Net Zero Oceanographic Capability - Scoping Study

WP6: Future Data Ecosystems

Work Package Leads: **John Siddorn, Alvaro Lorenzo Lopez**

Contributing Authors: **Justin Buck, Fiona Carse**

Date: **August 2021**

In 2019, UKRI (NERC) commissioned the National Oceanography Centre to identify the options for developing a world-class oceanographic capability with a reduced carbon footprint by presenting a range of options for transitioning to low or zero carbon capabilities. 6 work packages were initiated to examine the science and policy drivers for a future research capability and the various technologies that could enable the capability. The findings of the 6 work packages and a number of independent reports commissioned under the NZOC banner were combined in the NZOC Summary Report that provides more information about the project.

This report covers the detailed findings of Work Package 6: Future Data Ecosystems.

NET ZERO OCEANOGRAPHIC CAPABILITY NZOC

National Oceanography Centre

UKRI Natural Environment Research Council

Contents

Executive Summary and Key Recommendations 3
Introduction 4
Review Scope 4
Net Zero Definition 5
Baseline Review 2020 and Gap Analysis 6
Expedition Workflow 7
Marine Facilities Planning 8
Inventory Management System 9
Observation Data Flow: collection to storage and use 9
Near Real-Time Data Flows 10
Autonomous and non-Ship Based Data Flow 11
Ship Based Data Ecosystem 13
Data Archiving and Delivery 14
Gaps and Requirements: The Science Community 16
Gaps and Requirements: Summary of Data Ecosystem Workshop Outcomes 17
Skills 18
Horizon Scan 2020-2035 20
An Envisaged Data Ecosystem 20
The Physical Layer (the Ocean) 22
Virtual Space 23
Data Analytics Layer 27
Data Management 27
Applications Layer 28
Infrastructure 30
Communication Network 30
Architecture and Technology to Meet Zero Carbon 32
Software 32
Ship based infrastructure 33
Edge Computing 34
Cyber security 35
Data processing 36
From Sensor to Data Lake 36
From Data Lake to Data Warehouse 37
Data Management 37
Data Sciences (AI/ML) 39

Known Unknowns 40
Equality, Diversity and Inclusion 41
Scientific Data Licensing and Intellectual Property Rights 41
Review of Commercial Priorities and Opportunities for Collaboration 42
Big Tech 42
Academia 43
Defence 43
International Activities 44
Marine Industry 44
Citizen Science 45
Carbon Calculation 45
Scope 1 – direct fuel costs 46
Scope 2 – indirect emissions 47
Scope 3 – supply chain and business travel carbon costs 47
Recommendations Summarised 48
Pilot Studies 48
Digital Skills 48
Best Practice 48
Hardware and Telecommunications 49
Data Management 49
Software 50
Data Sciences and Modelling 50
Collaboration 50
References 51
Annex: Cybersecurity Report 52

Executive Summary and Key Recommendations

The long-term vision for NZOC data ecosystem is for the Research Expedition of the Future to be supported through a data ecosystem that allows seamless transfer of data from a multitude of sensors through to data repositories where it can then be accessed by a broad community of users. The data management process will incorporate quality and metadata control that is often presently missing and will ensure data repositories are able to offer a wide range of users' data with full provenance, creating significant additional value from the observations.

This data ecosystem will support the research expedition technicians and scientists to get the most out of the expedition. It will guide platform deployment and allow fast turnaround scientific analysis of the data being collected to allow on-the-fly decision making using Digital Twin technologies. The data ecosystem will support improved interaction with the research expedition data ashore and on board, allowing increasing engagement with scientists ashore and supporting enhanced science participation in the expedition whilst allowing reduced science presence on board, allowing reduced associated carbon consuming travel.

By 2030, data collected from ship or from autonomous platforms will be collected and transformed into digital information in real-time and transmitted with appropriate meta-data that allows them to be fully traceable from collection to use with the minimum of manual intervention and with the shortest possible latency.

To realise this ambition a phased approach needs to be taken with pilot projects started within the early years of the project to allow readiness for full implementation on the 2030 timescales:

Year 1 – 5: [2022 – 2027]

Key recommendation 1: Develop a pilot study or studies that demonstrates the capability of a digital twin ecosystem to improve decision making on a research expedition. The pilot(s) will include data management, data processing, data analytics and hardware and software developments following recommendations from the NZOC data ecosystem report, to inform the future NZOC data ecosystem.

Timescales: Building components iteratively within an agile framework, delivering value to users throughout years 1 – 5. Proof of concept demonstration around year 3 (aligned with a programmed expedition).

Partners: EPSRC (to support the data skills), big tech (for the computing infrastructure), industrial partners, Met Office, RN / DSTL and NERC partners.

The pilot study or studies will need to be defined in detail but should follow the recommendations in this report on pilot studies with the aim of defining the long-term data ecosystem (see below, Years 6 – 15).

The aim of the pilot studies will be to develop;

- an end-to-end data management culture,
- the Hardware and Telecommunication,
- the Data Management,
- the Software, and
- Data Science approaches.

The pilot studies will inform, and follow, best practice for scientific observing data ecosystem and support the growth of collaboration needed to have the required data ecosystem for the NZOC of the future.

Key Recommendation 2: Commission a skills audit and create a skills plan for the future generation of the NZOC workforce to support Digital Twins, including consideration of cross-Council skills fertilisation.

Timescales: By month 6, and at month 66 learning from pilot studies.

To meet the needs of the NZOC data ecosystem there will need to be investment in increasing the digital skills of the ocean science community and an increased focus on Research Software Engineer careers.

Key recommendation 3: UKRI programmes on digital twins follow international guidance on FAIR, CARE and TRUST[1] principles for data delivery and stewardship.

Key recommendation 4: Sensor / platform development should include the end-to-end data management as an intrinsic, and where needed funded, part of the design from the outset.

Year 6 – 15: [2028 – 2037]

Key Recommendation 5: Develop the full NZOC data ecosystem, informed by the pilot study/studies conducted in previous years, in time to inform and be implemented for the RRS James Cook replacement.

This NZOC data ecosystem will evolve in the detail of how it operates, the technologies it uses and the scope as the pilot studies inform requirements and best practice, and as the science need evolves. However, the work done in this study has highlighted some fundamental characteristics of the data ecosystem that we need to have in place for the RRS James Cook replacement to support Net Zero operations:

- The Data Ecosystem will include a Digital Twin that allows command-and-control of autonomy from both a ship (if deployed) and from shore;
- This Digital Twin will also support the Research Expedition Scientists in planning and undertaking the data collection, again from ship and from shore;
- The Data Ecosystem will include automatic processing of data (including the using AI where appropriate) and deliver the research data to data centres fully FAIR compliant

The above will enable a reduced crew Research Expedition, as well as enhancing the value of the data collected on that expedition through increased utility of and accessibility to the data.

To enable this a computer architecture that supports distributed (edge) computing from the sensor to the ship to the shore with high bandwidth telecommunications infrastructure is needed. Irrespective of developments made to push computing to the “edge”, reducing bandwidth, communication infrastructure will be a limiting factor and developments to this critical enabler will be needed to support the system.

As with any data management system, alongside the infrastructural developments there will need to be advances in the data management system to support the FAIR (and TRUST and CARE) needs of the system. This needs to be an intrinsic, embedded and cross-cutting part of the NZOC capability from sensor / platform to ship to data centre. This data management system must be part of a broader information management framework that supports interoperability across diverse thematic areas.

Introduction

Review Scope

This review outlines the present data ecosystem that supports National Marine Facilities capability (moored and autonomous sensors and crewed research ships) and explores the future Net Zero Oceanographic Capability data ecosystem that will meet stakeholder needs in a Net Zero context.

[1] FAIR (Findable Accessible Interoperable Reusable), CARE (Collective benefit, Authority to control, Responsibility, and Ethics), TRUST (Transparency, Responsibility, User focus, Sustainability and Technology)

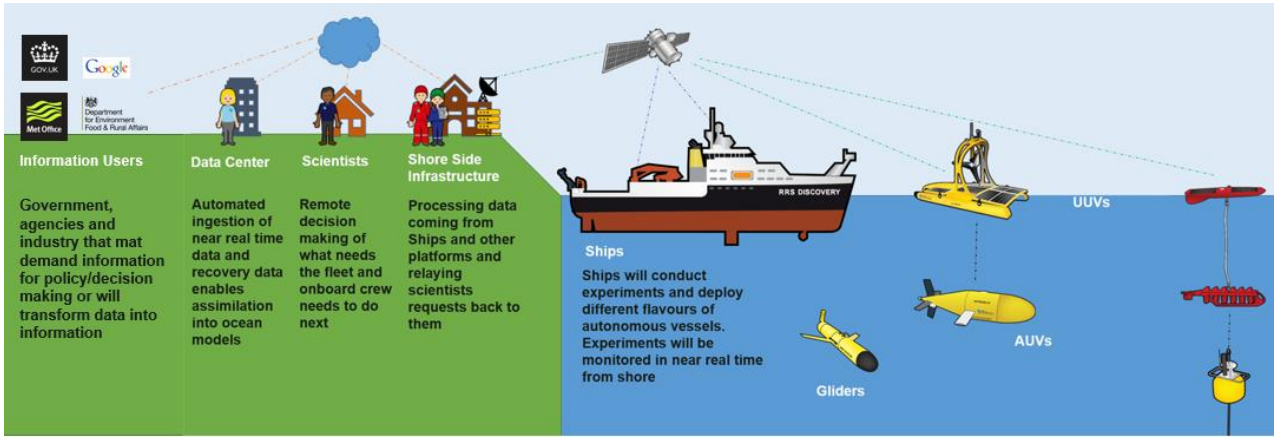


Figure 1: Schematic of a potential future Data Ecosystem for NZOC

This report will identify the present approach to a data ecosystems from collection through to processing and dissemination and identify how to develop the future data ecosystem to meet the requirements of the NZOC users / stakeholders, whilst meeting the Net Zero (defined below) ambition of the NZOC programme.

The ambition for the NZOC project is to maintain, and enhance where possible, the current capabilities whilst ensuring the carbon cost meets the NZOC ambitions of net zero.

There are many areas it may be possible to enhance the current capability, including the use of predictive methods to optimise the deployment of research infrastructure, which will be explored here. Following the description in the national data strategy[2] a data ecosystem encompasses the cyber infrastructure, documentation, and methodologies needed to build the end-to-end delivery of high-quality scientific data, including the planning of operations, data storing and data processing. In the context of the Net Zero Ocean Capability project, and the Data Ecosystem supporting the Research Infrastructure of the future, that is taken to include the digital infrastructure to directly support the data collection from ships and autonomous vehicles.

This data ecosystem includes:

- the scientific software and architecture on board the ship that facilitates research data collection on the NZOC
- The data infrastructure and communication between AUV, ship and other data systems (including, but not necessarily exclusively, archiving facilities) that support research data collection on the NZOC
- The data infrastructure and communications to allow scientific and other users, onshore and onboard, to engage with the observation collection process.
- Data infrastructure and software to support public good and commercial (where appropriate) re-use of NZOC data and capabilities

Net Zero Definition

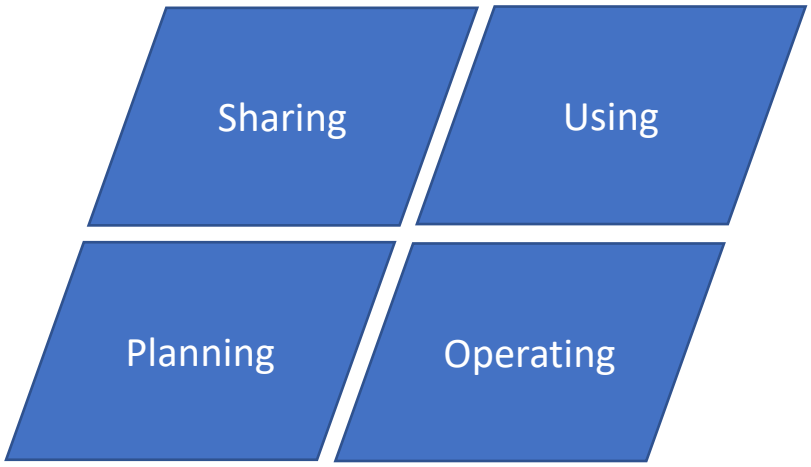
The ambition of the data ecosystem activity in the NZOC programme is to ensure that when planning for newly commissioned oceanographic capability the data ecosystem is Net Zero, or as close to it is practicable whilst continuing to enable delivery against our research and innovation mission.

[2] https://www.gov.uk/government/publications/uk-national-data-strategy/national-data-strategy

The data ecosystem has three significant roles to play in determining whether NZOC is able to meet its Net Zero ambitions; the carbon cost of the data ecosystem itself (scope 2, or scope 3 emissions if bought in services are used), the role of the data ecosystem in supporting the reduction of business travel (scope 3 emissions) and the contribution of the data ecosystem to ships' fuel emissions, including considerations around the amount of ship travel per unit useful research data produced (scope 1 emissions).

Baseline Review 2020 and Gap Analysis

The extent of the data ecosystem, and the interfaces that it needs to support, are defined through considering the different phases of a research expedition. The concept of the 4-screens supporting the Research Expedition is helpful in this context:



1) **Planning**: the data ecosystem needs to allow access to ocean data required to support effective expedition planning through tools such as the National Marine Facilities and Ships and Marine Equipment portals[3]. It does not however, include, these portals.
2) **Operating**: the data ecosystem needs to include the on-board (ship and autonomous vehicle) processing and data communications required to effectively manage scientific data processing and data flows on and between ships, autonomous vehicles and shore to guide the platforms, creating a smart command and control capability.
3) **Using:** the data ecosystem needs to develop the systems and processes required to for real-time interaction with the data whilst the research expedition is underway. This will include viewing and manipulating the data, as well as making predictions / simulations. The data ecosystem will ensure a user interface is available both from the ship and from facilities ashore to allow scientific users to interrogate the data as it arrives to allow on-going and improved decision making based on current information.
4) **Sharing:** The data ecosystem needs to deliver the scientific data to an environment where it can be used for its intended purpose and be shared widely for use and reuse. That requires the capture of appropriate metadata to maximise the use and to ensure appropriate long-term archiving. The data ecosystem needs facilities to process data to a level that data archiving is possible with minimal manual intervention. It will also include the physical infrastructure, for example telecommunications connections, to transmit the data to the data archive facilities.

[3] Marine Facilities Planning

The NZOC data ecosystem does not itself encompass the long-term archive facilities required for keeping the data in prosperity but needs to include storage and backup facilities to manage the data management required during the Research Expedition.

