Name: _____

Introduction to Data Management Best Practice

User Handbook

Ocean Country Partnership Programme

This project was funded with UK International Development from the UK government.

The Ocean Country Partnership Programme (OCPP) is a bilateral technical assistance and capacity building programme that provides tailored support to countries to manage the marine environment more sustainably, including by strengthening marine science expertise, developing science-based policy and management tools and creating educational resources for coastal communities. The OCPP delivers work under three thematic areas: biodiversity, marine pollution, and sustainable seafood. Funding is provided through the overarching Blue Planet Fund (BPF) by the UK Department for the Environment, Food and Rural Affairs (Defra).

The authors would like to thank and acknowledge Defra on behalf of the UK government for the funding: project number GB-GOV-7-BPFOCPP in funding this work.

UK International Development

Partnership | Progress | Prosperity

Contents

1. Overview

This handbook gives an overview of best practice guidance for establishing an effective data management process. This handbook follows all stages of the data life cycle, from planning to storage/publishing, to highlight key points to take into consideration to promote good data practice.

The handbook has two sections:

1. **Key data concepts**
   Clarifying data terminology and key considerations in data management.
2. **The Data Life Cycle**
   Step by step guidance on data management, from planning phase to reflecting on lessons learnt.

The handbook also includes a best practice checklist, for use as a tool to follow in your daily work in data management, and standard templates to adapt as needed for data collection. Finally, additional resources and guidance materials are summarised for further learning.

The examples given throughout this handbook focus on environmental data, but the general principles of data management apply to many data types.

2. Key Data Concepts

2.1. Data management

Data management is the practice of **collecting, keeping and using data in a secure way to enhance their value.**

Having a data management process in place is very important to ensure that the data are accurate and therefore they can be used for making effective decisions.

The data management process helps us to answer the following questions:

- What data do we have and how can we use it?
- Where did we get data from and when?
- Are there licences/restrictions in place?
- How can we ensure that data are secured and properly stored?
- How can we mitigate possible issues?

2.1.1. The importance of data management

Ineffective data management can limit the use of data, as users may not know a dataset exists or users may not be able to find the results of a useful research project. This can lead to lower productivity, unnecessary additional work or duplication of effort, and potentially damage the reputation of an organisation.

Therefore, it is important to have a process in place that defines rules and structures to be followed to ensure a high quality data are produced, which supports robust decision-making in conservation policies and strategies. Following data management best practice helps to realise value for money, enabling us to "**collect once, use many times**".

2.1.2. Key factors of good data management

1. Ensuring tidy and organised systems by having a set of rules in place to give guidelines on how to effectively catalogue data and ensure consistency.

2. Ensuring consistency by using standards: an agreed, repeatable way of doing things. This would help others to understand what has been done, avoiding misinterpretation of data, and makes it easier to track and correct errors.

3. A process in place to make it clear how data can be accessed and used appropriately and within their restrictions (i.e. having a licence).

2.2. Data

The word datum is derived from Latin, and it means "**something given**". Data, which is the plural for datum, are **facts**, for example relating to an object, event, or process.

Data are used to help organisations make effective decisions, and it is therefore crucial to ensure their quality. However, data have less meaning and value if users do not **also have information about what the data relate to**. For example, an image has less meaning or value without the knowledge of when, where and how it was taken. This is why it is important to always have the background information (referred to as metadata) for the data.

Additional notes:

2.3. Metadata

Metadata is data that provides information about a file / data set / resource, or simply "**data about data**". It is **structured reference** data that helps to sort and identify key information. By searching for a key term, or particular element (or elements) within the metadata, it is much easier to locate and use a specific file / data set / resource. **To follow data management best practice, data and their associated metadata cannot exist independently.**

2.3.1. Key principles of metadata

Metadata should be:

1. **Accurate:** the metadata correctly and precisely describes the resource in question.

2. **Beneficial:** the metadata contains information that is useful to the user and doesn't have lots of unnecessary information.

3. **Clear:** the metadata should be easily understandable by a non-technical user and should be unambiguous. Write for readers not robots!

4. **Distinctive:** the metadata contains information that allows it to be distinguished from other, potentially similar, resources.

2.3.2. Benefits and considerations of metadata

Benefits:

- **Enhanced data management:** Data are easier to manage and access, as metadata allows data to be catalogued/stored to enable the searching of the datasets using key characteristics (e.g. key words, date published, area covered).
- **Easier data sharing:** By making data more easily findable and accessible, metadata makes it easier to share data.
- **Increased use of data:** Increase in use of data because it can be more easily found, shared, and understood by users.

Considerations:

- **Cost:** Costs may be involved in maintaining metadata effectively, for example training staff to create or access metadata.

- **Complexity:** There are many metadata standards to choose from, and learning how to use these may require training. It is important to choose a simple standard/template to follow that works for you and your team.

2.3.3. Metadata management

As discussed, metadata is a labelling system that can be read by humans and computers and allows search engines to locate data using defined metadata fields. Managing metadata can be done in 2 main ways:

1. **Passive**: provides a basic identification system using technical information but does not offer significant context and does not change in-line with the source data, unless it is directly updated each time there is an alteration to the source data.

2. **Active**: provides greater context and is 'dynamic', meaning that the metadata is automatically updated whenever the source data are altered. Systems that use active metadata management promote the continuous updating of metadata being used for ongoing projects and real-time customer service.

2.3.4. Types of metadata

There are various types of metadata that are commonly used. Each type has different benefits as detailed in the figure overleaf:

TECHNICAL

Commonly used with passive metadata and includes data type, indexes referencing the data.

- **File formats**
- **File names**
- **Data sources**
- **Geographic locations**

OPERATIONAL

Includes information about when and how the data was transformed or created.

- **Dates of updates**
- **Loading date**
- **Lineage**
- **Data's status**

STRUCTURAL

Provides information about the physical organization of a data – the relationships, types, versions, and other characteristics.

- **Data element types**
- **Table names**
- **Record size**

ADMINISTRATIVE

Provides information to manage and establish data credibility and controls who and how the data may be used.

- **Copyright info / license agreements**
- **Technical data on rights management**
- **User restrictions**
- **Access control info**

PROVENANCE

Tracks the data's origin and any changes over time. It provides data traceability, improving Data Quality.

- **Authority**
- **Change logs**
- **Ownership records**
- **Versioning records**

PROCESS

A subdivision of operational metadata which provides details of the process of loading data into storage. This type of information is useful in case of a problem.

- **Errors logs**
- **Job execution logs**
- **Audit results**

BUSINESS

Provides definitions, rules and restrictions on the data's use.

- **Timelines**
- **Business requirements and models**
- **Business process flows**
- **Metrics**
- **Business terminology**

SOCIAL

Provides information and context on how people use data. Enables organisations to decide whether to alter outreach / productivity.

- **Author information**
- **Most queried tables**
- **Frequency of use**

Examples of information that should be saved within metadata include:

- Description
- Licence type
- Image description
- Copyright
- Licence limitation
- Species scientific name
- Survey code
- GPS coordinates on where the image was taken
- Data when the image was taken
- Survey date
- Survey location
- Species count
- Feature name
- Data owner
- Type of survey
- Method of collection
- Output type
- QA method

As just explored, there are many types of metadata. For survey data, it is common to focus on discovery metadata and content metadata:

Discovery metadata

Information at a top level of the survey to simply answering the following question: who, when, why, what and how. This metadata is associated with a survey when data are archived to help find it by using keywords.

