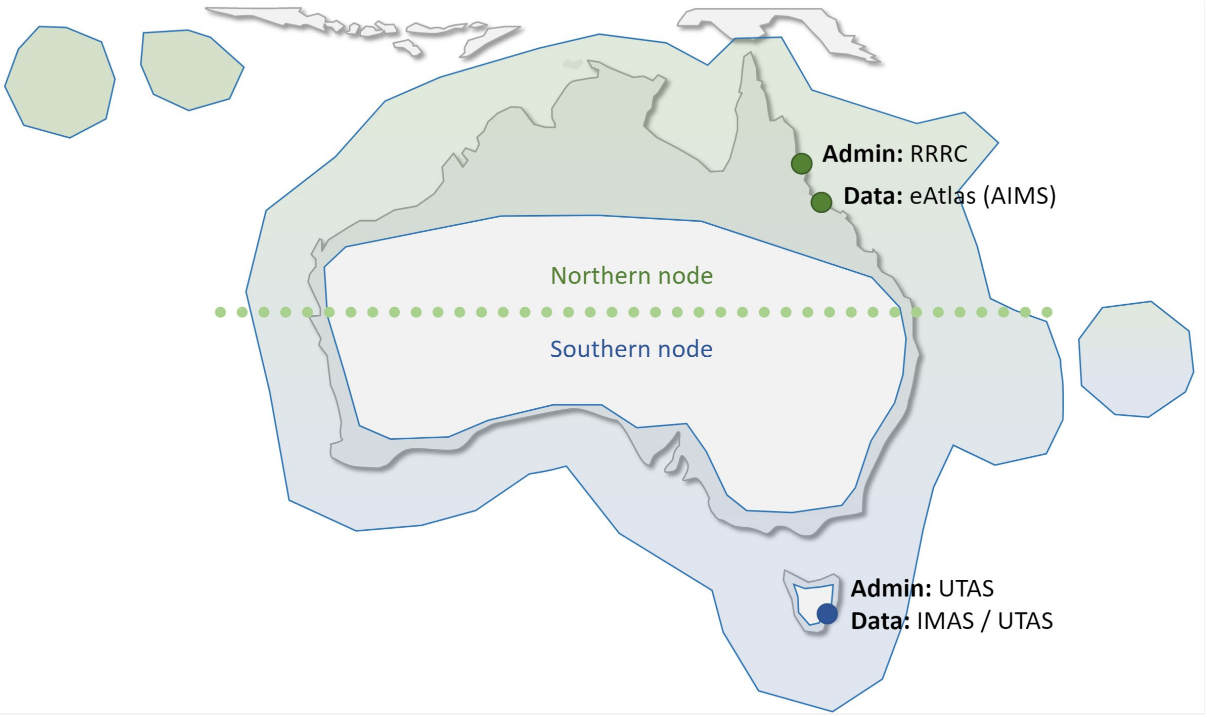


Figure 1. The projects in the NESP Marine and Coastal hub are divided into northern and southern node projects. Each node has its own project administration and Data Wranglers however they work collectively to provide consistent project management across the hub.

The MaC Hub uses a "distributed" data model, wherein data may be published through a number of different repositories depending on the type of research. All data is centrally discoverable through the Australian Ocean Data Network (AODN) catalogue via a single “Hub-level” metadata record.

By default, both raw and processed data from research projects should be made publicly available in an enduring manner. Where there are sensitivities associated with the data being gathered then researchers should discuss the most appropriate level of access with their associated hub Data Wranglers.

Project data life cycle

A key aspect of research is the development, management, and publication of its data. This data provides a foundation for new future research and is an important legacy of the NESP. NESP projects are expected to publish all data they generate, including raw and derived data products (e.g., modelled or aggregated data). By default, data should be published to an approved public, enduring data repository. Sensitive data should be prepared as if for publication and lodged in an approved secure data repository with suitable access controls.

> Projects should actively plan their data products, thinking about how others can use them to solve new research questions.
>
> Data discussions with the hub Data Wranglers can help develop this plan and strategies for maximising reuse.

It is therefore important for each project to plan what datasets they will develop and publish. Having a clear vision of these project outputs will provide a clear goal for project delivery.

The goal of the Hub Data Wranglers is to help projects meet their NESP data management obligations by providing advice on data management issues, planning for data delivery to maximise discoverability and reuse, and where to publish the data. The Hub endorses two primary data repositories (eAtlas and IMAS data repository) for publication of its data, but other institutional or discipline-specific repositories that meet the Hub's requirements for FAIR and enduring access to data may also be used (see Hub-endorsed repositories).

The Data Wranglers will organise **data discussions with the project team**, throughout each research project to:

1. Identify and plan what data, if any, will be created by the project.
2. How that data will be managed during the project lifecycle, including methods of sharing data with project collaborators prior to publication.
3. How sensitive data will be managed and curated.
4. Where relevant, determine whether discussions regarding data ownership and custodianship have taken place (including Indigenous groups where appropriate).
5. What record-keeping of third-party data and implications of its licensing is needed.
6. Where required, establish a reasonable data publication embargo period.
7. Where the data should be published and what supporting documentation needs to be prepared.

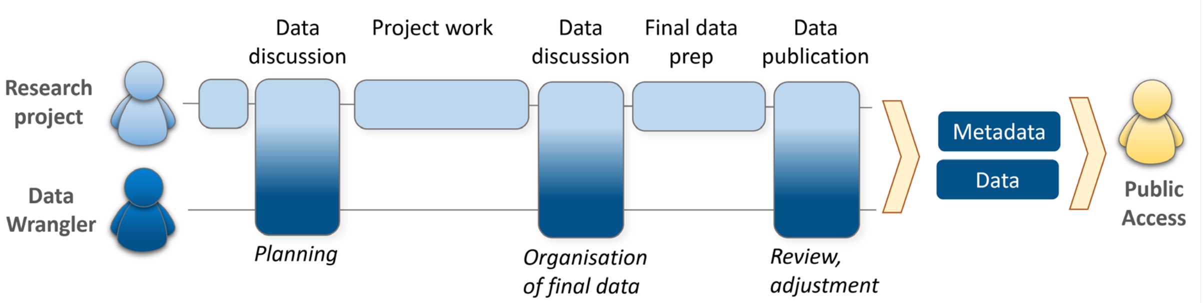


Figure 2. This shows an overview of the data management workflow for a typical research project. Each project has multiple data discussions with the project's Data Wrangler to help plan management of the project's data.

