CSIRO

Australia's National Science Agency

BLUE PAPER

Towards a fairer future: making COR imagery data FAIR

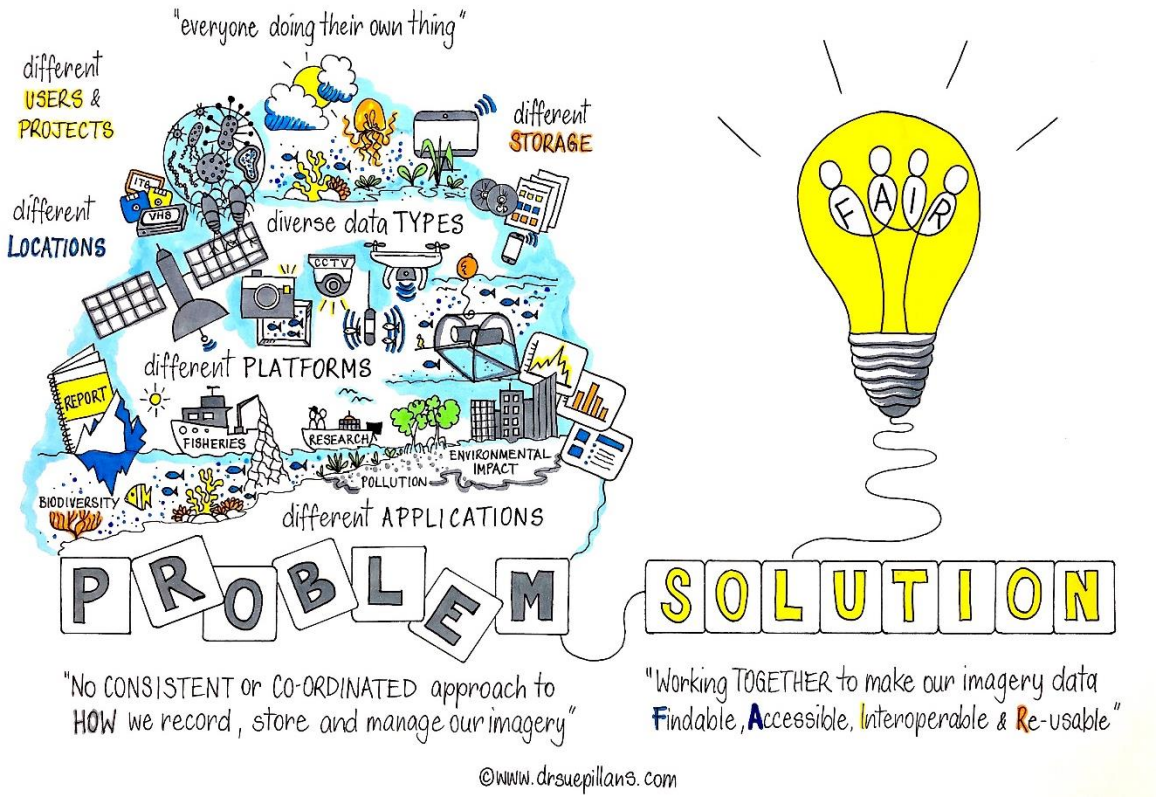
Ben Scoulding, Toni Cannard, Franzis Althaus, Candice Untiedt, Kylie Maguire, Mike Fuller

INTRODUCTION

Technological advances in cameras, sensors, and electronic control systems over the last decade has enabled new applications of photographic image data to marine research questions. The possibilities have been embraced enthusiastically in **Coasts and Ocean Research (COR)**, with cameras being deployed on many specialised platforms (e.g., tow cam, BRUVS, AOS, PLAOS, AUV, ROV, longline systems), added to others (e.g., coring platforms, pots and traps), and used as a key element of e-monitoring of catch on commercial fishing vessels. Imagery is also captured to record desktop specimens and for media purposes. As a result, COR are presently acquiring large volumes of image data of diverse types. In some cases, the rate of data acquisition has outstripped the ability to archive, retrieve and effectively use the data already collected. It is envisioned that this problem will worsen unless the necessary action is taken to address existing and future data needs.