Content metadata

Information providing specific context around a collection of data, for example where a specific sample was taken, and with what gear. In multi-level or hierarchical data, each level is generally used as the metadata for the data that sits underneath it.

2.3.5. Metadata standards

A **metadata standard** is a set of guidelines or specifications that defines how metadata should be structured, formatted, and used. These standards ensure consistency, and quality in the description of data across different systems, disciplines, and organizations. Metadata standards typically specify:

- **Elements**: The specific pieces of information to be included in metadata (e.g., title, creator, date).
- **Format**: How those elements should be formatted (e.g., data types).
- **Best Practices**: Recommended practices for creating, managing, and sharing metadata to enhance usability and discoverability.
- **Compatibility**: Guidelines that facilitate the exchange and integration of metadata between different systems and applications.

By following a metadata standard, organisations can improve access to their data, improve data sharing and ensure the metadata is in the same format no matter who created it.

Some key environmental metadata standards and concepts are:

- ISO 19115: This is an international standard for geographic information that defines how to describe geographic datasets, including environmental data. It covers metadata elements like identification, quality, spatial reference, and distribution.
- Dublin Core: A simple and widely used standard for describing resources, including environmental datasets. It consists of 15 core elements, such as title, creator, subject, and date.
- EML (Ecological Metadata Language): This is specifically designed for ecological and environmental data. EML allows for detailed descriptions of data, including methodologies and protocols used in data collection.
- INSPIRE Directive: In the EU, this directive aims to create a European Union spatial data infrastructure, facilitating the sharing and access of environmental data across member states.

ISO 19115 and EML standards are more complex standards that use an XML format to make the metadata readable by humans and machines. XML format is an eXtensible Markup Language which is designed for simplicity, generality, and usability across the internet. INSPIRE Directive aims to create a spatial data infrastructure and is used extensively across the EU. It sets out the technical guidelines that set out the creation and maintenance of this metadata.

The Dublin Core metadata standards is the most widely used metadata standard for environmental data as it is a simple and flexible standard that describes a wide range of resources including both digital and physical resources. It contains 15 core elements making it easy to implement. The elements include title, creator, subject, description, publisher, date, type, format, source and rights. The standard can also be extended by adding additional elements as needed.

2.3.6. Metadata examples

Figures below are an example of discovery metadata (Figure 1) and content metadata (Figure 2), accessed through the Marine Environmental Data & Information Network (MEDIN), it is a partnership of UK organisations that work together to improving access to UK marina data. MEDIN provides a MEDIN metadata standard, with guidelines that

provide a list of information to include alongside your data to maximise their re-used in future as well as it ensures consistency in the vocabulary used by having data standard in place. For more information, please visit the official MEDIN website.

These example metadata entries highlights some key elements of metadata, for example the date(s) the dataset relates to, the geographical coverage, the lineage of the data, the data owner etc.

Metadata: 2021, JNCC, Offshore Wind Evidence and Change Programme, Offshore Wind Environmental Evidence Register

Abstract:

he OWEER draws together the key evidence gaps around offshore wind environmental impacts as well as the research projects recently completed, in progress, and in planning, that are relevant to reducing these evidence gaps. The primary purpose of this register is to assist in prioritising funding for the Offshore Wind Evidence and Change (OWEC) programme of strategic research but it has also been made publicly available for wider use to highlight priority evidence gaps, increase understanding of the breadth and scope of the research field, reduce project duplication, foster collaboration and disseminate project findings.

Contact OWECEvidenceRegister@jncc.gov.uk for any feedback, comments or questions on the OWEER.

Data holder:

The Crown Estate

Click on the button to access resources online: Click on the button to access resources online

Use constraints:

Publically Available

Other details	
Internal code	7903
Title	2021, JNCC, Offshore Wind Evidence and Change Programme, Offshore Wind Environmental Evidence Register
File Identifier	d41ef53f7a540e0f78d93375bbc0aa70
Resource Identifier	OW-MDE-100-3480
Resource type	series
Start date	2020-01-01
End date	2021-12-31
Spatial resolution	inapplicable
Frequency of updates	biannually
Abstract	he OWEER draws together the key evidence gaps around offshore wind environmental impacts as well as the research projects recently completed, in progress, and in planning, that are relevant to reducing these evidence gaps. The primary purpose of this register is to assist in prioritising funding for the Offshore Wind Evidence and Change (OWEC) programme of strategic research but it has also been made publicly available for wider use to highlight priority evidence gaps, increase understanding of the breadth and scope of the research field, reduce project duplication, foster collaboration and disseminate project findings. Contact OWECEvidenceRegister@jncc.gov.uk for any feedback, comments or questions on the OWEER.
Lineage	This register was populated using information requested from relevant organisations (please see list in ReadMe tab and in OWEER content). Requests for information to populate the OWEER were made throughout April and early May 2021. The "Research" tabs cover current and recently completed research covering the last 18 months only. For this version of the OWEER, this covers the period of 2020 - May 2021. This register currently covers three receptor groups: ornithology, marine mammals, and benthic receptors. It is intended that future iterations of the OWEER will include fish as an additional receptor group. It is also intended that relevant research outside of the UK (probably limited to the EU) will also be included. There may be multiple entries for similar evidence gaps or similar entries that have been submitted by different organisations. The evidence gaps noted in tabs 1, 2 and 3 have been prioritised ny JNCC spcialists using the process outlines in the "Prioritisation process" tab. This gives a total scoring of between three and 15, which are shown in the relevant tabs and graded by colour for ease of interpretation.

Related keywords	
Keyword	Marine Environmental Data and Information Network
	Species distribution
	Zoobenthos generic abundance
	Cetacean abundance
	Bird behaviour
	Fish behaviour
Geographical coverage	
North	64
East	4
South	47
West	-16
Regional sea	unknown

Responsible organisations	
Role	pointOfContact
Organisation name	The Crown Estate
Position name	Marine Data Advisor
Phone	+44 020 7851 5000
Delivery point	16 New Burlington Place
Postal code	W1S 2HX
City	London
Email	enquiries@thecrownestate.co.uk
Role	distributor
Organisation name	The Crown Estate
Phone	+44 020 7851 5000
Delivery point	16 New Burlington Place
Postal code	W1S 2HX
City	London
Email	enquiries@thecrownestate.co.uk
Role	originator
Organisation name	The Joint Nature Conservation Committee (JNCC) - Aberdeen Office
Individual name	Marine data JNCC
Phone	+44 (0)1224 266550
Fax	+44 (0)1224-896170
Delivery point	Inverdee House, Baxter Street
Postal code	AB11 9QA
City	Aberdeen
Email	Oliver.Crawford-Avis@jncc.gov.uk
Role	custodian
Organisation name	The Crown Estate
Position name	Marine Data Advisor
Phone	+44 020 7851 5000
Delivery point	16 New Burlington Place
Postal code	W1S 2HX
City	London
Email	enquiries@thecrownestate.co.uk

Resource locators	
Locator URL	https://www.marinedataexchange.co.uk/details/18/summary
	https://www.marinedataexchange.co.uk
Dataset constraints	
20 Limitations on Public Access - Access constraints	otherRestrictions
20 Limitations on Public Access - Other constraints	Publically Available
21 Conditions for Access and Use - Use limitation	https://www.marinedataexchange.co.uk/content/info/faq
Available data formats	
Data format	Documents
Version info	
Date of publication	2021-06-10
Date of last revision	2022-01-20
Harvest date	2024-11-25
Metadata date	2022-01-20
Metadata standard name	MEDIN Discovery metadata standard
Metadata standard version	2.3.8

Figure 1. Example of discovery metadata, which includes headline information for a survey (marked by the blue headers) as follows: related keywords, geographical coverage, responsible organisations, resource locators (provides the link to where the dataset can be downloaded, in this case through the marine data exchange website) dataset constraints, data format and version information. This discovery metadata, is accessible via the MEDIN portal.