Typically, data discussions will occur multiple times throughout the project; one early in the project, one near the end of the project, and throughout the project as required. This is the minimum expectation, but more frequent data discussions can be arranged with the Data Wranglers if your project requires them. These discussions are an opportunity to receive help from the Data Wranglers on data issues within your project and an opportunity for the Data Wranglers to better understand your project data and track progress. Data discussions can be conducted face-to-face, via video / phone meetings, or email exchanges.

Managing data during the project

It is the responsibility of the NESP research teams to properly manage the data they collect and produce during the life of the projects, ensuring generated data is appropriately stored and managed. Ideally the project data should be sufficiently well organised that the project could be accessed and understood by new researchers if needed. This is a good standard to aim for as it promotes documentation and data cleaning during the project, which simplifies the final data publication processes.

> The research team is responsible for the data management during the life of the research project. They must ensure that project data is backed up, and that documentation is kept on the dataset sources and methods used to produce the data.

Some important considerations during the research project are:

1. **Is your project Indigenous co-designed or led (Tier 1 & 2 per the NESP three-category approach)?** If so, it is important to identify any data that incorporates indigenous knowledge or otherwise be considered of Indigenous interest and ensure that you discuss data custodianship and access with relevant Indigenous groups. You should discuss what can be made publicly available, and what is sensitive and needs restrictions and access control. See section on ***Managing Indigenous Data*** for more information.

2. **Consider if your data is sensitive?** Determine if any of the project data is sensitive. If so, consider what access controls are required to protect the data. To understand what constitutes sensitive data and how it should be managed, see ***Managing sensitive data***.

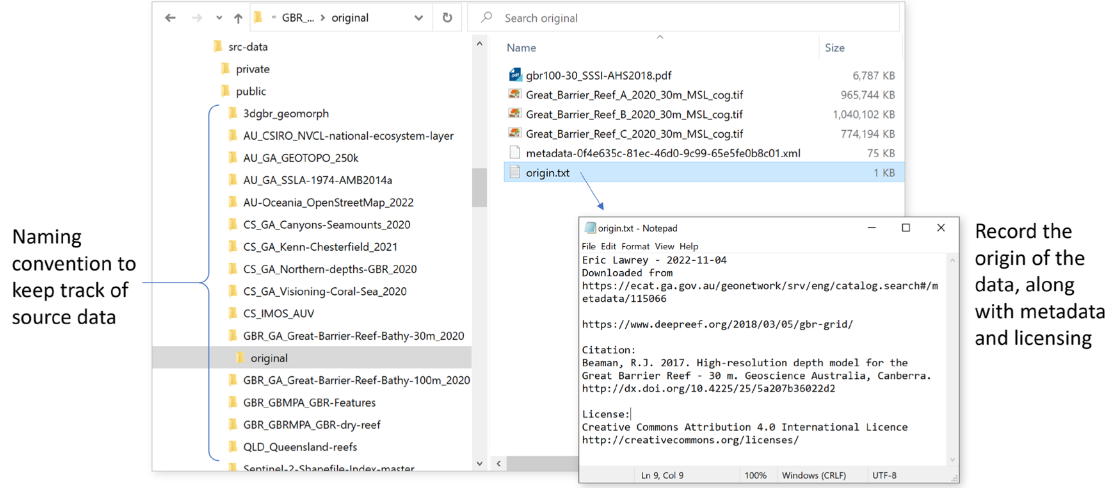
3. **Do you have a robust backup for your data to ensure no data is lost?** How this is best achieved depends on your institutional resources, institutional policies, and the nature of your data. Your backup should be regular enough to prevent data loss due to disk failure or human error. Your back should also have multiple versions to allow you to resolve problems that are not immediately noticed. Ideally your data should be stored in multiple physical locations / mechanisms (e.g., external hard drives, cloud storage) to protect against local disasters or limitations of each back up strategy. You should also consider how your backup strategy would handle a ransomware attack, where malware encrypts all your local files and backups that are attached or synced from your machine. Cloud storage services such as Dropbox, OneDrive/SharePoint or Google Drive can provide a partial backup solution, however they should not be your only backup, as local problems (e.g., mistakes, ransomware) aren’t always recoverable.

4. **Are you collaborating across organisations?** Do you have a suitable way for sharing data across team members and collaborators so that everyone can work together efficiently? This might be through Google Drive, Dropbox, OneDrive/SharePoint, Teams or even shipping of hard drives when the data is very large. Choose a tool that works for your team and is aligned with your institutional policies. If you have sensitive data, cloud-based options may not be suitable.

5. **Are you analysing existing data?** When producing research outputs from existing datasets, it is important to carefully record the origin and licensing of each source dataset. The nature of this source data, its quirks and limitations, and possible licensing or use restrictions, becomes part of any derived data. It is therefore important to track this information and to document what processing was used to generate derived data. Maintaining this provenance

information ensures that further details of the source data can be tracked down if needed (Figure 3).

It is also vital to record and consider the licensing associated with any third-party data used in your analysis. The licensing of any derived data must be compliant with the licensing of the source datasets, and so the incorporation of non-open data (e.g., supplied under exclusive use agreement) may restrict the usability of your final project data.