From the description above it is, we hope, clear where the NZOC data ecosystem and the broader ocean science and services data ecosystems start and stop. It is critical that the interfaces to the NZOC data ecosystem are well-designed and agreed, but it is not within the scope of this work to develop, for example, the data repository services or the science tools needed for the broader scientific function of NERC, at least in part because there are a broad range of repositories and capabilities that the NZOC data ecosystem needs to interface to and the scope of this report needs to be constrained to a manageable problem. It is, however, within the scope of this work to ensure that the wider needs of the community for the data as it is passed from the NZOC data ecosystem to the broader community are considered and these requirements are recognised and actioned. This is critical in ensuring the value of the NZOC data is maximised and is where the greatest value of a well-designed data ecosystem will be realised.

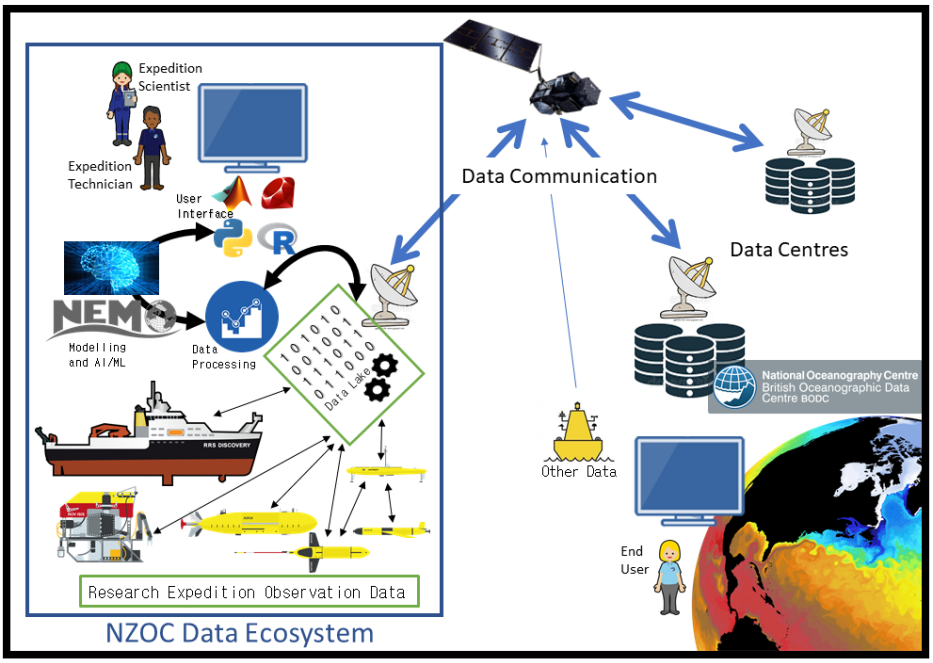


Figure 2: Schematic of the NZOC data ecosystem in the context of the broader data environment.

This chapter reviews the present state of the Research Expedition data ecosystem, and assesses where the gaps are at present in fulfilling the needs of a future NZOC data ecosystem,

Expedition Workflow

The NZOC Data Ecosystem should support the end-to-end workflow for information and data products required for and produced by a scientific research expedition, as described in the top-level schematic in Figure 3.

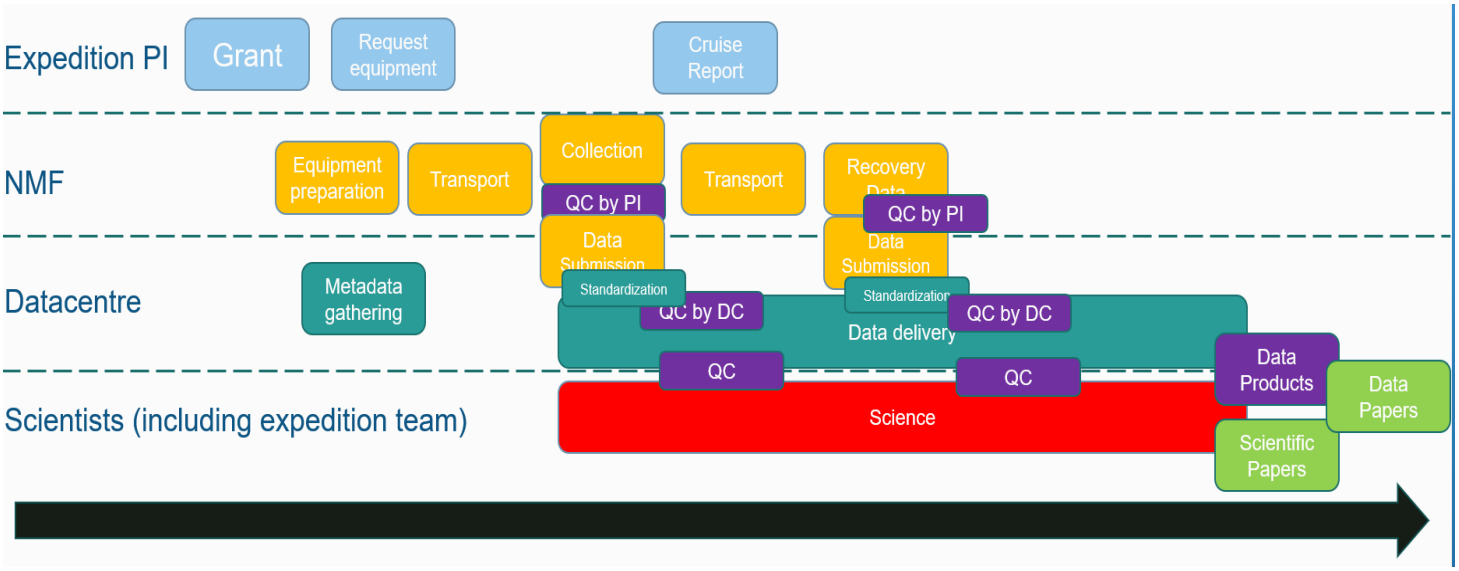


Figure 3: Expedition workflow as of 2020

The process from the grant application to the requisition of required kit, the expedition itself through to the data archiving and beyond is well-described in workflows but these are largely manual. This has significant implications:

1. A significant overhead on PIs, cruise scientists and NMF technologists to provide and access cruise planning data.
2. The likelihood of "lost" data either through poorly linked metadata or through a lack of following of protocols for logging data.
3. Inefficiencies during the whole research project workflow, making difficult for the different actors to understand what the situations is without reaching to individuals in charge of different tasks. This problem is accentuated by the fact that different activities are performed by different departments or institutions.

THE NZOC Data Ecosystem can improve significantly on the baseline through an end-to-end design of the Expedition Workflow with integrated software linking the different elements of the workflow.

Gap 1: Expedition Workflow is a standalone system with manual input needed to transfer metadata/information in and out of it.

Marine Facilities Planning

A Marine Facilities Planning (MFP) portal exists, allowing the science community to request assets time and technical expertise from the National Marine Equipment Pool (NMEP), including the two large research vessels RRS Discovery and RRS James Cook, the national pool of sensors, and the fleet of operated and autonomous robots. The MFP portal is the tool used by NMF staff to plan the physical infrastructure required for expeditions.

The MFP portal is one of the 2 tools of the NMEP ecosystem (the second one is the Inventory Management System, and it will be covered in the next section). This ecosystem is standalone with no further automatic connections with other parts of the delivery chain. All metadata needs to be communicated to different actors "manually" or, in the best-case scenario, the system send emails to people involved in the execution of the workflows that are managed by the system. No machine-to-machine interfaces are available to generate

automation or to extract the data and metadata stored into the system and that is required to generate rich and usable datasets at the end of the data chain. This requires the human intervention of many people across NMF and the British Oceanographic Data Centre to transfer the data and metadata to different tools that manage the data delivery, with the subsequent waste of resources and the potential loss of information. The good news is the MFP is a modern information system developed by a very capable software team and adapting it to achieve the required automation and exposure of data will be achievable with the right strategical investment.

Gap 2: The Marine Facilities Planning Portal needs machine-to-machine interfaces to be interoperable and allowing the integration on the NZOC data ecosystem.

Inventory Management System

Alongside the MFP Portal exists the Inventory Management System (IMS). This tool is part of the same suite of products, developed by the same company and connected to each other. The IMS is used by NMF as an inventory system of all sensors, vehicles and parts connected to vehicles and sensors. The original and main task of the IMS is to allow NMF to track the different items for customs purposes; the kit used by NMF is of high monetary value and all the movements of that equipment needs to be carefully monitored when travelling in and out of the country (which is very common).

As a by-product of the tracking of all the items, the IMS contains valuable information (metadata) for the data assembly: which sensor or device is used in which campaign, what calibrations are used etc. The IMS is also fully integrated with the MFP Portal, so the information between both applications is well synchronized, but this also makes the IMS to inherit the same shortfalls of the MFP portal, including the lack of interoperability with external tool (at the time of writing this report). It is worth to note that those interoperability issues are not endemic of the UK NMF and in fact IMS containing the whole of the NMEP assets presents a unique opportunity for the UK community; the IMS and MFP are an ecosystem developed following modern software engineering techniques, solving the issues of creating interoperability between them and other system is within reach for NOC, allowing the creation of a very consistent data delivery chain.

Gap 3: The Inventory Management System Portal needs machine-to-machine interfaces to be interoperable and allowing the integration on the NZOC data ecosystem.

Observation Data Flow: collection to storage and use

There are presently a range of different data chains depending upon use and observation platform and timeliness.

It is common to categorize the timeliness in two different groups:

1. Near real-time data (NRT): observations that reach users while platforms are deployed with limited delay / latency. Examples of NRT data sources include Argo floats, cabled observatories, many moorings and autonomous platform or a ship data sent using satellite communications.
2. Recovery data: many platforms can't send all the data (or any data) in NRT due to lack the hardware to transmit the observations or having insufficient bandwidth to send back the sampled data. This data only becomes available on "recovery" of the sensor. There may be a significant delay between the recovery of the platform and the data being available to the end-users for a variety of reasons.

Near Real-Time Data Flows

Near Real-Time data flows send the data back to shore while platforms are still in the water. The platforms described here are the ones managed by NMF and BODC; Ship underway data, gliders, Argo floats, PAP mooring and sea level instruments. The data coming from these platforms data go both to data archiving facilities and on to be used by real-time data consumers with automated QC and file conversion but without manual intervention. The present NRT data chain is described in Figure 4.

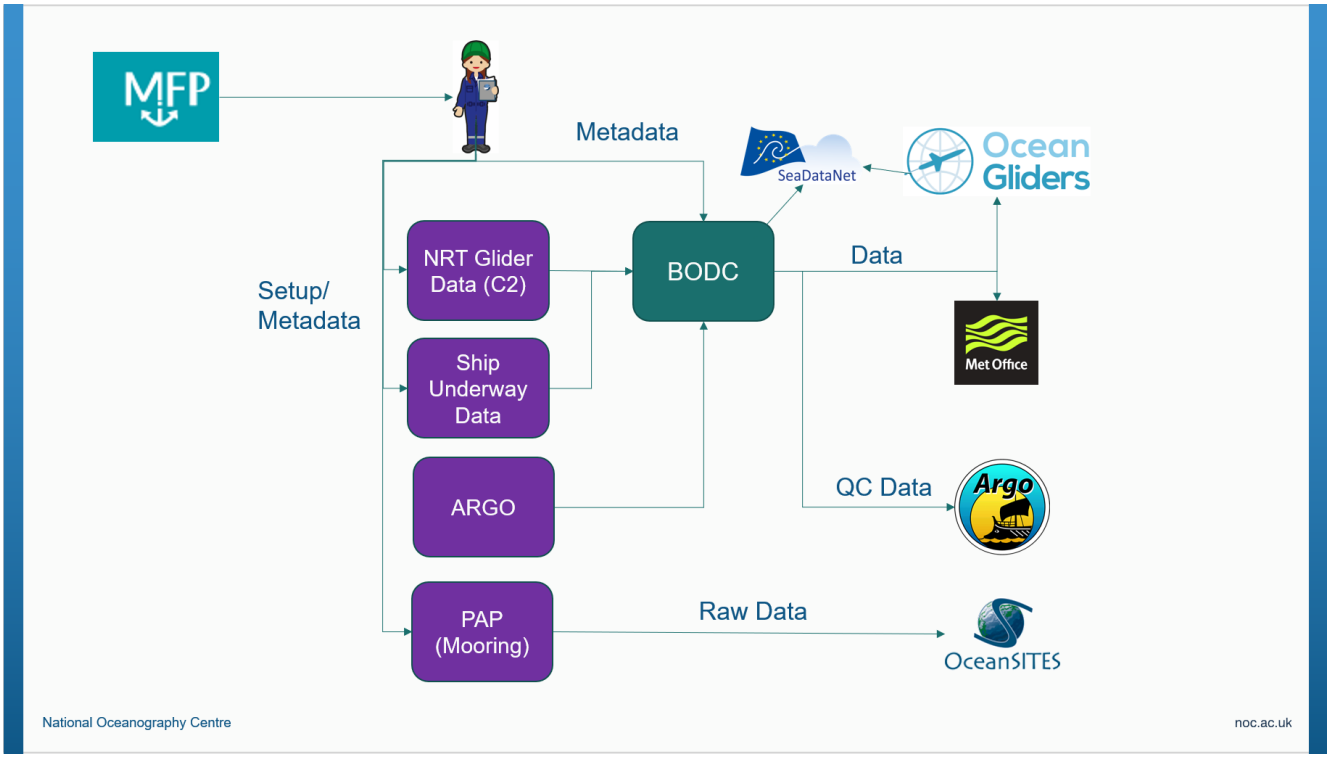


Figure 4: Schematic of NRT data chains

The Argo system has been designed and implemented as part of the international Argo program and Euro-Argo, currently offering the most mature of all the NRT chains. The glider component has been developed under the Oceanids program, modernizing many of the underlying systems. Those new developed structures will be used in the future to build new data chains for Ship underway data and any new platform.

Finally, for completion we show the NRT workflow for the Sea Level data in Figure 5.