MNCR: S684.31.14 1

← View all results →

Resource tools

File information | Options

Original JPG File
6315 × 4015 pixels (25.35 MP)
53.5 cm × 34 cm @ 300 PPI
1.4 MB

Download

2

Low resolution print

Download

2000 × 1272 pixels (2.54 MP)
16.9 cm × 10.8 cm @ 300 PPI
342 KB

Add to collection
Share
Image tools
Edit with AI (beta)

Full screen preview

Resource details

RESOURCE ID
12362

ACCESS
Open
3

CAPTION
<i>Coryphella lineata<i>.

DESCRIPTION
4
Digitized 35mm slide from JNCC's MNCR collection. <i>Coryphella lineata<i>. Taken at site: Cave N E of Cambir, Hirta, St Kilda.

ATTRIBUTION
JNCC

COPYRIGHT
5
(c) JNCC

LICENSE
View License

USE LIMITATIONS
Available under the UK Open Government License v3
6

NOTABLE TAXA (SCIENTIFIC/LATIN)
Coryphella lineata
7

LOCATION NAME
St Kilda, Hirta, Cave N E of Cambir

COUNTRY
Scotland

SURVEY CODE
8
S684

STATION CODE
31

IMAGE NUMBER
14

STILL REFERENCE
S684.31.14

GPS LONGITUDE
9
8.6106848329611

GPS LONGITUDE REF
W

GPS LATITUDE
57.837972673249

GPS LATITUDE REF
N

DATE
10
08 July 97

Figure 2. Example of content metadata: an image and its associated metadata found in the JNCC image catalogue.

The annotations are as follows: 1: Image reference number; 2: How to download the image; 3: Licence type; 4: Image description; 5: Copyright; 6: Use limitations under licence; 7: Species Latin name; 8: Survey code; 9: GPS coordinates of image location; 10: Date of image.

Additional notes:

Additional notes:

2.4. Data ownership

It is important to identify who owns the data in a dataset, as well as who has responsibility for managing the data, so it should be clear who is responsible for the data your organisation holds. This could be helped by having an inventory of data to establish lines of accountability.

Data owners have overall accountability for the meaning, content, quality, distribution, and use of a dataset. They are empowered to ensure that the data is designed for its intended purpose.

Data owners are accountable for:

- Understanding who owns the data in a dataset. This is particularly important when data has been collected from local communities or indigenous people's communities as data managers must respect their rights in line with Free, Prior, and Informed Consent (FPIC) principles.
- Understanding how and by whom their data are used, and ensuring that the flow of the data to users is controlled properly
- Ensuring an appropriate retention schedule to outline storage periods which should be reviewed regularly
- Limiting access to data as appropriate, respecting FPIC agreements with communities.

Free, Prior, and Informed Consent (FPIC)

Every local community or indigenous people's community has a diversity of rights that need to be respected: one of those is free, prior, and informed consent.

Free, Prior, and Informed Consent is a process that ensures Indigenous Peoples have a say in decisions that affect their rights, lands, and resources. It's stipulated in Convention 169 of the International Labor Organization on Indigenous and Tribal Peoples, Convention on Biological Diversity and United Nations Declaration on the Rights of Indigenous Peoples (UNDRIP).

"Consent should be sought before any project, plan or action takes place (prior), it should be independently decided upon (free) and based on accurate, timely and sufficient information provided in a culturally appropriate way (informed) for it to be considered a valid result or outcome of a collective decision-making process."

FPIC is a process that can lead to mutual agreements between indigenous peoples, local communities, governments, and companies. FPIC allows communities to give or withhold consent to a project that may affect them, their land or sea. Communities have the right to withdraw their consent at any stage without penalty. FPIC enables communities to negotiate the conditions under which a project is designed, implemented, monitored, and evaluated.

FPIC is important in the Pacific because it can help ensure that indigenous peoples are included in decision-making processes that affect them. For example, in the context of deep-sea mining, FPIC can help ensure that indigenous peoples are included in decision-making roles, even if the mining activities aren't directly on their land.

The principles of FPIC are:

Free: Consent must be given without coercion, intimidation, or manipulation.

Prior: Communities must be given enough time to review information and provide feedback before the activity begins.

Informed: The information provided must be detailed and presented in a language and format that the community can understand. It should also emphasize both the potential positive and negative impacts of the activity.

2.5. Data licencing

A data licence is **a legal arrangement between the creator of the data and the end-user** specifying what users can do with the data. The licence information is a requirement when sharing data to inform which restrictions (if any) are in place, who can use the data and how. This can be done at an entire survey (series) level or for each specific dataset or data type held within, in the cases where data types have differing sensitivities. Be aware that there might be different restrictions between different data outputs within a single project or survey, and data should be managed accordingly. As a rule of thumb, discovery metadata records should be created *at least* for each dataset or group of datasets that fall under specific license conditions. There are different licences, for example in the UK the open government licence (OGL) is a standard open licence in place for British institutions which enable open data sharing.

To demonstrate how licences work, some examples of licences are summarised below, outlining pros and cons for each.

Creative Commons No rights reserved licence (CC0):

- This licence allows users to: copy, modify and distribute the data without attribution or any other conditions.
- Pros: The lowest barrier licence, this underpins truly open data without concern from the licensee or licensor as to additional stipulations.
- Cons: Aside from general open data concerns that the organisation may have, the lack of attribution required under this licence may prevent an organisation from assessing the impact that their data holdings are having, as all usage will be silent.
- Link to licence information

Creative Commons licence with attribution (CC-BY):

- This licence allows users to: Copy, adapt, and distribute the data and use the data commercially.
- With this licence you must: include any attribution statement, provide a link to the licence and indicate if and what changes were made.
- Pros: Widely compatible with most international data repositories and other public sector licences, maintaining attribution whilst still remaining low-barrier.
- Link to licence information

Creative Commons with attribution non-commercial (CC-BY-NC):

- The data can be used similar to the CC-BY licence, however data cannot be used for commercial purposes.
- Pros: May prevent misuse of data under commercial organisations, specific to organisations requirements if they do have a non-commercial policy.
- Cons: May restrict opening up the true value of the data unnecessarily if the organisation has a true open-data policy.
- Link to licence information

Public Data Licence Agreement under SPREP and the Pacific Environment Data Portal:

- Based on CC-BY-NC, and currently used by the Vanuatu Environment Data Portal
- Link to licence information