GBR_... > original
Search original
src-data
private
public
3dgbr_geomorph
AU_CSIRO_NVCL-national-ecosystem-layer
AU_GA_GEOTOPO_250k
AU_GA_SSLA-1974-AMB2014a
AU-Oceania_OpenStreetMap_2022
CS_GA_Canyons-Seamounts_2020
CS_GA_Kenn-Chesterfield_2021
CS_GA_Northern-depths-GBR_2020
CS_GA_Visioning-Coral-Sea_2020
CS_IMOS_AUV
GBR_GA_Great-Barrier-Reef-Bathy-30m_2020
original
GBR_GA_Great-Barrier-Reef-Bathy-100m_2020
GBR_GBMPA_GBR-Features
GBR_GBRMPA_GBR-dry-reef
QLD_Queensland-reefs
Sentinel-2-Shapefile-Index-master
origin.txt - Notepad
File Edit Format View Help
Ln 9, Col 9
100%
Windows (CRLF)
UTF-8

Name	Size
gbr100-30_SSSI-AHS2018.pdf	6,787 KB
Great_Barrier_Reef_A_2020_30m_MSL_cog.tif	965,744 KB
Great_Barrier_Reef_B_2020_30m_MSL_cog.tif	1,040,102 KB
Great_Barrier_Reef_C_2020_30m_MSL_cog.tif	774,194 KB
metadata-0f4e635c-81ec-46d0-9c99-65e5fe0b8c01.xml	75 KB
origin.txt	1 KB

```
Eric Lawrey - 2022-11-04
Downloaded from
https://ecat.ga.gov.au/geonetwork/srv/eng/catalog.search#/m
etadata/115066

https://www.deepreef.org/2018/03/05/gbr-grid/

Citation:
Beaman, R.J. 2017. High-resolution depth model for the
Great Barrier Reef - 30 m. Geoscience Australia, Canberra.
http://dx.doi.org/10.4225/25/5a207b36022d2

License:
Creative Commons Attribution 4.0 International Licence
http://creativecommons.org/licenses/
```

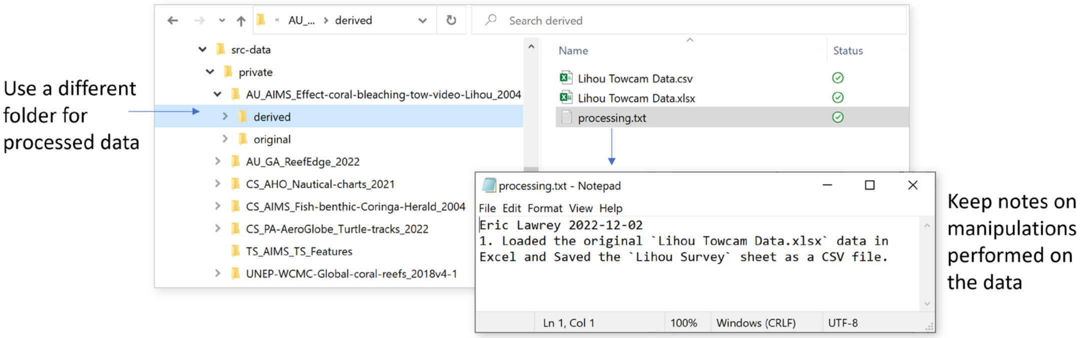
Naming convention to keep track of source data

Record the origin of the data, along with metadata and licensing

Figure 3. Tip: Keeping track of the source data in your analysis can be as simple as a folder storing the original downloaded data and a text file recording where the data came from, when you got the data, how to cite the data and its licensing.

If your project is planning to use data from a third-party source with access restrictions, you should try to negotiate whether the source data and/or derived data products can be made publicly available. This liberation of previously restricted data will benefit the broader research community and is a performance metric of the NESP. Using data with restrictions that will affect your ability to publish derived data products should only be considered in the absence of any other alternatives. Even if the data you use has restrictions on reproduction, you may still be able to publish metadata.

6. **Keep track of processing steps:** As you work on analysing project data, ensure that any processing steps used to transform the source or raw data through to the final analysed product, is recorded. This information is often vital for understanding the provenance and nuances of a processed dataset. Where possible, perform your analysis using code and publish this code alongside the dataset. If you are using a mixture of automated tools and manual steps, you should endeavour to record these steps in a file kept with the data.


AU_... > derived
Search derived
src-data
private
AU_AIMS_Effect-coral-bleaching-tow-video-Lihou_2004
derived
original
AU_GA_ReefEdge_2022
CS_AHO_Nautical-charts_2021
CS_AIMS_Fish-benthic-Coringa-Herald_2004
CS_PA-AeroGlobe_Turtle-tracks_2022
TS_AIMS_TS_Features
UNEP-WCMC-Global-coral-reefs_2018v4-1
processing.txt - Notepad
File Edit Format View Help
Ln 1, Col 1
100%
Windows (CRLF)
UTF-8

Name	Status
Lihou Towcam Data.csv	✓
Lihou Towcam Data.xlsx	✓
processing.txt	✓

```
Eric Lawrey 2022-12-02
1. Loaded the original `Lihou Towcam Data.xlsx` data in
Excel and Saved the `Lihou Survey` sheet as a CSV file.
```

Use a different folder for processed data

Keep notes on manipulations performed on the data

Figure 4. Tip: Keep notes on processing that you perform on source data. This log can be as simple as a text file, as in this example, or could be a script, or a Word document with screenshots. It is also a