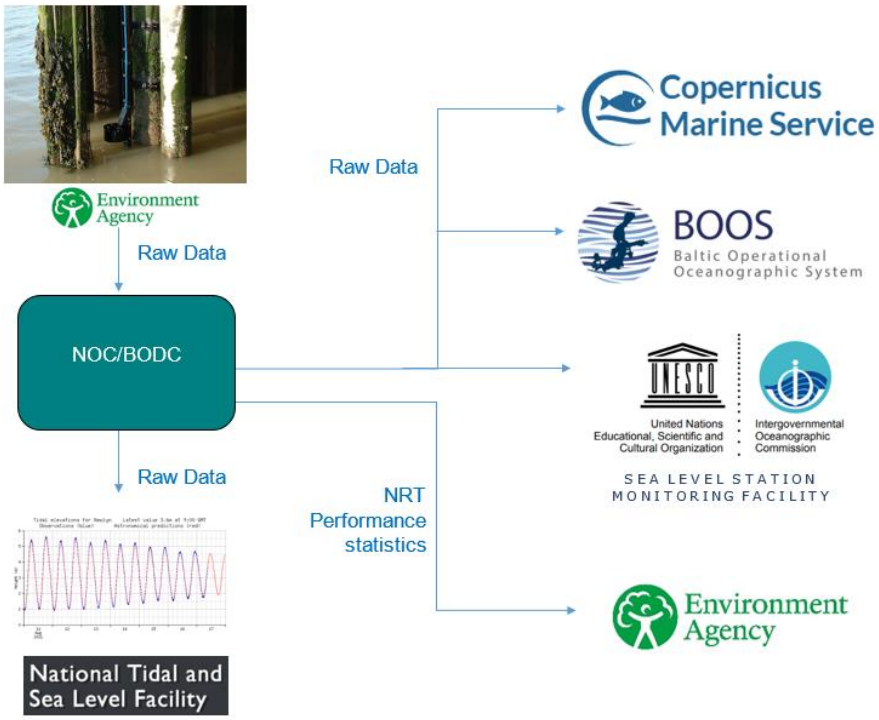


Figure 5: Sea Level NRT workflow

The Environment Agency's (EA) current telemetry system polls the tide gauge data loggers every 15 minutes and pulls data to their server. The EA then push NRT tide gauge data as xml files to NOC. This system is being reviewed as part of the EA's Future of National Telemetry (FoNT) project and is likely to move to a pull model rather than the current push. Data are loaded into an NOC database and then displayed on the National Tidal and Sea Level Facility (NTSLF) website. Data are also placed on the NOC FTP site for networks such as CMEMS, NOOS and the IOC Sea Level Station Monitoring Facility to pick up and display on their portals. The xml files do not use standard controlled vocabularies and the NRT data may be a combination of more than one type of sensor.

While almost all the NRT data chains are in an operational state, there is a problem of technological coherence; they have been developed at different times, with different purposes, making it difficult to discover them from a single point of entrance and using the same tools.

Gap 4: NRT data chains are incoherent between each other making difficult to access from different communities not familiar with that particular data chain.

Autonomous and non-Ship Based Data Flow

Several separate data chains have evolved for different autonomous or non-ship-based platforms.

For the big AUVs (Figure 6: Recovery Data flow for big AUVs), no data (or very limited, usually just telemetry) is sent back in NRT. Once the AUVs are recovered, the vehicle storage is downloaded, and copies of the data are provided to the PI. Some of the data gets stored by NMF for operational purposes.

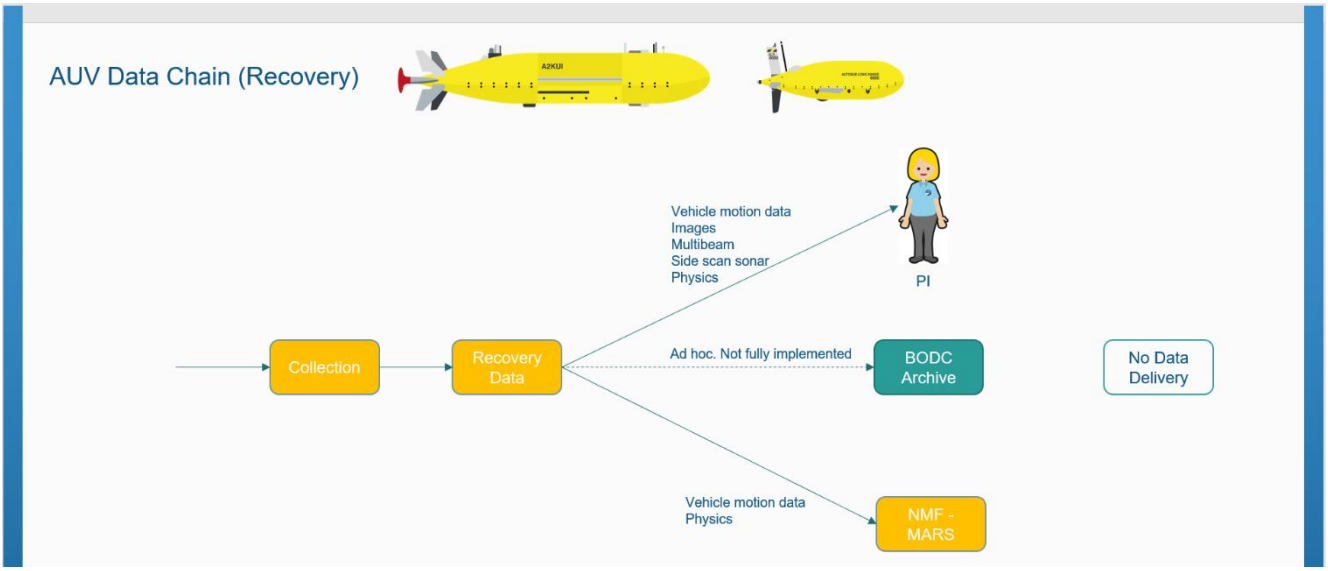


Figure 6: Recovery Data flow for big AUVs

Some gliders collect more data than it is possible to send in NRT. Once the gliders are back, the data is downloaded and given to the PI, archived in the BODC and stored by NMF for operations purposes. Data delivery mechanisms are currently under development.

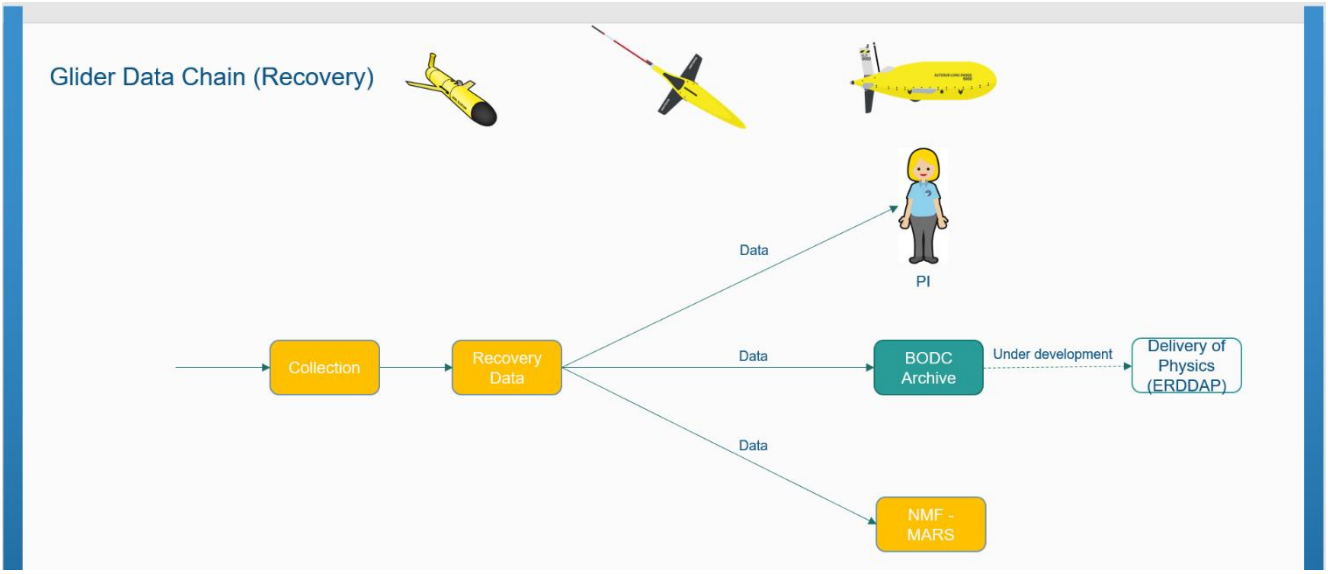


Figure 7: Recovery Data flow for gliders

World class ROVs generate very big quantities of data that are usually not easy to send in NRT. All data is copied when the research expedition has finished, providing that data to the PI and archiving it on BODC. NMF keeps a subset of the data (engineering data) for operational purposes. Data delivery mechanisms are currently under development.

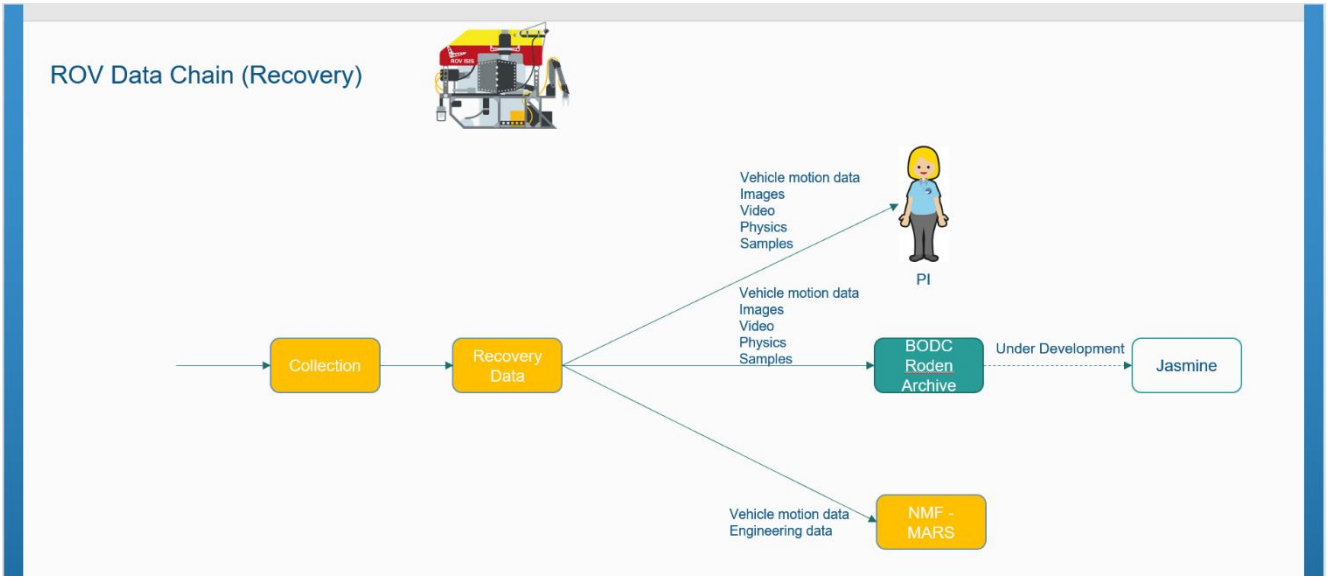


Figure 8: Data flow for ROVs

Ship Based Data Ecosystem

The current generation of ships are complex systems that include multiple workflows involving sensors and different sampling techniques (coring, trawling etc). This ecosystem can be seen as a real natural ecosystem in terms that has grown with time in an organic way serving multiple stakeholders and needs. Ships also integrate many ad-hoc sensors and activities in every new expedition, making it difficult to model the ecosystem at any moment in time. The task of integrating different instruments has ranging difficulty, with some of them offering very poor interoperability and documentation.

The NMF Ship Systems team has made very important steps in generating a ship-based infrastructure to allow the unification of systems and implement operational workflows, increasing the cohesion of the outputs. At the time of writing this document there are continuous cross-developments between the NMF Ships Systems team and the data centre to integrate data and metadata coming from the ships into the NRT workflows developed under Oceanids (see Near Real-Time Data Flows).

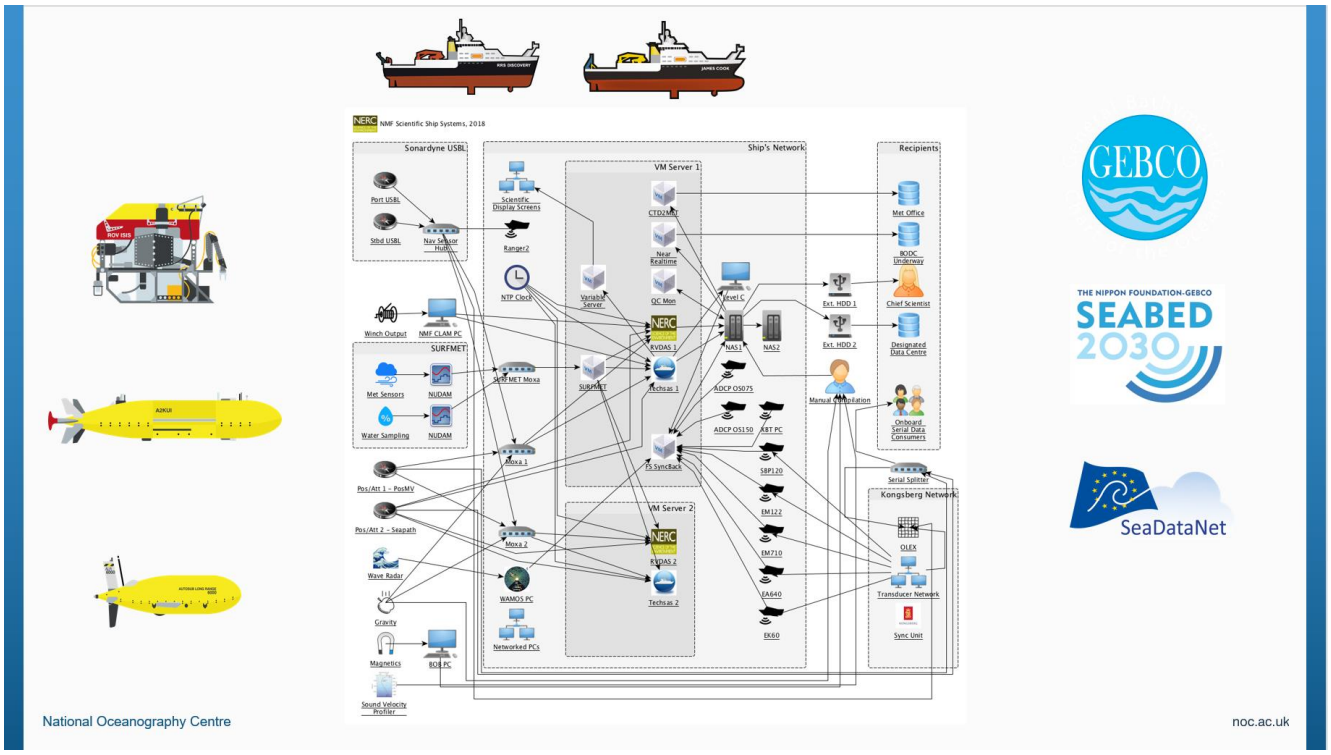


Figure 9: Combined data flow for recovery data, including the role of ships

Data Archiving and Delivery

Under NERC project terms, and following international best-practice, data collected during NMF expeditions should be collected and stored for prosperity. Under the NMF workflow for expeditions it is the Expedition's PI that has the responsibility to ensure all data collected is lodged with an appropriate data centre. For marine data this is BODC and BOSCORF for sediment cores, both hosted by NOC. For non-NERC data, or data outside the remit (or capability) of BODC and BOSCORF there may be other data centres that need to be provided the data.