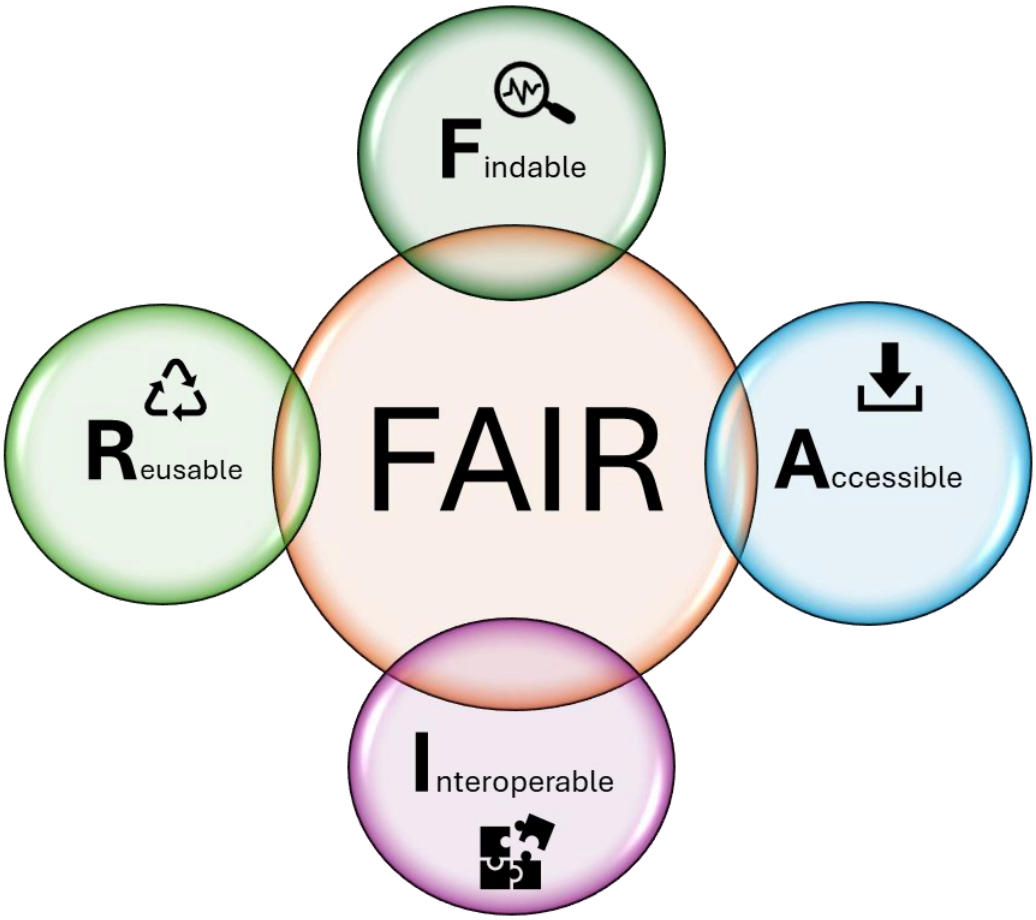
Despite the different licences that could be available, maintaining a standard and consistent approach to licensing across public sector bodies within a sovereign state allows for much lower friction data sharing, both externally to the public sector, but more importantly internally, providing cost savings and lower overheads, which would be required for licence management, negotiation and data sharing agreements.

Additional notes:

2.6. The FAIR data principles

The FAIR data principles are general guidelines to produce high quality data. These principles define the 4 most important rules in data management which must be referred to at each stage of the data life cycle. These rules are set to make sure the data are ready to be shared at the end of a research project. This promotes the “spirit” of having scientific data as openly accessible as possible, which is beneficial at a larger scale as existing data can be re-used to reduce duplicating efforts.

According to the **FAIR** principles, data must be: **F**indable, **A**ccessible, **I**nteroperable and **R**e-usable.



Findable:

The FAIR principles highlight the importance that both metadata and data need to be **easily findable by users**. Finding data can be a very time-consuming process, and it is essential to have a **unique identifier** (for example a reference number, URL, or digital object identifier “DOI”) associated with the data to make this process easier. To help the user to understand if the data fulfils their needs, data required to be associated with rich metadata, including information on the data format, collection details, and analysis methodology. It would be beneficial to have a unique data format across different data product to be machine-friendly (can be read by machine) to allow data searches.

How to make data findable, summary points below:

F

1. Assigning a unique identifier for both data and metadata (URLs or DOI)
2. Data are associated with complete metadata
3. Metadata clearly has the unique identifier of the data it describes
4. Being able to use keywords to find the metadata and data

Accessible:

Once the data have been found, it **needs to be clear how the dataset can be accessed** and used. There are some challenges in accessing data that should be taken into consideration such as:

- Can the data be downloaded? If yes, is the format accessible or does the format need to be converted?
- Can metadata be downloaded? If yes, is the format accessible or does the format need to be converted?

Data products need to be easily accessible by a common tool such as a web browser. The metadata also need to be accessible and complete to avoid misinterpretation of the dataset. After the data and metadata have been downloaded another possible barrier to keep in mind it is the format. It is useful to have a format that it is both human and machine readable. For instance, both Microsoft Word and Excel formats are user-friendly for humans, but they are not machine readable (for example when using statistical software or GIS with geo-packages). Therefore, having a comma separated files (.csv) could be a better option, as this format is both human and machine readable. The same principles apply to the metadata, which should be provided under an open file format that does not require licenced software (for example, PDF or HTML formats are more preferable than Microsoft Word format). Finally, having data products in one location (sometimes referred to as a data archive centre) is a good strategy to avoid users to have to search in different locations. Ideally, everything would be accessible from one place and datasets can be catalogued to enable search functions, making them easy to find.

How to make data accessible, summary points below:

A

1. Metadata and data can be retrieved by their identifier using an open tool
2. The tool needs to be free, universally implementable
3. Metadata needs to be accessible even when the data no longer exist
4. The format needs to be accessible and readable and
5. All in one place

Interoperable:

Data need to be able to **integrate with other data/applications** for both storage and processing to promote efficient joint working. For data to be interoperable they need to use a **common language** to allow data users to understand them easily and to enable data to be combined. Finally, where possible it would be beneficial to link data to relevant other data/metadata to make it easier for users to find additional information that can be combined for further analysis or presentation.

How to make data interoperable, summary points below:

I

1. Metadata and data need to use a common language/vocabulary to represent the data
2. Metadata and data should be linked to other relevant metadata/data

Re-usable:

The final goal of the FAIR principles it is to **optimise the re-use of data**. To do so, it is important to have **detailed metadata** so that future users are able to **understand the information easily** to avoid misinterpretation and incorrect use. Also, the methodology needs to be well described to enable users to repeat the collection in case further investigation is required (for example, when it is wanted to look at changes in the environment over time). The data/metadata need to have a clearly stated licence so users know how they can be used (what can/cannot be done with the data/metadata). This means that FAIR data are **not necessarily** open data and vice versa. Finally, it needs to be clearly stated how analysed data/metadata were created (for example, by linking the raw data and analysed data products together using a unique identifier).

How to make data re-usable, summary points below:

R

1. Metadata and data need to be accurate and thoroughly described
2. Metadata and data require a clearly stated licence
3. Metadata and data have a clear description of where they originated from
4. Metadata and data should meet community standards

FAIR data are **always managed** and shared respecting any legal restrictions.