PURPOSE

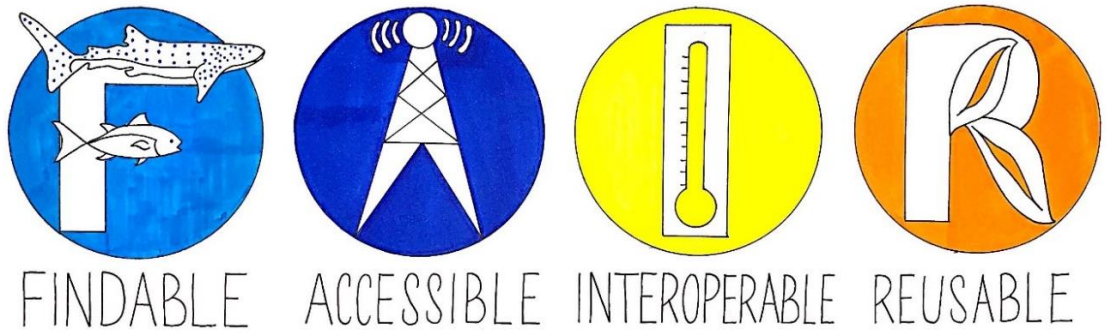
This blue paper is intended to introduce the **FAIR principles** (Findable, Accessible, Interoperable, Reusable), discuss the implications, benefits, and challenges to being FAIR, and present a **strategic roadmap** for making COR imagery data FAIR and **the ask** - what we need to make this happen.



FAIR PRINCIPLES

The **FAIR principles** (with their 15 underlying facets) are a set of standards intended to make data **F**indable, **A**ccessible, **I**nteroperable and **R**eusable. They are designed to facilitate knowledge and data sharing by humans and machines, support data and knowledge integration, promote standardising processes and outputs, and encourage sharing and reuse of data. They help to make data and metadata 'machine readable', thereby enhancing/supporting new discoveries through the harvest and analysis of multiple datasets and outputs.

FAIR is a set of nationally and internationally recognised principles which are endorsed by many governmental and academic organisations worldwide. Since being published in 2016 the FAIR principles have provided an important framework for documenting and sharing data and outputs in a way that maximises use and reuse (www.ardc.edu.au). The FAIR principles refer to three types of entities: (1) **data** (or any digital object), (2) **metadata** (information about that digital object), and (3) **infrastructure**. The following text is taken from https://www.go-fair.org/fair-principles/.



FINDABLE

The first step in (re)using data is to find them. Metadata and data should be easy to find for both humans and computers. Machine-readable metadata are essential for automatic discovery of datasets and services, so this is an essential component of the **FAIRification** process.

Facets:

F1. (Meta)data are assigned a **globally unique** and **persistent identifier**
F2. Data are described with **rich metadata** (defined by R1 below)
F3. Metadata clearly and explicitly include the identifier of the data they describe
F4. (Meta)data are registered or indexed in a searchable resource

ACCESSIBLE

Once the user finds the required data, they need to know how they can be accessed, possibly including authentication and authorisation.

Facets:

A1. (Meta)data are **retrievable** by their identifier using a standardised communication protocol
- A1.1 The protocol is **open**, **free**, and **universally implementable**
- A1.2 The protocol allows for an **authentication** and **authorisation procedure**, where necessary

A2. Metadata are **accessible**, even when the data are no longer available or if the data are confidential

INTEROPERABLE

The data usually need to be **integrated with other data**. In addition, the data need to interoperate with applications and/or workflows for analysis, storage, and processing.

Facets:

I1. (Meta)data use a **formal**, **accessible**, **shared**, and **broadly applicable language** for knowledge representation
I2. (Meta)data use **vocabularies** that follow FAIR principles
I3. (Meta)data include **qualified references** to other (meta)data

