BLUE ECONOMY
COOPERATIVE RESEARCH CENTRE

FINAL REPORT

DATA INFRASTRUCTURE DESIGN FOR THE BLUE ECONOMY

JUNE 2025

Australian Government
Department of Industry, Science and Resources

Cooperative Research Centres Program

Citation

Bruce, L, Walsh, P, Busch, B, Hart, C., Ghauri, A. et. al (2025). Data Infrastructure Design for the Blue Economy, 4.22.001 – Final Project Report. Blue Economy Cooperative Research Centre.

The Blue Economy CRC

The Blue Economy Cooperative Research Centre (CRC) is established and supported under the Australian Government's CRC Program, grant number CRC-20180101. The CRC Program supports industry-led collaborations between industry, researchers and the community. Further information about the CRC Program is available at http://www.business.gov.au.

Disclaimer

Copyright Notice

Table of Contents

Executive Summary 5

1. Introduction 8

1.1. Rationale 9

1.2. Objectives 9

2. Methodology 10

2.1. Terminology 10

2.2. Design Team and Stakeholders 11

2.3. Workflow summary 12

2.4. Review of Data Infrastructure 14

2.5. Review of CRC Data Requirements 15

2.6. Workshops 16

2.7. Review of data sharing and analytics platforms 17

2.8. Document findings and recommendations 17

3. Results 17

3.1. Design Context 17

3.2. Open Data Review 19

3.2.1. Commonly used data platforms 19

3.2.2. Survey insights 22

3.3. CRC Data Requirements 24

3.3.1. Online survey results 24

3.4. Data workflow 26

3.5. Functional Requirements 29

3.6. Non-functional Requirements 31

3.7. Review of data sharing and analytics platforms 32

3.7.1. BMT Deep: Data Exploration and Analytics Platform for Marine and Environmental Applications 32

3.7.2. Advanced Technology Integration 33

3.7.3. Key Features and Advantages 38

3.7.4. Australian Agricultural Data Exchange (AADX) 41

4. Discussion 42

4.1. Data Infrastructure Review 42

4.2. CRC Research Project Data Needs 42

4.2.1. Data Types and Collection Methods 43

4.2.2. Data Management and Storage 43

4.2.3. Data Analytics and Processing: 43

4.2.4. Data Sharing and Accessibility 43

4.2.5. Challenges and Solutions 43

4.2.6. Desired Features for Data Infrastructure 43

4.2.7. Future Vision and Aspirations 44

4.3. Analysis of Requirements 43

4.4. Blue Economy Data Needs 45

4.5. Characterising Data Infrastructure 45

4.6. Federated, Secure Integration and Identity Management 47

5. Conclusions and Recommendations 48

5.1. Data lifecycle for data generated by CRC research projects 49

5.2. CRC Data Inventory 50

5.3. CRC Standards Based Catalogue 51

5.4. Publish CRC data on open source platforms 51

5.5. Rules of CRC data exchange 51

5.6. CRC Interoperable Infrastructure Prototype for a Blue Economy Case Study 52

5.6. Conclusion 53

6. Acknowledgements 53

7. References 53

CONTENTS

List of Figures

Figure 2-1. Design workflow approach 12

Figure 2. Frequency of use of online data platforms 22

Figure 3. Type of data available in identified online platforms 22

Figure 3-4. Classification of data utilised in BE CRC projects 24

Figure 3-5. Variety of data formats utilised across BE CRC projects 25

Figure 3-6. Storage mediums, storage volumes and data retention periods for BE CRC projects 25

Figure 3-7. Data workflow 27

Figure 3-8. SEAF multi-zonal architecture 36

Figure 3-9. Overview of Dataflow Architecture within BMT Deep 39

Figure 3-10. Example features in BMT Deep 40

Figure 11: AADX powered by Eratos 41

Figure 12. Data infrastructure characterisation 46

Figure 13: Data Space components 47

Figure 14. Concept example showing CRC data infrastructure design for federated, integrated platforms using an identity service, secure protocols and other ISDA standards to connect environmental approval, fisheries and wind farm proponent data 48

Figure 15. An example data journey decision tree for CRC data 50

Figure 16. Federated Identity Management for Open and Secure Industry data 52

List of Tables

Table 1. Terminology used in the data infrastructure design 10

Table 2. Methods in the design process using the DADR process 13

Table 3. Platforms Identified in the distributed existing data portals & repositories survey and workshop 19

Table 4. List of Functional Requirements for CRC Data Infrastructure 29

Table 5. List of Functional Requirements for CRC Data Infrastructure 31

Table 6: Policies, roles and users explained 51

Executive Summary

Effective data and knowledge sharing between research programs and users is essential for the Blue Economy CRC (CRC) to remain successful and achieve its full potential. The "Data Infrastructure Design for the Blue Economy (DIDBE)" project brought together a team of digital experts, data scientists, strategists, engineers, and engagement specialists to design a framework to manage CRC knowledge sharing and data.

The main aim of the project was to develop a scalable data infrastructure architecture to support data sharing for the CRC and marine community more broadly. The data infrastructure architecture needed to be capable of capturing, storing, analysing and visualising data generated by CRC Research Projects as well as enable access to data sourced from outside the CRC including but not limited to research, industry and government. The architecture needed to provide users with **simple, intuitive and easy to use interfaces to find and retrieve relevant data to "underpin the growth of the Blue Economy".** Additionally, the data infrastructure architecture had to be robust, scalable and sustainable to last the duration of the CRC and beyond.

A series of surveys, interviews, meetings and workshops with project partners, stakeholders, and end-users were held with the design team to co-design a fit-for-purpose data infrastructure architecture. This approach ensured suitable functionality and relevance to end-users. Central to the philosophy behind the design was to reuse existing infrastructure wherever possible. To facilitate this, the DIDBE project involved two initial review studies. The first study reviewed existing open data platforms developed for marine, maritime and other blue economy data in Australia and internationally. This study concluded that existing data platforms were fit for the purpose of long-term storage and access of blue economy data generated by the CRC. The second study reviewed the data management plans and requirements for CRC research projects. This study found that there was a diversity of data types, standards for collection and curation, storage methods, analysis techniques and tools, and data management practices across the research projects.

The study found common challenges and future aspirations for data usage employed across various research programs. It concluded there would need to be flexibility in the data infrastructure design to respond to specific project data needs.

Two main challenges were identified during the design process:

1. **How to design a single infrastructure to support the range of requirements across a broad range of research?**
2. **How to support an open data philosophy while enabling data providers to retain data ownership and security to protect commercial intellectual property and culturally sensitive information?**

Meeting these challenges required a governance framework capable of demonstrating both flexibility, adaptability, and security. A further review of existing data infrastructure guidelines at three levels of governance (viz. national, international and CRC) was undertaken by the design team. Three existing data analytics platforms currently used to support Blue Economy industries were reviewed in detail as case studies to gain understanding and insights regarding how they could be leveraged in the design of the CRC data infrastructure.

The DIDBE project found that the best way to maximise value and ensure longevity of CRC data was to adopt an internationally recognised standards-based approach to data governance that enables interoperability between existing and future data analytics platforms. This was preferred over investing time and funding to develop and support the maintenance of a bespoke CRC data and analytics platform. Alongside open data standards proliferated by organisations like the International Oceanographic Data and Information Exchange (IODE) and the Australian Ocean Data Network (AODN), the most viable standards-based approach that includes secure, managed access to data has been developed by the International Data Spaces Association (IDSA) and is gaining significant interest in the Australian research data community. These standards are presented as the preferred solution for the CRC data governance framework.

The following recommendations set out the next steps in developing a robust data governance framework and service provision for the CRC:

1. Include mandatory requirements in the CRC Data Management Policy (DMP) for Research Programs to ensure standards for data collection, storage, sharing identity and where IP appropriate, and an ultimate resting place within a list of approved open data platforms.
2. Conduct an inventory of data generated from all past and current CRC Research Projects.
3. Contribute CRC generated data to a standards-based data catalogue that aggregates to the Australian Ocean Data Network (AODN).
4. Commit past and current CRC data to an approved data platform.
5. Establish the rules of CRC data exchange through interoperable data platforms to ensure security, trust and collaboration using a standard based approach.
6. Develop a prototype data analytics platform for a blue economy case study with interoperability to a regulatory platform and freely available cloud-based research computing service. e.g. NeCTAR.
7. Deliver a Capacity Building Program for the use of blue economy data and analytics tools targeted at Blue Economy growth and awareness, including a set of on-line training modules and in-person training workshops.

The DIDBE project has laid the groundwork for a scalable, flexible, and robust data governance framework to support the CRC's mission of fostering sustainable growth in the Blue Economy.

By leveraging existing data platforms and focusing on interoperability through standards based services, the recommended approach will achieve long-term usability, security, and adaptability to diverse data needs. When coupled with appropriate policies and guidelines, comprehensive, searchable data inventories, and targeted capacity-building programs, CRC participants will be able to effectively manage and utilise marine and maritime data, driving evidence-based decision-making and maximising the value of CRC research for the Blue Economy's future.

The solution proposed was designed to respond to changes in the CRC requirements as the CRC matures and adapts to changes in the marine data landscape, and as it progresses to greater digitisation of data sharing in a trusted and secure collaborative space.

Acronym	Meaning
AADX	Australian Agricultural Data Exchange
ADCP	Acoustic Doppler Current Profiler
AIMS	Australian Institute of Marine Science
AODC-JF	Australian Ocean Data Centre Joint Facility
AODN	Australian Ocean Data Network
API	Application Programming Interface
ARDC	Australian Research Data Commons
BE	Blue Economy
BOM	Bureau of Meteorology
CRC	Cooperative Research Centre
CSIRO	Commonwealth Scientific and Industrial Research Organisation
DCCEEW	Department of Climate Change, Energy, the Environment and Water
DIDBE	Data Infrastructure Design for the Blue Economy
DMP	CRC Data Management Policy
ESCC	Earth Systems and Climate Change
ELT	Extract Load Transfer
ETL	Extract Transfer Load
GA	Geoscience Australia
GBRMPA	Great Barrier Reef Marine Park Authority
GIS	Geographic Information System
GOOS	Global Ocean Observing System
IaC	Information as Code
IDS	Intrusion Detection System
IDSA	International Data Spaces Association
IMAS	Institute of Marine and Antarctic Studies
IMOS	Integrated Marine Observing System
IOC	Intergovernmental Oceanographic Commission
IOOS	Integrated Ocean Observing System
LLM	Large Language Model
NCRIS	National Collaborative Research Infrastructure Strategy
NERP	National Environmental Research Program
NESP	National Environmental Science Program
NOAA	National Oceanic and Atmospheric Administration
OBIS	Ocean Biodiversity Information System
OCED	Organisation for Economic Co-operation and Development
ORCID	Open Researcher and Contributor ID
PNNL	Pacific Northwest National Laboratory
QCIF	Queensland Cyber Infrastructure Foundation
RWSC	Regional Wildlife Science Collaborative
SAC	Scientific Advisory Committee
SPC	Pacific Community
SPREP	Secretariat of the Pacific Regional Environment Programme
UNESCO	United Nations Educational, Scientific and Cultural Organisation
DOE	U.S. Department of Energy
UTas	University of Tasmania
UWA	The University of Western Australia

1. Introduction

The Blue Economy Cooperative Research Centre (CRC) was established to undertake industry focused research and training to support the growth of the Blue Economy in Australia and New Zealand. The CRC brings together 43 industry, government, and research partners from ten countries with expertise in aquaculture, marine renewable energy, maritime engineering, environmental assessments and policy and regulation.

Over a period of 10 years the CRC seeks to undertake research projects targeted towards developing innovative, commercially viable and sustainable offshore developments and new capabilities. Effective knowledge sharing and collaboration between research programs and across the various project, data, and technology outcomes is essential for the CRC to achieve enduring success.

The Data Infrastructure Design for the Blue Economy (DIDBE) Project was proposed to undertake a comprehensive review of the data needs of the CRC partners and consult with various stakeholders to come up with a set of design requirements and recommendations to build data infrastructure for both the CRC and the community it represents beyond the life of the CRC.

The design team comprised a group of digital experts, data scientists, strategists and engineers, and engagement specialists. The challenge for the design team was to come up with a framework for data management and collaboration to effectively address the complex and often contrasting relationship between the main CRC stakeholders within the research, industry, and government sectors.

The CRC's focus on two main growth areas in the blue economy, offshore aquaculture and renewable energy production is centred on promoting the principles of sustainable development prioritising respect for Traditional Owners' rights and connections to Sea Country, protection of natural marine ecosystems and the preservation of existing industries such as commercial fishing. By centering the design requirements on the CRC's sustainable growth goals when developing the data governance framework, the design can be viewed as a catalyst for interdisciplinary collaboration, innovation, and responsible stewardship of our oceans.

1.1. Rationale

Knowledge and data sharing through effective data capture, storage, analytics and visualisation underpins each of the CRC's target outputs and milestones and is central to the vision of sustainable growth through collaboration and innovation.

Without a scalable, standards based and interoperable data architecture, there is risk of incompatibility and duplication of effort related to data hosting costs, software, server maintenance and end-user support. The adoption of a data architecture that supports the FAIR data principles (Findable, Accessible, Interoperable, Reusable) using a standards based approach ensures a reliable and trusted data supply chain and integration with other data across social, economic and environmental domains.

Data (often disparate in nature) is essential for informed decision making in investment and planning, environmentally sustainable development, and economically viable blue economy operations. It is critical that data is readily available through a standards based and domain relevant data framework and compliant infrastructure. It must be scalable and sustainable and address the needs of a diverse range of end users including government, industry, research and the general community. Moreover, it must be easy to use and targeted to meet specific requirements so all users can access information and realise benefit from investment in the CRC and beyond.

The CRC Data Management Policy requires that every CRC project has a designated data manager responsible for ensuring that new data collated and/or generated through the CRC is properly archived. In addition, CRC research projects have or are proposing to identify gaps in knowledge data requirements for advancing investment, compliance, operations and regulatory standards in the blue economy. Through the Scientific Advisory Committee (SAC), the CRC enables the process of supporting new research and development projects aimed at generating data to fill these gaps. As the diversity, size and volume of CRC data grows, there is an immediate need to identify current and future data requirements for the CRC and ability to tap into the fast growing national and international marine data networks. This understanding will drive further research and development to bridge the gaps between stakeholder requirements, available data sources and derived data products.

1.2. Objectives

To ensure the legacy of data generated by the CRC into the future and maximise the potential for the proposed data architecture to continue serving the blue economy, this project aimed to identify existing data resources and only recommend new infrastructure where a clear need is identified.

The purpose of the DIDBE project was to design a data infrastructure to support access to data and to support evidence-based decision making for the Blue Economy CRC that:

1. **Is capable of capturing, storing, analysing, and visualising data generated by CRC Research Projects and the blue economy sector more generally beyond the life of the CRC.**
2. **Connects through APIs to a federation of data sourced from research, industry and government.**
3. **Provides simple, intuitive and easy to use access portals for CRC partners to retrieve and use relevant data and data analytic tools to support the Blue Economy.**
4. **Is robust, scalable and sustainable to last the duration of the CRC and beyond.**

Core to the design process was engagement with the CRC Executive Board, Research Program Leads, industry partners, project teams and end users throughout the project to validate their requirements and ensure that the design would be fit-for-purpose.

2. Methodology

A strategic methodology was necessary to achieve a fit-for-purpose data architecture to meet the needs of participants and ease the data management burden of the CRC.

The approach adopted by the design team involved engaging with research partners and stakeholders to ensure the blue economy industry can meet future economic and environmentally sustainable goals and can invest with confidence in targeted research to support innovation, technology and operational efficiency measures.

2.1. Terminology

Following initial workshops at the commencement of the project it became evident that a clear set of terminology definitions around data infrastructure design were needed to guide the process. While the project set out to design data infrastructure to include a proposed architecture solution, the project became much more in scope to include the process of deriving a CRC data management framework. Core definitions established and agreed on by the project team included terms for data framework, infrastructure and architecture (Table 1).

Table 1. Terminology used in the data infrastructure design.

Term	Description	Context	Examples
Framework	A conceptual structure that provides a set of guidelines, standards, and tools for designing, building, and managing data systems.	To establish best practice for data collection, storage, processing, sharing and archiving, aligned with the Blue Economy's vision for sustainable offshore industry. Based on the FAIR principles (Findable, Accessible, Interoperable, and Reusable) and compliance with relevant standards (e.g., ISO, Open Data standards).	Establish a data workflow specifying the steps from data collection or extraction and storage through analytics and visualisation with standards for metadata including authorisation, shared between organisations.
Infrastructure	The physical and digital resources required to support the data lifecycle, from acquisition to analysis and access.	Includes the **hardware, software,** and **networking** components that enable data storage, analytics, processing, and access.	**Hardware:** High-performance servers, cloud storage solutions, and devices for ocean data collection (e.g., sensors on buoys). **Software:** Data management platforms, numerical modelling and analytics tools, and APIs for accessing marine datasets from outside the CRC. **Networking:** High-speed data pipelines connecting offshore research sites, cloud systems and project participants.