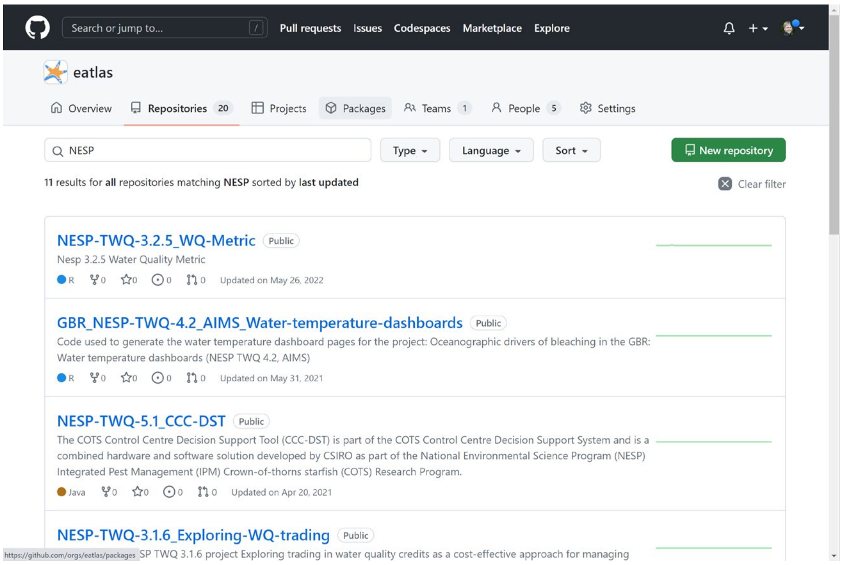
good idea to keep your original raw data untouched, saving any processing separately so you can always check the original data.

7. **Plan what datasets you will publish:** Often it is difficult to decide on the best way to package research data, particularly when multiple data sources and experiments are involved that are partially dependent on each other. As early as possible in the project, you should start to think about how to organise data into downloadable packages in a way that will promote reuse. There is no 'one size fits all' approach, but you should aim to strike a balance between clustering all the data from a project into a single dataset and splitting it into many independent datasets. Clustering ensures all dependent components are kept together, but this can reduce the discoverability of parts of the data, as the title and description will generally be more focused on the whole package. Splitting the data into too many parts can lead to a loss of context around the data, or repetition of common elements between disparate components of the dataset. Ultimately, the choice should be guided by what will maximise discoverability and usability in perpetuity for research topics not necessarily envisaged by your project.

8. **Use version control for code:** If your analysis can be implemented using code, the use of version control software such as Git and GitHub to record improvements over time is strongly encouraged. This will also allow you to publish the analysis associated with the research, significantly enhancing the reproducibility of the data analysis. At the end of the project this repository can be copied to the eAtlas or IMAS GitHub account or another suitable institutional account for archiving (Figure 5). This code can then be crosslinked with the matching metadata records associated with the project. The following is an example of a dataset with some scripts and some manual processing documented in GitHub: GBR_AIMS_eReefs-basemap.

   It is generally not a good idea to include data in your code repository, unless the data is very small (couple of MB). Git repositories don't manage non-text files (binary files) well. Every change in the data will result in another copy being recorded, resulting in potentially a very large repository. Public Git repositories should be no more than 1GB in size (this includes every version of the files stored in the repository) and GitHub blocks files larger than 100 MB.

   Important note: If your code includes connecting to remote services it is vital that you don't include any passwords or any form of credentials in your code.



11 results for **all** repositories matching **NESP** sorted by **last updated**

NESP-TWQ-3.2.5_WQ-Metric Public
Nesp 3.2.5 Water Quality Metric
R 0 0 0 0 Updated on May 26, 2022

GBR_NESP-TWQ-4.2_AIMS_Water-temperature-dashboards Public
Code used to generate the water temperature dashboard pages for the project: Oceanographic drivers of bleaching in the GBR: Water temperature dashboards (NESP TWQ 4.2, AIMS)
R 0 0 0 0 Updated on May 31, 2021

NESP-TWQ-5.1_CCC-DST Public
The COTS Control Centre Decision Support Tool (CCC-DST) is part of the COTS Control Centre Decision Support System and is a combined hardware and software solution developed by CSIRO as part of the National Environmental Science Program (NESP) Integrated Pest Management (IPM) Crown-of-thorns starfish (COTS) Research Program.
Java 0 0 0 0 Updated on Apr 20, 2021

NESP-TWQ-3.1.6_Exploring-WQ-trading Public
SP TWQ 3.1.6 project Exploring trading in water quality credits as a cost-effective approach for managing

Figure 5. The eAtlas maintains a team account for research code published by NESP researchers. Code repositories can be transferred at the end of the project for permanent hosting not reliant on an individual's GitHub account.

Towards the end of the project

By the end of the project all data should be published openly or archived for sensitive data.

All datasets (open and sensitive) should be described in a published metadata record with sufficient detail to allow reuse of the data.

Where necessary, embargos can be used to align data publication with research publication.

By the end of each project, all data from NESP-funded activities should be published or suitably archived (for sensitive data). The Hub Data Wranglers, as part of scheduled data discussions, can help navigate the preparation and publication of project data.

The following are some of the issues to consider and discuss:

1. **Logical grouping:** Divide your data into datasets based on the logical structure of the research. Group data by experimental methods, data processing methods, or intended use. Where data can be considered independent, publish it as a stand-alone dataset. The Hub Data Wranglers can provide advice on how best to structure your published data.

2. **Sensitive data:** Consider if there are sensitivities associated with the data. If so, you will need to plan how this data will be archived, curated, and governed. See section on *Managing sensitive data* for more information.

3. **Suitable data repository, long-term storage, and accessibility:** Work with the Data Wranglers to determine the most suitable data repository for your datasets. Datasets can be published via either of the Hub's primary repositories - **eAtlas** for northern node projects and **IMAS** for southern node projects, or other enduring public data repositories that comply with the Hub's standards. The choice of repository should be guided by what will maximise

discoverability and use of the data. In many cases this will be a thematic national data repository. See *Data repositories* for more information and for Hub-endorsed repositories.