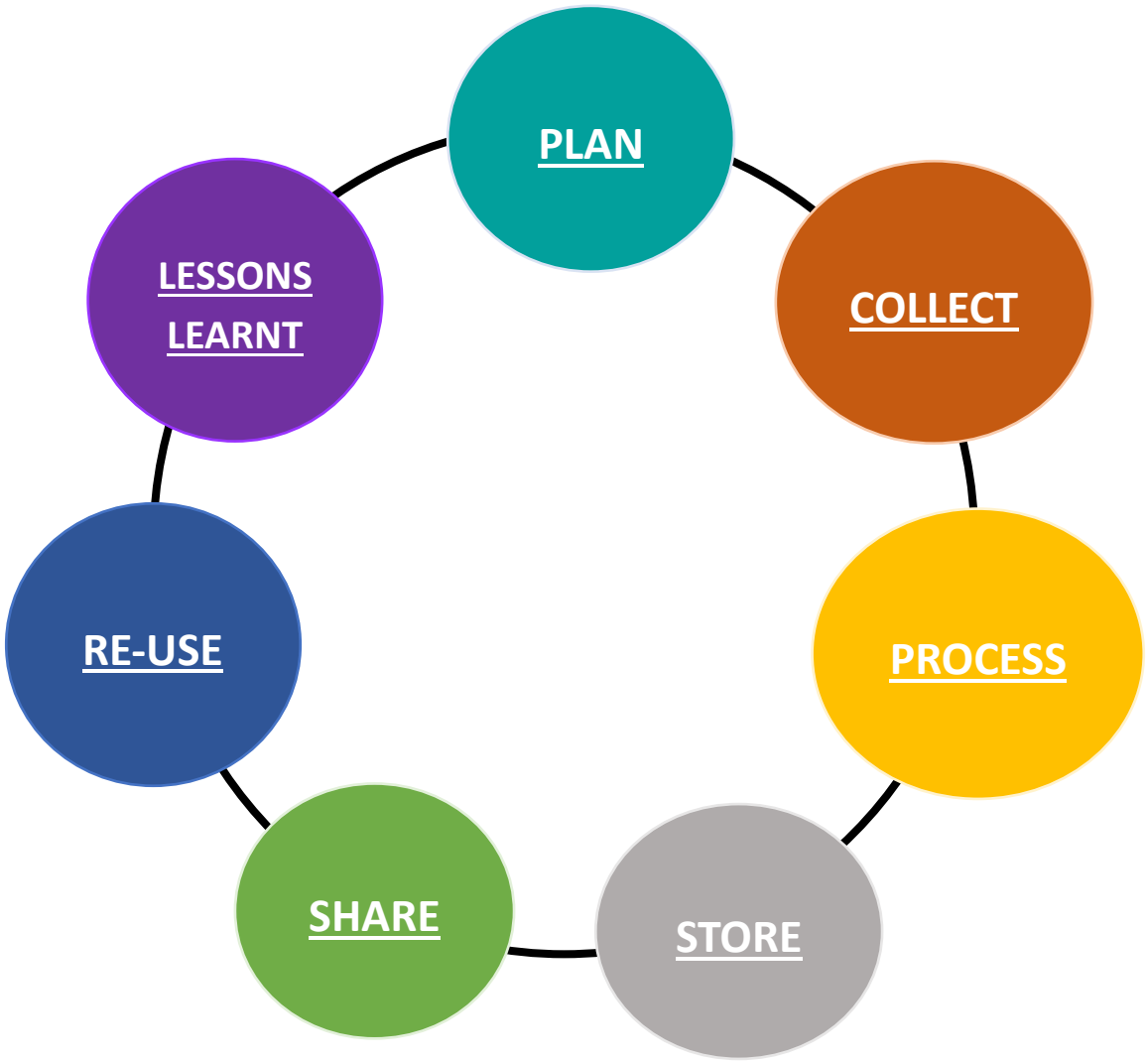
Using FAIR principles, it is not a goal but a process.

Additional notes:

Additional notes:

3. The Data Life Cycle

In this section we will talk about the data life cycle to explore the flow of the data all the way from collection to publication. Understanding each step is essential to appreciate the importance of having good data practice. Each stage of the life cycle will be explored to highlight the different requirements/procedures for ensuring high quality data and will include helpful tips on good data management. Effective data management throughout the data life cycle helps us to understand the resources required to fill data needs, realise the full value of the data collected, and ensure their long-term use.



3.1. Plan

The first stage of the life cycle is the **planning phase before data are collected**. Having a clear data management plan in place that defines the expectations of collection, using and managing the data is essential. This will be a guideline to refer to during a project to **promote consistency and integrity of the data** produced. The plan can be updated at each stage of the life cycle to reflect changes (for example in methodology used or data that were effectively collected) and data management process as they happen.

What to consider during the planning phase:

What to collect/goals:

What is it that you want to know? Think about what information you need to collect to answer your question or to fill an evidence gap. Create a brief description of the survey and its goals. Have protocols/checklists to guide you in the collecting phase to ensure the data collected are verified, replicable and understood by others to maximise their future use. Doing this at the early stages will give enough time to understand the resources needed and to make changes without losing the integrity of the data quality.

How to collect:

Tools to collect data can be mobile applications, computer programs, physical equipment (e.g. instruments or measuring devices) or pre-made identification guides (for example waterproof ID guides).

What to consider when picking for the right equipment:

- The difficulty of using a specific equipment and if training is required (either online or in person)
- The cost
- Risk of damaging/losing the equipment
- Limitation of the tools (e.g. do you need GPS or Internet signal? Do you need a waterproof camera?)
- Ensure you have considered Health and Safety and have completed the required procedures

How to track progress:

It is useful to have a document to track progress to make sure nothing is left behind. This can be a simple Excel spreadsheet.

Permissions required:

Contact the relevant authority to obtain an access agreement and to receive approval to visit the site of planned data collection. When collecting data from local communities or indigenous communities, ensure that Free, Prior, and Informed Consent (FPIC) guidelines

are followed. Make sure to record and save written permissions and complete any mandatory forms prior going to a site to collect your data. This is essential in order to collect/use and publish the data.

If there are personal data, make sure to follow appropriate guidelines/permissions (if needed) in place in case of sharing of either personal information or being in photographs.

Metadata:

Metadata are the information about the data collected and are essential in providing the context which ensure comparability and repeatability (e.g. who owns the data, restrictions to the data, unique identifier for the dataset, when the data were collected). When a dataset is published a metadata record must go alongside it. To ensure high quality metadata it is important that they follow metadata standards, which are a set of rules for formatting the data to make them consistent across datasets.

Data archive:

During the planning phase keep in mind where the data will be finally stored and which requirements they need to have (e.g. example having a unique identifier). Assign who will be responsible of the maintenance of the data once archived in case minor changes are needed or questions are asked by third parties wanting to use the data.

Summary tips for planning:

Category	Tip
Pre survey	If needed, have a paper where you can ask authorisation from people to be in images before collecting the data.
Templates	Use the provided checklists (or similar) to help going through all data life cycle Use the provided spreadsheets template (or similar) for data collection

Additional notes:

3.2. Collect

Collecting the data is the second stage of the life cycle. It is crucial that all parties involved are aware of what information needs to be collected, how it needs to be collected and who is responsible for assigned tasks. The method decided during the planning phase should be as simple as possible to promote understanding amongst data collectors and should follow a data collection protocol to promote the quality of the data. In addition, having consistency in the collection protocols amongst different survey promote the FAIR principles as the data are interoperable.