Architecture	The high-level design and organisation of the infrastructure components, specifying how they interact to achieve the system's objectives	Outlines the blueprint for data flows, identity networks, system integration, and interoperability.	**Data Architecture:** Describes how marine datasets (e.g., satellite imagery, sensor data) are ingested, transformed, stored, and queried. **System Architecture:** Specifies how different components (e.g., open access data and secure analytics platforms) are integrated. **Security Architecture:** Details measures to protect sensitive marine data, including identity protocols, licensing, encryption, access controls, and compliance with privacy regulations.

2.2. Design Team and Stakeholders

The Core Design Team (CDT) was selected based on their knowledge and practical experience in in data infrastructure design, management and strategy, stakeholder engagement, surveys and workshops, and project management skills.

The CDT comprised five Project Participant organisations:

BMT — **Lead Organisation**

Griffith University

OceanPixel — fluid . energy . intelligence

University of Tasmania

The University of Western Australia

A leadership structure was established to ensure collaboration and communication between Project Participants with a Project Manager (PM) and Technical Lead (TL) assigned to each of the main project tasks and deliverables.

Navigating the complexities of designing data infrastructure for the CRC was compounded by the need to respond to a diverse set of stakeholders, each with specific needs, experiences and expertise. Balancing often divergent user requirements and managing stakeholder expectations was a challenge throughout the design process.

The following stakeholder groups were foreseen as potential users of the CRC Data Infrastructure and as such included in the considerations for design.

- CRC: Board, Scientific Advisory Committee, Data Manager, Project Participants (research, government and industry partners)
- Additional academic researchers (e.g. postdoctoral researchers, PhD students; anyone working in marine ecology/ conservation/restoration/engineering/ aquaculture)
- Additional blue economy industry members (e.g. new proponents and consultants in the aquaculture or offshore renewable space)
- Government (Local, State and Federal, e.g. for management and monitoring, permits, incident response)
- General public (e.g. for education, communication)

An Advisory Committee was set up to provide regular feedback to the DIDBE project team to ensure the design outcomes were fit for purpose and made best use of existing data infrastructure, availability and governance structure in the marine data ecosystem:

- BE CRC Research Director, or the BE CRC Manager, Research and Partnerships.
- Research Program Leader or representative from Program 4 - Environment and Ecosystems.
- CSIRO representative
- IMOS representative
- Geoscience Australia representative

A Stakeholder Reference Group (SRG) was established comprising a wider team of research partners, stakeholders and end-users. The SRG was formed to participate in workshops to contribute the interests of the wider research community and industry representatives to the design team. Engagement with this wider team was facilitated by experienced engagement specialists to ensure we maximised input and maintained productive and engaged participation throughout the project.

2.3. Workflow summary

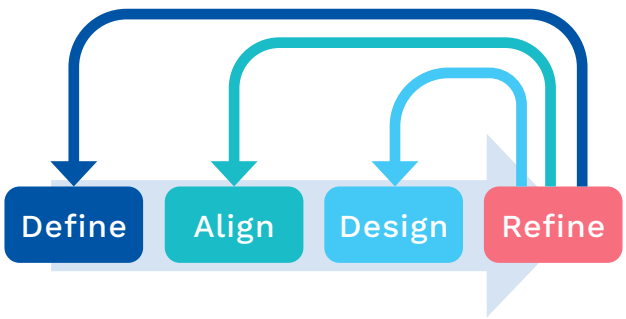
The original project plan was broken into five distinct Work Packages (WP1-WP5), each with a defined outcome and set of deliverables to be undertaken in chronological order:

- **WP1** Undertake a systems and information architecture study for marine and maritime data based on existing open data platforms relevant to the Blue Economy.
- **WP2** Undertake an audit, review and study of the past, current and future data generated by the CRC through the Research Projects and data management requirements of these data.
- **WP3** Conduct a series of workshops involving CRC stakeholders to conceptualise the design requirements and provide feedback on the design process.
- **WP4** Document the CRC data infrastructure framework including Basis of Design, Functional Specifications and Concept Design, design demonstration mock-up, infographics, and architecture blueprints.
- **WP5** Final report (summarising findings of WP1 4) to incorporate the outcomes of the workshops and final project deliverable.

As the project progressed it became clear that a more flexible approach to project management was required as the design took on an iterative process based on the principles of Define, Align, Design, Refine (Figure 2-1) generally applied to software design.

Rather than complete each Work Package before proceeding to the next, the design workshops continued throughout the project and shaped the review of existing data infrastructure and CRC data needs to aid the design process.

Figure 2-1. Design workflow approach.



Different methods were deployed to aid in the design process including surveys, workshops, presentations and feedback meetings. A summary of methods used related to the DADR process and mapped to the five WPs is presented in Table 2, preliminary work done in the refinement of the project proposal are annotated as WP0.

Table 2. Methods in the design process using the DADR process.

Process	Meetings	Surveys	Workshops	Presentations
Define	WP0 - Met with representatives from CRC Management Team to define the project aims and objectives.	WP0 – 2022 CRC Participants Workshop Poll. WP2 - Surveys of CRC project leads and data managers to determine the CRC data needs, data types, storage options, collaboration and CRC data lifecycle and use.		
Align	WP1-5 – Advisory Group Meetings	WP1 – Review of existing marine data portals to canvas the user experience in data access and ease of use.	WP0 – Workshop at the 2022 CRC Participants Workshop	WP0 – Presentation at the 2022 CRC Participants Workshop
Design			WP3 - Design workshops to determine the data infrastructure to include in the data federation and operability through APIs. WP3 - Stakeholder workshops to brainstorm the preliminary list of design requirements and classify as non-negotiable, what to avoid and nice to have. WP4 – Preparation of a UX mock-up for visualisation for stakeholder feedback. WP3 – Workshop generation of User Stories to aid in design communication.	
Refine	WP4 - Presentation of functional and non-functional requirements to the CRC Data Infrastructure Advisory Committee for feedback and further refinement of design.		WP3 - Design workshops to refine the list of functional and non-functional design requirements. WP3 - Demonstration of recommended infrastructure at participant workshop. WP5 – Writing workshops for documenting the project findings into a Final Report.	Participant Workshop

2.4. Review of Data Infrastructure

A review of current, best practise open data access portals relevant to the CRC with a focus on marine and maritime data was undertaken.

The aim of the review was to ensure that the data infrastructure design would meet the needs of multiple users by:

- Identifying design specifications and components of an Open-Data Sharing Ecosystem from which to build a data federation for the CRC;
- Identifying barriers to success and lessons learnt from similar open data models; and
- Identifying measured of success for robust, scalable and sustainable data infrastructure design.

These objectives were met by conducting a survey of existing online data platforms and repositories distributed survey, seeking feedback in design workshops and reviewing case studies for refined design parameters.

A survey was constructed by the CDT with the aim to collate information regarding existing data portals utilised by CRC stakeholders. This survey was the distributed to all CRC Participants and attendees at workshops from outside the CRC. Once participants completed the survey, the survey was closed, and information collected was reviewed and findings were recorded.

Project meetings at CRC Participants Workshops were open to all participants as a chance to get additional survey participation to compile a list of open data portals identified as useful to the CRC for data access. This list of portals was used as a basis to create a database to incorporate information regarding API accessibility, access protocol, standard operating procedures, metadata standards, access standards, scale, origin, function, users, theme, data type and partners.

Results and findings from both the survey and workshop discussions were combined to compile a list of challenges, considerations and must haves applicable to the proposed design of the CRC data infrastructure.

2.5. Review of CRC Data Requirements

An inventory audit of existing CRC project data and review of data needs was undertaken to understand the current and anticipate the future data requirements of the research projects undertaken by the CRC.

The review comprised the following tasks:

1. Identification of project leads and stakeholders

- Project leads and stakeholders relevant to the CRC were identified and engaged in the inventory audit process
- A collaborative approach was adopted to ensure representation from diverse research programs and industry partners.

2. Survey distribution:

- A survey form was developed by the design team to collect information relevant to the metadata associated with projects within the CRC
- The survey was distributed to project leads across various research programs within the CRC
- Upon completion of the survey period, collected data was reviewed, analysed, and synthesised to identify commonalities and patterns in the survey.

3. One-on-one interviews:

- Project leads who volunteered to share detailed information regarding the data being used, analysed, or generated in their projects were invited for one-on-one interviews
- Interviews gave deeper insights into specific data requirements, challenges, and opportunities within each project
- Information gathered from the interviews was collated and cross-referenced with survey findings.

4. Stakeholder Collaboration:

- Close collaboration between the core design team, project leads, and stakeholders was maintained
- Stakeholder inputs and feedback were integrated to ensure alignment with the overall objectives of the DIDBE project.

5. Data analysis and synthesis:

- Data collected from surveys and interviews was systematically analysed to identify trends, patterns, and gaps in existing data infrastructure and requirements
- Findings were synthesised to develop recommendations for enhancing data management practices and optimising the data infrastructure framework.

6. Documentation and reporting:

- A technical report summarising the methodology, key findings, and recommendations of the CRC data needs review was prepared for incorporation into the DIDBE project Final Report.

2.6. Workshops

A series of workshops were conducted throughout the design process with the intent to enable a co-design approach to ensure long-term usability and relevance to users including to:

- Inform CRC participants of the DIDBE project, aims and objectives;
- Gather feedback and information from industry, researchers and end-users to help inform and support the design process;
- Select and refine the CRC data infrastructure scope and requirements;
- Interact with web interface design mock-ups to create interface design features; and
- Generate User Stories to socialise and test the design.

A series of ten design workshops were conducted with a broad range of stakeholders to determine data needs, identify opportunities and barriers to data sharing and establish an infrastructure framework. Each workshop was scheduled to accommodate multiple time zones and repeated so that those who missed one could attend the other. Workshops were delivered as a hybrid of face-to-face and on-line to ensure inclusivity for all potential attendees from industry and research partner organisations. Workshops iterated through invited attendance between the CDT, stakeholder reference groups and potential end-users, Advisory Group and report authors for the writing workshops.

The workshops followed an iterative approach, with the aim of progressively refining a design solution to meet existing and future data needs (covering both CRC-generated and external data sources) and overcome a myriad of data challenges faced.

Each workshop had a set of objectives, preparatory material sent through to the participants prior to the workshop, and outcomes and deliverables related to the WP4 System Design Outcomes.

The outcomes from each workshop will be documented in a technical report. When collated together, these reports were then used to document the design and development journey for the data infrastructure framework as it was progressively refined and continually influenced through the consultation and engagement process.

Deliverables from the workshops included:

- Workshop presentation slides
- Workshop design diagrams, minutes, process flow diagrams and findings
- Data product mock-up (“Design Demonstration Tool”)
- Reports on workshop outcomes for each workshop.

The aim of the final workshop was for the CDT to reach a design framework for the data infrastructure solution to meet the aims and objectives of the project.

2.7. Review of data sharing and analytics platforms

Following the initial design process including the review of existing data infrastructure and sharing portals it was concluded that the key to design would be to adopt an internationally recognised standards-based approach to data governance and infrastructure and a service to CRC participants that enables interoperability between existing and future data storage and analytics platforms. This would enable access to robust and fit-for-purpose data and analytics platforms that could be used by CRC for data management and access, and analytics and visualisation tools. An in-depth review of three data analytics platforms all developed with marine data at the core but serving vastly different purposes to meet research, industry and government needs was conducted. This aim of this review was to identify existing features that could address the design requirements of the CRC and what work would be required to address gaps to meet the CRC needs for an interoperable system.

2.8. Document findings and recommendations

The findings for the project were presented at the 2024 Participants Workshop with a series of User Stories to socialise the design for stakeholder feedback then documented in a Final Report.

3. Results

3.1. Design Context

During the preliminary stage of the project a document detailing the design context was developed to provide an overview of:

- Key design considerations and principles.
- Design requirements (functional and non-functional).
- Relevant guidelines and standards.
- The philosophy and approach for designing and implementing an effective data infrastructure specifically tailored to support CRC and the ongoing needs of the Blue Economy.

The context of data infrastructure for the design context documentation was predominantly focussed on the standards, applications and the broad technical environment required to support end-user data and information needs.

Key design considerations and principles affecting the design included:

- The data architecture must respond to the needs of both open access and sensitive data requiring access restrictions.
- Abiding by the FAIR data principles ensuring data and metadata are Findable, Accessible, Interoperable and Reusable (Wilkinson et al. 2016). Where possible, open access to data will be highly encouraged. For sensitive data requiring restricted access, metadata assisting discovery and appropriate access will be highly encouraged.
- A well-defined data governance structure.

Design requirements were divided into functional and non-functional defined as:

- **Functional** – These requirements define what a product must do and what its features and functions are. They serve as the blueprint detailing the essential characteristics and capabilities of the data infrastructure. For example, secure and scalable storage hardware.
- **Non-functional** – a description of how the infrastructure will operate including a list of rules that govern the functional requirements and performance measures to ensure success in design. These requirements speak to the quality of the infrastructure design. For example, data management policy rules and guidance.

Existing instruments identified as relevant to the CRC data infrastructure design approach included:

- The National Marine Science Plan (sections on baselines and monitoring and the Australian Ocean Data Network);
- The EPBC Act review (highlighting data needs);
- The Australian Government Public Data Policy Statement;
- The Data Availability and Transparency Act (2022);
- The Digital Economy Strategy (2021);
- Environment Information Australia (Department of Climate Change, Environment, Energy and Water; and
- The Intergovernmental Agreement on Data Sharing (2021).

Existing guidelines and standards identified included:

- The Shared Analytic Framework for the Environment (SAFE 2.0);
- Open Geospatial Consortium (OGC) Standards for data access and integration;
- ISO19115 for metadata.

Design strategy included consideration for:

- Federated architecture, allowing for a broad range of data sources from Government, universities, industry, etc.;
- Reuse of existing infrastructure where available (e.g. the Australian Ocean Data Network and other long term NCRIS facilities, Australian Bureau of Statistics, Bureau of Meteorology);
- Reuse of existing, highly relevant architecture (which may include cloning existing infrastructure used in Australia or overseas);
- Use of open-source solutions when available;
- Access to data through download, API or visualisation services;
- Access to metadata conforming to common schema (e.g. ISO19115) that include expert contacts, well described data collection methodologies and quality assurance processes;
- Access to data products that may be an amalgamation and/or summary of sourced (potentially harmonised) data to meet specific end uses;
- Secure data storage and platforms to ensure restricted access to sensitive data and limited risk associated with hardware failure or malicious attack;
- Ease of maintenance and sustainability; and
- Compliance with relevant Government laws and policies.

3.2. Open Data Review

To assess current use of existing online data platforms and repositories which contain data related to the CRC Research Projects, an anonymous online survey was developed and distributed to all participants in the CRC currently comprising 43 partner organisations located in 9 countries.

From this group of potential stakeholders 16 individuals completed the survey. The survey was conducted using the Microsoft Forms platform created on 21 February 2023. The survey consisted of 23 questions and took an average of 14 minutes to complete.

Due to the low response rate and lack of representation about the CRC participants the review of open data platforms was completed in the DIDBE workshops.

3.2.1. Commonly used data platforms

The list of commonly used data platforms generated by survey respondents and workshop participants were dominated by Australian data with some internation and global coverage (Table 3).

The most commonly used platform as identified in the survey was SeaMap Australia with other common responses including Australian Ocean Data Network (AODN) and Bureau of Meteorology (BOM). Most of the platforms were hosted by government agencies, had open accessibility and were used frequently by survey participants.

Table 3. Platforms Identified in the distributed existing data portals & repositories survey and workshop.