4. **Metadata for each dataset:** Ensure that each published dataset has comprehensive metadata to describe the data. Clear metadata will make your datasets more discoverable and usable by others. The metadata should provide an overview of what the data contains, how it was produced, a description of the attributes and codes in the data, licensing conditions, links to associated documents that provide additional information, and an assessment of the limitations of the data. A metadata record should contain all contextual information that a third party would need to discover, understand, and re-use your data.

   All metadata records, regardless of where they are published, must be crosslinked with the metadata record for the NESP MaC project that produced the dataset. This ensures that a complete catalogue of all NESP MaC data can be compiled and accessed from a central Hub-level record. Notify the Data Wranglers upon publication of any metadata to allow them to make this crosslinking.

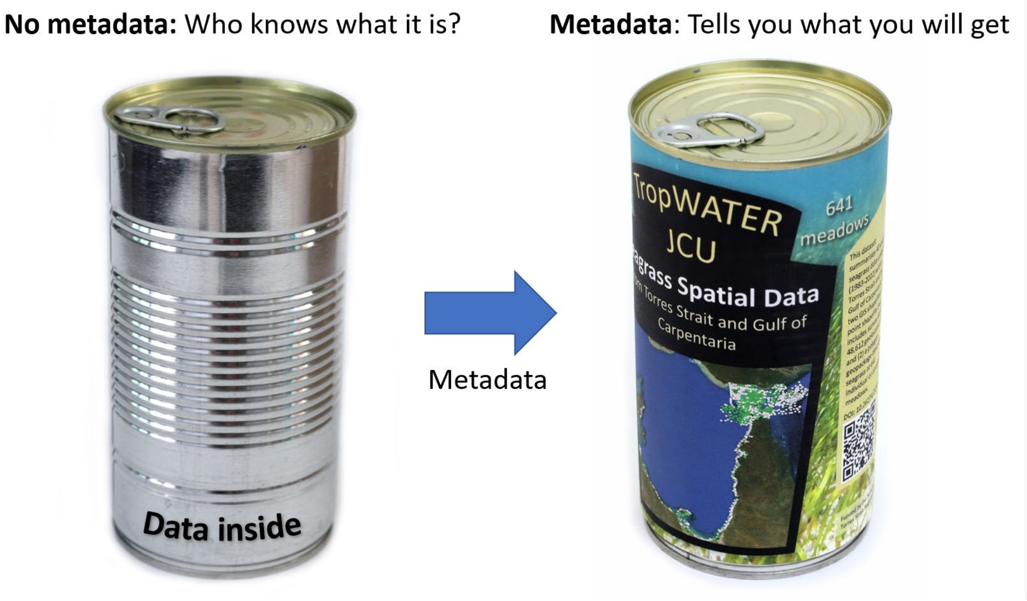

No metadata: Who knows what it is?
Metadata: Tells you what you will get
Data inside
Metadata
TropWATER
JCU
641 meadows
Seagrass Spatial Data
Torres Strait and Gulf of Carpentaria

Figure 6. Metadata is vital for making datasets discoverable and understanding their meaning. Without sufficient metadata those using the data must make educated guesses about its interpretation. This can result in incorrect use, potentially leading to spurious research outcomes. It is the responsibility of all researchers to ensure their data is sufficiently documented to allow its reuse.

5. **Consistent formatting and data cleaning:** Standardise the format, structure, and naming conventions for your datasets. This will improve the usability and interoperability of your data, making it easier for others to access and analyse. Use logical nomenclatures for files and folders and where possible convert the data into open data formats to ensure that propriety software isn't needed for reading the data. If your dataset consists of many files in a complex folder structure, use a readme.txt file to explain the organisation of the files.

6. **Data provenance:** Track and document the origin and processing history of your datasets, including any data cleaning or transformation steps. This information will allow users to assess the quality and reliability of your data and to reproduce your results. This can be reported in the metadata, or as a supplemental document linked to from the metadata.

7. **Data licensing:** Clearly state the licensing terms for each dataset. The default licensing for NESP datasets is Creative Commons Attribution 4.0 International (CC BY 4.0) license (https://creativecommons.org/licenses/by/4.0/). For datasets involving third-party data,

determine the most open license that is compatible with the source data licenses. Describe any access and use restrictions that may be necessary.

8. **Metadata review:** To ensure consistent quality of data archiving, the Data Wranglers should be given the opportunity to review metadata and provide feedback on all datasets regardless of where they are published. The Data Wranglers will review whether the dataset documentation is sufficient to allow reuse of the data and provide guidance where required. Datasets published through Hub's default data repositories will automatically be reviewed. If you are publishing with other repositories, please discuss with a Data Wrangler how this review will be achieved. Projects are expected to correct or improve their data documentation if requested by the Data Wranglers.

9. **Publication embargos:** Where there is a strong case that publication of the data will compromise research outcomes, such as for PhD students, it is possible to apply an embargo of up to 12 months on datasets. After the embargo period has lapsed, Data Wranglers will contact the researcher as to whether an extension is needed. If an extension is not required or the researcher is no longer available, the data will be published. Any data that is held under embargo must be fully prepared and ready for publication in the appropriate data repository (or supplied to the Data Wranglers directly). Note that the metadata for datasets under embargo will be published, this is to foster collaborations and controlled access to the data during this period.

10. **Dataset Digital Object Identifiers (DOI):** DOIs provide an enduring link to digital resources such as journal papers or datasets. In many cases journals require datasets to be accessible via a DOI link. Datasets published through the default Hub data repositories (eAtlas and IMAS) can be issued with a DOI to support these journals. If a DOI is required, this should be requested at the time of data submission as DOIs are not issued by default.

11. **Publish early if possible:** Publish your data as soon as it is ready; don't wait to the end of the project. Early publication of data maximises its potential uptake, encourages collaboration, and reduces potential duplication of effort. If necessary, datasets can be versioned to accommodate improvements over time.