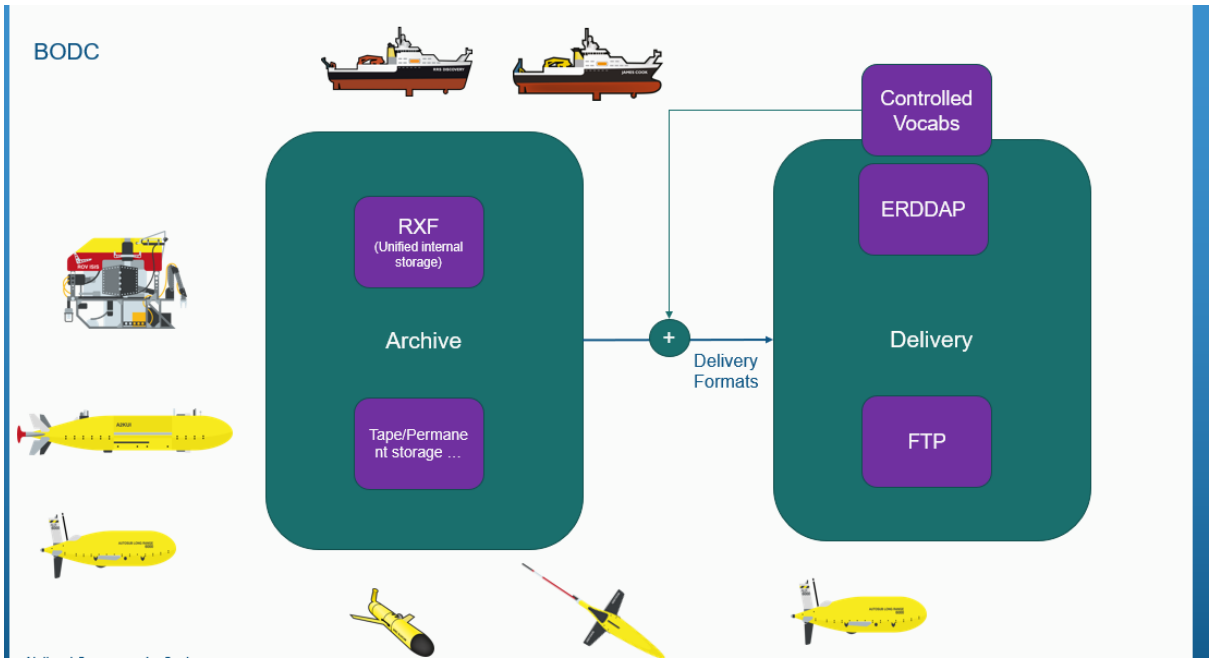


Figure 10: Schematic of the role of BODC in the data management process

Figure 10 shows the current (August 2021) generic BODC data flow for autonomous platforms. The data flow was originally design for autonomous platforms (in particular gliders), but the generic and modular nature of the design allows the expansion of the system to new platforms and sensors.

The workflow can simplistically seen as 2 main blocks, and a third sub-block:

1. An archiving section, where all data arrives and get transformed into standard data structures. For timeseries and point observation measurements this is done using a NetCDF derived format called RXF. It is under the Archive block all traditional data management activities are considered, like permanent storage (BODC follows a 3 copies policy, 2 on-site and 1 off-site). The data on the Archive is not directly accessible to users, and this has become a problem from the data centre as times has changed and the demand of data usage has increased. To solve that we will need to explain the right-hand panel in Figure 10.
2. Data delivery. All the activities that aim to provide data to end-users fall in this section. BODC has been providing data to users for a long-time using technologies like FTP, but in recent times the need of producing time in NRT using many different output formats has push the datacentre to implement new delivery methods like the community standard ERDAPP. The archive and delivery activities are linked using transform functions from the long-time archiving (RXF and others) to the community focus delivered outputs. Most of this transform functions reformat the data, but also do standardizations using probably the most important BODC system the Controlled Vocabs.
3. Controlled Vocabs. They can be considered the crown jewels; "*Controlled vocabularies are used by data creators and data managers to standardise information. They are used for indexing and annotating data and associated information (metadata) in database and data files. They facilitate searching for data in web portals. They also enable records to be interpreted by computers. This opens up data sets to a whole world of possibilities for automated data workflows, computer aided manipulation, distribution, interoperability, and long-term reuse*."

Gaps and Requirements: The Science Community

The NZOC WP1 Future Science Needs report does an excellent job of highlighting the requirements that the scientists have of the data ecosystem of the future. These requirements will not be repeated in detail here, but the key highlights are presented to frame the gaps analysis and the horizon scanning in future sections.

Data access is a considerable concern for the scientists, which, given the considerable cost of gathering data, must be the primary consideration for the data ecosystem of the future. Improving data access through adherence to the FAIR principles (Findable Accessible Interoperable Reusable) is highlighted as a priority.

Gap 5: Data access to the observations, and limited adherence to the FAIR principles

The science community anticipate increasing volumes of data that being collected in the coming decades highlighting the need for the data ecosystem to be scalable to the larger data volumes.

Gap 6: The data ecosystem must scale to manage the rapidly increasing data volumes anticipated.

There is an increasing emphasis on the need for real-time access of data by a variety of users, including, but not exclusively, operational users. Real-time data transmission also reduces the risk of data loss through platform.

Gap 7: Real-time transmission of data from vulnerable environments needs to increase to improve data safety. In general, latency in data availability to the user should be reduced

Machine learning (ML) or artificial intelligence (AI) development and integration with automated platform, sensor and sampling technologies for mission planning both prior to an expedition and in real-time.

Gap 8: Improved data management and data workflows are required to for the use of AI/ML both in the data collection workflow and in eventual data use.

Mission planning prior to an expedition and in real-time would benefit from the use of data sciences approaches. Adaptive sampling methods could be developed for intelligent sampling, where the vehicle has some awareness of its scientific mission.

Gap 9: Improved data management and data workflows are required to for the use of AI/ML and adaptive sampling methodologies in mission planning and during operations.

The use of models to inform research expedition planning, and the integration of data streams (observations and models) making them more end-user friendly.

Gap 10: The integration of modelling and observations both for mission planning and for improved access and use of data (observations and models).

AI/ML ready data is required as an output of the NZOC data ecosystem to allow for more effective mining of data.

Gap 11: The data management workflow must enable AI/ML ready data for the data sciences of a broad range of users once the data has been made available to the community.

The need for interconnectedness between different technologies, including autonomous vehicles, ships and potential "hubs" (with both charging and data caching potential) has been raised.

Gap 12: A fully interconnected data communications network linking a diverse range of vehicles, the ship and / or other system nodes.

The science community also noted the need to grow the range of skills related to the management of data and data sciences in the ocean observations, which will be addressed in the section on skills below

Gaps and Requirements: Summary of Data Ecosystem Workshop Outcomes

How will data be accessed?

1. Following the principles of FAIR and TRUST is important.
2. Coming up with a standard data format was thought likely to be very difficult, interoperability of the data sets may need to be achieved by alternative means.
3. The adoption of data standards is important.
4. Cloud-based storage, perhaps in partnership with existing data centres, will be important.
5. Sustained funding for all data ecosystem components (including data centres) is needed.
6. Consider whether data will be served to users by files or by streaming.
7. Not all data are numbers – consider other formats including images and video.
8. Discoverability and reusability are vital for reducing carbon footprint (carbon cost per data use) – data must be easy to access and use.

What is the shape of the future data ecosystem?

1. From a user perspective, the data ecosystem should contain observational data (both raw and 'improved' versions), model products and re-analyses, and associated metadata. Data should confirm to standards (possibly global standards), but it was recognised this would be difficult, and perhaps impractical.
2. The data ecosystem should be extendible and flexible, to adapt to as-yet-unknown future needs. Use a workflow and agile development approach.
3. Cloud storage and federated data services will be important.
4. Edge computing should be used to minimise data transfer from observing platform to shore.
5. Consider how the data ecosystem will operate in 'no internet' regions. Mesh networks such as relays of moorings and floats are used, e.g., in Antarctica. No standards exist yet for this.
6. Data should be accessible in a seamless, interoperable way that is agnostic to its physical location.
7. Avoid duplication in processing (human effort) and duplication of storage (carbon cost).
8. Virtual research environments could be used to ensure visibility and accessibility (e.g. Pangeo, earth engine)
9. Engage with UK stakeholders / future users asap to ensure their needs will be met by the data ecosystem.
10. Historical data / long time-series observations are important and often overlooked – how to include these?
11. Citizen science data: how to include it?
12. Data gathered by industry / non-public-funding – can it be included, and should it be?
13. How to handle intellectual property, both for commercial and scientific interests? Will it cost money to access data (perhaps for a subset of users)?

What services we cannot implement now because the data ecosystem is not "good enough"? The concept of ecosystem should allow growth including new functionality and services in the future.

1. Images and video – storage, access and standardisation of format (especially video) needs development.
2. Lagging infrastructure at data centres can struggle to keep pace with changed in technology: platforms, instruments, sensors etc..

Does the Net Zero just depends on the hardware manufacturers and energy providers to develop clean technologies? is there anything that can be done at the design point to minimize the carbon footprint?

1. Need to know the carbon costs of the current system to help focus on where we can reduce.
2. Transfer of data has a carbon footprint – reduce the amount transferred. This means both after initial acquisition of the observations and subsequent movements en route to the user.
3. Consider how much data needs to be stored readily accessible and whether some could be stored in a less carbon-intensive way such as tape archive. Carbon price of storage could drive innovation and encourage users to only 'get' the data they need.

Do we need more infrastructure/software on board the ships to run things locally? (models, experiments …).

1. Yes! Investigate what can be achieved with edge computing. This would reduce data transfer and could also reduce the data processing burden on data centres.
2. In areas with no internet, consider what capabilities can be used on the 'edge' to create a mesh/relay for data transfer.
3. Consider physical samples.
4. Consider images, both raw and annotated.

Software nowadays is sometimes an afterthought in research, does people think about it and do they have ideas on how software needs to be developed, maintain, and used in the future?

1. Software is a key component of all aspects of the data ecosystem and should be considered in the context of the whole of the system.
2. Investment in skills is needed throughout all phases of 'designing a research programme', particularly at postgraduate level (data management, software engineering, etc.).
3. Can machine learning / artificial intelligence / quantum computing help. A note of caution: ML/AI need large training datasets that may not be available in all cases.

Skills development/investment in training will be critical, both to create the data ecosystem and for users to use it to its best potential.

Skills

Realising the data ecosystem vision presented here will rely on a workforce having the necessary skills. The UK ocean research community is in many ways extremely well-placed in some of the key skills required for the development of the NZOC data ecosystem, although there are some obvious areas where there are shortages of skillsets in key areas.

The National Digital Twin Programme (NDTp) Skills and Competency framework[4] outlines the critical roles needed at an organisational level to support the integration of data into a digital environment and has an excellent set of resources that provides insight into the kinds of skills that will be required. Figure 11 gives a

[4] 010321cdbb_skills_capability_framework_vfinal.pdf (cam.ac.uk)

schematic representation of skills identified, each of which comes with associated competencies. A number map well to the skills already required of data science practitioners and are already well-represented in those working on the research data presently connected. However, there are clearly some skills that are not presently well-represented in the NERC and NOC communities and will increasingly be in demand as the digital dependency of our science infrastructure increases. It is also clear that there is a growing emphasis within government on developing the people with skills to meet the AI challenge[5] and it is therefore incumbent on our community to ensure we have the skills pipeline needed. It is It is therefore essential that the NERC and NZOC community plans well for the skills needed in the future and pre-emptively recruits and/or trains appropriately in the digital space. The latest skills review by NERC[6] resulted in a skills framework created in 2010 and updated in 2012, which noted the shortage of skills in areas of direct relevance to the data ecosystem. It seems unlikely, given the growing emphasis on data over the subsequent decade, that these skills gaps have been fully resolved and are likely to become increasingly in demand in the future.

Recommendation 1: Do a skills audit to assess present gaps in critical digital skillsets, and plan for training the future generation of the NZOC workforce

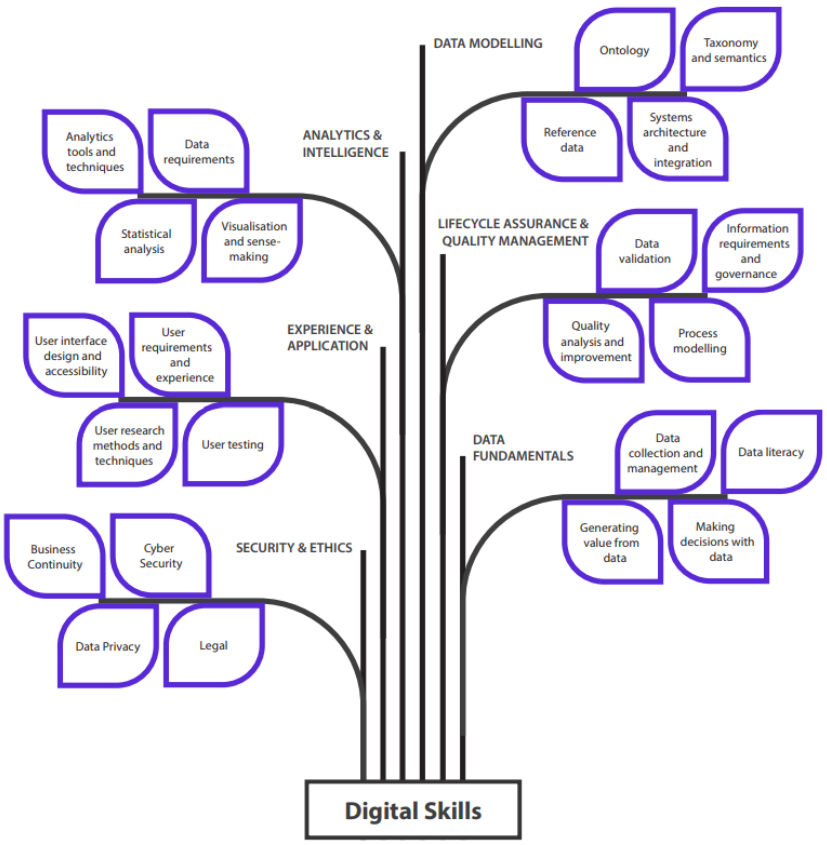


Figure 11: The Digital Skills required for the NZOC data ecosystems, from the NDTp Skills Framework

[5] AI Sector Deal - GOV.UK (www.gov.uk)
[6] OUTPUT 3 - POSTGRADUATE SKILLS NEEDS FRAMEWORK, JULY 2010 (ukri.org)

Horizon Scan 2020-2035

None of us can be entirely clear exactly what form the data ecosystem will take in 15 years' time but given the planning timelines it is important we set out the fundamental ambitions and form that we need to develop in the intervening years. The concept of Digital Twins is clearly gaining momentum and has an important role to play in shaping the NZOC data ecosystem. The NOC already have a limited form of a Digital Twin in development as part of the autonomy programme in the so-called C2 activities, with autonomous vehicles collecting data that is then gathered and presented to the command-and-control centre to provide the tools to allow adaptation to the glider tasking. The Digital Twin concept also creates a framework in which a lot of the component parts of a data ecosystem can be conveniently viewed so an envisioned Digital Twin is used to frame the data ecosystem of the future NZOC.