Digital Platform	Website	Coverage	Host	Description
Allen Coral Atlas	https://allencoralatlas.org/	Global	Arizona State University	Maps world's coral reefs and monitors threats
Aqualink	https://aqualink.org/map	Global	Aqualink	Ocean data from sensors, models, satellite observations, surveys, images, and videos for reefs.
ARMADA	https://www.cmar.csiro.au/data/armada/	Australia	CSIRO	Marine data summaries
Atlas of Living Australia	https://www.ala.org.au/	Australia	CSIRO	Australian biodiversity
AusSeabed	www.ausseabed.gov.au/	Australia	Geoscience Australia	Seabed mapping data
Australian Marine Parks (AMP) Science Atlas	https://atlas.parksaustralia.gov.au	Australia	DCCEEW	Marine and social research data for Australian Marine Parks
Australian Ocean Data Network Portal	https://portal.aodn.org.au/	Australia	IMOS	Australian marine and climate data
Australian Research Data Commons	https://ardc.edu.au/about_us/	Australia	NCRIS	Research data infrastructure facility
Blue Pacific 2050 Dashboard	https://blue-pacific-2050.pacificdata.org/oceans-and-environment	Pacific Region	Pacific Data Hub	Ocean and environment data for 2050 Strategy for the Blue Pacific
Bureau of Meteorology (BOM)	http://www.bom.gov.au/marine/	Australia	BOM	Marine and ocean data (wind, wave, swell, tides)

Digital Platform	Website	Coverage	Host	Description
Copernicus Marine Service Access Data	https://marine.copernicus.eu/	Europe/ Globa	European Union	Physical, ice and biogeochemical data for Blue Economy
Digital Coast	https://coast.noaa.gov/digitalcoast/data/home.html	USA	NOAA	Coastal data (bathymetry, climate, imagery, socioeconomics)
Digital Earth Australia	https://maps.dea.ga.gov.au	Australia	CSIRO/ Geoscience Australia	Satellite data
Earth Systems and Climate Change Hub (ESCC)	https://nespclimate.com.au/category/resources/	Australia	NESP	Data, tools and software for earth and climate change
eAtlas	https://eatlas.org.au	Australia	AIMS/NESP	Environmental research and reference data Great Barrier Reef and terrestrial tropical ecosystems
EcoCommons	https://www.ecocommons.org.au/	Australia	QCIF/ARDC	Biodiversity data and analytics tools
eReefs AIMS Visualisation Portal	https://ereefs.aims.gov.au/	Australia	AIMS	Visualisations of the eReefs Hydrodynamic and BioGeoChemical models of the Great Barrier Reef
eReefs Data Explorer	https://portal.ereefs.info/	Australia - GBR	CSIRO	Various data sources for Great Barrier Reef
EU Blue Economy Observatory	https://blue-economy-observatory.ec.europa.eu/dashboard-0_en	Europe	European Commission	Economic indicators for activities related to oceans, seas, and coasts
Geoscience Australia Portal	https://portal.ga.gov.au/	Australia	Geoscience Australia	Geoscience Australia data, other publicly available data sources and analytical and multi-criteria assessment tools
Global Archive	www.globalarchive.org	Australia	UWA/Greybits Engineering	Marine and freshwater fauna data with a focus on stereo techniques.
IMAS Data Portal	https://data.imas.utas.edu.au	Australia	IMAS	Data relating to temperate marine, Southern Ocean, and Antarctic environments
IMAS Metadata Catalogue	https://metadata.imas.utas.edu.au/geonetwork/srv/eng/catalog.search#/home	IMAS	Australia	Searchable catalogue of meta data for Australia
IMOS Ocean Current	http://oceancurrent.imos.org.au/	Australia	IMOS	Sea surface currents and temperature
Marine Scotland	http://marine.gov.scot/	Scotland	Scottish Government	Marine data and maps.
National Offshore Petroleum Information Management System (NOPIMS)	www.ga.gov.au/nopims/	Australia	Geoscience Australia	Data relating to Australian Offshore Petroleum wells

Digital Platform	Website	Coverage	Host	Description
National Map	https://nationalmap.gov.au	Australia	Geoscience Australia	Location-based data including topography and bathymetry.
NOAA New Blue Economy	https://www.noaa.gov/blue-economy	USA	NOAA	Access to data for Blue Economy (climate, fisheries, satellite, sanctuaries)
North-West Atlas	https://northwestatlas.org/	Australia	AIMS/NESP	Marine data from northwest Australia
Ocean Biodiversity Information System (OBIS)	https://obis.org	Global	UNESCO	Marine biodiversity
OCED Sustainable Ocean Economy Database	https://www.oecd.org/ocean/data/	Global	OECD	Ocean related datasets and indicators
Open Researcher and Contributor ID (ORCID)	https://orcid.org/	Global	ORCID	Researcher identification data enabling interoperability through API
Pacific Data Hub	https://pacificdata.org/	Pacific Region	Pacific Community (SPC)	Central platform for data about the Pacific region
Pacific Environment Data Portal	https://pacific-data.sprep.org/	Pacific Region	SPREP	Regional and national environmental data conditions and trends
Portal and Repository for Information on Marine Renewable Energy (PRIME)	https://openei.org/wiki/PRIMRE	USA/ Global	PNNL/DOE	Marine energy data and information (e.g. power performance data, environmental monitoring reports, device testing guidance and software code)
Reef Life Survey (RLS)	https://reeflifesurvey.com/	Global	IMAS	Underwater Visual Survey reef biodiversity
Regional Wildlife Science Collaborative for Offshore Wind (RWSC)	https://database.rwsc.org/	USA Atlantic	RWSC	Wildlife and marine ecosystems to support offshore wind power development.
Seamap Australia	https://seamapaustralia.org/	Australia	IMAS	Seafloor habitat spatial portal and state of knowledge tool.
Squidle+	www.squidle.org	Australia	IMOS/Greybits Engineering	Marine Image
State of the Environment Report Australian Government	https://www.dcceew.gov.au/science-research/soe	Australia	DCCEEW	2021 State of the Environment Report
Tethys	https://tethys.pnnl.gov/monitoring-datasets-discoverability-matrix	USA/ Global	PNNL/DOE	Documents, information, and resources about the environmental effects of marine energy and wind energy
World Bank Data Catalog	https://datacatalog.worldbank.org/home	Global	World Bank	Economic and energy data

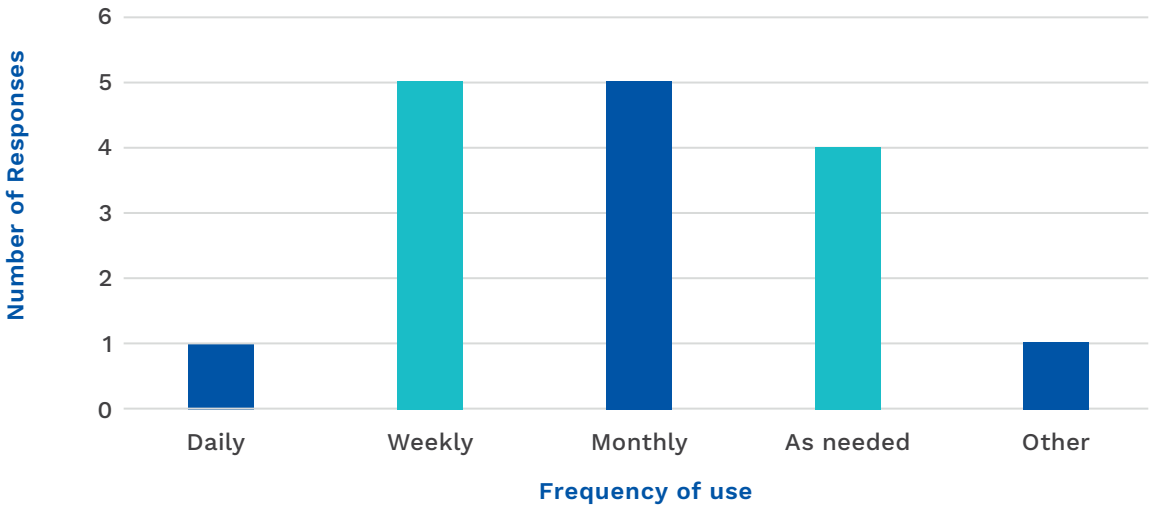
3.2.2. Survey insights

While the number of survey respondents were insufficient for any meaningful representative analysis the statistics do provide some insights that were useful in design.

Participants primarily used the data platforms for consulting or industry with only a quarter using for academic/research purposes. It is possible that academics have easier access to locally stored or collaboration data than industry or consultants acting on behalf of industry.

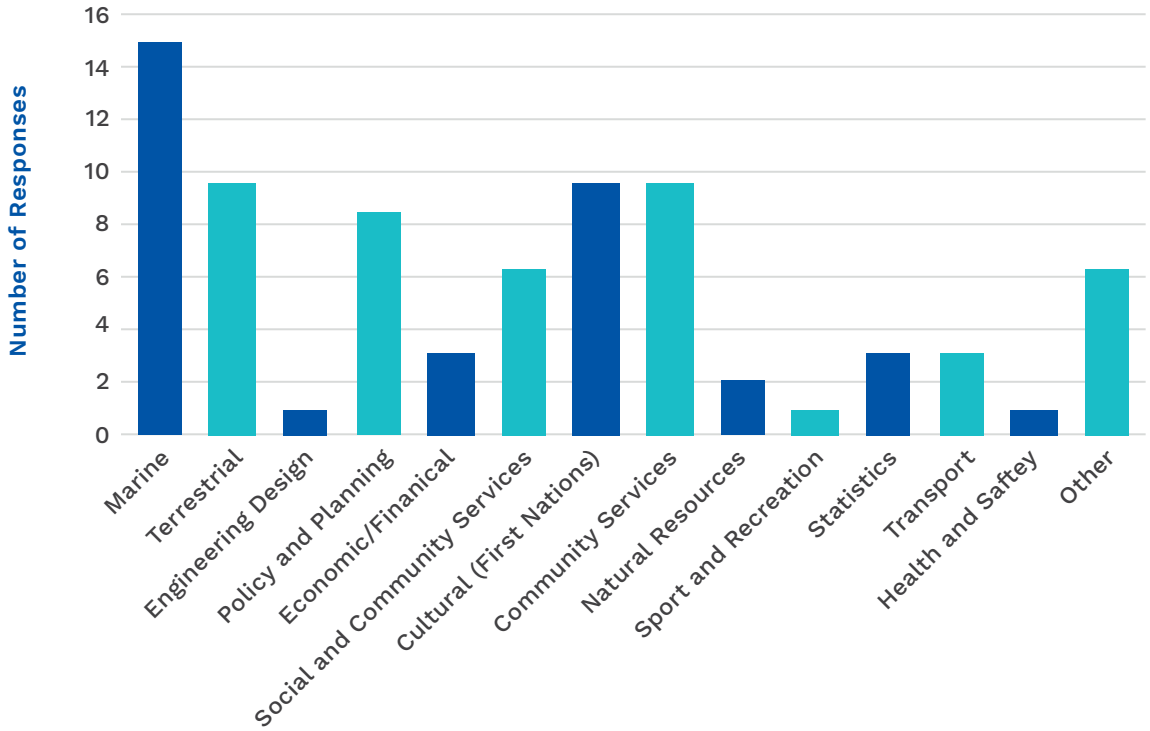
The frequency of use of data platforms was relatively evenly distributed between weekly, monthly or as needed with only one survey respondent accessing daily (Figure 2).

Figure 2. Frequency of use of online data platforms.workshop.



Survey responses to the type of data accessed from online data platforms were relatively mixed with as expected marine having the greatest number of responses and terrestrial, cultural (first nations), policy and planning and natural resources all commonly used.

Figure 3. Type of data available in identified online platforms.



61% of the online data platforms identified in the survey were used to download data only, 23% of platforms were utilised for data analysis and visualisation tools, and the remaining 15% of platforms were utilised to download and upload data. The file types which individuals downloaded were mostly CSV/Excel and spatial data. While other file types downloaded include PDFs, maps and Images/videos. Of the data that was used for data analysis and as visualisation tools maps most file types were PDFs and spatial data.

The most common positive features listed by survey participants were: easily accessible, use of map to allow filtering of region, intuitive navigation, and ability of filtering of data by type.

Other common responses included: advanced search bars, ability to upload spatial data to identify areas of interest and providing various types of spatial data to download. While only one survey responder selected platform support for integration and machine to machine protocols, this was interestingly a common feature listed by workshop participants as something they required for their research, government and/or industry applications.

Common negatives of existing data platforms identified in the survey included how data filtering is not intuitive, and how platforms are slow to perform, clunky and contain too many layers of data to navigate efficiently. Less common responses include how the platforms are difficult to access, have security which is difficult to use and the absence of a search function, this is probably because most of the platforms listed by survey participants already had these features.

Participants identified various facets that they would like to see included in the CRC data infrastructure. The most common requests included downloadable data, ability to search data and layers. While other responses include inclusion of spatial data, categorisation of data, interest for search and areas on maps.

Features which participants identified to avoid in the design of the BE CRC data infrastructure include:

- Spatial layers that are unclear or confusing layers;
- Unpublished datasets;
- Absence of information and/or metadata; and
- Use of multiple variable names.

3.3. CRC Data Requirements

3.3.1. Online survey results

Covering aspects such as project details, data collection methods, storage mediums, challenges faced, and desired features of data infrastructure, the survey provided a holistic understanding of requirements. Additionally, it explored the volume of data generated, data retention needs, and the accessibility of project data to the public. Through these questions, the survey sought to identify opportunities for optimising data management processes and informing an effective data infrastructure design.

Respondents were asked to describe the type of data they collect, generate, or analyse, and for what purpose. This question aimed to understand the diversity of data types within the ecosystem and their intended uses. The formats in which data was/is collected, method of collection/generation and the source of the collected data were queried. This information provided the basis of assessment for the current practices related to data management and identifying the gaps for future improvements.

Figure 3-4 illustrates the classification of data utilised across BE CRC projects, based on the responses received. Figure 3-5 showcases the variety of data formats utilised across BE CRC projects.

Figure 3-4. Classification of data utilised in BE CRC projects.

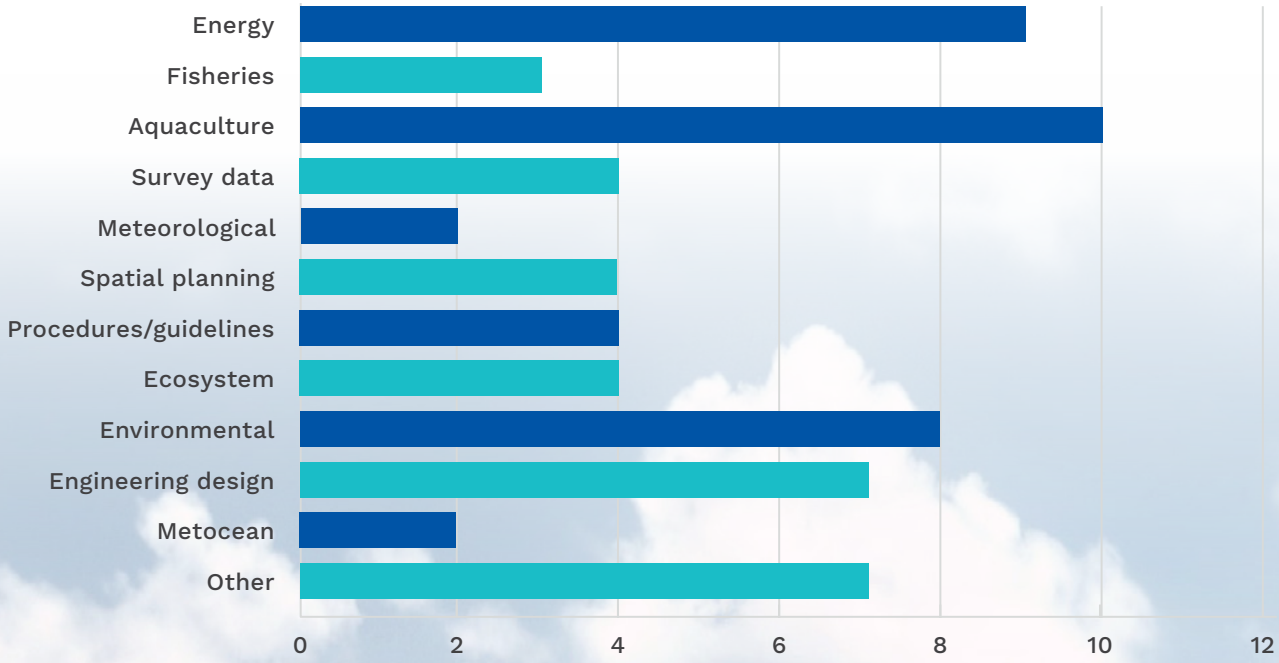
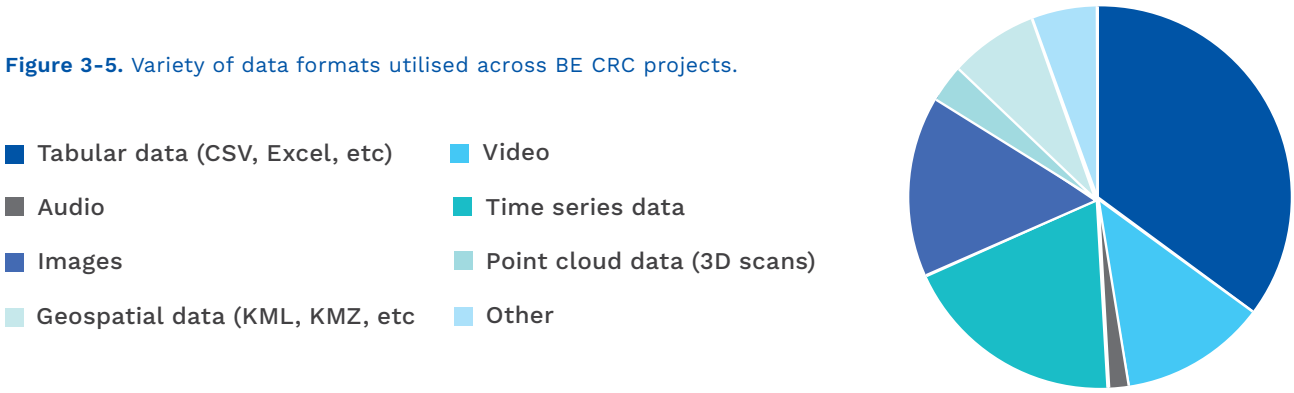


Figure 3-5. Variety of data formats utilised across BE CRC projects.