REUSABLE

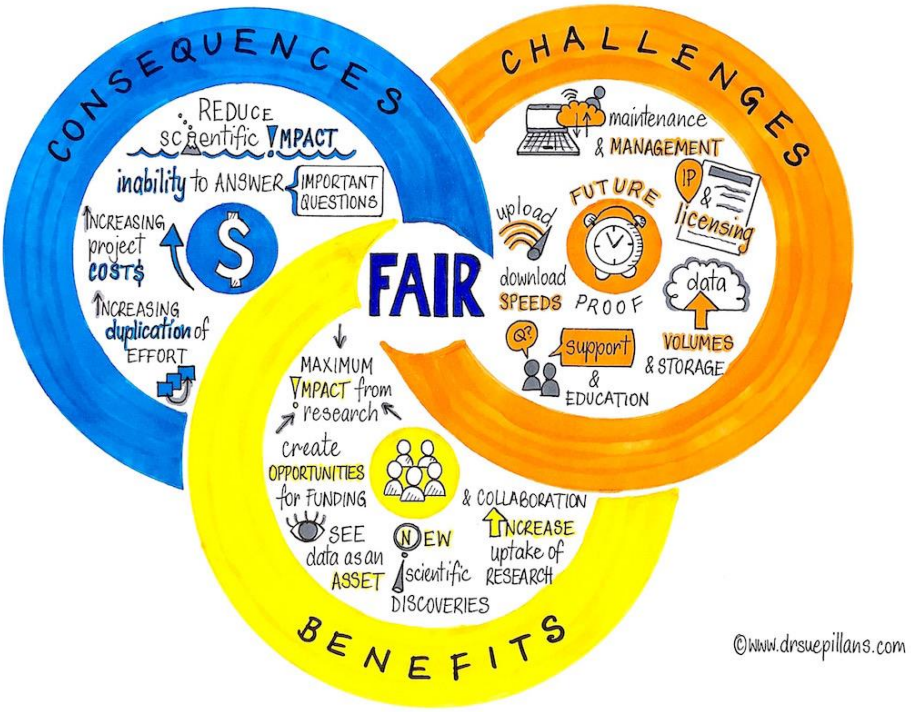
The goal of FAIR is to **optimise** the reuse of data. To achieve this, metadata and data should be **well-described** so that they can be replicated and/or combined in different settings.

Facets:

R1. (Meta)data are richly described with a plurality of accurate and relevant attributes.
R1.1. (Meta)data are released with a **clear** and **accessible data usage license**
R1.2. (Meta)data are associated with **detailed provenance**
R1.3. (Meta)data meet **domain-relevant** community standards

IMPLICATIONS AND CHALLENGES

Whilst making data FAIR has obvious benefits, translating the FAIR principles into practice is challenging. The resources (people/time/money) needed are not trivial. So why invest the time and effort in to being FAIR?



BENEFITS

Making research data FAIR provides a range of benefits to researchers, research communities, research infrastructure facilities and research organisations, including (www.ardc.edu.au):

- achieving maximum impact from research
- appreciating and recognising the value of the data – allows us to see data as an important asset
- enabling new scientific discoveries
- increasing uptake of the research
- creating a standardized set of tools for data storage, retrieval, and reporting
- creating opportunities for funding and collaboration
- improving efficiencies
- increases the visibility and citations of research
- creating recognizable formats thereby making data more meaningful and usable
- staying aligned with international standards and approaches
- adherence to increasing publication requirements to provide FAIR data

CONSEQUENCES

If data are not FAIR there can be consequences to researchers, research communities, research infrastructure facilities and research organisations, including:

- reducing scientific impact
- increasing project costs
- increasing duplication of effort
- failing to meet international standards and approaches
- staying relevant while future proofing
- inability to answer important research questions
- missed opportunities/collaborations