Managing sensitive data

Some Hub projects will generate data that might be considered sensitive. This includes datasets with potential social privacy issues, environmental datasets that could result in species exploitation or environmental pressure, confidential data (including "commercial in confidence"), those containing traditional cultural knowledge and those based primarily on third party data with restrictive licensing. It is the researcher's responsibility to ensure these are appropriately managed, curated, and have suitable access controls. Managing sensitive datasets requires additional care, as compared with open data, to ensure they are stored securely and are used appropriately. When dealing with sensitive data the following should be considered:

> Projects should Identify any sensitive project data and develop a plan for its archiving, access control and governance. This should consider what will happen with the data once the original project team is no longer available.
>
> Private data must still be described with a public metadata record.

1. Identify sensitive datasets early in the project and develop a plan to securely work with the data during the life of the project.
2. Discuss the plan with your hub Data Wrangler during data discussions.
3. Plan for how the data will be managed after the project finishes including determining a suitably secure repository. Develop an access control policy that outlines who can access the data and under what conditions. A long-term custodian of the data should be identified and its governance. It is important to consider what will happen with the data once the original research team is no longer available. In many cases the sensitive data can be housed in internal institutional data repositories. The long-term archival location of the data should be described in the public metadata record.
4. Look for ways to publish a desensitised form of the data. This will ensure that the project can maximise on delivery of data suitable for public use, whilst protecting the sensitive data. This might include removing personally identifiable information, aggregating to larger spatial regions, or dithering the positional information so that only coarse spatial information is available. The goal is to develop a data product that is directly useful and helps other researchers to understand the scope of the underlying dataset and allow them to determine if they should pursue obtaining access to the full dataset. The Data Wranglers can provide advice or help to generate appropriately desensitised public forms of sensitive datasets, which can be published as a normal open dataset.
5. If the data is indigenous knowledge, then its management and use should be determined in consultation with the indigenous community. See *Managing Indigenous Data* for more detail.
6. Sensitive data that is not public must still have a public metadata record developed to ensure the work is discoverable. This metadata should document the data in the same level of detail as a public dataset and should contain details about the access control constraints of the dataset. This metadata should be published publicly through a suitable repository. Talk with your hub Data Wrangler about the most appropriate place to publish this record.
7. Some sensitive data is subject to legislation or organisational policies. Make sure you have identified the Intellectual Property (IP) owner of the data, considered policies of the IP owner with regard to data classification and access. Also consider seeking legal advice if in doubt of legislative requirements.

Managing Indigenous data

> Data that incorporates traditional knowledge should be considered an asset of the indigenous community.
>
> How the data can be used (published, archived, etc.) must be determined through discussions with the relevant indigenous community.

NESP Projects are expected to apply CARE principles for indigenous data governance and to respect Indigenous cultural intellectual property (ICIP) rights. This means projects collecting or analysing data that incorporates indigenous knowledge should be treated as an intellectual and cultural asset of the relevant indigenous community. Projects must establish a **Free, Prior and Informed Consent** plan for its collection, storage, access control and publication. This means that projects should discuss with relevant indigenous groups what data they intended to collect and produce. They should also determine what data can be published openly and what data needs to be managed as sensitive data. For sensitive data, a plan should be established for how and where the data will be stored, appropriate access controls, and how the indigenous communities can access and benefit from the data.

What are the CARE principles?

The CARE principles are:

Collective benefit: Data ecosystems shall be designed and function in ways that enable Indigenous Peoples to derive benefit from the data.

Authority to control: Indigenous Peoples' rights and interests in Indigenous data must be recognised and their authority to control such data be empowered. Indigenous data governance enables Indigenous Peoples and governing bodies to determine how Indigenous Peoples, as well as Indigenous lands, territories, resources, knowledges, and geographical indicators, are represented and identified within data.

Responsibility: Those working with Indigenous data have a responsibility to share how those data are used to support Indigenous Peoples' self-determination and collective benefit. Accountability requires meaningful and openly available evidence of these efforts and the benefits accruing to Indigenous Peoples.

Ethics: Indigenous Peoples' rights and wellbeing should be the primary concern at all stages of the data life cycle and across the data ecosystem.

For more information about the CARE principles see https://www.gida-global.org/care

What is Indigenous Cultural Intellectual Property?

Indigenous cultural intellectual property (ICIP) refers to the human rights of Indigenous communities to protect their traditional knowledge and cultural expression. Australian intellectual property laws such as patents, trademarks, designs, and copyright, provide only limited legal protections for ICIP as they do not recognise any communal rights. These communities hold highly valuable knowledge about place, natural resources, culture, and history.

NESP researchers working with Indigenous communities must ensure their ICIP knowledge is adequately protected and used in a way deemed fit by the community the information came from. This means that **Free, Prior and Informed Consent** should be obtained surrounding the collection and use of traditional knowledge. Projects working with ICIP should develop with the relevant indigenous groups a mutually agreed upon plan on what data will be gathered, how it will be stored and curated, what access control is necessary and how it can be published.

Indigenous data discussions

Each NESP project is categorised into three levels of indigenous engagement depending on whether the focus is on communicating (Category 3), collaborating (Category 2) or project co-design (Category 1). Category 1 and 2 projects will almost certainly produce datasets that are informed by indigenous knowledge and so must discuss and establish an agreement with the relevant indigenous groups as to what will be done with the project data. The occurrence of these discussions should be communicated with the Data Wranglers and will be reported as a NESP metric. For Category 3 projects the focus is on communicating research results to relevant indigenous groups so they can benefit from the knowledge gained from the research. In the case of a Category 3 project analysing third party data that incorporates ICIP, discussions about the use of the derived data must be had with relevant indigenous groups.