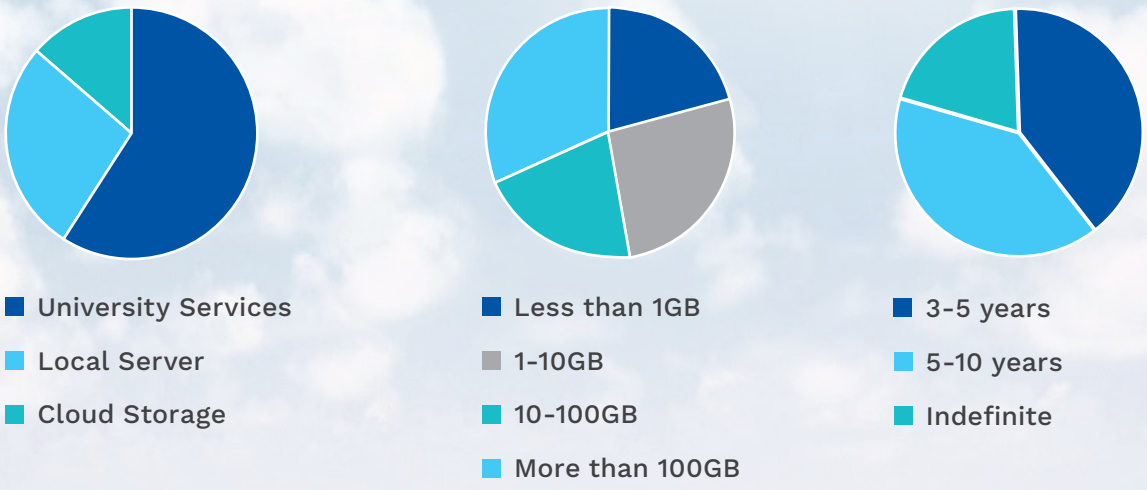
The survey identified the data storage mediums, processing techniques, and digital tools employed by respondents. This effort to gather insights into the storage solutions and processing methods will assist in understanding the future infrastructure's strengths and limitations. Additionally, understanding the volume of data generated or collected, along with the anticipated duration of data storage, yields valuable insights into the software and hardware requisites of the platform.

Based on the survey responses, University/Organisation servers emerged as the preferred storage medium, followed by local storage with 15% using Cloud Storage (Figure3-6).

In terms of data volume, most projects reported generating or handling less than 100 gigabytes (GB) of data. Five projects fell within the 1–10 GB range, while four projects reported data volumes between 10–100 GB. Four respondents generated/collected less than 1 GB of data. Six respondents dealt with data exceeding 100GB (Figure 3-6).

The survey found that a data retention period of 3–5 years was the most popular choice, with eight respondents selecting this option. This was closely followed by a retention period of 5–10 years (seven respondents), and only three respondents expressed a need for indefinite data storage. While it is difficult to tell from the responses given the reason behind the retention period, it is envisioned that most data collected by CRC research projects will be transferred to an open database such as AODN within the course of the project, hence the 3-5 year period. Industry has their own archiving procedure also where data greater than a certain period is placed in archive and no longer required for retention by project participants.

Figure 3-6. Storage mediums, storage volumes and data retention periods for BE CRC projects.



A variety of data processing techniques were identified across different research projects, and were dominated by analytics, insights and visualisation tools. The software used by respondents for these processing techniques were dominated by Microsoft Excel, MATLAB and R Studio. Other Software identified by respondents included ArcGIS, CAD, Python, ANSYS and Power BI.

The data sharing and accessibility survey questions centred on the transparency of project data, to understand how much of the data is made publicly available, as well as the sharing methods employed. The responses demonstrated a mix of data accessibility and sharing within and outside of the CRC, as well as the sharing capabilities expected from the data infrastructure platform. Surprisingly only a few projects identified data that would be shared under open data licence with the bulk of responses indicating data sharing would be restricted to project participants and partners only.

Researchers were asked to write in detail the challenges and limitations they face while accessing and managing data. The responses from the survey revealed significant challenges and considerations for data management across various research projects, highlighting issues related to ethics, accessibility, data integrity, and collaboration. A recurring theme is data availability and reliability, with concerns over the cost and verification of data sets and the varied accessibility of data from third parties. Moreover, the issue of data size and the logistical hurdles of managing extensive datasets without adequate resources are pointed out, emphasising the need for efficient data management strategies that extend beyond the capabilities of individual researchers or immediate project teams.

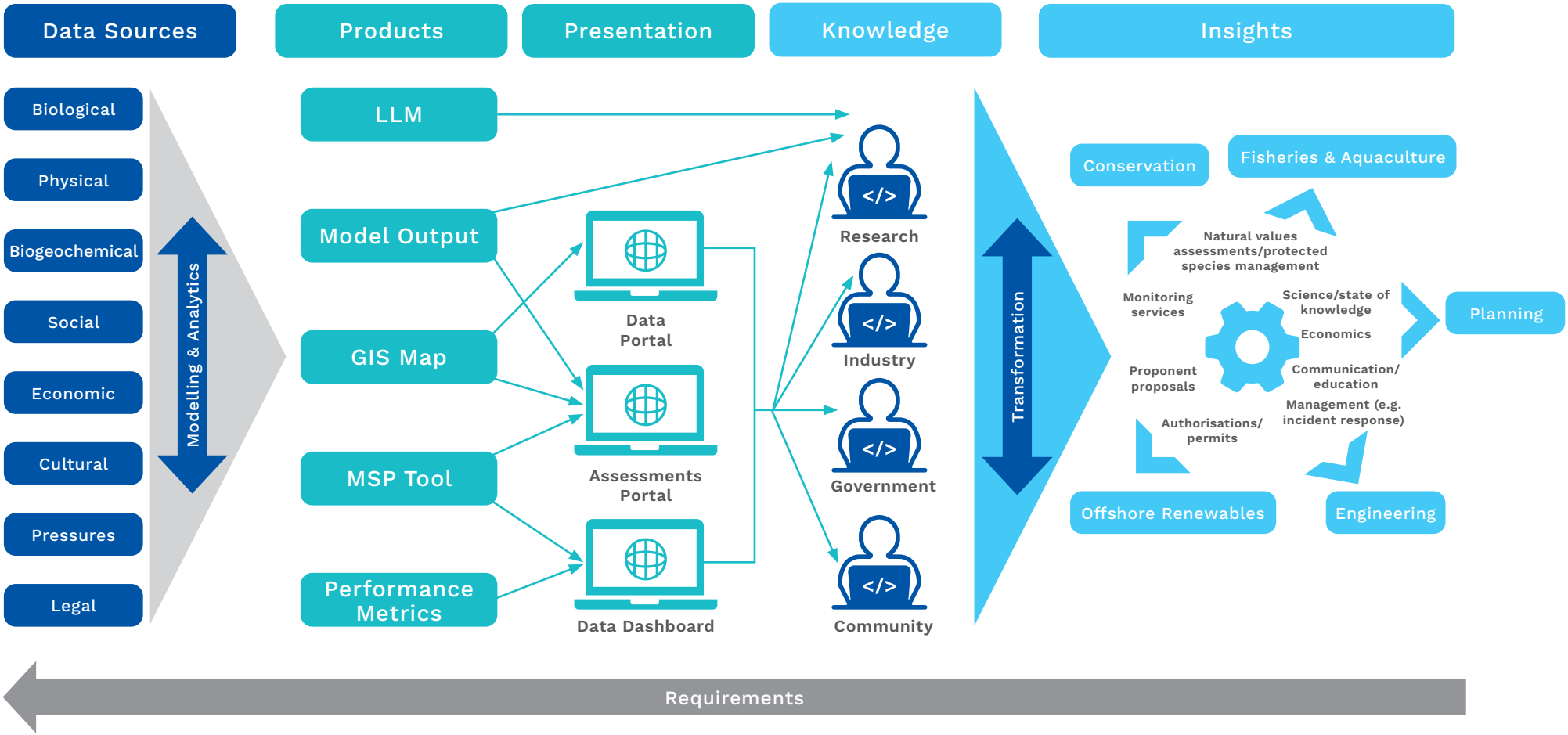
Furthermore, the qualitative feedback highlights a pressing need for improved data sharing mechanisms and metadata quality. Respondents expressed frustration over the lack of detailed metadata and the reluctance of data collectors to share their information, which hampers the effective use of publicly accessible data. The lack of a centralised database for storing, finding, and analysing project data further exacerbates these challenges, indicating a fragmented approach to data management across different projects and partners.

The survey concluded by asking respondents to envision features they would like to see included in data infrastructure. Key recommendations put forth by respondents, such as standardising ethics approval processes and establishing minimum metadata standards, hold the promise of enhancing the design's international credibility and usability. Additionally, the endorsement of robust cloud solutions with adequate backup mechanisms reflects a shared commitment to ensuring sustainable data accessibility aiding seamless collaboration among researchers globally.

3.4. Data workflows

A data workflow diagram was developed to illustrate the path of data from source through the infrastructure towards actionable insights and decision making (Figure 3 7). This figure was used in design workshops as a point of reference to lead discussions on design requirements to achieve the project objectives.

Figure 3-7. Data workflow.



Data and analytics progress through the workflow from source to insight through seven stages in the data infrastructure design (Figure 3 7). These stages and the associated requirements for the CRC data infrastructure are listed below.

1. Data Sources

- Data is sourced from a diverse range of inputs from CRC projects as well as other marine databases in Australia and globally.
- The data collected may include oceanographic sensors, satellite imagery, marine mammal, bird and fish surveys, bathymetry, benthic habitat maps, engineering standards for offshore structures and support vessels, and stakeholder engagement surveys.
- Data sources may include various open data providers, closed systems and using on-premise or cloud storage.
- The aim is to collect and collate comprehensive and high-quality data using consistent methods (documented as part of published Standard Operating Procedures) that can be used to drive and monitor sustainable growth while supporting ocean governance and stewardship measures.

2. Analytics

- Analytics involves the processing and examination of raw data to extract meaningful patterns and insights. This step utilises a range of analytics from statistical analysis of survey data through to machine learning algorithms used to analyse large high frequency ocean datasets.
- Methods of analysis are broad and include (but not limited to) trend analysis, risk assessments (environmental, economic, and/or social or cultural license to operate), and predictive analytics for decision making.

3. Products

- The results derived from analytics are transformed into products that can be used by stakeholders in the CRC for decision making. E.g. regulatory, management or general community interests.
- These products might include GIS maps, predictive model outputs, risk assessment matrix, marine spatial planning and other decision support tools.

4. Presentation Tools

- Presentation tools are employed to visualise and communicate the products effectively.
- These tools may encompass a variety of formats such as graphs, maps, infographics, and reports.
- The objective is to present complex, preferably near live data in a clear and understandable manner, enabling stakeholders to grasp key information quickly and make informed decisions.

5. Knowledge

- This step focuses on the generation and dissemination of knowledge gained from the data analysis and products.
- The infrastructure requirements for this step encompass knowledge management systems which may be utilised for publications, workshops, and training programs to share findings with a broader audience.
- The emphasis is on fostering a deeper understanding of the marine environment and blue economy and promoting the application of data-driven insights in policy and practice.

6. Transformation

- Transformation involves the integration and application of new knowledge to drive change within the blue economy.
- This step may include the development of new policies, the implementation of sustainable practices, and the adoption of innovative technologies.
- The goal is to translate insights into tangible actions towards growth of the blue economy within a sustainable framework.

7. Insights

- Insights are the ultimate outcomes of the data infrastructure, reflecting the deeper understanding gained through the entire process.
- These insights can inform strategic planning, resource management, and policy development.
- They represent the cumulative knowledge and foresight that enable stakeholders to address challenges and invest in sustainable blue economy development for growth.

3.5. Functional Requirements

The functional requirements define what the infrastructure must do and what its features and functions are. They serve as the blueprint detailing the essential characteristics and capabilities of the data infrastructure.

The functional requirements define what the infrastructure must do and what its features and functions are. They serve as the blueprint detailing the essential characteristics and capabilities of the data infrastructure.

Table 4. List of Functional Requirements for CRC Data Infrastructure.

Functional Requirement	Description
Data Storage and Management	» Standard metadata requirements based on existing federated systems (e.g. IMOS, AODN, Geoscience Australia). » Scalable and secure data storage systems capable of handling different types of data. » Assurance of data integrity, redundancy, and backup mechanisms. » Design features to ensure custodial rights and sharing are compliant with relevant laws and agreements, licensing, data origin and purpose. » Ability to manage different privacy and accessibility options that can change through time, e.g. data embargos that can lift after set amount of time. » Data security to allow for the ingestion, protection, and managed access for sensitive Industry and cultural data.
Data Upload and Connection	» Mechanisms to collect diverse data types, including oceanographic, biological, environmental, cultural and socio-economic data, from multiple sources such as but not limited to sensors, satellites, research vessels, reports and external platforms. » Manual data input via user profile with ability to update through time. » Minimum metadata requirements for data ingestion. » Tag data uploads for ease of search in data library
Data Processing and Analysis	» Develop data processing pipelines and algorithms to clean and process raw data to BE CRC prescribed standards » Incorporating machine learning and artificial intelligence techniques for automated data interpretation and anomaly detection. » Data analysis hub to connect data from multiple sources to fit-for-project decision making » Ability to combine and/or analyse data to create new data products as required by end users. » Tag new data products for ease of search in data library
Data Visualisation and Reporting	» Utilise interactive and intuitive data visualisation tools to present complex data sets in a user-friendly manner, enabling stakeholders to gain insights and make informed decisions. » Develop customisable reporting features for different user groups.
System Management Workflows	» Develop mechanisms to allows BE CRC to easily maintain and upgrade the system. » Including reporting metrics highlighting usage trends, upkeep costs and logging change.

Functional Requirement	Description
User-Friendly Interface	» Easy, intuitive navigation through user portals/dashboards » Navigation options depending on user preferences e.g. BE CRC projects or industry use cases » Novel ways of displaying data on landing page, e.g. with 'news' channel » Responsive design (can access from multiple devices e.g. computer or mobile)
Data Search and Filtration	» Data fully searchable with multiple key terms connected to metadata standards » Ability to incorporate rules from federated data in search function if mechanisms allow » Filtered by theme, topic, geographic region and data format » Use of dropdown filters, maps » Data discoverability matrix to include availability/security/permissions/ restrictions, provenance and citation/credit » Use of custom trained LLMs to aid in search functions
Training Resources	» Provide training materials/resources aimed a wide group of stakeholders to navigate and make greatest use of the data infrastructure. » Help section including user manuals, training videos, cheat sheets, FAQs, example workflows. Chatbot to answer FAQs and link to a support service and community forum. » Use of custom trained LLMs to guide users and make their use of time efficient.
Community Function	» Establish user profiles with entry data (name, company, background, BE CRC projects involved in etc) and derived data (what projects interested in) » A community forum/collaborative space for collaboration, discussions, Q&A, data sharing, news. » Community guidelines for membership that all members must sign up to » Automated and manual removal for postings/questions/comments (identify tone, trigger words) » Pre-approved community for responses » Metrics collected for BE CRC to guide future projects, gap analysis, funding, strategy and direction. » Feedback loop from community forum to regulate responses (downvote/ upvote) » Customisable incentive features to encourage participation (e.g. badge for frequent responses, number of postings, data uploads). » Cross-links between relevant data and projects highlighted for users automatically.
Project Visibility	» To be easily able to find past projects and their summaries and associated meta data. » Link from the project to the associated researchers/collaborators profiles in the Community Function.

3.6. Non-functional Requirements

Non-functional requirements describe the general properties of a system. They are also known as quality attributes.

These requirements speak to the quality of the infrastructure design and describes what success looks like in the infrastructure design.

Functional requirements identified through the project have been listed under general themes in Table 4.

Table 5. List of Functional Requirements for CRC Data Infrastructure.

Non-Functional Requirement	Description

Security Requirements	» Security mechanisms need to be sufficient to mitigate the chances of cyber security incidences and make users of the platform assured the platform is safe.
	  » Multi-layered security measures.
	  » Stringent Penetrative testing through third-party engagement.
	  » Use of encryption to safeguard personally identifiable data even in the case of data breach.
	» The security of the platform must be strong to safeguard data while still allowing researchers to seamlessly integrate with broader data and analytical federations.

Agility and Scalability	» The platform ought to foresee periods of heightened demand and dynamically allocate resources to prevent users from encountering sluggish performance or disruptions. Utilising reporting metrics, workflows will be optimised to address the long-term requirements of personnel, software, and hardware. » Additionally, the platform should possess adaptability to seamlessly incorporate novel and emerging analytical techniques stemming from the Blue Economy research community. This flexibility ensures that the platform remains relevant and capable of evolving alongside advancements in research methodologies within the Blue Economy domain.
Availability and Portability	» Excepting scheduled and communicated downtime due to system maintenance, the platform will meet the maximum uptime deemed achievable by the product owner. The platform is to be compatible with major hardware, operating systems, and applications.
Maintenance	» The platform should experience minimal critical failures when these occur, troubleshooting and return to normal operations should be easy and rapid.
User Experience	» Data sharing and collaboration will be encouraged by establishing data sharing agreements, data portals, and virtual collaboration and human networking tools. The platform will foster a culture of open data and knowledge sharing while respecting proprietary and sensitive information.
Compatibility	» Compatibility with other data systems such as IMOS. Define standards and protocols for data integration and interoperability across different data sources, platforms, and stakeholders. Implement data exchange formats, APIs, and data ontologies to facilitate seamless data sharing and integration.
Governance	» Data governance frameworks are to ensure ethical data use, compliance with data protection regulations, and proper data access controls. Elements of design and data sharing arrangements are to comply with laws and government policies for Australian, New Zealand and other countries that connect through the data infrastructure.

3.7. Review of data sharing and analytics platforms

A review of four existing and well developed data analytics platforms helped to identify existing features that could address the design requirements of the CRC. The four platforms were either developed with government regulation, research or industry needs core to the design to enable a comparison of platforms to meet the CRC data infrastructure needs and design requirements identified by CRC stakeholders.