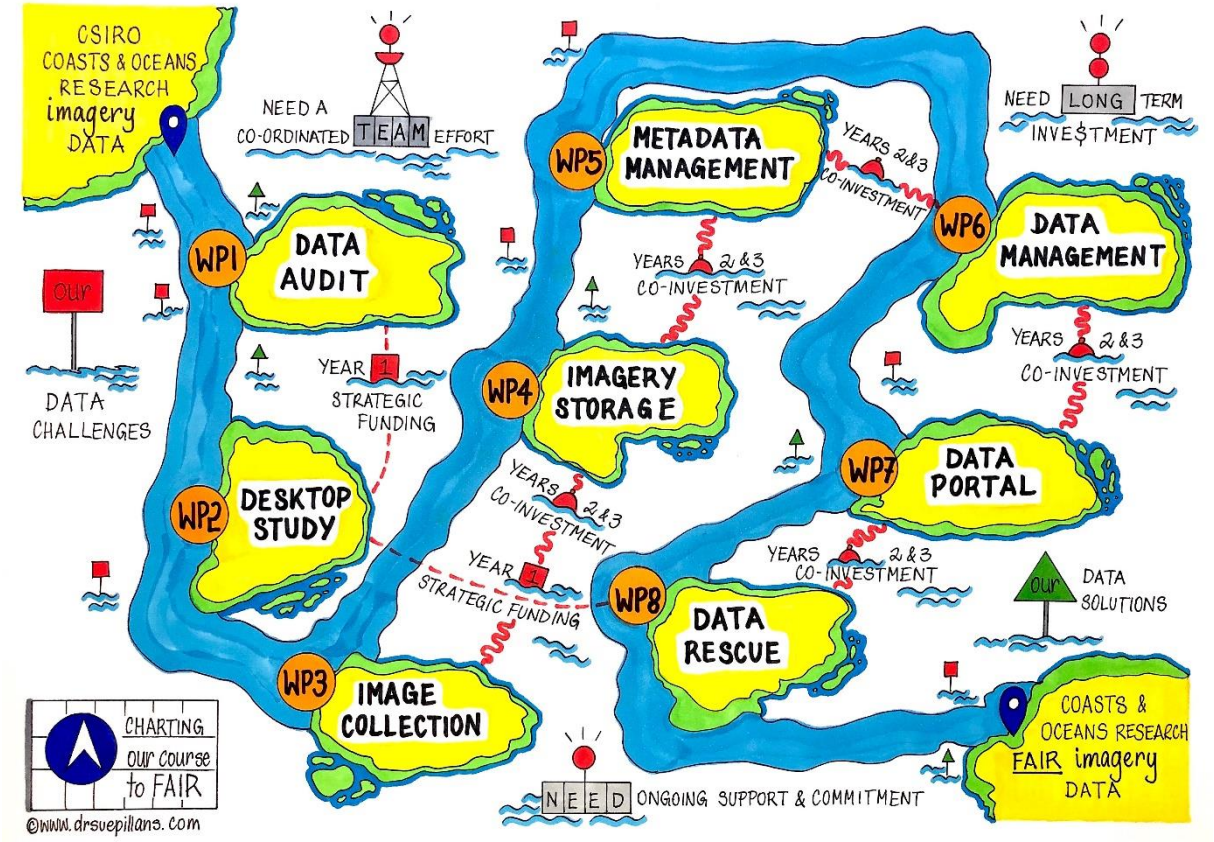
CHALLENGES

The greatest challenge in transitioning data to FAIR are the availability of (or lack of) resources and infrastructure that allow for systematic change. Agreed standards and vocabularies must occur at every stage of the (meta)data pipeline, including acquisition, post-processing, annotation, analysis, storage, and publication processes. Moreover, these 'procedures and protocols' should also align with global standards. Naturally, adoption of FAIR data principals will take time and require solutions at incremental scales. Some challenges include:

- data volumes and data storage
- databases
- unique identifiers
- reliability of ancillary data
- licensing and IP
- upload/download speeds
- standardisation (e.g., metadata, vocabularies, data collection, and storage)
- maintenance and management
- support and education
- staying up to date, e.g., with changing technologies- future proofing
- encouraging individuals/teams/projects to abide by the FAIR principles

STRATEGIC ROADMAP

Addressing the challenges described above will require a coordinated team approach comprising of diverse CSIRO people (e.g., from O&A, NCMI, IM&T) and will need appropriate strategic investment. Here we present the beginnings of a strategic roadmap for transitioning COR imagery data to FAIR. The roadmap is broken up to manageable work packages (WP) with clearly defined outcomes. A platform (e.g., TEAMS or a Community of Practice (CoP)) will be established to assist staff to tackle each WP.



WP1: Data audit

Objective: Carry out a comprehensive audit of COR imagery data (data types and formats, storage locations, existing workflows, documentation, naming conventions, custodians).
Approach: Develop a targeted and detailed questionnaire to send to data managers regarding imagery data collected as part of past and present projects. Extract existing metadata descriptions.
Outcome: A complete list of COR imagery data assets and all associated data and metadata (including identification of existing workflows) to feed into **WP5** & **WP6**.

WP2: Desktop study

Objective: Identify how other disciplines and other institutions have approached the challenges to being FAIR (existing standards, vocabularies, ontologies).
Approach: Desktop study and literature review searching for imagery and image data collections and relevant documentation and vocabularies used in metadata records.
Outcome: A report of existing standards and approaches (including available resources) to feed into **WP5** & **WP6**.