What to consider before/during/after collection:

Survey method:

The method used depends on the survey type. For example, transect or quadrat surveys could be used to assess coral reef species composition, biodiversity surveys could explore fish community composition or monitoring the abundance of crown of thorns starfish, and fish biomass surveys could be used to track the health of fish communities. Make sure to use a standard protocol for the method, which can be obtained from existing comparable surveys, and use standard terms and definitions to promote consistency.

Ensuring validity:

Depending on the survey type, it could be useful to have a checklist to make sure all the mandatory information were collected (for example GPS coordinates).

Equipment:

To produce high quality data the equipment will need to be standardised and calibrated. Having a briefing on how to record the data (for example, ensuring collectors are clear on animal/plant taxonomy and identification) could be useful to make sure there is consistency on how different surveyors record the data. The type of equipment used will depend on the data that are meant to be collected. As common procedure, a photograph of a clapboard with information of the site is taken prior collecting that site.

Here a list of possible tools:

- Paper recording materials e.g. notebooks, pre-designed worksheets.
- Digital recording materials e.g. mobile phones, apps, computer programs such as Microsoft Excel, cameras.
- Measuring tools e.g. measuring tapes, GPS devices, quadrats, transect lines, diving equipment.
- Specimen collection tools e.g. nets and traps.
- Guides e.g. identification guidebooks or identification software e.g. WoRMS

Summary tips for collecting:

Category	Tip
Pre collection	Have a training session before collection to make sure everyone will record in a consistent way and using the same “language" when recording information and design a standardised data collection sheet. Ensure at the pre-collection stage that everyone involved is familiar with the FPIC approaches required as appropriate.
During collection	Where using paper data collection sheets, take photographs of the sheets to ensure the data are not accidentally lost. When taking photographs/videos to collect data at a site, use a “clapperboard” with the site details (e.g. location, date of visit, survey ID) and take a photograph of the clapperboard before collecting the images so it is clear which site the images refer to. Have waterproof handbook for reference in the field could be useful.
During the whole survey	Assign a data manager during the survey to be responsible for successful implementation of the data management plan, coordination of the activities during survey and ensuring validity of the data and their metadata.

Additional notes:

3.3. Process

After data collection, the data should be verified, standardised, quality assured and validated to ensure high data quality, promoting the re-use of data in line with the FAIR principles.

What can compromise data quality?

- **Input values**: Human errors arise during data collection and processing, and it is easy to accidentally import the wrong values into a data spreadsheet.
- **Inconsistency**: This could arise from inconsistency of information between different spreadsheets used in the survey. For example, the location of data collection may be called by a local community name in one spreadsheet and called a different name in another, or date formats could be different between spreadsheets.
- **Lack of context**: This could arise from accidentally missing mandatory metadata fields.
- **Duplication**: This could arise by having duplications of data records.

How to improve data quality?

To minimize potential errors that can arise at any time during the life cycle there are 2 processes that should be in place: quality control (QC) & quality assurance (QA).

Quality control (QC)

Quality control is a series of pre-defined checks to ensure the data are accurate and reliable. Therefore, it is to physically check that the data are as they should be. Having a quality control process can help identify potential issues like missing mandatory information/data, errors made during data transfer, duplicated records, corrupted files, or wrong GPS coordinates. Quality control is a part of the quality assurance process which is discussed below.

There are three important components of quality control:

1. **Metadata**

   Ensuring detailed metadata accompany the data record/dataset is an important step in promoting high quality data. See the Metadata section for more detail.

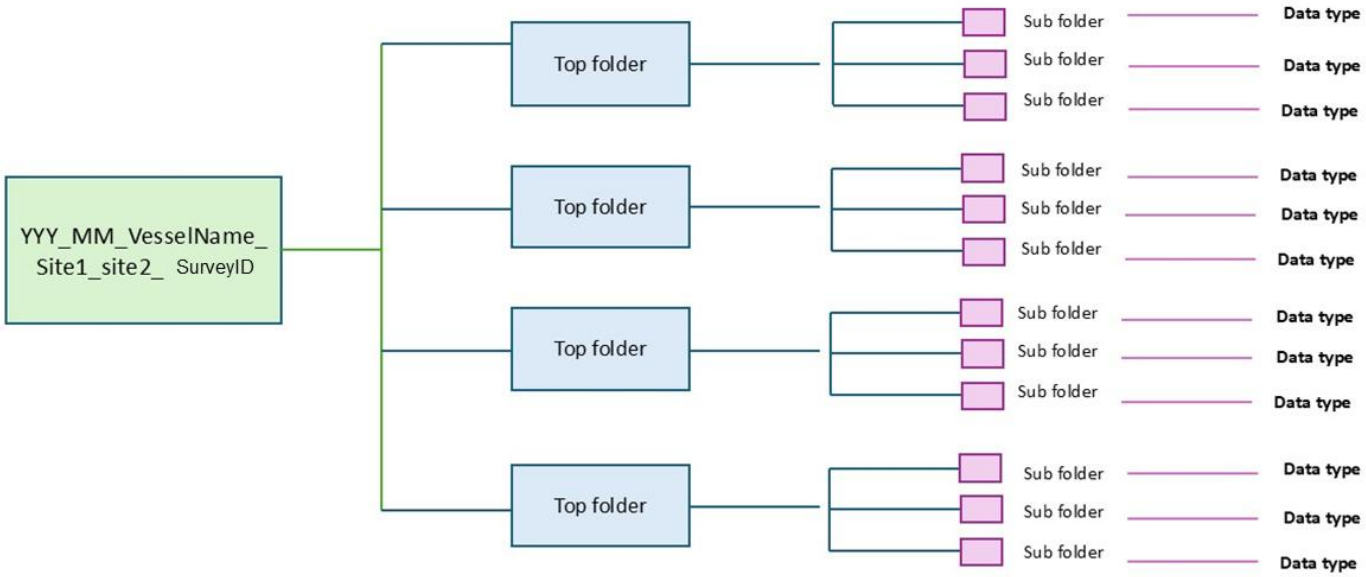
2. **Data standard**

   Data standards are a set of rules for formatting data, including the data structure and terminology to use, to apply to ensure consistency across datasets. Applying a data standard is extremely important to ensure the data can be re-used by others in future, without users needing to contact the data provider for clarifications. Using a data standard reduces the risk of misinterpretation, saves time and resources, maximising efficiency of data collection. Data standards should be set prior to

collection, which will help to identify in advance what data fields are mandatory (e.g. survey location, survey data, sampling tool) based on the type of survey.

Data standards may include:

- **A standardised folder structure to store data.**
  Make sure to have in place a documented folder structure to help organise where the data are stored within your organisation network. By following a standard folder structure, it will be easier to find data. See an example folder structure for a marine survey below.



- **A file name convention.**
  Making sure to set name conventions for files, spreadsheets, reports, metadata spreadsheets etc. makes it easy to understand what information each file contains and if any updates/changes were made by having different versions (referred to as “version control”). A version control can be a simple ReadMe file within the folder that describes the changes made.

- **Standard data collection forms.**
  It is best practice to have template data collection forms, such as spreadsheets, ready for use. Using standard forms ensures consistent mandatory fields are applied across forms to make sure all required information is entered. It is recommended to include a “ReadMe” tab in the form that explains the content of the spreadsheet and how to use.