Australian Research Data Commons

3.7.1. ARDC Nectar Research Cloud

The ARDC Nectar Research Cloud (Nectar) is Australia's national research cloud, specifically designed for research computing.

Launched in 2012, Nectar provides Australia's research community with fast, interactive, self-service access to large-scale computing infrastructure, software and data. A powerful platform for collaboration, it allows researchers and research support staff to access compute resources, software and data from their offices and homes and easily share them with collaborators at other institutions.

Nectar is a federation that is co-designed and receives co-investment from universities across Australia. The federation enables cross-institutional research collaborations to deliver research computing services at a national scale.

Nectar is a versatile cloud computing infrastructure that can be used in many ways to support research, such as a virtual desktop for a single researcher, or as a powerful computational server that can be shared by researchers in Australia and internationally.

With Nectar, researchers can:

- connect to a suite of advanced research computing resources directly from your desktop at home or in the office
- access large-scale computing including large memory machines and GPUs
- run Jupyter Notebooks or virtual desktops
- deploy and run software, workflows or platforms using containers and container orchestration
- use a flexible, scalable and innovative world-class service that allows you to customise your computing infrastructure to meet the requirements of your research project
- collaborate nationally and internationally with easy shared access to compute, software and data to ensure your research meets the highest standards
- access and rapidly deploy and share research software tools and data to easily collaborate with peers
- benefit from centralised support and expert knowledge and continual development of a cloud designed for Australian researchers, offering the support you need to succeed in academia and beyond.

ARDC Services Powered by Nectar

The ARDC hosts important services for researchers on Nectar, which can save time, boost your productivity and give you added power to conduct ground-breaking research.

Here are some ways we use Nectar's versatile cloud infrastructure to provide services to support research:

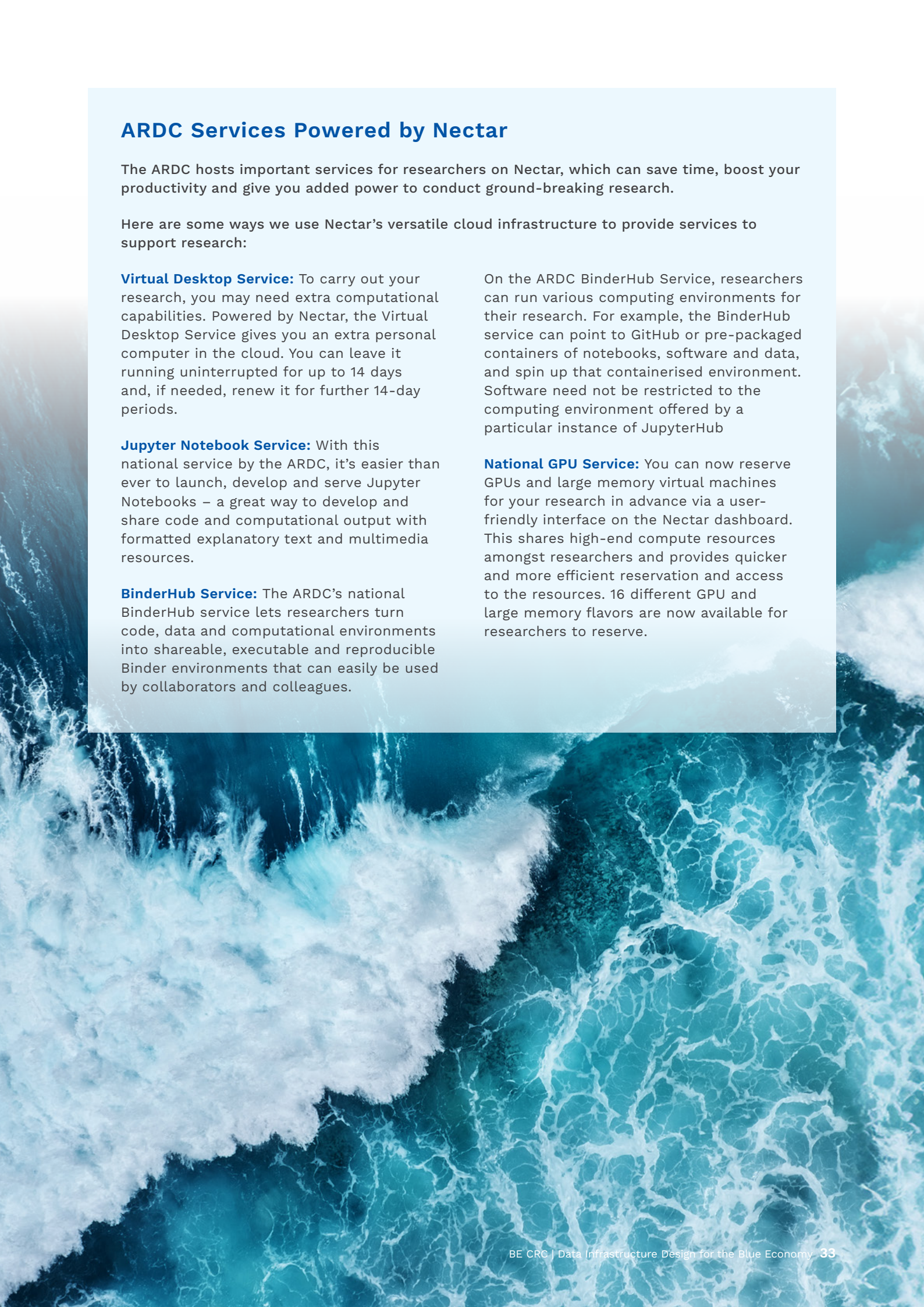
Virtual Desktop Service: To carry out your research, you may need extra computational capabilities. Powered by Nectar, the Virtual Desktop Service gives you an extra personal computer in the cloud. You can leave it running uninterrupted for up to 14 days and, if needed, renew it for further 14-day periods.

Jupyter Notebook Service: With this national service by the ARDC, it's easier than ever to launch, develop and serve Jupyter Notebooks – a great way to develop and share code and computational output with formatted explanatory text and multimedia resources.

BinderHub Service: The ARDC's national BinderHub service lets researchers turn code, data and computational environments into shareable, executable and reproducible Binder environments that can easily be used by collaborators and colleagues.

On the ARDC BinderHub Service, researchers can run various computing environments for their research. For example, the BinderHub service can point to GitHub or pre-packaged containers of notebooks, software and data, and spin up that containerised environment. Software need not be restricted to the computing environment offered by a particular instance of JupyterHub

National GPU Service: You can now reserve GPUs and large memory virtual machines for your research in advance via a user-friendly interface on the Nectar dashboard. This shares high-end compute resources amongst researchers and provides quicker and more efficient reservation and access to the resources. 16 different GPU and large memory flavors are now available for researchers to reserve.



3.7.2. SEAF

The Shared Environmental Analytical Facility (SEAF) represents a comprehensive, cloud-based analytical ecosystem that unifies environmental data and modelling resources across various research and industry partners. Established through the collaboration of the Western Australian Marine Science Institute (WAMSI), the Western Australian Biodiversity Science Institute (WABSI), Microsoft, IXUP, ARINCO, and others, SEAF is designed to address environmental assessment complexities by creating a centralised, secure platform that supports cumulative impact analysis and regional insights.

Hub-and-Spoke Model

SEAF's architectural framework is built on a **Hub-and-Spoke model**, a strategic design that balances centralised control with regional flexibility. This model is particularly well-suited to SEAF's mission, as it ensures robust security and policy adherence across all environments while allowing individual spokes to innovate with specialised tools tailored to regional scientific needs.

1. **Centralised Security and Governance in the Hub:**
    - **Unified Security Standards:** The central Hub establishes and enforces a consistent security policy, covering areas such as data encryption, user access management, and compliance with both federal and state cybersecurity guidelines. This unified security approach, supported by Azure's robust security features, simplifies governance across all spokes, minimising vulnerabilities while ensuring regulatory compliance.
    - **Standardised Data Governance Policies:** The Hub manages global data governance frameworks, such as the Five Safes model from IXUP, ensuring that all spokes follow the same protocols for data privacy, sharing, and access auditing. This includes the enforcement of privacy-preserving analytics, federated identity management, and data ownership clarity, which fosters trust across stakeholders.
    - **Automated Policy Deployment via Azure DevOps:** Using Azure DevOps and infrastructure-as-code (IaC) practices, SEAF's Hub deploys standardised policies, network configurations, and security patches across all spokes. This automation streamlines updates, reduces operational overhead, and ensures that each spoke remains compliant with the latest security policies.

2. **Spoke Autonomy for Regional Innovation and Customisation:**
    - **Localised Tool Development for Unique Challenges**: Each spoke is empowered to develop and deploy bespoke tools and workflows to address specific environmental challenges within its region. For example, the Cockburn Sound spoke can implement targeted hydrodynamic models to assess cumulative environmental impacts unique to its ecological context, drawing on datasets relevant to the region's marine biodiversity.
    - **Flexible Data and Compute Integration:** Each spoke is equipped to integrate with regional data sources and utilise compute resources, such as those available through partnerships with local institutions (e.g., Pawsey S3). This autonomy allows spokes to work with unique data sets and computing environments, facilitating advanced scientific analysis that aligns with regional goals.
    - **Boutique Tool and Model Development:** The model allows spokes to create customised, boutique analytical tools and machine learning models that address novel scientific inquiries or industry needs. This capability is critical for developing solutions tailored to specific research goals, such as unique cumulative impact analyses or specialised ecological modelling that may not be relevant in other regions.

3. **Inter-Spoke Data Sharing and Collaborative Workflows:**

   - **Secure Data Exchange Channels:** Each spoke is connected back to the Hub, enabling secure data exchange across regions. IXUP's homomorphic encryption allows for data collaboration without compromising confidentiality, fostering inter-regional collaboration on complex issues such as ecosystem-wide impact assessments.
   - **Shared Analytical and Machine Learning Templates:** SEAF's Hub provides shared templates and best practices for analytics, machine learning, and data processing, which can be utilised by any spoke. These resources allow regional spokes to adopt tried-and-true methodologies while still customising their applications based on local scientific and regulatory needs.

4. **Efficient Resource Utilisation and Scalability:**

   - **Cost Optimisation and Shared Infrastructure:** The Hub consolidates essential infrastructure and resources (e.g., Azure services, centralised logging, and monitoring) to reduce costs, allowing spokes to focus their budgets on local data acquisition and tool development. This shared foundation enhances cost efficiency and makes SEAF more scalable as it expands to additional regions.
   - **Dynamic Resource Allocation:** The Hub-and-Spoke model facilitates dynamic scaling, allowing spokes to allocate additional compute or storage resources as their projects evolve. This flexibility ensures that each spoke has the capacity needed to meet scientific and data demands as they grow over time.

The Hub-and-Spoke model is instrumental in SEAF's ability to deliver secure, scalable, and region-specific analytics. It combines the benefits of centralised governance and economies of scale in the Hub with the innovative potential of decentralised, autonomous spokes. Each spoke can respond nimbly to local scientific and regulatory challenges, developing tools and analytics that drive impactful, actionable insights while remaining within SEAF's overarching framework of security and data governance.

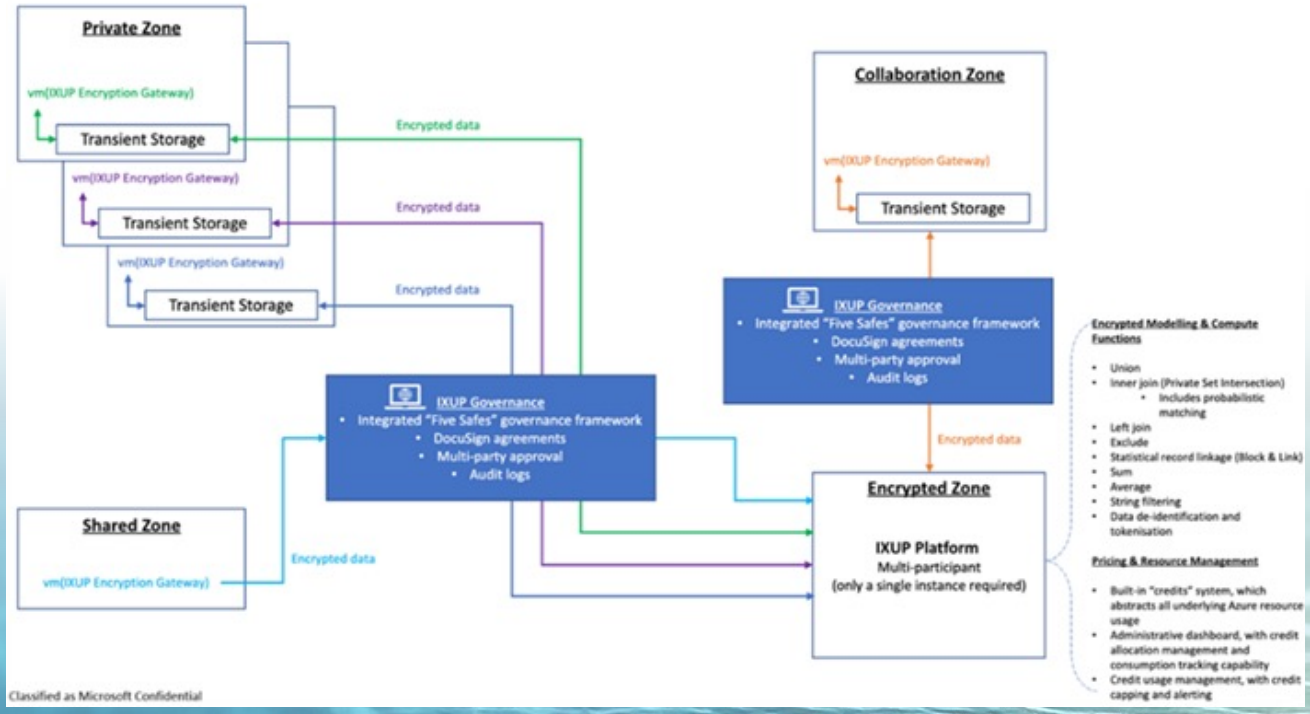
Key Architectural Features: Multi-zonal Design

SEAF leverages a robust, multi-zonal architecture on Microsoft Azure, designed to facilitate secure data exchange, analytics, and collaboration across different organisations. The architecture supports:

Isolated, Customisable Private and Collaboration Zones: Each zone can be tailored to an organisation's specific security requirements, allowing for the development of secondary and tertiary data products without compromising data confidentiality.

Encrypted Zones for High-Security Needs: Sensitive data processing is conducted in encrypted zones governed by the IXUP Secure Data Engine, which applies advanced homomorphic encryption to enable analysis without data decryption. This setup allows multiple data custodians to collaborate securely without direct data exposure.

Figure 3-8. SEAF multi-zonal architecture.



Data Security and Compliance

Data within SEAF is safeguarded using industry-leading governance and security frameworks:

Homomorphic Encryption with IXUP: IXUP's platform ensures data is encrypted at rest, in transit, and in use. This system supports secure analytics by allowing computations on encrypted data, thus facilitating insights without data exposure. IXUP's Privacy Analysis Engine, based on the Five Safes governance framework, enables secure, real-time collaboration monitoring.

Federated Identity and Access Management: SEAF utilises Azure Entra ID, with role-based access control (RBAC) and multi-factor authentication, ensuring that access to sensitive resources is strictly managed.

Regional Impact and Expansion

The SEAF platform is tailored to address the unique challenges of environmental assessment in regional contexts. It has already facilitated groundbreaking work in Cockburn Sound, where it assists in integrating and analysing data related to ecological health and regulatory impact requirements. As the platform evolves, it is positioned to expand to additional regions. The planned deployment of additional regional spokes aims to create a nationwide network of trusted environmental information supply chains.

Future-Proof Infrastructure and Operational Efficiency

In line with Microsoft's Cloud Adoption Framework, SEAF incorporates resilient infrastructure designed to support scalability and disaster recovery:

Azure DevOps Integration for Automated Workflows: SEAF leverages Azure DevOps to automate deployment, monitoring, and management of resources, supporting a GitOps-driven approach that minimises manual intervention and enhances operational efficiency.

Comprehensive Monitoring and Compliance: With Azure Defender and centralised logging through Azure Monitor, SEAF's infrastructure is continuously assessed for security vulnerabilities, providing compliance assurance across cloud resources. Azure policy enforcements further ensure adherence to governance and cost management controls.

The SEAF platform represents an innovative, scalable approach to environmental analytics, blending advanced modelling, data science, security, and governance to support sustainable decision-making across research, regulatory, and industrial landscapes.

3.7.3. BMT Deep: Data Exploration and Analytics Platform for Marine and Environmental Applications

BMT Deep is an interactive data platform developed by software engineers, data scientists, and subject matter experts, designed to provide deeper insights for data-driven decision-making. Initially created for asset performance management in the offshore oil and gas industry, it now also manages environmental assets, including marine ecosystems and aquaculture operations.

Advanced Technology Integration

Utilising cloud computing, big data, the Internet of Things (IoT), and Artificial Intelligence (AI), BMT Deep efficiently stores, manages, integrates, analyses, and visualises vast datasets. The platform captures data from high-frequency points to annual monitoring data, enabling decision-makers to track assets in near real-time, from immediate updates to long-term trends.

Key Features and Advantages

Comprehensive Data Management:

- Efficient storage, quality control, management, integration, and post-processing of vast datasets.
- Lakehouse architecture supporting structured, semi-structured, and unstructured data.
- Federated data management across distributed systems.