An Envisaged Data Ecosystem

This overview is based on descriptions of a digital twin, which provide a useful framework in which to describe a future NZOC data ecosystem, include all the components needed for the NZOC ecosystem, and if fully implemented will maximise the value in our observing capability.

A digital twin is a virtual representation of an object or system that is updated from real-time data and uses simulation and machine learning to help decision-making.

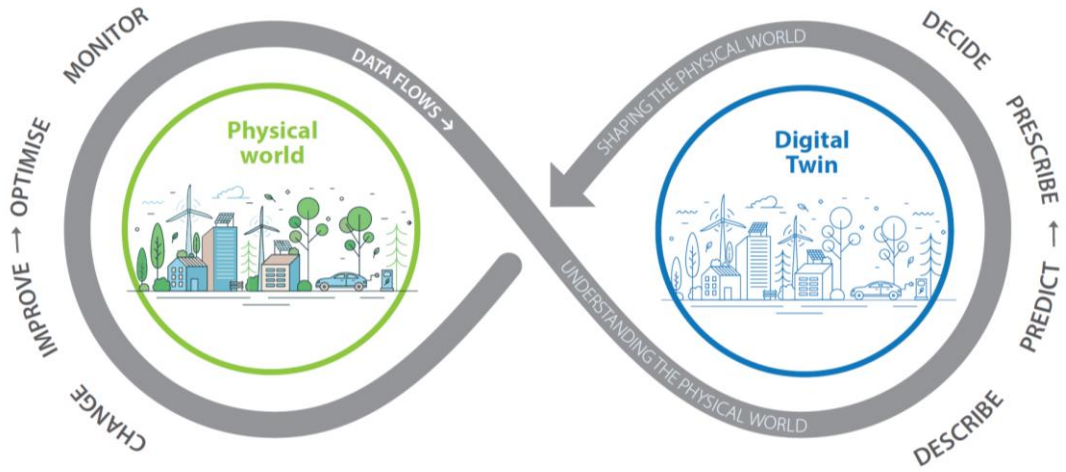


Figure 12: Schematic of the Digital and its relationship with the physical world. Reproduced from Figure 1 of the Centre for Digital Built Britain Digital Twin toolkit[7].

It is closely related to the concept of the Internet of Things (IoT), which can be described "as the linkage between the physical world on the one hand and the cyberworld on the other with the help of items or objects that possess sensing abilities and transmit the measured results via a network to achieve a purpose"[8]. This is clearly a description that applies to much of the present ocean sensing capacity, and this will undoubtedly increase on the timeframes of relevance to this report.

The UKRI document 'National Digital Twin Programme' defines Digital Twin as a "computational representations of specific physical entities, assets or processes with a flow of data between the 'physical' and

[7] Digital Twin Toolkit - Community Resources - DT Hub Community (digitaltwinhub.co.uk)
[8] Muntjir, M., Rahul, M., Alhumyani, H.A.: An analysis of Internet of Things (IoT): novel architectures, modern applications, security aspects and future scope with latest case studies. Int. J. Eng. Res. Technol. 6(2), 422–427 (2017)

the 'digital' twin, often in real time, as part of a dynamic process with the ability remotely to interact with, or control, the physical twin". There is increasing impetus behind a national strategic approach to Digital Twinning and the Centre for Digital Built Britain (CDDB) has begun the process of setting out principles on which a National Digital Twin can be developed, focused on the built environment but with the expectation this will grow to include other domains, including the natural environment, more explicitly. It seems sensible (and probably essential) that any work in the ocean domain on data ecosystems should follow the same principles and learn lessons from the broader built environment digital activities.

The CDBB describe the four fundamental elements of a digital twin as follows:

- **Physical Twin**: in our case the ocean and surrounding environment
- **Data**: the link from the physical twin to the digital twin, collected by the research infrastructure itself, but also from other sources (satellites, moorings)
- **Digital Twin**: data storage, model, analysis and insights to support decision-making
- **Intervention**: the link from the digital to the physical, in this case scientific insight and mission planning decisions, and longer-term the changes to human behaviour that support a sustainable ocean.

Alongside the UKRI activities there are also other important initiatives at European level (DestinationEarth) and internationally (the UN Decade Programme DITTO and the G7 FSOI and related coordination efforts) that will formulate plans anon d protocols for Digital Twinning. Any NZOC data ecosystem should be developing in line with national and international best practice.

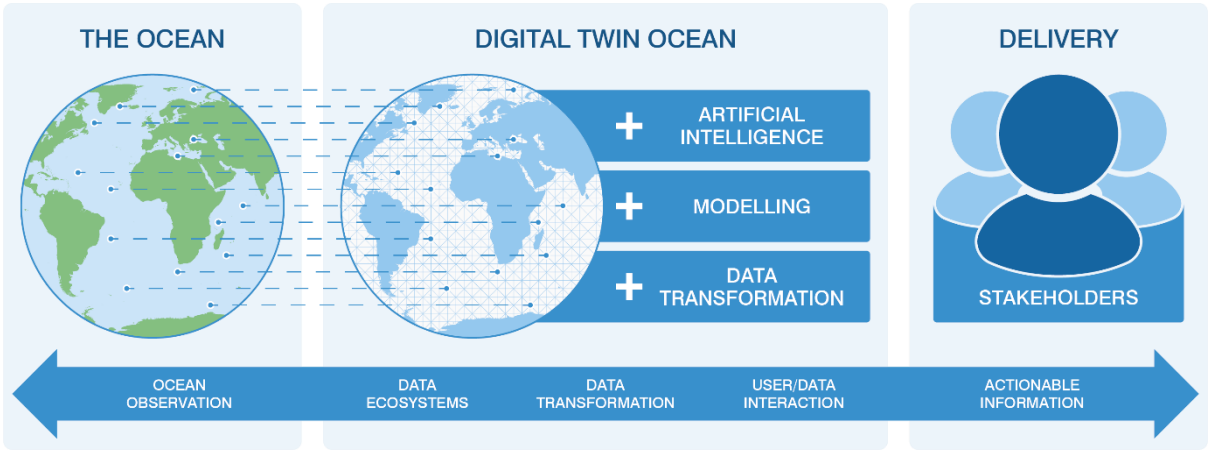


Figure 13: A schematic of a Digital Twin Ocean from the G7 Digital Twin scoping paper, showing the relationship between the ocean and the Digital Twin, to provide actionable information.

Recommendation 2: Activities in the NZOC data ecosystem are developed following national and international best practice, and particularly following the guidelines laid down with the National Digital Twin Programme (NDTp)

The Digital Twin itself can be broken down into some key constituent parts:

- A **data infrastructure** that frees up access to ocean observations through data communication and management.
- A **virtual space** on which marine observed data, and other data resources, are aggregated and made accessible along with the computing capacity to add value to these data. This may include model data, either pre-computed or regenerated live as new observations become available.

- A **data analytics layer** that provides AI / ML tools to access, manipulate and analyse marine information to maximise the understanding and value from these data.
- An **interactive layer** allowing users to visualize, interact and tailor the data, scenario and models to meet their needs.

This is illustrated in more detail in the concept model of a Digital Twin shown in Figure 14.

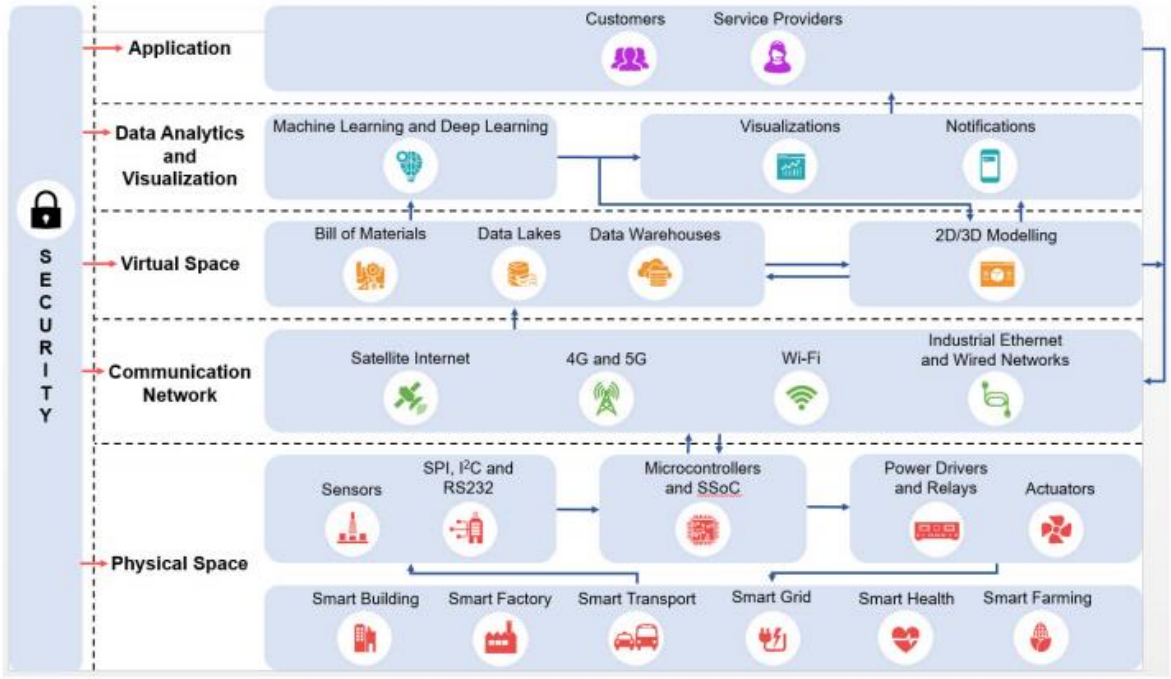


Figure 14: A model of a digital twin, from Al-Ali et al. (2020)

Although this model is not of the ocean, and therefore not all of the components directly map to the ocean use case, it provides a framework in which to break down the requirements of the data ecosystems. The physical space (the ocean) is measured using sensors, which transmit and receive information via a communication network to a Virtual Space on which transformations can be made to the data, either via modelling or Machine Learning. The data ecosystem for the purposes of this work will be taken to be those elements from the Communication Network to the Data Analytics and Visualisation layer, upon which the applications can be built.

The Physical Layer (the Ocean)

This report is not directly looking at the ocean itself, or the platforms / sensors that are making measurements in the ocean. But it is important to co-design the instrumentation and the computational infrastructure around them (including on the ships where applicable) with the data ecosystem being developed. The concept of Edge Computing is gaining increasing traction and is particularly important when considering the communications layers described below. Sensors are already required to do a certain amount of processing locally. The extent to which processing is pushed to the "edge" on an autonomous platform or on another part of the physical research infrastructure (the ship or a third-party processing facility) will determine the requirements for the communications layer.

Recommendation 3: Map the computational architecture on the research ships, autonomous platforms and potential 3rd party communication architecture required for an NZOC Digital Twin

Virtual Space

The Virtual Space is an environment in which diverse datasets can be collected together, aggregated and transformed into structured data that can then be processed into usable information and in many cases transformation by tools available through the data analytics layer. The data aggregation layer collects in-network sensor data from the underlying sensors layer for further storage and processing. It also collates (and transforms to appropriate forms) data from outside the immediate sensor environment for use within the analytics layer.

Data Lake

A data lake is defined as a central location that holds a large amount of data in its native, largely raw format. Compared to a data centre which stores hierarchical data in a data warehouse, with data in files or folders, a data lake uses a flat architecture and object storage to store the data. Object storage stores data with metadata tags and a unique identifier, which makes it easier to locate and retrieve data across regions, and improves performance. By leveraging inexpensive object storage and open formats, data lakes enable many applications to take advantage of the data.

Data lakes were developed in response to the limitations of data warehouses. While data warehouses provide highly-performant and scalable analytics, they are expensive, proprietary and can't handle the modern use cases most companies are looking to address. Data lakes are often used to consolidate all of an organization's data in a single, central location (big lake), where it can be saved "as is," without the need to impose a schema (i.e. a formal structure for how the data is organized) up front like a data warehouse does. This makes them ideal for low latency requirements for the real-time research expedition activities. Unlike most databases and data warehouses, data lakes can process all data types — including unstructured and semi-structured data like images, video, audio and documents — which are critical for today's machine learning and advanced analytics use cases.

The architecture on which the data is hosted for the Digital Twin to allow the data analytics to function is therefore almost certainly going to be a data Lake of some form. It should be noted that unlike a data centre, where the ambition is to store data for prosperity and for a broad range of users, the use of the data lake in this case is simply as a staging post between collection and transferal to the data centre, with transformation of the data on the data lake required for two purposes – firstly, to ensure the appropriate post-processing of the observations takes place as close to source as possible to minimise the communications overhead and secondly to create the information required by the NZOC data ecosystem users.

The NZOC Data Lake must scale easily at a reasonable cost as it will be required to provide fast storage of massive amounts of data. The information in the data lake will only be structured once they need to be used for further operations such as analysis. Data lakes are usually a combination of different technologies such as locally hosted databases and cloud storage depending upon the architecture design requirements.

Recommendation 4: Scope a scalable data lake architecture to receive NZOC data, developing pilot projects to develop expertise in managing the data architecture across cloud, ship and shore-based infrastructure.

As part of a pilot study, deploy a hybrid high-availability & high-performance data-driven computing infrastructure. This subtask is responsible for the design, provisioning, and exploitation of the infrastructure (integrating Cloud, High Performance Data Access, HPC and GPU technologies) enabling computing such as on-demand modelling, execution of what-if scenarios, and a JupyterHub environment. System requirements will be defined in with potential use of well-known technologies (e.g. kubernetes, S3 object storage) and should include:

- The Data Lake (AWS S3-interoperable object storage) and non-object storage services;
- Computational services (HPC, VMs, containers, GPUs);
- Networking services;
- Other cloud services (snapshots, etc.)

Data Centres

Data Centres (referred to as Data Warehouses in Figure 14) are the final resting place for the data of long-term value from the Digital Twin, and will also be the source from which the twin receives a range of external data, including reanalysis and forecast model products run outside the NZOC data ecosystem (see Section: Modelling Layer). The data centres will manage the highly curated data that serves as the central version of the truth.