Additional resources

Carroll, S, et al. 2020. The CARE Principles for Indigenous Data Governance. Data Science Journal, 19: XX, pp. 1–12. DOI: https://doi.org/10.5334/dsj-2020-042

Delwyn Everard (2022) What is Indigenous Cultural Intellectual Property and how does it protect traditional knowledge? Accessed 29 Apr 2023. https://www.winyama.com.au/news-room/what-is-indigenous-cultural-intellectual-property

Moran A. (2020) What is Indigenous cultural intellectual property and copyright and how can I respect it? https://www.abc.net.au/news/2020-05-11/what-is-indigenous-cultural-intellectual-property-and-copyright/12150308

National Environmental Science Program 2021, Indigenous partnership principles, Department of Agriculture, Water and the Environment, Canberra. https://www.dcceew.gov.au/sites/default/files/env/pages/2f561690-b47e-4bf2-b028-d18739b3486f/files/nesp-indigenous-partnerships-principles.pdf

Indigenous partnerships strategy. National Environmental Science Program Marine and Coastal Hub. (2021) https://www.nespmarinecoastal.edu.au/wp-content/uploads/2022/09/NESP-MaC-Indigenous-Partnerships-Strategy_May-2022.pdf

Terri Janke (2019) Aboriginal Cultural and Intellectual Property Protocol. Aboriginal Affairs NWS Government. https://www.aboriginalaffairs.nsw.gov.au/media/website_pages/our-agency/staying-accountable/aboriginal-cultural-and-intellectual-property-acip-protocol/AANSW-Aboriginal-Cultural-and-Intellectual-Property-ICIP-Protocol.pdf

Three-Category Approach, Communicate, Collaborate, Co-design, https://nespurban.edu.au/3-category-workbook

United Nations (2007) United Nations Declaration on the Rights of Indigenous Peoples. https://www.un.org/development/desa/indigenouspeoples/wp-content/uploads/sites/19/2018/11/UNDRIP_E_web.pdf

Delivering Indigenous Data Sovereignty (2019) Presentations from the 2019 National Indigenous Research Conference. https://aiatsis.gov.au/publication/116530

Metadata checklist

In this checklist we focus on the most common mistakes or omissions from metadata records. Including these items in your metadata will significantly raise its quality, making your dataset easier to discover and reuse.

Does your metadata have a title that is short but descriptive and includes project codes and institutions? The title of the metadata record is important for the discoverability of the dataset. This should be kept concise as some systems will truncate the title after ~10 words, so the most important keywords should be used early in the title. Another consideration is that it is beneficial to include the program and institution abbreviation in the title. For example: \`(NESP MaC 1.30, IMAS, JCU and DPIRD)\`. Metadata records are aggregated, harvested, and made discoverable through multiple outlets, such as the Australian Ocean Data Network, Research Data Australia, Data.gov.au and Google Search. Including the institution in the title ensures credit for the institutions regardless of how the metadata is presented. Including the project code and the hub abbreviation helps keep track of NESP datasets.	□
Does your metadata provide sufficient details to understand the data? This includes the methods and processing that was used to produce the data. The metadata should provide sufficiently detailed descriptions (or link to a publication containing these descriptions) so that another researcher can fully understand the data. Without this the data is ambiguous and of limited value. If the methods are well described in publicly accessible technical reports or papers then these should be linked to from the metadata record.	□
Does your metadata include some kind of preview of the dataset (an image, map, or graph)? Having an image that represents the dataset, such as a photo of the experiment, a map of the data or a graph of the results, makes it quicker for researchers to determine if a dataset is relevant for their research. This helps them understand the scope and scale of the data without having to download and generate data previews. Multiple images can be associated with a single metadata record.	□
Does the metadata describe all the attributes and codes used in the dataset? Metadata records should include a description of all the attributes (column names in data tables) along with the units of that data. This is also known as a data dictionary. Often data is encoded into shorthand codes for efficiency reasons. The metadata should list what each code means. Consider using established vocabularies where variables relate to broadly used classifications (e.g., CATAMI, AODN vocabularies). Describing the data attributes also makes your dataset more discoverable as the text description of the attributes is searchable. Data dictionaries are rarely included in technical reports or papers associated with the data and so it is important to include this detail in the metadata. Data dictionary example: BOTANAL CSV headers: - ID: Unique identifier for sample - TRAN: Transect number (numbered from nearest to gully) - SP1 to SP5: Species name abbreviated using first two letters of genus and first three of species name e.g., Bothriochloa pertusa = boper. Species recorded in order of highest biomass represented. - YIELD: Total biomass estimate for quadrat in kg/ha - HARD: Soil hardness (after Tongway et al – landscape functional analysis). Categorical: 1=Easily broken, 2=Moderately hard, 3=Very hard, 4=Sand, 5=Self mulching.	□
Does the metadata describe the limitations of the dataset? All data has limitations and caveats. The metadata should highlight likely potential misuses of the data, or important caveats. Some common limitations that should be highlighted include:	□

Sampling biases – Sampling might have focused on a particular season or habitat and so might not be representative for a different research question. Data quality – Instrumentation measurements through to evaluations from expert elicitation all have errors in their data. The metadata should discuss the data accuracy. Data quirks – Long term datasets often have subtle changes in methodology or sampling strategies leading to potential pitfalls in how they should be analysed. Data misuses – Highlight key ways the data should not be used and why. An example: *This dataset should NOT be used for vessel navigation. It is a Digital Elevation Model (bathymetry) where some areas have poor levels of detail. Even where the source data is detailed the pixels represent an average of the depth over a 100 x 100 m area and small shallow reef navigation hazards are smoothed out and not visible in this dataset.*	
Does this dataset incorporate indigenous knowledge? If so, then the metadata should record what aspects of the data constitute indigenous knowledge and which indigenous groups it came from. It should document key agreements on the use, constraints and access control and governance associated with the data as determined by discussions with the relevant indigenous groups. The purpose of this is to track the Indigenous Cultural Intellectual Property associated with the dataset and its expected use. This will allow future researchers and data curators to understand how the data should be managed and which indigenous groups should be contacted should there be a new potential use of the data. Where the dataset is made openly available then having this in the metadata indicates that the relevant Indigenous groups have agreed to its public release.	□

Data repositories

NESP MaC Hub researchers are encouraged to consider which repository is the most appropriate to host their project datasets. In many cases, discipline-specific repositories are likely to have the most significant impact, but these must also adhere to established best practices for data publishing.