Advanced Analytics and Visualisation:

- Interactive and intuitive online interface for exploring single or multiple assets.
- Powerful tools for data analysis and visualisation, including real-time tracking and historical trends.
- Analysis Studio, powered by Databricks, for scalable and diverse analysis.

Customisation and Support:

- Fully customisable platform tailored to specific application needs.
- Continuous technical support with data quality verified by analysts and consultants.
- Custom solutions for creating fit-for-purpose applications.

Robust Security and Governance:

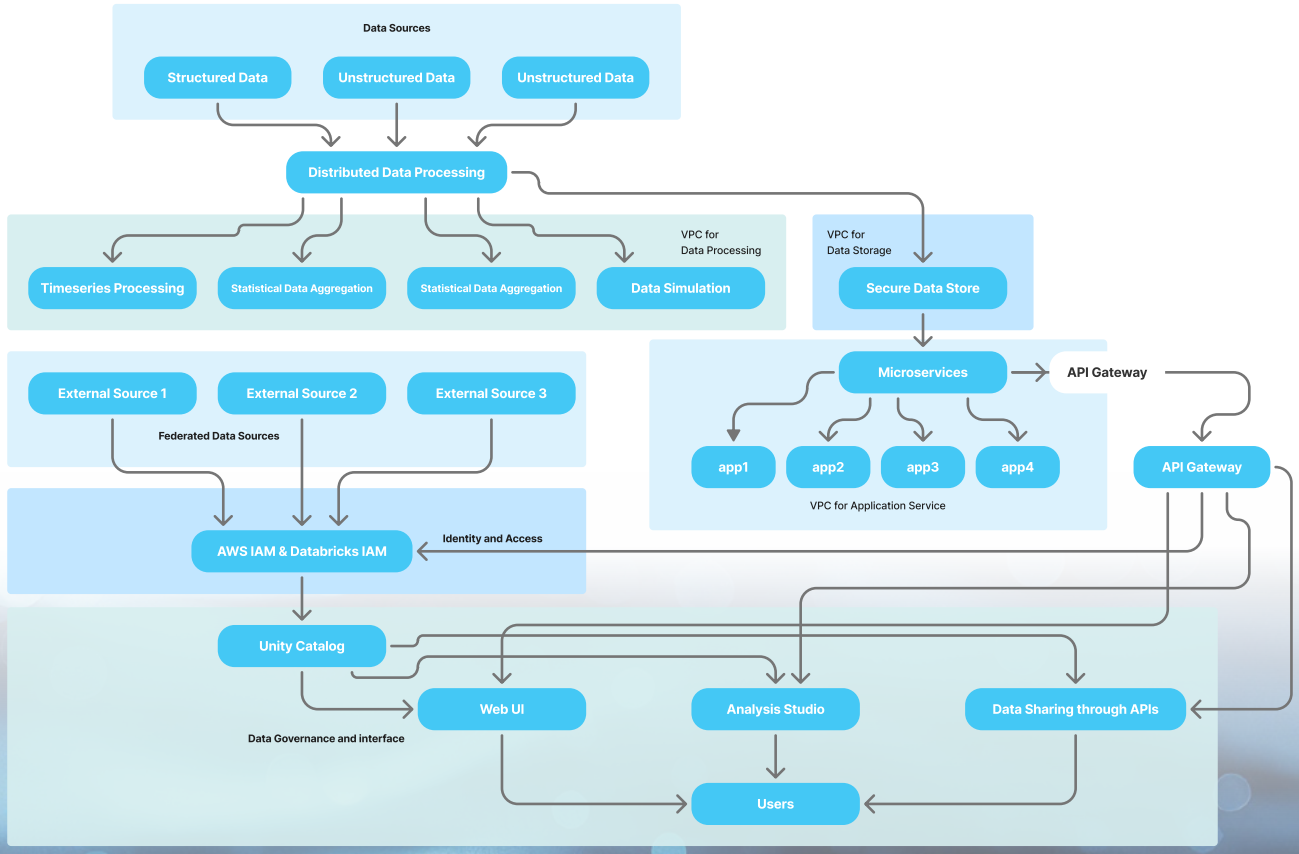
- Built-in advanced security features, including Single Sign-On (SSO) and SOC2 compliance.
- Comprehensive data governance with Unity Catalogue and robust user management.
- Secure data storage, management, and processing.

Global Collaboration and Deployment:

- Cloud-based platform enabling global collaboration and simultaneous data access.
- Deployed within AWS Australia Cloud for secure, reliable, and scalable infrastructure.
- Ensures data handling and complex analytics remain within the country.

The platform's modular design allows users to fulfill current operational needs while planning for future upgrades and advancements in marine and environmental technology. A high-level data flow architecture with consideration of the CRC data requirements works through the data source through analytics in a web based user interface (Figure 3 9). shows flexibility in design for various data sources, use of a Virtual Private Cloud (VPC) environment for data processing and cleaning, storage and analytics.

Figure 3-9. Overview of Dataflow Architecture within BMT Deep.



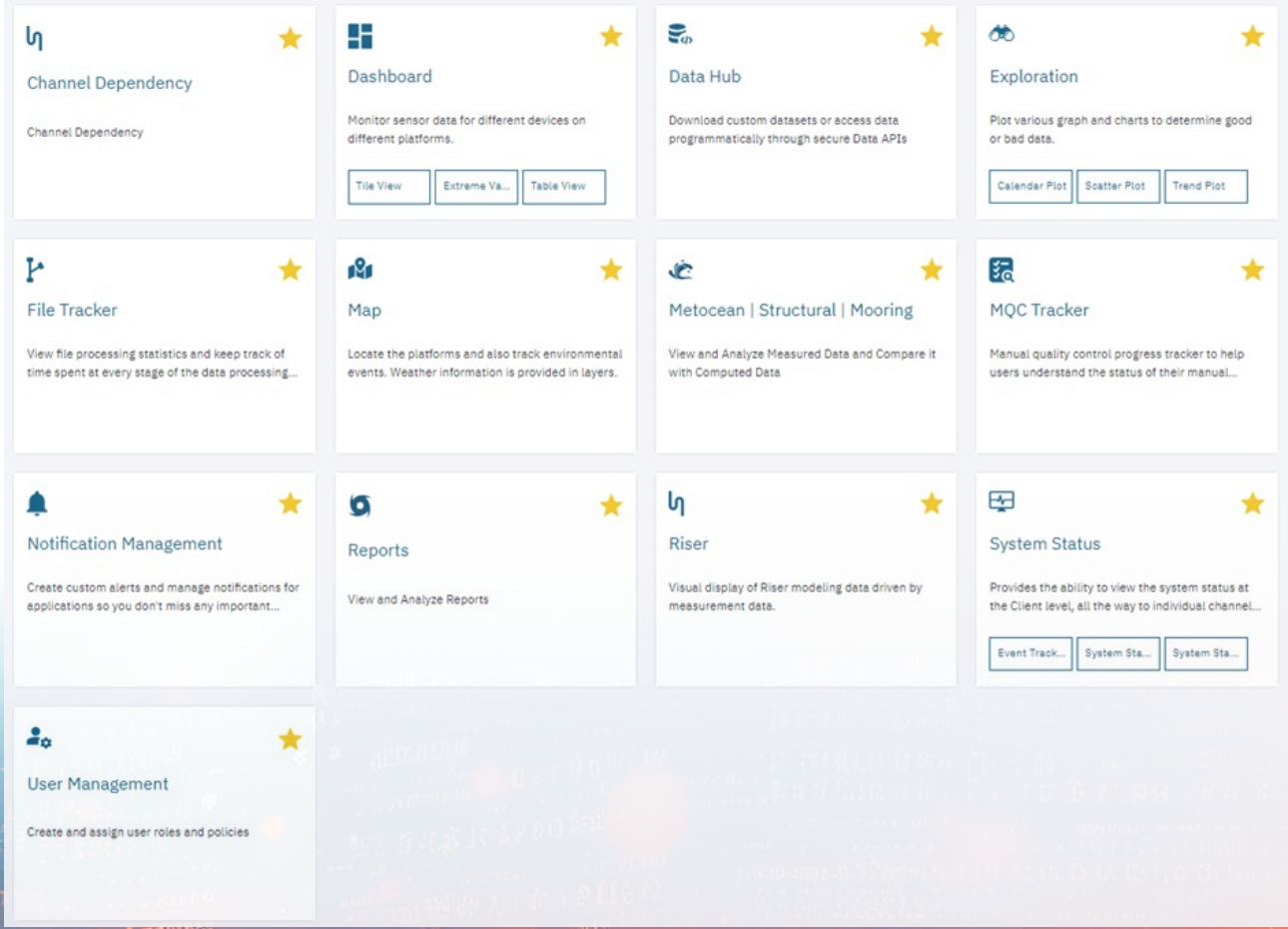
Deployment and Infrastructure

Deployed within AWS Australia Cloud, BMT Deep leverages secure, reliable, and scalable infrastructure, ensuring data handling and complex analytics remain within the country.

Proven Success and Continuous Improvement

Supported by over 20 years of practical in-field experience, BMT Deep has been used in diverse projects for a decade, from deep-water offshore facilities to environmental compliance monitoring in Australia. These projects demonstrate BMT Deep's capability to manage complex environmental data effectively and provide actionable insights for sustainable operations. The platform undergoes rigorous testing and continuous improvement to maintain high performance and reliability standards. BMT Deep offers comprehensive, customisable solutions for marine and environmental data management. Its advanced technology, secure infrastructure, and proven record of accomplishment make it an ideal choice for informed decision-making and optimised operations in marine and coastal environments. (Figure 3.2).

Figure 3-10. Example features in BMT Deep.



3.7.4. Australian Agricultural Data Exchange (AADX)

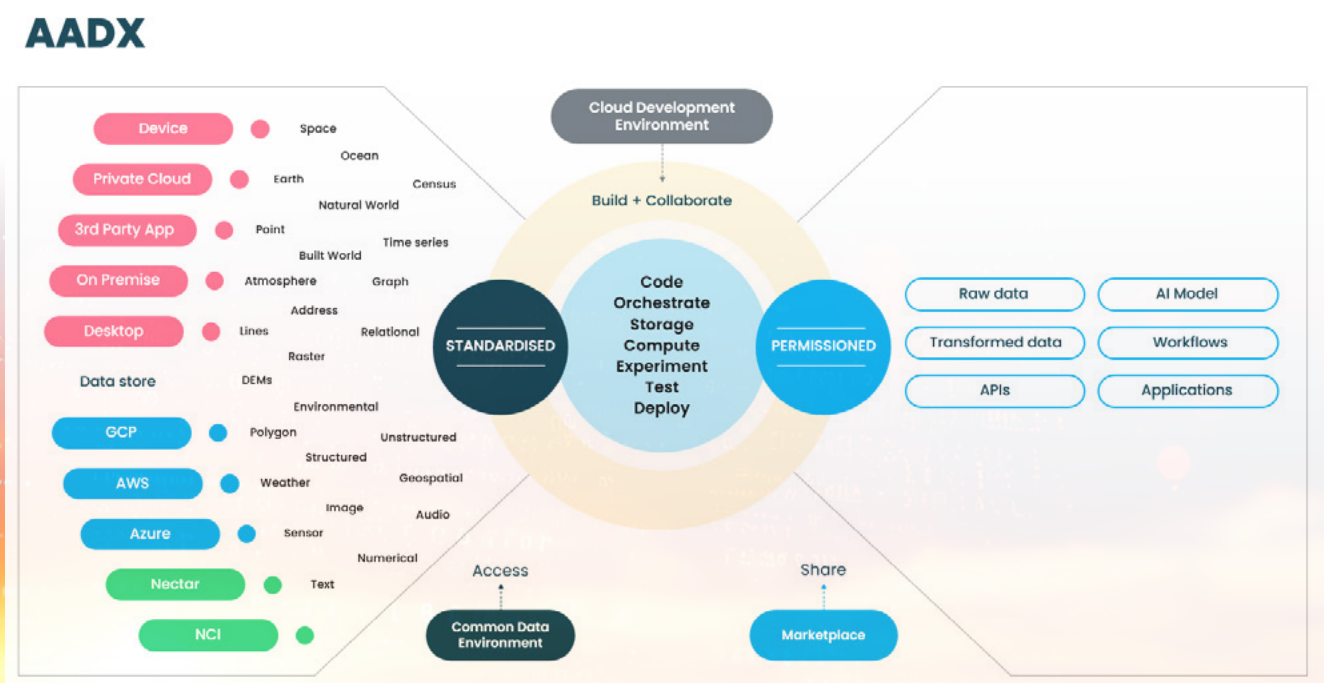
The Australian Agricultural Data Exchange (AADX) is a joint initiative by a partnership of agricultural, fisheries and research organisations to develop a platform for sharing data from disparate sources in a secure environment. Users can establish their own private data exchange, publish datasets to a shared data catalogue (also called a data marketplace), and integrate and interoperate with other technologies for processing, analysing and visualising data.

AADX will be integral to data management for the blue economy as it is rapidly evolving to be the platform of choice for access to fisheries data across Australia through support from the Fisheries Research and Development Corporation (FRDC). While in its infancy, FRDC are deploying several case studies as proof of concept for ongoing development, aggregation and harmonisation of data across federal and state jurisdictions.

The AADX supports integration across platforms and are currently partnering with Eratos as a solution provider. Eratos have developed data Infrastructure that solves the way organisations and individuals make informed decisions about the world we live in. Eratos simplifies and automates finding and accessing disparate, often sensitive, data at scale, with its interoperability and reusability tools.

Using Eratos, users can access, build, collaborate and share data, models and solutions. Users have typically included research, education, federal and state government, agriculture, fisheries, and financial services. This has enabled a collection of relevant, isolated and disparate data sets and research models to be assembled onto a single platform.

Figure 11. AADX powered by Eratos



4. Discussion

4.1. Data Infrastructure Review

The Australian marine science community have a long history of planning and building federated data infrastructure from the establishment of the Australian Ocean Data Centre Joint Facility (AODC-JF) in 2004 through to the formation of the Integrated Marine Observing System (IMOS) in 2006 and the merger between the AODC-JF and the IMOS data facility to establish the Australian Ocean Data Network (AODN) in 2016.

The AODN is now the data management facility of IMOS and the aggregation point providing discovery and access for Australian marine data more generally. Partners of, and contributors to, the AODN conform to a set of standards and associated infrastructure design for their platforms to integrate and be interoperable with the AODN and each other.

This has led to a well-defined backbone for sharing and integrating marine science data in Australia and remains the most viable approach to integration and interoperability of open marine data sources in Australia. It provides much of the data required for both the CRC and the Blue Economy generally and there is broad commitment from data providers to comply with its requirements. However, while uptake of these standards amongst marine data providers is generally broad, it is not universal and doesn't allow for managed access to secured, sensitive data. Data security was identified as a considerable challenge in the establishment of the CRC data infrastructure requirements.

The main challenges identified in the consideration of the CRC infrastructure design include the substantial amount and variation of data generated by the CRC that may not be catered for in existing marine data infrastructure in Australia. Further, the analytics features of existing infrastructure may not cater for the wide range of analytics tools generated that access is required for. Finally, the useability of the infrastructure must be simple yet contain all required aspects to ensure the value of the data.

To overcome these challenges, learnings from similar projects in the past can be implemented to ensure longevity and relevance of the CRC data. Several challenges to consider:

- Should users be able to access data externally or should this data be duplicated/restored in the infrastructure?
- How will the infrastructure handle and dictate the use of secured and free accessible data simultaneously?
- How security of infrastructure will protective private/sensitive data from misuse?
- How the infrastructure will utilise protocols to ensure it meets data requirements and has unified workflows?
- Consider how to establish comprehensive governance policies which outlines data accessibility, ownership, guideline and compliances requirements?
- How to ensure data is validated and standardised to meet an appropriate, consistent standard?
- How will the infrastructure promote collaboration with stakeholders such as providers, developers, researchers and the community?
- How will the infrastructure manage metadata to describe, catalogue and organise data?
- How will the longevity of the project be sustained?

4.2. CRC Research Project Data Needs

The integration of insights obtained from both the survey forms and the one-on-one interviews with project leaders has provided a comprehensive understanding of the data management needs and challenges within the BE CRC.

Combining these sources of information allows for nuanced analysis and synthesis of key takeaways, facilitating the formulation of informed strategies for the development of a robust data infrastructure. Here, we discuss the results derived from the survey forms and interviews, highlighting their significance and implications for the design and implementation of the CRC Data Infrastructure ecosystem.

4.2.1. Data Types and Collection Methods:

The survey forms revealed a diverse range of data types being collected across various research programs, including tabular data, time series data, images, and spatial data. Interviews further corroborated this finding, emphasising the need for flexibility in accommodating various data formats within the infrastructure.

4.2.2. Data Management and Storage:

Survey responses indicated that data storage requirements vary widely, from megabytes to terabytes in volume, with storage mediums ranging from university servers to cloud storage. Interviews underscored the importance of scalable storage solutions to accommodate future data expansion, especially for projects dealing with real-time operational data.

4.2.3. Data Analytics and Processing:

The survey highlighted the use of a variety of tools for data analysis, including MS Excel, MATLAB, and R. Interviews provided additional insights into the analytical techniques employed, emphasising the importance of integrating analytical tools within the data infrastructure to support ongoing research activities.