WP3: Data collection
Objective: Develop standard operating procedures for collecting and uploading imagery in COR.
Approach: In collaboration with those responsible for collecting imagery data (e.g., Field Instrumentation (NCMI)), develop workflows, standards for unique identifiers, metadata (**WP5)** and data fields (**WP6**) as imagery is uploaded to the CSRO network (**WP3**).
Outcome: An agreed workflow for pre-processing imagery data for uploading to the CSIRO network, ensuring *unique identifiers* and basic *metadata* descriptions are captured at the image collection stage.

WP4: Data storage
Objective: Develop a data storage infrastructure for COR imagery data and imaging platforms – consider access/retrieval (internal and external), security and maintenance.
Approach: In collaboration with IM&T identify long-term solutions for storage and dynamic retrieval of images with centralised access control possibilities.
Outcome: A robust storage infrastructure capable of handling future COR needs with an appropriate support network.

WP5: Metadata management
Objective: Develop standards for describing and documenting imagery data collected in COR.
Approach: In collaboration with the Information & Data Centre (NCMI) develop easy to use standard forms for data descriptions (metadata) implementing agreed ontologies and vocabularies based on outcomes of **WP1** & **WP2**.
Outcome: An agreed standard for documenting imagery data, including easy to use metadata templates specific to describing imagery. This results in making the image collection *findable* for both humans and machines.

WP6: Data Management
Objective: Develop a solution for making the variety of image databases more accessible.
Approach: Based on information gathered in **WP1**, **WP2** & **WP5** identify common/essential data fields in existing image databases and work towards reaching compatibility in the formats used between databases for essential data fields.
Outcome: Extracts/views targeting common fields from different databases can be accessed, extracted, and combined contributing to the *interoperability* of image data sets to feed into **WP7**.

WP7: Data portal
Objective: Develop views for application processing infrastructure (API) style delivery of data and links to images (or thumbnail views) to internal (+ external) portals and data consumers (e.g. AODN)
Approach: Building on **WP5** & **WP6** and collaborating with the Information & Data Centre (NCMI) and administrators of web-portals to create views from the metadata and databases that can be harvested by the portals.
Outcome: Internal and if the data type allows external *findability* and *accessibility* of imagery and image data through web services.

WP8: Data rescue
Objective: Digitisation of imagery that is stored in non-digital formats.
Approach: Digitise and document non-digital data assets identified in **WP1** that have valuable metadata. Sufficient metadata is defined as complying with having associated data to populate standard metadata forms (**WP5)** and data fields (**WP6**).
Outcome: Making historical image collections *FAIR*, allowing for re-use and re-analyses of historical imagery for potential time-series analyses. Safeguards against loss of image assets due to out-dated technology and deterioration of the storage media.

The ask

To deliver the benefits of the FAIR framework we need **significant long-term investment** in **allocation** and **operating**. This will require an **ongoing commitment** from CSIRO leadership. To succeed we will need **co-investment** and **support** from IM&T and NCMI. We propose a two staged approach to funding:

Year 1: Strategic funding to tackle **WP1**, **2** and **8**
Year 2 and 3: Co-investment across by O&A, NCMI and IM&T to tackle **WP3**, **4**, **5**, **6** and **7**

Once year 1 is funded we will explore mechanisms to fund year 2 and 3 including leveraging of existing infrastructure. We will identify opportunities for broader conversations with O&A, NCMI and IM&T and the creation of a data management working group. We will identify how this project aligns with O&A digital strategy and how the project fits within IM&T and NCMI. Together we will identify a collaborative pathway to impact. We propose to use COR imagery data as a test case for finding a solution to a shared data management problem.

This Blue Paper is the result of the **FAIR for Imagery (the other four-letter F word)** workshop convened on the 14th April 2021.

2021 © Dr Sue Pillans www.drsuepillans.com

All graphics were produced by Dr Sue Pillans (www.drsuepillans.com)

As Australia's national science agency and innovation catalyst, CSIRO is solving the greatest challenges through innovative science and technology.
CSIRO. Unlocking a better future for everyone.

For further information
CSIRO Oceans & Atmosphere
Ben Scoulding
ben.scoulding@csiro.au
csiro.au/O&A