For a data repository to be endorsed by the Hub, it must commit to the permanent preservation of data and make it available for public download. If the facility also hosts metadata, it must provide sufficient detail to enable confident reuse of the data and utilise an internationally recognized machine-readable schema such as ISO 19115. For projects intending to use a data repository other than the Hub primary repositories (eAtlas and IMAS) – including ones listed in the table below – this should be discussed with the Data Wranglers who can provide advice and guidance.

The Hub's "distributed" data model encourages and fosters institutional data publishing capability and empowers researchers to manage their data from collection to publication. All Hub data is centrally discoverable via the Australian Ocean Data Network aggregator and is linked to a single central Hub-level record, allowing navigation between all project- and dataset-level records in a hierarchical tree.

The portals and repositories listed below are recommended by the Hub due to their endorsement by the research community. This list is not intended to be exhaustive and may change through time. Please contact the Data Wranglers to discuss any options not included below.

Data repository/portal	**Research discipline**	**Repository?** (can data be *stored* here?)	**Endpoint?** (is this a *visualisation platform?)*	**Note and contact**
IMAS data catalogue (UTAS)	Various	✓	✗	Primary repository for southern node hub projects, national focus. IMAS.DataManager@utas.edu.au
eAtlas	Various	✓	✓	Primary repository for northern node hub projects, northern Australia focus. e-atlas@aims.gov.au
eCAT (GA)	Various	✓	✗	Used by Geoscience Australia collaborators and various seabed data products hosted for AusSeabed. clientservices@ga.gov.au
MarLiN data catalogue (CSIRO)	Various	✓	✗	Used by CSIRO collaborators. hf-data-requests@csiro.au
Seamap Australia	Benthic habitat data, spatial data relevant to planning and management	✓ (habitat data)	✓	National repository for benthic habitat data. Visualisation platform and tool for management for other spatial data types with regional to national focus. IMAS.Seamap@utas.edu.au
AusSeabed	Bathymetry data	✓	✓	Processing pipeline and national repository for bathymetry data. ausseabed@ga.gov.au
SQUIDLE+	Still seafloor imagery annotation data	✓ (annotations)	✓	Platform for viewing and analysing seafloor imagery. Provides API for access to images and a repository for annotation data. Relies on remotely stored imagery. ariell@greybits.com.au
GlobalArchive	Fish video annotation data	✓ (annotations)	✓ (deployment info)	National repository of fish video annotation data. Flexibly ingests annotation data in a range of common formats. Does not provide storage for raw video data. tim.langlois@uwa.edu.au
Atlas of Living Australia (ALA)	Biodiversity observational data	✓	✓	Data is aggregated nationally and not attributable to a single project – data should be published to an alternate primary repository and

				ALA used as a second contribution point. support@ala.org.au
AAD Data Centre	Antarctic-relevant	✓	✗	aadcwebqueries@aad.gov.au
Australasian Right Whale Photoidentification Catalogue (ARWPIC)	Right whale imagery	✓	✓ (all images are public)	Hosted by the Australian Marine Mammal Centre ammccoordinator@aad.gov.au
Institutional GitHub account	Code	✓	✗	Code must be associated with an institutional or team account, not an individual.
eAtlas GitHub account	Code	✓	✗	Code can be transferred to the eAtlas account for long term publication.
IMOS Animal Tracking Facility	Animal tracking data	✓	✓	Data may be subsequently contributed to international animal tracking repositories.

Roles and Responsibilities

Researchers/projects

- Consider what datasets will be developed during the life of the project and plan for their development and final publication early in the project. (Consider evolving projects).
- Follow good data management practices during the life of the project (backing up data, keeping track of source data and licensing, recording processing that was performed to create the published data)
- Engage in data discussions with hub Data Wranglers throughout the life of the project.
- Prepare datasets for publication including metadata preparation, data cleaning and documentation of the production of the data. Documentation needs to be sufficiently detailed that datasets can facilitate their correct interpretation.
- Publish the datasets through an approved public, enduring data repository. Have the required discussions with your Hub Data Wrangler to help you chose this repository, and to arrange publication with this repository.
- Address any issues that arise from the data review process.

Data Wranglers

- Conduct data discussions with each project to help identify and resolve any data related issues and help determine which datasets might be published by each project.
- Track data related project milestones and report these to the Hub management.
- Assist projects to publish their datasets. If the data is published via IMAS or the eAtlas, ensure that the relevant data and metadata services are created.
- Create a project metadata record for each project based on information provided in the project proposals. Datasets created by each project will be linked to their project record and these will be linked with a record representing the Marine and Coastal hub, forming a tree of linked records. Where datasets are published in third party repositories (not IMAS or eAtlas), request that repository administrators add links (provided) to Hub metadata records to ensure cross linking between datasets and their associated project metadata records in centralised aggregators (e.g., AODN).
- Prepare, curate, and publish map data that represents the activities and area of relevance for each project. The boundary of the map should be refined based on input from the project researchers.
- Perform a review of project datasets, focusing on assessing the completeness of the dataset documentation. Check that key aspects of the documentation are complete, that the dataset aligns with the documentation, and that key limitations of the dataset are highlighted and documented. Send any issues raised during the review process to projects to address.

Key Contact Information

For up-to-date contact information see the NESP MaC website.

Team	Contact	Role (in NESP context)
IMAS (UTAS)	Emma Flukes Emma.Flukes@utas.edu.au Tel: 0408 901 952	Data Wrangler (south)
eAtlas (AIMS)	Eric Lawrey e.lawrey@aims.gov.au Tel: 0402 900 580	Data Wrangler (north)