4.2.4. Data Sharing and Accessibility

Survey responses indicated a desire for improved data accessibility and sharing mechanisms, with most data intended to be shared with project partners. Interviews further emphasised the importance of data-sharing abilities within the infrastructure to facilitate collaboration among stakeholders and researchers while ensuring data confidentiality and privacy.

4.2.5. Challenges and Solutions:

Both survey forms and interviews identified common challenges in data management, including data availability, reliability, and accessibility. Solutions proposed included implementing collaborative data management practices, standardised protocols, quality assurance measures, and efficient project management tools within the infrastructure to address these challenges effectively.

4.2.6. Desired Features for Data Infrastructure:

Survey responses and interviews highlighted the importance of user-friendly features such as visualisation tools, data filtering, and soft sampling in the design of the data infrastructure. Additionally, there was a strong emphasis on the need for flexibility and adaptability especially with regards to data sharing and accessibility to accommodate evolving project needs and stakeholder requirements.

4.2.7. Future Vision and Aspirations:

Interviews provided valuable insights into project leaders' future aspirations and potential uses of the data infrastructure, ranging from long-term data management to predictive analytics and educational purposes. These aspirations underscored the importance of designing a scalable and sustainable data ecosystem that can support the evolving needs of the blue economy.

Through the design process the team came across two main challenges, how to design a single infrastructure to support the range of requirements across all five research programs and how to support an open data philosophy while enabling CRC partners to retain data ownership and security to protect commercial intellectual property and culturally sensitive information. Key to meeting these challenges was to come up with a governance framework capable of demonstrating both flexibility, adaptability and security. A further review of existing data infrastructure guidelines at three levels of governance, national, international and CRC was undertaken by the design team. Four existing data analytics platforms currently used to support Blue Economy industries were reviewed in detail to gain understanding and insights as to how they could be leveraged in the design of the CRC data infrastructure.

4.3. Analysis of Requirements

An analysis of the requirements formulated during workshops were used to broadly identify characteristics of a data architecture required to fulfill the needs of end users. Unsurprisingly, those requirements display the need for a complex and diverse set of infrastructure elements. However, the CRC neither wants nor needs to develop all the required infrastructure itself.

In terms of data access, much of what is required exists as part of a complex, publicly operated and accessible data ecosystem provided through Government investment. Examples relevant to the CRC include IMOS and the AODN, Geosciences Australia, DCCEEW, AIMS, CSIRO, IMAS and a range of other institutional and Government departments both Federal and State.

The following elements outline the broad infrastructure needed to meet the requirements analysis outlined in sections 2.3 and 2.4:

- The ability to be part of, and contribute to, a federated network of standardised and interoperable data infrastructure for providers and consumers.
- The ability to publish data meeting the FAIR data principles and using standards-based approaches to metadata, data access and formats.
- The ability to secure and manage access to sensitive data demonstrating trust (including industry and Indigenous data and knowledge).
- The ability to operate with a broad range of platforms delivering data processing, analysis, visualisation and reporting for knowledge transfer and decision making.
- The ability to provide sustainable and reliable infrastructure.

4.4. Blue Economy Data Needs

Data essential to understanding and managing the blue economy can be characterised across and within domains including oceanographic, biological, environmental, cultural, and socio-economic. Broadly defining those data is a key step to ensuring a viable multi-disciplinary approach for aggregating and harmonising data resources, enabling interoperability, and cumulative representation and understanding. In turn, this has implications for infrastructure design and governance.

The National Marine Science Committee have established a national marine baselines and monitoring program defining a base set of data required for long term environmental monitoring. It includes variables across the following categories:

- Physical
- Biogeochemical
- Biological
- Ecosystem
- Pressures

This provides a strong set of base data for mapping ecosystem state and cumulative impact. The needs of the blue economy will require this to be supplemented by other data including:

- Climate and weather
- Economic
- Cultural and Indigenous
- Engineering
- Policy and planning

Some of these data are collected, processed, managed, and delivered by various entities using a range of infrastructure that may include databases, APIs for access to the data and websites for displaying or visualising the data. For example, in Australia, IMOS produce and make available sea surface temperate and phytoplankton biomass; Geosciences Australia collect, store, aggregate and make available bathymetry data holdings; and the Australian Bureau of Statistics provide a wide range of socio-economic data. While these organisations deliver these data reliably using well established, standards based infrastructure, other data critical for managing the marine estate require further work to become established sustainably and their data to be fully liberated, accessible, and interoperable.

4.5. Characterising Data Infrastructure

Data infrastructure is a broad term representing the hardware and software environments supporting the data and information lifecycle. Based on design workshops, surveys, stakeholder workshops and discussions with AGC, four classifications were identified. These relate to the hardware and software environment used, the data itself and the functions required to meet end user needs (see Figure 12). Defined as:

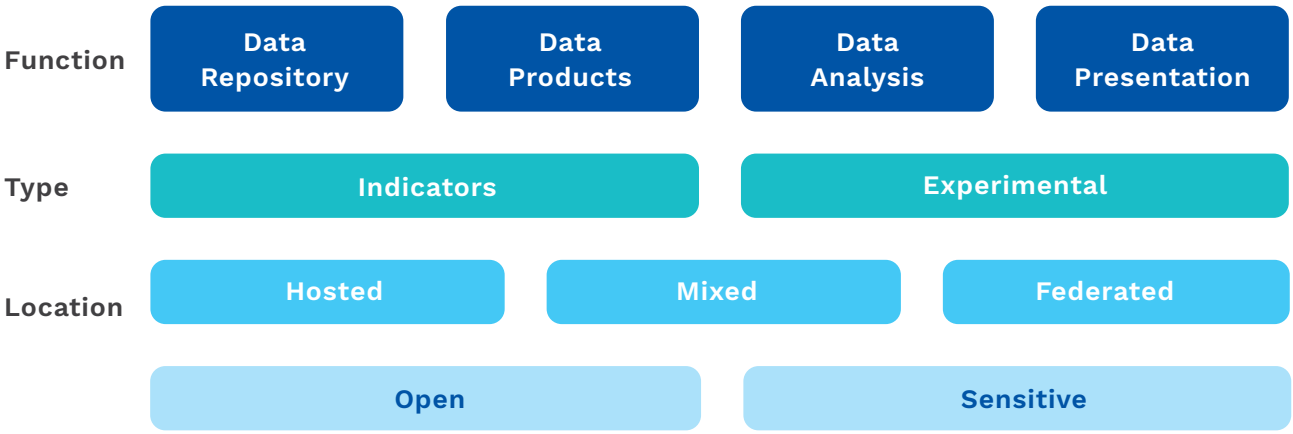
Type: indicators (such as those defined under the NMSC Baselines and Monitoring Program) or experimental data (data collected, usually at a small scale, to address a specific and immediate problem that is not a long-term monitoring indicator). In most cases, data for the blue economy will be indicators, but there will be some requirement for managing experimental data.

Location: defined as either locally hosted data, data accessed remotely through a federated solution (e.g. the AODN) or a mixed approach using both local and federated data access.

Function: generally refers to how the infrastructure will be used for delivering the data and includes data repository, data products, data analysis platforms and software and data presentation (such as dashboards, mapping and other reporting tools).

Classification: refers to whether the data is publicly available open data or some form of sensitive data with commensurate access restrictions (which may include time limited embargoes).

Figure 12. Data infrastructure characterisation.



Individual data platforms can provide one or more elements of each classification.

Some examples:

- The Australian Ocean Data Network Portal supports **open** data only, is a **mixed** environment (being both the aggregation point for a **federated** network of data portals and **hosting** data from IMOS facilities), has a primary function to provide **indicator** data, and is primarily a **data repository.**
- AusSeabed (Geoscience Australia's public seabed mapping portal) supports **open** data, **hosts** its data locally, provides **indicator** data (bathymetry) and is a **data repository** with some **data presentation** services (a mapping interface).
- Seamap Australia is the national repository for benthic habitat data and the National Benthic Habitat Layer. The **data** it hosts is open, includes **indicators** (such as seagrass and macroalgae cover), and is a **data repository** providing **data products and presentation** through its mapping interface and state-of-knowledge tools.
- The Shared Environmental Analytics Facility (SEAF) **hosts sensitive** industry data along with **open** data from other data sources. It uses mostly **indicator** data and is primarily a **data analysis** platform generating **data products** and **data presentation** (mainly for its own internal use, but also some publicly available).
- BMT Deep is a **data product** designed to **store, analyse** and **visualise** marine and maritime data designed for comprehensive data management, governance, and analytics. It supports hosting and federated management of **sensitive, mixed, experimental,** and **open** federated data, ensuring flexibility and scalability.

Characterising the data, its access and use, and required deliverables allowed us to understand and review the processes required for these data supply chains to meet end user needs.

Managing sensitive data effectively while still ensuring it is used for maximum benefit represents new challenges moving forward. Within the confines of CRC, and the requirements for a sustainable blue economy more generally, sensitive data falls within the following major categories:

- Industry data that has a commercial-in-confidence component and/or competitive intelligence.
- Indigenous knowledge and data that needs to be managed under the CARE principals and ICIP.
- Spatial data for Threatened, Endangered and Protected Species (TEPS).
- Research data prior to publication or under temporary IP protection.

A federated data solution to meet the requirements of the CRC and the blue economy needs to include access and interoperability for both open and sensitive data. This is not currently available in institutional data repositories mentioned above that have focused on open data initiatives. However, there are a number of commercial offerings available, usually based on a subscription model, for storing, accessing and/or providing analysis tools for closed, sensitive data (examples include BMT Deep, AADX and Eratos).

These commercial solutions offer a means for securely managing sensitive data while ensuring its access where needed. However, interoperability between the platforms remains unsolved to date. It is also likely government and institutional research repositories will need to provide their own in-house solutions for interoperable, sensitive data management in the future.

4.6. Federated, Secure Integration and Identity Management

Enterprise security is of paramount importance to all the CRC participants, where the balance between security and timely and efficient access to data is a critical consideration. While each organisation may have its own approach to security, the central point of collaboration must ensure a robust and seamless environment.

The Five Safes framework is a set of principles designed to ensure safe and secure access to data for researchers (Desai et al. 2016; Ritchie 2017). It consists of five elements: safe people, safe projects, safe settings, safe data, and safe output. Federated identity management (FIM) is a system that brings in the principals and addresses the challenges identified in the Five Safes framework and enables users to access multiple data sources – both open and secure with the same access credentials (Chadwick 2007; Shim et al. 2005). FIM involves the use of Identity Providers (IdPs) that store and manage user credentials, allowing users to access data from various Service Providers (SPs) without providing their credentials each time. This approach is based on mutual trust agreements between the IdP and the SPs. Federated identity allows for single sign-on (SSO) across different systems and applications, streamlining the authentication process for users.

The International Data Spaces Association (IDSA) are working to establish standards and protocols for trusted infrastructure organisations can use to share data with full control over access and use. This brings together services for data catalogues, vocabularies, governance, identity management and data exchange between disparate platforms. The IDSA Reference Architecture Model is gaining significant traction globally with an extensive list of partners including the Australian Research Data Commons.

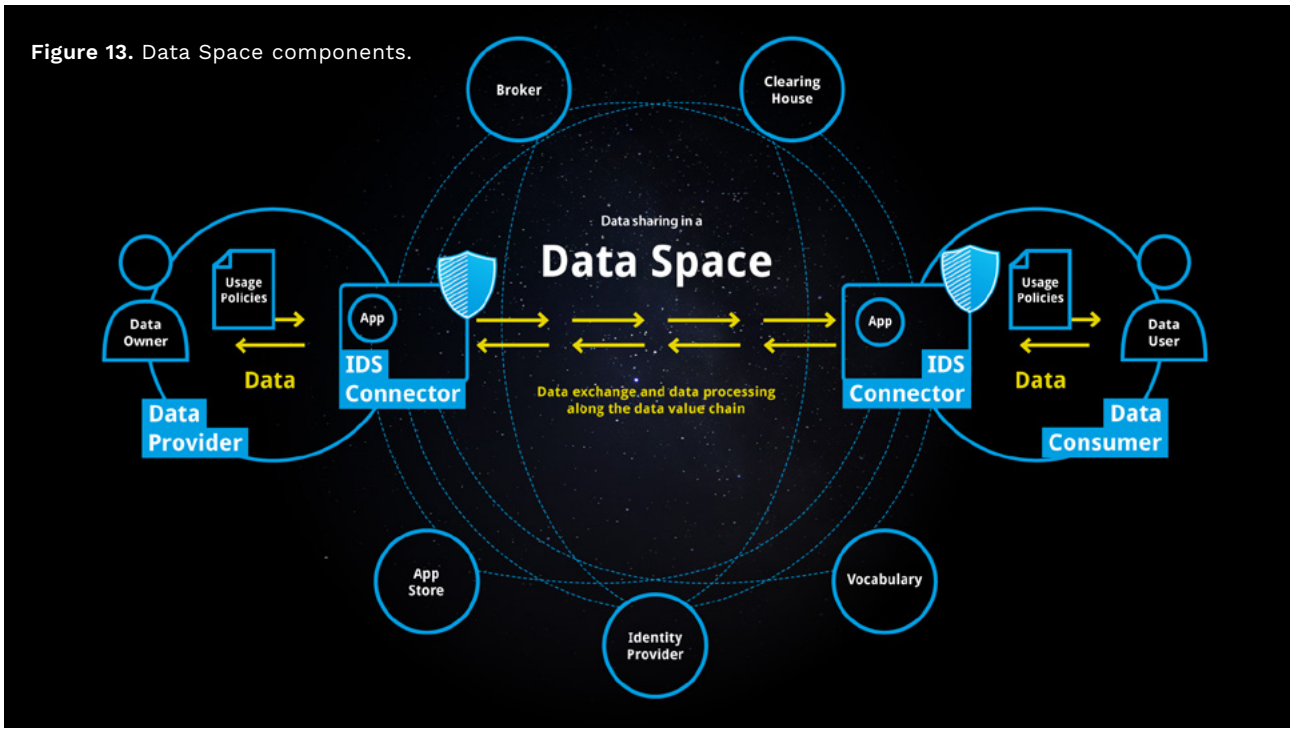


Figure 13. Data Space components.

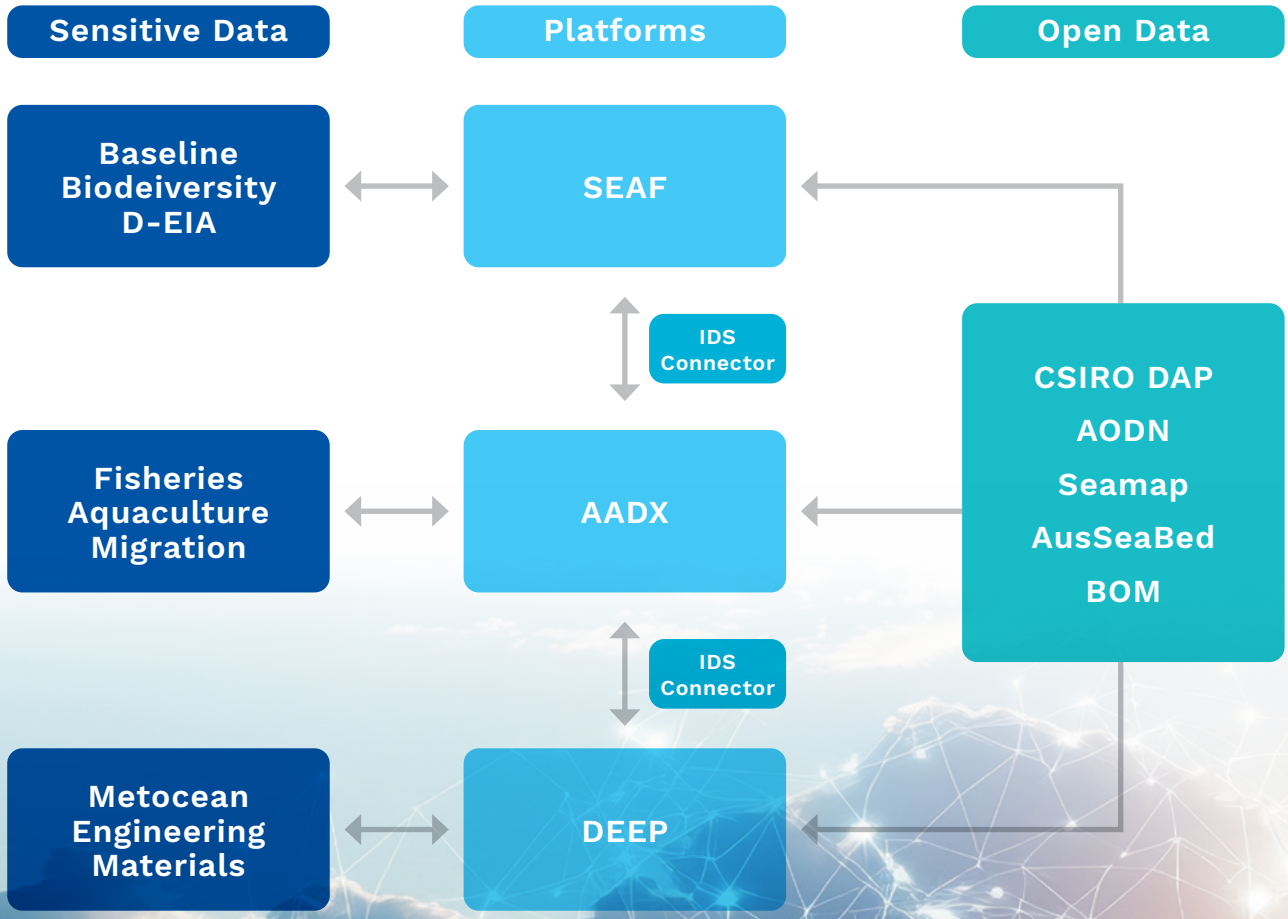
5. Conclusions and Recommendations

At the end of the design process, it was determined that the best way to maximise the value of and ensure longevity of CRC data was to adopt an internationally recognised standards based approach to data governance and a service to CRC participants that enables secure interoperability between existing and future data analytics platforms.