Characteristics	**Data Warehouse**	**Data Lake**
Data	Structured and quality-controlled data with well-described characteristics and meta-data	Structured and unstructured data from IoT devices (smart gliders etc.), GTS, Eumetcast and API access to diverse data warehouses
Schema	Designed prior to the DW implementation (schema-on-write)	Written at the time of analysis (schema-on-read)
Data Quality	Highly curated data that serves as the central version of the truth	Any data that may or may not be curated (i.e. raw data)
Users	Policy, science, climate services	Data scientists, data developers and research expedition PIs
Analytics	Predefined reporting, visualisations, data extractions	Machine Learning, predictive modelling, data discovery

Table 1: The characteristics of a Data Lake and a Data Warehouse in the context of NZOC

The British Oceanographic Data Centre, alongside the other NERC Environmental Data Centres, will be both the source of much of the environmental data required for the twin as well as the final destination for the data from the Research Expeditions (so considered as data warehouses in the context of NZOC). The Data Lake will need to be well connected to the BODC data centres solutions, as well as several other key data repositories.

The BODC data repository, and other Data Centres, are not part of the NZOC Data Ecosystem itself, and so are out of scope in terms of the NZOC data ecosystem design. However, there needs to be a synergistic relationship between the key data centres of relevance to oceanographic research data and the NZOC Data Ecosystem. Given the vast quantities of data potentially available as a result of future research data collection activities, the cost involved in collecting the data and the obvious societal benefit of the data both in real-time and in delayed mode, communication between the NZOC data ecosystem needs to be front and centre of the design of both. The reality is likely to be a complex array of data centres and data provision mechanisms that provide data to, and in many cases receive data from, the NZOC data ecosystem, something that needs to be accounted for in the communications capability. There are likely to be a number of research digital twins built upon the environmental data collected upon the Research Expeditions, but these are not part of the NZOC Digital Twin and ideally will receive the data via a data centre rather than directly to ensure that the data is

appropriately managed and traceable. All the same, the ambition to use the data collected on the Research Expeditions as effectively as possible obligates us to manage the flows of the data and metadata from the Data Lake to BODC (and / or other appropriate data centres) with these other applications in mind. The interface between digital twins and data centres needs to be defined. Current data centre data delivery developments are focusing on FAIR this which may need enhancing to facilitate digital twins and the NZOC data ecosystem. The NZOC Data Lake, and the transformations applied upon the data within it, unusually for a Digital Twin, therefore serve two purposes, with the local application of the data for the Digital Twin being important but perhaps subservient to the need to make sure the data is transformed in such as a way as to optimise the reuse within other data environments.

Data Aggregation Layer

The data aggregation function involves the aggregation of the operational and environmental sensor data with the other forms of data required by the NZOC Digital Twin.

Hosting the data lake on the ship will allow large quantities of the research expedition data to be included in the data lake, but relies on the ship being part of the research expedition. It also requires the movement of large amounts of externally sourced data into the ships' environment, traditionally a bandwidth limited environment. Hosting the data elsewhere (in the cloud, or at some centralised location convenient to the processing of the data such as the NOC) has benefits; shore-based communications will allow a larger volume of data to be incorporated from external sources. A balanced approach to the location of the data aggregation layer, taking into account edge computing concepts, data communication costs and robustness of access to the data is important. Decisions on how the aggregation layer will be structure needs to consider the role of a ship (or indeed if there is a ship at all) and / or any other offshore data hubs to balance the communications costs and dependencies from the sensor to the ship and/or to the shore, and similarly from the ship to the shore. Additionally, it is important to understand the primary working location of users of the data ecosystem. It is likely that there will be communication breakdowns between the ship and the shore and the importance of access to the data layer means the primary location of the users would be the ideal location for the aggregation infrastructure, but noting that there are likely to be users in multiple locations, and it would seem sensible to plan for a mirrored approach with an infrastructure aboard the ship and ashore allowing access to the data and tools (see Recommendation 3).

Modelling Layer

A digital twin of any object requires a model of the object, an evolving set of data relating to the object, and a means of dynamically updating or adjusting the model in accordance with the data. Modelling, and the approaches to integrate the modelling with updated observational data, is therefore indispensable to an integrated picture of the physical twin within the digital twin. In oceanography we are most used to dealing with predictive models such as those developed at NOC using, for example NEMO, ERSEM or FVCOM. However, there is increasing work being done to develop data sciences approaches to modelling the ocean, either as part of a predictive modelling framework or as a self-contained modelling tool. Some of the models will be based on gridded physical models with several subcomponents, while others will make use of innovative artificial intelligence algorithms exploring the trade space between speed, resolution, and accuracy. It is therefore important to remain open minded within this space about the tools available to develop the models required to inform the decision making of the research expedition PI or technical support.

It is conceptually important to differentiate the methods used within this modelling layer to the methods used within the data analytics layer. This modelling layer is delivering the integrating and extrapolating capability to turn the dispersed (in space and time) information available from the sensors into actionable information. The data analytics layer will employ data sciences approaches to help the human user interacting with the data make appropriate and timely decisions based upon the digital information available to it. Despite this

difference, many of the tools required to deliver the capability will be in common and are discussed in the Section: Data Sciences.

The modelling layer will need to function on two levels. For the far field / large scale context the system will access operational ocean and weather model data produced by providers such as Copernicus Marine Environmental Modelling Service and the Met Office. These services already provide API interfaces to their data, allowing the NZOC data ecosystem to download the environmental contextual data for the region of interest as the research expedition moves. Driven by these input model data products a high-resolution modelling system will provide the integration of the observations collected by the research system, through assimilation, to create a 4D representation of the local oceanography, including forecasts. The need to develop a model that enables real-time access to consistent multi-dimension, multi-variable, and high-resolution description of the ocean, to enable the generation of insight for decision making and strategic plans. This digital framework will be able to closely replicate reality to respond to known issues as well as to uncover new questions for further advancements in marine science and policy.

In the context of the NZOC data ecosystem this implies that there needs to be at the minimum a physical ocean model of the immediate environment being researched. This is not to be confused by off-line (i.e., precomputed, without the input of the new data collected by the NZOC sensors) ocean model data which will also form some of the Digital Twin input contextual data.

The modelling component of the digital twin is therefore fundamental and must incorporate the information collected by the NZOC sensors within the realisation of the Digital Twin. Ocean forecasting services have now for many years been able to create simulations of the ocean environment through the assimilation of observations, and it is sensible to start with these capabilities and develop them further for the NZOC context. This makes sense particularly given the close relationship between the NOC (and PML), where the ocean (and biogeochemistry) model development takes place, and the Met Office where the assimilation systems are developed.

First and foremost, the modelling component of the digital twin would become a data assimilation instrument that continuously cycles real-time, highly detailed, high-resolution Earth system simulations and ingests observational information from all possible instruments. Ideally it should rely on an ensemble of models, and ensemble information from models, to allow decision making to be based on probabilistic information.

Recommendation 5: Develop a modelling capability that can dynamically assimilate observations in a moving frame (ship-following) at a resolution relevant for observation collection decision making, using modelling tools (e.g. NEMO, ERSEM, NEMOVAR) already well-established in the community.

It is worth noting this sort of capability has wide applicability, from ship, or other marine infrastructure, operations to Defence applications, and is already well-advanced in some of the component parts, although would need investment to get to the stage of robustness required for research expedition support.

DestinationEarth[9] will provide a series of Digital Twins with a foundation in the global-scale, coupled modelling systems based largely on the existing Numerical Weather Prediction systems at ECWMF. This system is relevant to the NZOC Digital Twin because it will provide a potential source of large-scale boundary conditions for a high-resolution local scale relocatable modelling capability developed within the NZOC Data Ecosystem. It

[9] Bauer, P., Stevens, B., & Hazeleger, W. (2021). A digital twin of Earth for the green transition. In Nature Climate Change (Vol. 11, Issue 2, pp. 80–83). Nature Research. https://doi.org/10.1038/s41558-021-00986-y

will also presumably continue to be developed using the European Centre for Medium-Range Weather Forecasting (ECMWF) modelling systems, which include NEMO as the ocean component of the system. The development of a NEMO-based software environment within DestinationEarth therefore has the potential to benefit the NZOC activities and should be closely followed and where possible complimentary and/or shared methodologies used.

Recommendation 6: Maintain engagement with ocean Digital Twinning activities (such as, but not exclusively, DestinationEarth).

Data Analytics Layer

Tools to transform the data in the data lake into information will be needed. This layer will provide the software and data sciences to transform the data in the data lake, to empower the technical and scientific users' and enable dissemination of knowledge through applications and services. Tools will need to be developed to integrate the data collected on the research expedition with external data sources (satellites, pre-computed modelling etc.). As well as tools many expert users are already familiar with such as Jupyter Notebooks, bespoke analytics software for the scientific and technical user will need to be developed following a more detailed understanding of the use cases. The deeper integration of Artificial intelligence in the NZOC data ecosystem will be key, with both automated and user defined actions likely to be required.

The development of better integrated data through technology advances that have started and will accelerate over the coming decades provide the opportunity to benefit from the integration of all marine, and related data, in a digital space**,** removing barriers to data use.

Data Management

Creating a Digital Twin requires a data management approach that follows an Information Management Framework (IMF) that allows data to be used interoperably across disciplines and through machine-to-machine interfaces without manual intervention. The Data Management approach for the NZOC Data Ecosystem is discussed elsewhere in this report (Section: Data Management). It is worth noting that a significant part of the challenge of developing a Digital Twin is in the data management, and ensuring that it follows commonly agreed protocols of a broader Information Management Framework including FAIR (Wilkinson et al., 2016, Tanhua et al., 2019), TRUST (Lin et al., 2020) and CARE[10] principles. It is therefore important that the NZOC data management approach is well-integrated with broader national and international initiatives, including the National Digital Twin, and that there is work done to ensure that national and international frameworks reflect the ocean Digital Twin needs for Data Management and vice versa. A key technical enabler for the National Digital Twin is the Information Management Framework (IMF). The IMF, currently under development, is a common language by which digital twins can integrate to improve decision making and enable better outcomes for people and nature. This offers a wealth of new and exciting career and training opportunities for people, regardless of their technical ability, to make our built environment a fairer, more environmentally friendly, and more productive through the National Digital Twin (NDT).

Recommendation 7: Engage with Information Management Frameworks to ensure the NZOC Data Management framework meets nationally and international interoperability requirements.

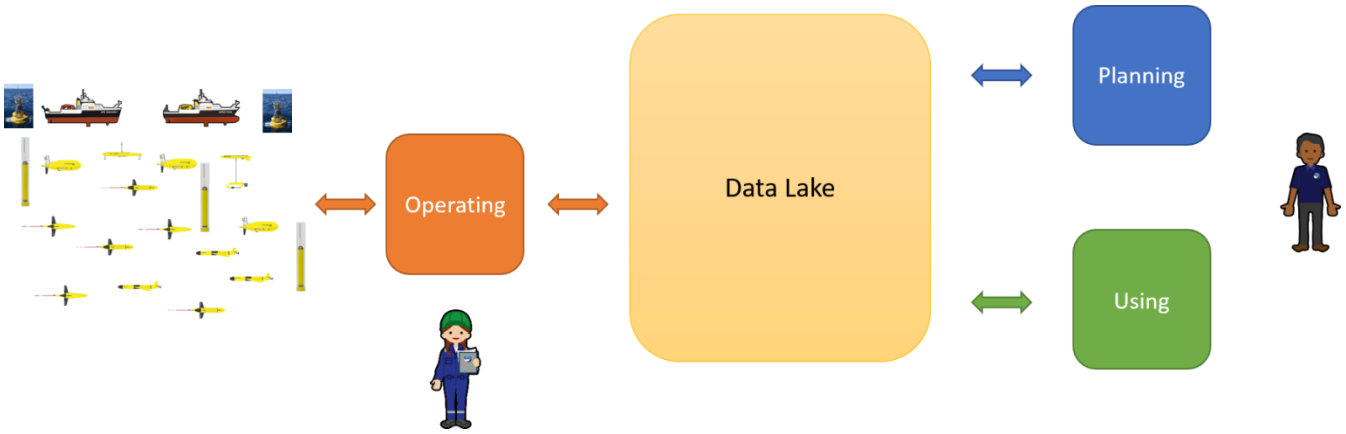
[10]https://static1.squarespace.com/static/5d3799de845604000199cd24/t/5da9f4479ecab221ce848fb2/1571419335217/CARE+Principles_One+Pagers+FINAL_Oct_17_2019.pdf

Applications Layer

The NZOC data ecosystem will need to develop a set of tools, including advanced viewing system, to enable users efficient viewing, manipulation, and mining of the data as, or soon after, it is collected. These tools will need to be organized in a digital framework enabling the creation of workflow to gather data, generate on demand modelling, activate what-if scenario, analyse the data (with classical or AI based tools), visualize the result in effective 3D and user-friendly environment.

Provision of robust back-end services allowing simple and optimised data ingestion pipelines, data streams and a wide variety of APIs for data/metadata access and publishing. It will implement and deploy back-end services for data transformation and cloud optimisation, feeding analysis-ready data into the Data Lake and enabling serverless access to data and metadata. Finally, for addressing needs critical to modelling and other processing activities will deploy a harmonised gateway to external data holdings, linking external brokers with comprehensive query validation and dynamic harvesting of metadata.

The Applications Layer is the mechanism by which the NZOC data ecosystem users interface with the NZOC data and data tools, and is designed to cater to three primary use cases, expedition guidance (prior planning for the cruise, “Planning”), command and control (autonomy piloting, “Operating”) and research Principle Investigator decision support (“Using”). These three use cases are shown below.



Planning: Expedition Guidance

Prior to any NMF Expedition the PIs and scientists involved will normally plan locations and timing of the observing campaign based on known oceanographic features, their location and the timing of processes or events of interest. A priori knowledge of such features is aided through access to historical, ideally gap-filled, information including climatologies, model simulations and previous observing campaigns. In some cases, forecasts may also prove useful in Expedition Planning.

After the Expedition has begun, eyes on the underwater world through real-time access to visualisation and manipulation of the observations being gathered is essential for fine tuning data collection. The availability of the broader context, ideally into the near-future, through model simulations constrained by the available local environmental information would significantly improve guidance to PIs if available. The types of models need defining through engagement with the PIs and engineers and will undoubtedly get refined through pilot studies and as the data collection methods and strategies evolve. However, it seems likely that models of the physical and biogeochemical state of the ocean at scales varying from the regional (~1000 km horizontal, 10 m’s in the vertical) to the turbulent (~100 m’s in horizontal, 10 cm’s in vertical), including air-sea fluxes and representation of local scale meteorology will be needed. Other modelling types such as benthic process models should also be considered.