- **Standard vocabulary.**
  It is important to clarify appropriate terminology and use this consistently. For example, confirm a set of site locations so that each collector/user is referring to a site using the same name. This helps to avoid confusion and

ensure data collectors/processors/users understand key concepts relating to data.

3. **Audit log**

   It is easy to lose track of issues found during quality control processes. Having an audit log could be useful to track any issues and their progress towards being resolved.

Quality Assurance (QA)

Quality assurance (QA) is about ensuring that the right **processes** are in place so that the data come out "correct" first time, it will include training, data structure standards, use of common vocabularies, collection standards. Quality assurance defines quality control procedures, for example quality assurance processes may include the list of quality control checks that need to be done on the data to ensure their quality. The main areas addressed during QA are spatial accuracy, taxonomic accuracy, methodology accuracy and temporal accuracy.

When quality assuring both the data and their associated metadata, a variety of tools could be used, such as Microsoft Excel formulae, taxon matching tools or AphiaID (e.g. WoRMS), coordinate converters, and geographic mapping software (e.g. QGIS or Google Earth) or metadata guidelines.

Summary tips for processing:

Category	Tip
Pre collection	Make a checklist at early stage to ensure you have all needed by following the one provided for the all life cycle.
How to display the content of a spreadsheet	Include “ReadMe” information in data collection and processing templates (e.g. in the first tab of and Excel spreadsheet) to explain the content and terminology used. Follow a protocol that defines which quality control and assurance checks need to be done on the data to ensure consistency in processing.

Additional notes:

3.4. Store

After processing, the data will need to be stored securely. This is important to ensure the data and metadata are findable and accessible to preserve their longevity, aligning with the FAIR principles. A data archive centre to store data digitally can be cloud-based or server-based. Utilising a clear folder structure is advised to ensure data are findable.

Data can be stored either digitally or physically:

Digital data:

Examples of digital data include spreadsheets, digital reports, databases, digital images or GIS layers. They can be uploaded into the archive centre in their original format but ensure the associated **discovery metadata** is uploaded alongside for accessibility (e.g. key words can be used to find the data).

Physical data:

Physical data can include paper survey sheets, record cards, printed/developed photographs, and notebooks. Physical data will **need to be digitised** as soon as possible to avoid loss/damage. However, since their importance in being in some cases the most persistent records, after being digitalised they should be securely archived.

Data can be stored for long or short term:

Short term:

This could involve backing up the data within the organisation's network while data processing is carried out. Data could be stored on hard drives or USB devices, as paper records, or in an online repository, such as SharePoint or Google Drive. Storing data on physical devices can be risky, as there is a higher chance of losing or damaging the devices, and the storage capacity is often limited. Many agencies actively discourage use of this form of data storage. With modern data types such as imagery, acoustics and genetic data, files can be very large, so cloud storage provides more flexibility with storage capacity.

Long-term:

Storage should be done via an archive centre (for example the pacific environment data portal) which benefit from effective data management, reduces the risk of losing physical copies of data and have the capability to archive large datasets consisting of multiple data types. This preserves the data which can be accessed by a wide range of future users. The short-term storage can be a starting point to store data before they are being processed and then having a long-term storage is the ultimate goal.

Having a retention policy is important to have to establish storage periods, for example:

- How long should the data be stored as short-term storage before migrating them into the long one and these rules should be regularly reviewed.
- To establish when the data should be move from being actively used to not being actively used something that it is refer as “Hot” (actively used data and so a swift access to the data is allowed) vs “Cold” (not actively used data and so it takes longer to retrieve the data) copies of data.

A data archive centre is designed to be able to publish and share the data by following a standardised format, including:

- Discovery metadata that describe the dataset (by having a quick abstract as well) and that helps find them by using keywords.
- Content metadata to give a depth description of the data. It could be helpful to have the report to go alongside the dataset when uploading them into the archive centre. This should be available under an open file format preferable both machine and human readable and should contain information as definitions of terms, data structure, origin of the data, methodology used and links to additional data products, such as the raw data used for analysis (this can be achieved by using the same identifier).
- A way to download both data and metadata under a readable format preferable for both machine and humans.
- Information about licence/ownership/data restrictions/data limitations (including FPIC considerations).
- An identifier which is unique for the survey. This is vital for finding the dataset and make it easily accessible. It can be versioned to reflect chances/updates in the dataset.
- Who to contact for further information (contact point)
- Who is responsible for keeping maintenance of the data in the archive centre (custodian)
- Who has collected the data (be mindful of joint data collection: 2 or more organisations collecting together the data which will need a copyright statement) (originator)
- Who is responsible to distribute the data (distributor)
- Publication date and version information in case of updates of the uploaded data
- Who published the data (published by)
- Publication of data products in searchable and index under one archive centre

Additional notes:

3.5. Share

Prioritising effective data sharing ensures data are accessible to a wider audience and are readily re-used, in line with the FAIR principles, leading to data-driven policymaking. Sharing data is the ultimate goal of the FAIR principles. To ensure effective data sharing, make sure data are findable, accessible, and interoperable. For example, data could be uploaded to a platform where they can be easily found through a key word search and downloaded (alongside their metadata and associated report) under an open data licence. For data that are suitable for public access, publishing the data can be done through an accredited archive centre (for example the pacific environment data portal).

Sharing data publicly is good data practice and highly supported within the scientific community. However, data cannot always be shared fully openly. Some data cannot be shared because they might have sensitive information, for example an image showing the location of a sensitive species or anthropogenic presence (e.g. shipwreck). It is also not advisable to share data that have not yet been through quality assurance processes, to avoid producing incorrect or misleading information publicly. To help maximise data sharing there are some tools that can be used, for example blurring the resolution of spatial data if sensitive (to give a large buffer area of a sensitive species/development) or by manually removing sensitive records from data and metadata.

In these cases, the advised method is to go through a specific and auditable data request process, where reasons for data access are investigated and recorded, and datasets are assessed for their ability to be shared on a case-by-case basis. In specific instances where two organisations commonly share a specific sensitive or restricted data type, data sharing agreements may be established to provide a more efficient mechanism under which data can be passed between organisations without individual data assessments.

Additional notes:

3.6. Re-use

Following the FAIR principles by ensuring data are findable, accessible, and interoperable will help to ensure the data are reusable. Other users could re-use data for purposes such as research, conservation, monitoring and to inform policy. Promoting the use of keywords/unique identifier for both the dataset and associated metadata, and clarifying the relevant licence/copyright information, improves the reusability of data.

Benefits of reusing data include:

- Maximising efficiency and increasing the Value for Money of the original data collection
- Preventing duplication of effort and thus reducing costs
- Improving decision making due to a greater availability of data
- Clarifying where data gaps might exist.
- Boosting collaborations within the scientific community
- Promoting advancement in methodologies
- Increased visibility of your work
- Promoting the principle “collect once, use many times”.

Additional notes:

3.7. Lessons learnt

Thinking about lessons learnt is the last stage of the data life cycle. It is dedicated to reflecting on how the life cycle was managed and what can be improved. This is essential to highlight areas that need more attention next time to ensure data management good practice and resolve any issues encountered.