This was in preference to investing time and funding to develop and support the maintenance of the infrastructure requirements for a bespoke CRC data and analytics platform. Currently, the most viable standards available are being developed by the International Data Spaces Association (IDSA) and is gaining significant interest in the Australian research data community. These standards are presented as the preferred adoption for the CRC data governance framework.

An example concept design for an interoperable systems linking the three data sharing and analytics platforms uses the IDSA standards, rules and protocols to enable data exchange between systems maintaining trust and secure access (Figure 14).

Figure 14. Concept example showing CRC data infrastructure design for federated, integrated platforms using an identity service, secure protocols and other ISDA standards to connect environmental approval, fisheries and wind farm proponent data.



The following recommendations were made setting out the next steps in developing a robust data governance framework and service provision for the CRC:

1. Include mandatory step by step guidelines in the CRC data management plan for Research Programs to follow to ensure standards for data collection, storage, sharing identity and, where appropriate, an ultimate resting place in an approved open data platform.
2. Conduct an inventory of data generated from past and current research projects.
3. Contribute CRC generated data to a standards based data catalogue that aggregates to the Australian Ocean Data Network (AODN).
4. Push past and current data to an approved data platform with embargo management as needed and a plan for when it will be released as open data if appropriate.
5. Establish the rules of CRC data exchange to ensure standards based, secure, trusted and timely and efficient collaboration.
6. Develop and expand on existing technology to prototype data repository and analytics platform(s) for a blue economy case study with interoperability to a regulatory platform and freely available cloud-based research computing services. e.g. NeCTAR.
7. Execute a Capacity Building Program for the use of marine and maritime data and analytics tools targeted at Blue Economy growth and awareness to including on-line training modules and in-person training workshops.

5.1. Data lifecycle for data generated by CRC research projects

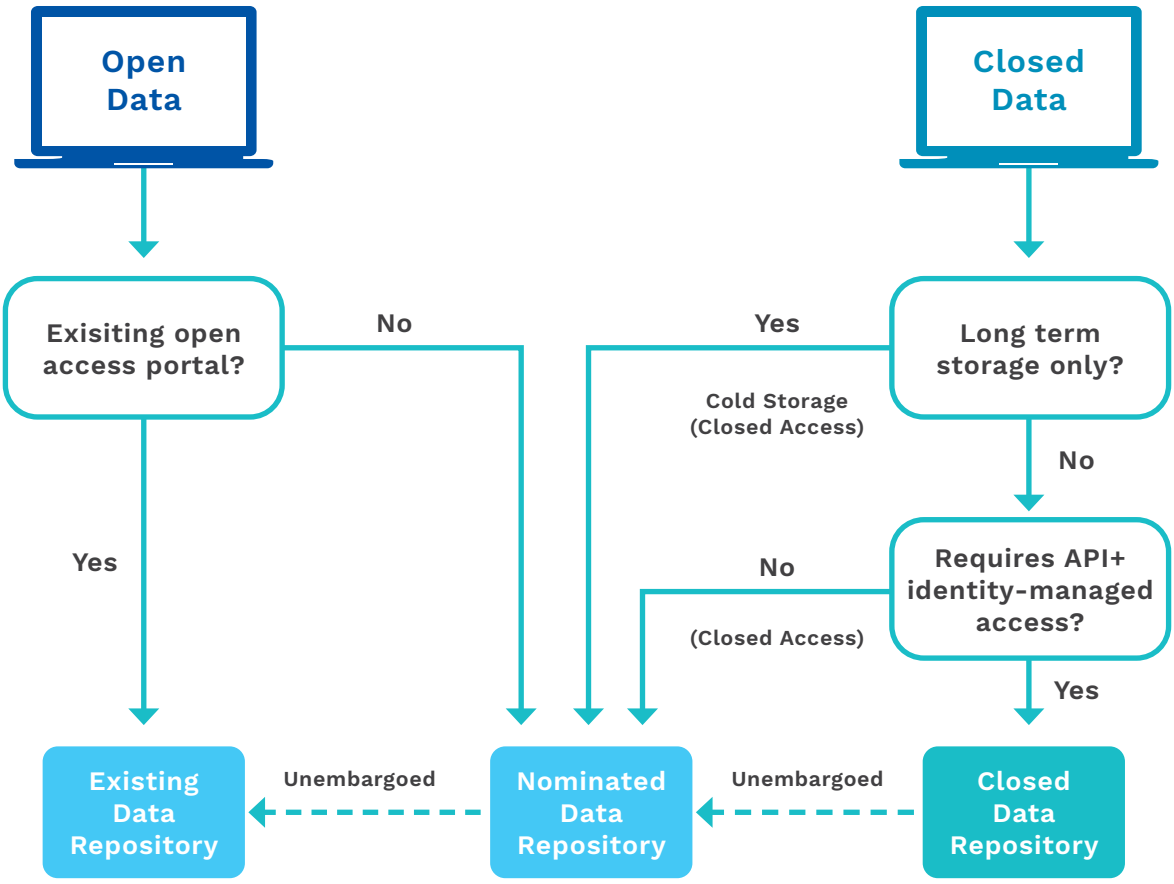
The CRC through its research projects will continue to generate data that will need to be managed and stored according to the life cycle required and the level of security needed. This may change through the life of the CRC. It is intended that as much of the CRC data generated will be made openly accessible at some stage of the project life cycle.

It is recommended that as part of the CRC Data Management Guidelines a decision tree for the data lifecycle for CRC generated data is developed. The decision tree will need to differentiate between secure to open data sources through data assurance steps to deliver data products, a range of data infrastructure to deliver data and associated products to a range of end users and use cases (Figure 15).

Processes for data archiving and sharing will be dependent on the type of data and will vary on a project-by-project basis as defined in the project Data Management Plan. However, as a general guide using storage on personal devices should be avoided. Where personal devices are used when offline, users will ensure that the data is backed up to secondary media on a regular basis and backed up to the appropriate storage facility at the earliest opportunity.

Open Data not under Embargo may be shared freely using appropriate mechanisms for licensing and use constraints. Where a Data User does not have access to compliant storage facilities, the Data Manager will provide advice on alternative facilities.

Figure 15. An example data journey decision tree for CRC data.



When well designed and implemented following the FAIR data principles, this end-to-end workflow would be capable of providing a homogenous, traceable, efficient and timely supply chain delivering data to meet both specific end use cases and more general user needs.

5.2. CRC Data Inventory

The CRC data inventory will need to include both past and current projects. It is recommended that the CRC Data Inventory will include information on datasets used or collected during the Project lifecycle including:.

- Data classification (Open Data, Sensitive Data, Embargoed Data)
- IP ownership
- Data use limitations for Sensitive Data and Embargoed Data
- Creative Commons' License applied to Open Data
- Data custodian and storage facility used
- Quality Assurance and Quality Control (QA/QC) procedures
- Methodology / Standard Operating Procedure used (where applicable)
- For Open Data (including data released from embargo), where will the Data be published
- Links to the Research Project Data Management Plan

5.3. CRC Standards Based Catalogue

At the conclusion of the inventory, metadata will be published for the CRC based on AODN and/or IODE standards where applicable. This catalogue of data will be made available through the AODN Portal.

Where applicable, Research Data file and storage formats will comply with established research methodologies.

Data storage and file formats used will be chosen and documented to ensure the likelihood of Data being readable and reusable in the future.

Where data includes code lists or vocabularies, preference will be given to vocabularies in established research methodologies or published in a relevant vocabulary service (e.g. Research Vocabularies Australia).

Where a vocabulary has been established as part of work undertaken by the BE CRC and the vocabulary has meaning or scope outside the BE CRC, consideration will be given to publishing the vocabulary through a relevant and publicly available vocabulary service (e.g. Research Vocabularies Australia).

Metadata will comply with a standard metadata schema compliant with the appropriate research discipline (e.g. ISO19115 for Data with a spatial context).

5.4. Publish CRC data on open source platforms

When CRC data is at the "Open Data" stage of the data journey where it can be made available through an open source commons licence a decision will be made to upload to the relevant database. The CRC will have a list of recommended data repositories (e.g. AODN) based on the data characterisation including type and intended use.

license to be used being the Attribution 4.0 International (CC BY 4.0) or its derivative.

Data repositories used must be able to aggregate metadata according to national discipline specific and generalised data portals (e.g. AODN Portal, Research Data Australia)

Preference will be given to data repositories providing interoperable, standards-based web service access to data (e.g. Open Geospatial Consortium (OGC) compliant web services).

5.5. Rules of CRC data exchange

It is recommended that the CRC adopt a FIM model to enable interoperability across multiple data sharing and analytics platforms for CRC participants. In the context of providing access to users to data, models, and applications within a Data Spaces framework, FIM can be implemented using policies, roles, and users to control access (see Table 6).

The roles and privileges of users are integrated securely across different domains, allowing for seamless and secure access to resources. This approach also involves distinct roles for Identity Providers and Relying Parties, with the former managing user identities and credentials, and the latter being responsible for authentication. Federated identity management alongside a Data Spaces framework provides a single point of integration and policy enforcement, reduces risk, and promotes secure sharing of information.

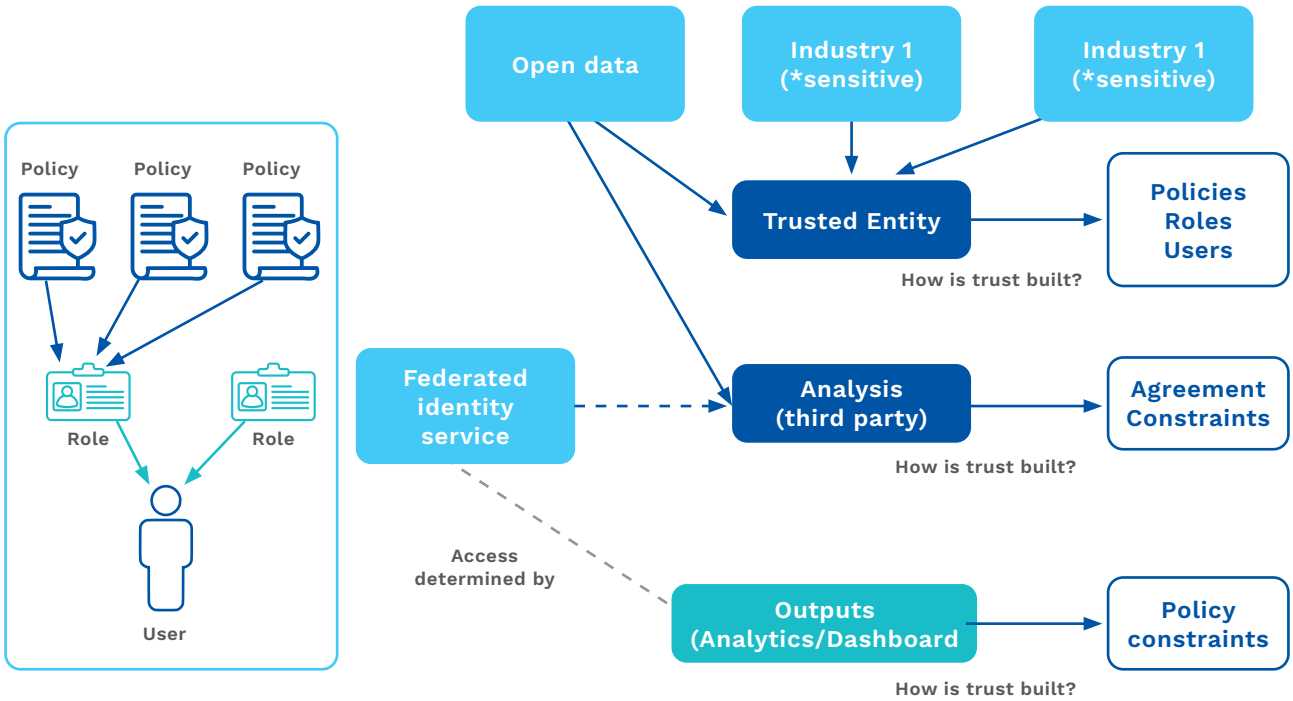
Table 6. Table 6: Policies, roles and users explained.

Policies	Policies define provide fine grained control over permissions for data, models, and applications. A policy can be comprised of access to one or more items – either data, models, or applications.
Roles	Roles are formed by combining one or more policies.
Users	Users are assigned roles, which in turn dictate their permissions and access.

Using policies, roles, and users within this framework provides a secure and efficient way to manage access to data, models, and applications across multiple domains and networks, while ensuring that users can access resources seamlessly and securely.

This kind of structure allows for fine grained control over the permissions assigned to users as depicted in the conceptual diagram in Figure 16 (following page).

Figure 16. Federated Identity Management for Open and Secure Industry data.



5.6. CRC Interoperable Infrastructure Prototype for a Blue Economy Case Study

The DIDBE project team recommends developing a data infrastructure prototype for a blue economy case study (BECS) to demonstrate the interoperability between data from government regulations, industry and researchers.

The BECS would be designed to test and evaluate the data workflow starting with the data collected from a CRC research project and adding the appropriate regulatory approvals and or standards reporting to a web-based data portal that enables a data analytics facility to aid in research projects as well as for industry access and community consultation.

The BECS infrastructure will be designed to manage the data, and serve applications on the cloud to comprise of several services:

- Storage and Databases
- Computing Clusters and Processing
- Analysis tools
- Third party applications and custom dashboards

The purpose of the prototype would be to incorporate as many of the “must have” design features identified by the CRC stakeholders as listed in Sections 3.5 and 3.6

5.7. Conclusion

The DIDBE project has laid the groundwork for a scalable, flexible, and robust data architecture to support the CRC's mission of fostering sustainable growth in the Blue Economy.

By leveraging existing platforms and focusing on interoperability through a secure Data Spaces service, the recommended approach ensures long-term usability, security, and adaptability to diverse data needs. These efforts, coupled with effective governance (including policies and guidelines), a comprehensive data inventory, and targeted capacity-building programs, will enable CRC participants to effectively manage and utilise blue economy data, driving evidence-based decision-making and maximising the value of CRC research for the Blue Economy's future.

The design put forward was designed to respond to changes in the CRC requirements as the CRC matures and adapts to changes in the marine data landscape as it progresses to greater digitisation of data sharing in a trusted and secure collaborative space.

6. Acknowledgements

The authors acknowledge the financial support of the Blue Economy Cooperative Research Centre, established and supported under the Australian Government's Cooperative Research Centres Program, grant number CRC-20180101.

The authors would like to thank all participants of the Data Infrastructure Design Workshops held from May 2023 to January 2024. Their valuable experience, insights, ideas, and contributions have shaped the results, discussions and recommendations listed in this report.

We also acknowledge and thank Kyaw Kyaw Soe Hlaing for providing information on the AADX and Eratos.

7. References

Chadwick, D.W., 2007. Federated identity management. In International School on Foundations of Security Analysis and Design (pp. 96-120). Berlin, Heidelberg: Springer Berlin Heidelberg.

Desai, T., Ritchie, F. and Welpton, R., 2016. Five safes: designing data access for research. Economics Working Paper Series, 1601, p.28.

Martínez-Vázquez, RM, Milán-García, J, de Pablo Valenciano J (2021) Challenges of the Blue Economy: evidence and research trends. Environmental Sciences Europe 33:1-17

Ritchie, F., 2017. The 'Five Safes': a framework for planning, designing and evaluating data access solutions. Data for Policy.

Sepúlveda, Joel. (2023). FROM BIG DATA TO SMART DATA. OPPORTUNITIES FOR ENTREPRENEURS USING DATA SPACE ECOSYSTEM APPROACH. Journal of Entrepreneurial Researchers. 1. 87-96. 10.29073/jer.v1i2.19.

Shim, S.S., Bhalla, G. and Pendyala, V., 2005. Federated identity management. Computer, 38(12), pp.120-122.

Tenopir C, Allard S, Douglass K, Aydinoglu AU, Wu L, Read E, Manoff M, Fram M (2011) Data sharing by scientists: practices and perceptions. PLOS ONE 6

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E. and Bouwman, J., 2016. The FAIR Guiding Principles for scientific data management and stewardship. Scientific data, 3(1), pp.1-9.

BLUE ECONOMY
COOPERATIVE RESEARCH CENTRE

Blue Economy CRC

PO Box 897, Launceston, Tasmania 7250

www.blueeconomycrc.com.au

enquiries@blueeconomycrc.com.au

Australian Government
Department of Industry, Science and Resources

Cooperative Research Centres Program

ISBN: 978-1-922822-25-3