This is a process done in person or following relatively informal methods of communication like emails; the scientific crew interchanges information using the available information channels, but this is an ad-hoc process that is not automated or recorded in a machine-readable format. To fully develop the NZOC concept, allowing a close loop between the scientific crew decisions and the executions of activities (either automatically or manually) a system to enable the distributed decision making and communication of the plans for Expedition guidance and objectives is needed.

Recommendation 8: Design an open planning ecosystem with well-defined standards and open APIs to allow shared planning between different actors of the system.

Operating: Command and Control

NZOC technicians will need to access tools and data available on the new data ecosystem during operations. The Oceanids command and control (C2) activities have made good progress in defining an autonomy digital backbone and the user interface to operate autonomous platforms over the horizon, providing a baseline system, these activities have been mainly focused on operations of long-range autonomous robots, and more complete requirements gathering exercise focusing on the future data ecosystem and the role of the ships and autonomy enhancing ship activities will be needed.

Recommendation 9: User requirements for the NZOC data ecosystem applications should be gathered, and iteratively enhanced following experience in developing the user interfaces.

Currently ship operations are self-contained in terms of guidance; decisions are made from the ship based on the different activities schedule for the campaign and based on the potential information gathered on site. Once the scientific crew is on board the ship all activities are quite isolated from the outside world, and while there are many examples of expeditions leveraging the usage of satellite and modelling products and coordinating with other ships, all this remains an ad-hoc process, tailored for specific projects and campaigns. The NZOC data ecosystem will develop a digital backbone enabling scientific crews to leverage enhance information quicker and in a more systematic way. There are several envisioned scenarios of the usage of this backbone:

1. To enhance the activities of the digital twin. Digital twins will “command” ships and un-crewed autonomous robots, both in the surface and under water to gather the measurements required for the digital twin operations.
2. A better data ecosystem should allow the planning and execution of exploratory science campaigns on ways that are unthinkable today, like the execution of part of the digital twin on-board enabling scientists to make decisions based on better information (high resolution models, quick data turnaround from the sampling point to provide a visualization).
3. The NZOC data ecosystem will enable distributed decision making, bringing the necessary information to people on board ships and in remote locations across the world, to facilitate the command of the ship and un-crewed platforms. There are already successful examples of this mode of operation. (Porto + Schmidt institute[11])

Requirements in this area are expected to be similar to the broader theme of guidance for Expedition planning and in operations, although the automation and interoperability demands may be more significant. Autonomous Vehicles and Autonomous planning and guidance a key part.

Guidance and Operating are two critical components of the expedition and are interconnected; plans generated by the Guidance will be sent to the Operating component to be converted into real actions for the

[11] https://schmidtocean.org/cruise/exploring_fronts_with_multiple_aerial-surface-underwater-vehicles/

different vehicles on the fleet, and vice versa, the executed plans and the vehicle feedback needs to be propagated to the Guidance "layer" to allow replanning of the whole system.

Recommendation 10: Invest in the design and develop open standards to enable coordinated command and control of autonomous vehicles and ships between different institutions (NOC, BAS …).

Recommendation 11: Develop a multidisciplinary/multi-council strategy to increase collaborations on the development of novel autonomous planning and optimization systems to maximize the usage of autonomous assets at sea (NERC-EPSRC crossovers)

Using: Research Expedition Scientist access to the Data

Scientists will need to access tools and data available on the new data ecosystem during operations. The NZOC report on the Science needs will provide some guidance on some of the requirements for accessing, manipulating, and visualising the data from the scientists but a more complete requirements gathering exercise focusing on the data ecosystem function will be needed (see Recommendation 9). A discussion on user processing of the data is given in Section: Processing by the User.

A priori, however, we expect the application layer to be accessed via a navigation page to guide the different users to the technical parts of the data ecosystem, which will need to be developed. This will need to include a viewer (the "screen") which will integrate a fully functional catalogue of capabilities to include free-text queries, data searches and favourites. It should support the wide range of datasets in the Data Lake as well as a curated list of datasets from external providers (for example via WEkEO[12]), which will need integrating / federating with the NZOC data lake, potentially dynamically and on-demand. It will also need to support accessing, and generating, datasets resulting from on-demand modelling. A subsetting attribute form, with attributes dependent on dataset-specific metadata, will be needed to allow the user control of the data they access. A comprehensive query validation engine, based on dataset-specific rules, will need to be implemented to support these on-demand activities.

The viewer will need to make it possible for PIs to analyse the data freely using tools such as Jupyter notebooks as well as giving data viewing capabilities. The spatial and temporal representation of non-gridded in-situ data with variations in all three spatial dimensions, for example from gliders, is a challenge for most even advanced data viewers which will need development.

Infrastructure

Communication Network

The main purpose of the communication network layer is to effectively transmit/receive the data collected by the sensors, and to bring in external data as required to support the functioning of the Digital Twin. The communication network must therefore:

- Provide access to external data sources for inclusion in the twin (e.g. real-time satellite and moorings data).
- Deliver data from the twin (perhaps at reduced volumes) sufficient to support shore-side operations of the research expedition.
- Move data from the platforms within the NZOC research expedition to the Virtual Space for analysis and transformation.

[12] https://www.wekeo.eu/

- Provide information from the Virtual Space to the sensors / platforms for updated guidance.

A digital twin is only possible if it is sustained by the communications network. Large amounts of real-time big data are being gathered from potentially hundreds of sensors at a time, a high bandwidth and high data rate are required to relay them all at once to the virtual space layer described below and in return, once analysis has been completed there needs to be a response back to the sensors to create the adaptive feedback mechanisms required of a Digital Twin. The communications capacity to support the NZOC data ecosystem will need to provide the capacity to move data into and out of the data lake. For a twin to be effective the data communications need to be in place to allow the flow of data through the data ecosystem, much of it in real-time, to support Vessel and Autonomy operations. A successful implementation of the digital twin therefore requires both high data volume transmission and low latency in that data transmission. The optimal configuration for the Digital Twin infrastructure will largely depend upon the cost benefits of the communications costs, and the configuration of the data flows through the system will need to be carefully scoped to optimise the data ecosystem.

The concept of Edge Computing is critical in understanding the data volumes required for the communication system to manage. The further to the “edge” the analysis can be completed the lower load upon the communications infrastructure.

Given the timescales on which the NZOC will be functioning the communications capacity will also need to be scalable to the (presumably growing) data volumes throughout its lifespan. Any growth in the sensor / platform network will therefore need to be accompanied by a corresponding growth in the communication capacity. Integration of sensors is a normal activity on research expeditions, and a something that currently is done by ships systems technicians. To achieve the NZOC data ecosystem the data from sensors will require that there will be communications capable of sending data from the sensor to an NZOC data lake where appropriate transformations and access by the research and technical users of the NZOC data ecosystem can be facilitated.

A number of ways of communicating data from the autonomous platforms to the central data lake are envisioned:

- Using direct satellite communications to transfer data in near-real time from the platform to the data lake architecture.
- Using other platforms, such as surface vehicles or ships, with satellite communications capability as a relay. For this to happen, the platform-to-platform communication will need to be developed and integrated within the NZOC Digital Twin architecture.
- Using telephone networks if vehicles are within reach of the land coverage (5G). This can be used for very coastal operations or for surface vehicles used as “mules” for other platforms.

For ship-to-shore communications the same arguments apply as in the autonomy to shore, with the only caveat that the ships may be able to provide more bandwidth given it would be able to host more powerful, and sizeable equipment, and use bigger antennas positioned higher.

Given the rise in Low Earth Orbit Satellite systems (e.g. Starlink, OneWeb) that are already providing fibre-like speed communications (~100 Mb/s) in remote areas we can expect some of the present ship-to-shore communications barriers to be reduced. Communications bandwidth of 100 Mb/s, for example, would be approximately sufficient to transmit today’s typical cruise full data complement of approximately 100 Tb during a cruise period. However, it is not likely that the projected increase in bandwidth will enable the full transmission of all data collected on a research expedition in real-time given the likely increases in data volumes, and a more detailed scoping of the profiling of bandwidth and data collection volumes of the future is needed.

Recommendation 12: Scope the data communication required to support communications at sea and integration within the NZOC Digital Twin.

Architecture and Technology to Meet Zero Carbon

Data centres account for 200 TWh yr^{-1}, or around 1% of total global electricity demand[13] While their energy usage has been stable in recent years as efficiencies increase, it may grow to between 15–30% of electricity consumption in some countries by 2030.

Compute resources and data storage in NOC data capabilities are potentially a signification contributor to carbon emissions. However, the global community is moving towards net zero or negative carbon emissions for computing infrastructure e.g. Microsoft[14], Apple[15], Amazon[16], IBM[17]. Further to this UKRI are committed to a net zero target in future infrastructure development[18].

A Nature paper[19] studying the hidden carbon cost of cloud infrastructures found:

"*The problem lies in the lack of transparency. None of the data required to calculate the greenhouse gas (GHG) emissions of an organization is available once they move to the cloud. Cloud vendors' aggregated global reports do not provide the data needed to understand the environmental footprint. If an organization is implementing reporting under International Organization for Standardization (ISO) 14064-1 then it will be unable to attribute its 'other indirect emissions', the ISO standard category equivalent of Scope 3 (ref.). Further, if an organization wishes to mitigate its emissions, such as through offsets, then the only option is to rely on the cloud vendor to do it for them."*

But the paper showcases there are good news on the transparency front with Microsoft releasing a new carbon calculator for enterprise users of their Azure infrastructure. Greenpeace, also highlights Apple and Google high transparency standards.

Recommendation 13: Evaluate the compute infrastructure needed for the data ecosystem, including the relative benefits of on-premise, JASMIN (or other common NERC/UKRI facility) and cloud compute.

Software

The Data Ecosystem data Aggregation layer, modelling layer, data analytics layer and applications layer are aggregations of different software packages and services that will form a cohesive software infrastructure, this will provide oceanographers and environmental scientists with a big data infrastructure to manipulate, visualize and in general interact with data and models. It is impossible to conduct research without software these days[20], and with the new NZOC data ecosystem many scientists may need to acquire a new set of skills

[13] Jones, Nicola. "How to stop data centres from gobbling up the world's electricity." *Nature*, vol. 561, no. 7722, 2018
[14] https://blogs.microsoft.com/blog/2021/01/28/one-year-later-the-path-to-carbon-negative-a-progress-report-on-our-climate-moonshot/
[15] https://www.apple.com/uk/newsroom/2020/07/apple-commits-to-be-100-percent-carbon-neutral-for-its-supply-chain-and-products-by-2030/
[16] https://sustainability.aboutamazon.com/environment/sustainable-operations/carbon-footprint
[17] https://newsroom.ibm.com/2021-02-16-IBM-Commits-To-Net-Zero-Greenhouse-Gas-Emissions-By-2030
[18] https://www.ukri.org/news/ukri-invests-to-upgrade-uks-world-class-research-infrastructure/
[19] https://www.nature.com/articles/s41558-020-0837-6
[20] https://scholar.google.com/scholar?cluster=4561094303174532329&hl=en&oi=scholarr

to interact with that data infrastructure or integrate new profiles of people on research teams to help leveraging the new software capabilities. Research Software Engineers [21] are people in a variety of roles who understand and care about both good software and good research, hence this role is likely to be critical in facilitating the uptake of the new available software features and increasing skill levels in environmental scientists

Recommendation 14: Invest in Research Software Engineer careers to deliver UKRI/NERC world class science.

Software will need to be developed with a clear strategy for protecting the Intellectual Property Rights of the software owners. The aspiration is to have an open data ecosystem to foster collaboration and to promote interoperability, although it is recognised this will not be possible in all circumstances.

Recommendation 15: Ensure there is a software licensing strategy that fosters collaboration and innovation, boosting the environmental sciences software ecosystem, whilst maintaining secure code.

Ship based infrastructure

Modern research vessels have capable on-board computing equipment to run the scientific required workflows during the expeditions, and to store the data gathered during operations (in the range of 25 to 100 TB depending on the vessel). These capabilities are racked mounted servers accessed using local area networks. The following table includes the capabilities of some of the most important research vessels in the world.

Ship	**Infrastructure type**	**Storage**	**Current bandwidth**	**Current average data per expedition**
RRS James Cook	Virtualization cluster	48TB	1.5 Mb/s	100 TB
RRS Discovery	Virtualization cluster	48TB	1.5 Mb/s	100 TB
RRS Sir David Attenborough	Virtualization cluster	250TB	1.5 - 3 Mb/s	100 TB
R/V Falkor	Virtualization cluster HPC (nVidia Tesla)	Unknown	>3 Mb/s	Unknown

Table 2: Current R/V computing hardware

From the table above, most of the computing clusters on-board, while very capable have been designed to run workflows around acquisition of data, and in some cases to provide VMs for experiments, but just one of the studied vessels (R/V Falkor) include HPC capabilities allowing scientist to run high precision simulations and models helping on the expedition guidance and execute very power intensive workflows. The Falkor allows VPN access from the shore-side, enabling sciences on remote to set up and run their own experiments.

It can also be seen the data storage, very variable from ship to ship, is mainly dedicated to storage the data gathered during the expedition. For the NZOC concept to be achieved, mini data-lakes will need to be

[21] https://rse.ac.uk/what-is-an-rse/

deployed on board, storing data required for local computation of models and to storage all the data gathered by the ship and the autonomous assets that will be operating around the ship and amplifying its capabilities.

It can also be seen that while the bandwidth has been increasing the most capable vessels in terms of communications can just send and receive data at a maximum speed of 3 Mb/s while at sea. With an average of 100 TB of data gathered by cruise, at the current peak speed it would take 386 days using the entire bandwidth to transfer all the data. Even if we aim to transfer 10% it would use too much bandwidth. While we can assume the bandwidth will increase (as noted above with Starlink and OneWeb ambitions), the likelihood is the data gathered will increase at an even faster pace. Processing of data on-board of ships will become increasingly important to enable the transfer of information to shore-side rather than data that requires more bandwidth and to take more informed decisions on-board and remotely.

Recommendation 16: Invest in computing infrastructure to deploy and run components of the data ecosystem on the ship, including HPC and data lake facilities

Recommendation 17: Commission a study to design and develop a seamless infrastructure between ships and on-shore facilities, allowing the deployment and running of models on ships, and data transfer between nodes, providing consistency between different parts of the system.

Edge Computing

The NZOC concept considers the ship and MAS platforms as nodes of data ecosystem that gathers and processes data. The ship can be considered as an extreme case of edge computing that encompasses very big capabilities, but that in any case needs to be treated as node. Some of the hardware requirements have been considered in the previous section.