Key points to consider include:

- What went wrong?
- What worked well?
- How could we improve?
- What do we need to be aware of next time? Are there any systematic issues (i.e something that keep coming up)

Additional notes:

4. Resources

- A1: (Meta)data are retrievable by their identifier using a standardised communication protocol - GO FAIR (go-fair.org)
- Atlan What is Metadata
- Barclay K., Mangubhai S., Leduc B., Donato-Hunt C., Makhoul N., Kinch J. and Kalsuak J. (eds). 2021. Pacific handbook for gender equity and social inclusion in coastal fisheries and aquaculture. Second edition. Noumea, New Caledonia: Pacific Community. 202 pp
- BGSSurveyDataManagementtHandbook_V11.pdf
- Citizen Science | Marine Conservation Society (mcsuk.org)
- Coordinate systems, map projections, and transformations—ArcGIS Pro | Documentation
- DASSH - Citizen Science Best Practice
- Data Ownership in Government (HTML) - GOV.UK
- Dataversity Types of Metadata
- DCC Guidance Metadata Standards
- Euro-BioImaging's Template for Research Data Management Plans (zenodo.org)
- FAO. 2016. Free prior and informed consent: An indigenous peoples' right and a good practice for local communities.
- FutureLearn Metadata and the FAIR Principles
- Go Fair: FAIR Principles
- JNCC Report No. 640: United Kingdom Terrestrial Evidence Partnership of Partnerships data products: improving opportunities for re-use
- Marine Environmental Data and Information Network | Working together to improve access to and stewardship of marine data (medin.org.uk)
- MEDIN guidelines | Marine Environmental Data and Information Network
- Operational Guidance on Free, Prior and Informed Consent, Accountability Framework 2019
- Projection basics for GIS professionals—ArcMap | Documentation (arcgis.com)
- Seasearch - Record
- Setting up new surveys and contributing data | JNCC - Adviser to Government on Nature Conservation
- Twenty spreadsheet principles | ICAEW
- WoRMS - World Register of Marine Species
- Women4Biodiversity: FPIC

Table of embedded links

Reference	Page	Link URL
ISO 19115	9	https://www.ncei.noaa.gov/sites/default/files/2020-04/ISO%2019115-2%20Workbook_Part%20II%20Extentions%20for%20imagery%20and%20Gridded%20Data.pdf
Dublin Core	9	https://www.dublincore.org/resources/metadata-basics/
EML (Ecological Metadata Language)	9	https://eml.ecoinformatics.org/
INSPIRE Directive	9	https://knowledge-base.inspire.ec.europa.eu/index_en
MEDIN	10	https://medin.org.uk/
Marine data exchange website	15	https://www.marinedataexchange.co.uk/content/info/what-is-the-marine-data-exchange
Example of discovery metadata, MEDIN portal entry	15	https://portal.medin.org.uk/portal/start.php?details=&tpc=015_d41ef53f7a540e0f78d93375bbc0aa70&step=_00002021%2C+JNCC%2C+Offshore+Wind+Evidence+and+Change+Programme%2C+Offshore+Wind+Environmental+Evidence+Regist
Example of content data, JNCC image catalogue	17	https://jncc.resourcespace.com/pages/search.php?search=!collection9018&order_by=&sort=&archive=&daylimit=&k=&restypes=Global%2C1%2C3
Creative Commons No rights reserved licence, licence information	20	https://creativecommons.org/publicdomain/zero/1.0/
Creative Commons licence with attribution, licence information	21	https://creativecommons.org/licenses/by/3.0/
Creative Commons with attribution non-commercial, licence information	21	https://creativecommons.org/licenses/by-nc/3.0/
Vanuatu Environmental Data Portal	21	https://vanuatu-data.sprep.org/
Public Data Licence Agreement under SPREP, licence information	21	https://view.officeapps.live.com/op/view.aspx?src=https%3A%2F%2Fpacific-data.sprep.org%2Fsystem%2Ffiles%2FPublic%2520Data%2520License%2520Agreement%2520%2528Creative%2520Commons%2529.docx&wdOrigin=BROWSELINK
WoRMS	31	What%20to%20consider%20before/during/after%20collection:
Pacific environment data portal	38, 41	https://pacific-data.sprep.org/index.php/

5. Data Life Cycle Checklists

PLAN

Survey information	
Site name/Survey ID	
Dates	
Initials of who did the survey /survey lead	

☐ What are our objectives? What information do we need to collect?

Comment:

☐ How are we going to collect this information? What method should we apply?

Comment:

☐ Is there a suitable data standard to apply? Or can we create a data standard to aid consistent collection (e.g. data collection template with mandatory fields)?

Comment:

☐ Identify tools required

Comment:

☐ Assess the costs associated and confirm suitability of the location

Comment:

☐ Do we have land access permission in place? If not, complete permissions/ authorisation process

Comment:

☐ Obtain Free, Prior, and Informed Consent (FPIC) from communities, and ensure consensus from participants for publication/social media (as needed)

Comment:

☐ Establish protocols for performing quality control and assurance

Comment:

☐ Data storage locations are defined, and data flows established

Comment:

☐ Outline any conditions for data sharing, access and re-use

Comment:

☐ Assign a team member to be responsible for the data in case of required corrections during the project and once the data are published.

Comment:

Additional notes:

COLLECT

Survey information	
Site name/Survey ID	
Dates	
Initials of who did the survey /survey lead	

☐ Confirm and outlined the data collection method with all involved, and provide any required training

Comment:

☐ Standardise/calibrate any equipment as needed

Comment:

☐ Test the equipment prior to starting collection to detect potential issues

Comment:

☐ Log any issues during data collection

Comment:

☐ Complete the data collection and metadata templates, ensuring digital copies are taken regularly, using the data standard as planned

Comment:

Additional notes:

PROCESS

Survey information	
Site name/Survey ID	
Dates	
Initials of who did the survey /survey lead	

☐ Quality Assurance (QA) procedures complete

Comment:

☐ Perform Quality Control (QC) procedures

Comment:

☐ Confirm data collection and metadata templates are complete according to the data standard agreed

Comment:

☐ Ensure consistent terminology applied

Comment:

☐ Ensure a consistent folder structure and file name convention is applied

Comment:

☐ Complete the audit log to track issues and resolutions

Comment:

Additional notes:

STORE

Survey information	
Site name/Survey ID	
Dates	
Initials of who did the survey /survey lead	

☐ Confirm storage location and retention period

Comment:

☐ Digitise any physical data

Comment:

☐ Ensure a consistent and clear folder structure is in place and ensure that versions of the files alongside top copy are clearly labelled and that older versions are archived/ deleted as appropriate

Comment:

☐ Confirm established data flow and data are findable for internal used (e.g. catalogue keyword search)

Comment:

Additional notes:

SHARE & RE-USE

Survey information	
Site name/Survey ID	
Dates	
Initials of who did the survey /survey lead	

☐ Confirm what data can be shared, or if there are any sensitive/restricted data

Comment:

☐ Ensure a consistent and accessible format is applied

Comment:

☐ Confirm relevant licence and outline user restrictions as needed

Comment:

☐ Establish data sharing process (e.g. data download, data request form)

Comment:

☐ Confirm established data flow and data are findable for external used (e.g. catalogue keyword search)

Comment:

Additional notes:

LESSONS LEARNT

Survey information	
Site name/Survey ID	
Dates	
Initials of who did the survey /survey lead	

Reflecting questions:

Questions	Answers
What worked well?	
What could be improved?	
Were there any issues encountered to be aware of next time?	

UK International Development

Partnership | Progress | Prosperity