Edge computing on the autonomous platforms

Autonomous platforms can be categorized under 2 big groups in terms of edge computing capability:

- Low power platforms like gliders or very small AUVs. These platforms have a very limited edge computing capability and will not be able to make big calculations on-board, but it is reasonable to think that the current trend of miniaturization and reduce power computing will allow these platforms to increase their autonomy and the ability of on-board calculations. An example of edge computing on gliders are the reporting of cetacean detections form passive acoustic data or the derivation of water currents form ADCP sensors.
- High power platforms, like high class AUVs and surface vehicles, are usually larger vehicles that can accommodate a larger energy budget and more powerful computers. It is conceivable to run ML systems on board of this platforms to process data. This has clear applications, including increasing the level of autonomy and decision making of an AUV when submerged or in cases of interpretation of high-density datasets such as image or video data (e.g. BIOCAM)[22].

Edge computing on moorings

Moorings have always been edge computing platforms (before the concept existed); platforms without connectivity or very low bandwidth, but equipped with computing capabilities.

[22] https://ocean.soton.ac.uk/biocam

Moorings will continue to play a critical role in observing systems as reliable sources of time series, and there are several trends to observe as they evolve:

- More computer power installed within the moorings, either as part of instruments or central hubs serving multiple instruments, allowing the execution of more complex in-situ data workflows (including ML).
- Better local connectivity, to communicate with nearby platforms, this will allow routine hoovering of data by autonomous robots and ships in the vicinity, reducing the time it takes for the data to reach end-users (traditionally this can take years).
- Increase the usage of satellite communications, leveraging newer satellite constellations and lower prices.

Recommendation 18: Consider the case of integration of edge computing into the NZOC data ecosystem.

Cyber security

The cybersecurity landscape keeps changing and is becoming more and more complex for organizations; it is not enough securing the internal organization network, but it is essential to also secure edge nodes than can be deployed in many locations across the world and are needed to send data back. Organizations are also exposing more digital services for the different stakeholders, exposing many more points of attack for cyber-criminals. Currently NOC cyber security efforts are based on the UK government cyber essentials framework[23]. All these concerns are valid for the NZOC digital twin, that is made up of a big cyber infrastructure, receiving data in near-real time and providing services to scientists. Considering the NZOC aspirations of including more assisted, or even automated, decision making for the operation of research vessels and MAS platforms, system integrity and trustworthiness is key.

During this project, an independent cybersecurity study by the company Marine AI has been commissioned. Their finding and recommendations cover the following areas:

- Cyber security on the ships, detailing the cyber systems that can be exposed and how.
- Security around swarms of vehicles, and the potential issues around many systems operating together but controlled from a central location. The report highlights the potential usage of ML systems to detect intrusions.
- An approach that secures the nodes (ship, MAS platforms and sensors) and the central infrastructure deploying ML detection systems.
- Industry best practices.
- Cyber security of shore side servers and infrastructure.

The report is available as an appendix.

Several cybersecurity areas of interest can be identified form the report findings relevant to the NZOC data ecosystem:

1. The central cyber infrastructure: this includes all the servers required to provide services, run models, and store data centrally. There can be a clear advantage of deploying all this infrastructure on a cloud provider that provides cybersecurity at scale.
2. The ship-based infrastructure; will need to be secured by the organization following industry standard best practices.

[23] https://www.ncsc.gov.uk/cyberessentials/overview

3. The edge nodes, including MAS platforms and deployed sensors; requires more specialized cybersecurity approaches due to the niche nature of the platforms.
4. The communication channels; usually controlled by third-party organizations but following industry standards when designing and deploying these platforms, like the usage of encrypted communications and encrypted storage.

Recommendation 19: Follow industry standard best practices on cybersecurity, and in particular align with the UK Digital Twins Programme on cybersecurity

Data processing

In the context of digital twins there are different elements of data processing which centre around the data lake. Namely from the sensor to the data lake, from the data warehouse to the data lake, by users of data within the data lake, and the curation of processed data of long-term value from the data lake to data warehouses. Each of these will be covered separately.

From Sensor to Data Lake

Timely delivery of data from sensors to the data lake is required to facilitate near-real-time applications and decision making. Thus, there is a need for data flows standardisation and automation. Two technologies to achieve such automation are OGC sensor web enablement[24] and Internet of Things[25]. These two options where available have the potential to provide the standards to readily ingest data in to the NZOC digital ecosystem. *However, t*o date uptake of such technologies has been limited in the marine domain and largely restricted to demonstrator projects (FP7 oceans of tomorrow projects[26]) or reference implementations (such the Spanish research vessels). The Oceanids C2 solution developed by NOC for autonomous platforms creates a command-and-control layer on top of disparate non-standardised platforms with different manufacturer specific interfaces and standardised data to a marine community NetCDF format. Further to this the UK digital twin programme is developing standards for digital twins and the concept of edge computing. Consequently, the transfer of data from the sensor to the data lake may need a pragmatic solution where community and international standards are used when available with bespoke tools/services to reformat and expose data to the data lake when no standards are available.

From data warehouse to data lake

The data lake will require additional data beyond NRT data feeds for purposes such as autonomous piloting and planning and data quality assurance. *This additional data will come, at least partly, from the data warehouse.* The research community is moving towards adoption of the FAIR data principles which define community specific protocols for accessing and describing data. As such the digital twin will need to be able to interact with FAIR implementations that may differ by domain e.g. the marine community use ERDDAP, the ecology community Darwin Core etc. Additionally, FAIR is a relatively new concept with data centres progressively moving towards implementation of the principles so the digital twin will need to support the data centres to develop interfaces that meet their specific requirements.

[24] https://www.ogc.org/node/698
[25] https://en.wikipedia.org/wiki/Internet_of_things
[26] https://op.europa.eu/en/publication-detail/-/publication/85b05ee8-7f0b-49ae-80ba-0bbb811de915

From Data Lake to Data Warehouse

The NZOC digital ecosystem will generate datasets of long-term value within the data lake which need to be curated by data warehouses for long term dissemination. Many data warehouses are developing tools to enable automated data submission with humans in the loop to triage data ahead of curation and subsequent dissemination (NOAA, NOC-BODC, PANGAEA, etc). To expedite the curation of datasets arising from the digital twin, collaboration is needed with data warehouses to develop automated data submission routines including the metadata and data formats to reduce the workload required or data to be curated.

Processing by the User

The NZOC data ecosystem should empower users to transform and interact with the data, empowering them to leverage the developed infrastructure for their own activities (environmental research, computer sciences, industry, decision making).

To do so, the NZOC data ecosystem will provide a flexible framework enable users to deploy their own workflows to access and transform data in sustainable and scalable way. This framework should be tied to training to enable users to gain the necessary skills to maximize its usage. The framework consists of the following parts:

- Software version control. All the software should be written using best practices like versioning. The framework should allow users to deploy code from different version control management solutions (for example github, gutbucket or gitlab), but will provide a default option fully integrated in the pipeline, allowing users to easily automate the testing and deployment of their pipelines.
- DOI Generation. To make science reproduceable the code use for experiments that will be published, needs to have an associated DOI. There are providers like ZENODO [27](CERN) doing this. This capability should be seamlessly integrated within the data ecosystem.
- Data pipeline deployments. How the code is deployed. It should be sandboxed, and scalable. The usage of containerized technologies is applicable here.
- Access to data. Users should have quick access to the required data inside the environment, enabling the code to do things.
- Build on the concepts of virtual research environments to enable users to work on large datasets within the digital twin environment.

Recommendation 20: A pilot use case be developed for the data ecosystem that embodies a range of use cases to allow the scoping of the software required to fulfil the processing needs of a future data ecosystem

Data Management

Effective data management is a requirement of the NZOC Data Ecosystem and best practices need to be adhered to. The importance of data management was highlighted in the OceanObs'19 community statement[28]. Further to this the Ocean Best practices system[29] has been created to facilitate the sharing and documentation of best practices[30]. Depending on the users' needs, based on the TPOS second report[31] produced by NOAA, it is estimated that 10% of research funding is required to support research with fit for

[27] https://zenodo.org/
[28] https://www.oceanobs19.net/wp-content/uploads/2019/09/OO19-Conference-Statement_online.pdf
[29] https://www.oceanbestpractices.org/
[30] https://www.frontiersin.org/articles/10.3389/fmars.2019.00277/full
[31] https://tpos2020.org/project-reports/second-report/

purpose and timely data management services. The two fundamental principles that need to be adhered to meet community best practices are TRUST and FAIR.

The FAIR guiding principles for data [32,33] describe how data need to be delivered to community standards:

- Findable
- Accessible
- Interoperable
- Reusable

Application of the FAIR principles ensure that data are accessible in a standardised way with all supporting metadata, unique identification (enabling data to be cited), machine actionable interfaces & formats, and data usage licences. To enable digital twins of the marine environment the application FAIR principles will need aligning with the national digital twin programme and the pathway towards an information management framework[34].

The TRUST Principles for digital repositories[35] are equally appropriate for the NZOC Information Management Framework:

- Transparency: To be transparent about specific repository services and data holdings that are verifiable by publicly accessible evidence.
- Responsibility: To be responsible for ensuring the authenticity and integrity of data holdings and for the reliability and persistence of its service.
- User Focus: To ensure that the data management norms and expectations of target user communities are met.
- Sustainability: To sustain services and preserve data holdings for the long-term.
- Technology: To provide infrastructure and capabilities to support secure, persistent, and reliable services.

The TRUST principles ensure irreplaceable environmental data are preserved for long term reuse enabling long term time series to be constructed and in turn climate science.

TRUST and FAIR are applicable in different contexts within the NZOC data ecosystem. Any data generated within the ecosystem that require long-term management need to be deposited in a TRUST accredited data repository. This does not necessarily apply to short-term data that are included in the data lake. The application of FAIR is in cases where data are delivered to users and apply at the interface between data warehouses and the NOZC data system and to data products created within the NZOC data ecosystem. Research is needed to develop and align FAIR with digital twins and application of the FAIR/TRUST principles will ensure services developed as part of digital twin activity will meet the NERC data policy[36] and EU H2020 data policy[37] for research data.

Effective interfaces to access and manipulate data that meet user needs are crucial in creating digital eco-system that engage users. Building a FAIR data service layer that numerous applications for specific users and

[32] https://www.nature.com/articles/sdata201618
[33] https://www.frontiersin.org/articles/10.3389/fmars.2019.00440/full
[34] https://www.cdbb.cam.ac.uk/files/the_pathway_towards_an_imf.pdf
[35] https://www.nature.com/articles/s41597-020-0486-7
[36] https://nerc.ukri.org/research/sites/environmental-data-service-eds/policy/
[37] https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-dissemination_en.htm

communities enable this. Expert users can access and manipulate data with tools like Jupyter Notebooks[38] or virtual research environments like Google Earth Engine[39]. Non-expert users need an intuitive interface that allows easy access to data and information for their specific requirements, an example of one such interface is DigitalOcean Ireland[40].

Ensure that data management is resourced for datasets and products arising from the NZOC data ecosystem enabling its deposit, curation and delivery by data warehouses to the broader community by TRUST certified repositories.

- Work with funders and projects to ensure datasets of long-term value generated by digital twins are included in data management plans and resourcing

Work with, resource and support data warehouses to provide FAIR data interfaces to the NZOC data ecosystem and digital twins.

- Collaborate with a resource data warehouses to augment their FAIR delivery to meet the needs of digital twins
- Collaborate with data warehouses to define and develop the FAIR protocols needed by digital twins

Ensure data within the data lake carries sufficient provenance and quality information to enable its utility for applications in the NZOC data ecosystem to be unambiguous.

- Collaborate with data warehouses to define the provenance and quality information required within the NOZC data eco system
- Support data warehouses to implement the agreed provenance and quality information within FAIR delivery
- It needs to be clear to users where the definitive version of a dataset resides, be it in the data lake and subsequently the data warehouse once it is archived.

Recommendation 21: Develop a data management infrastructure that allows effective deposit, curation and access to NZOC datasets from TRUST accredited facilities, with FAIR interfaces and carrying sufficient provenance and quality information.

Recommendation 22: Sensor / platform development should include the end-to-end data management as an intrinsic, and where needed funded, part of the design from the outset.

Recommendation 23: The NZOC digital ecosystem enables unambiguous data provenance including data versioning.

Data Sciences (AI/ML)

Data science covers a broad spectrum of activities and is something that is part and parcel of traditional science. Using data science, you can uncover patterns in data that help you understand the data and improve its quality and value. But in principle data science doesn't need to use AI approaches, or indeed computers at all. The NOC, and oceanographers in general, have been doing data science for a long time. Artificial Intelligence most broadly is the use of computers to allow us to make the most of our data. It can be thought of as the ability that can be imparted to computers which enables these machines to understand data, learn

[38] https://jupyter.org/
[39] https://earthengine.google.com/
[40] https://www.digitalocean.ie/

from the data, and make decisions based on patterns hidden in the data, or inferences that could otherwise be very difficult for humans to make manually, either due to the volumes or nature of the data, or the speed at which the inferences need to be made. AI also enables machines to adjust their "knowledge" based on new inputs that were not part of the data used for training these machines. AI relies on having training data and so works best when data volumes are high – as datasets become bigger and less manageable using traditional methods (Big Data) then increasingly sophisticated methods are needed to manage the size of the data. Machine learning is not only an opportunity but given this rapid growth in its use it is critical that the access to and understanding of Machine Learning methods is developed within the NZOC community. The scope and speed of developments create challenges and to make data sciences a fundamental part of the NZOC data ecosystem we will need to:

- increase data science skills within the workforce
- implement appropriate AI/ML appropriate software and hardware infrastructure
- undertake software retooling of the workflows to include machine learning

Recommendation 24: Create a community of support around AI/ML in the NZOC data ecosystem

Recommendation 25: Develop a skills base in the NZOC community in AI/ML to support developing the data ecosystem workflows

Data sciences has a role to play in several parts of the data ecosystem:

- Quality control of observation data, both prior to use in the data applications within the NZOC data ecosystem and prior to delivery to the data centres
- Processing of the observations from raw data from the sensors to the data on the data lake. This will include automatically undertaking data transformations such as aggregating, averaging and extrapolating the data before it is exposed to the user.
- Exploration, and categorisation, of high-volume data such as satellite retrievals or ocean acoustics from fibre optic cables
- Data analysis for process understanding
- In support of predictive modelling, either through combining AI and predictive modelling to improve the quality or computational efficiency of the models, or through transforming the model outputs through post-processing to improve the model skill

Known Unknowns

A number of known unknowns will need to be addressed in preparing for the future NZOC data ecosystem, and are largely addressed within the body of this report but are included here for completeness:

- Data volumes within the data ecosystem
- Data volumes to be transferred in and out of the data ecosystem via satellite telemetry or other communications methods
- Offshore infrastructure for compute, data storage and power ("hubs") and whether they will become available on the NZOC timeframes

It has also been hard to give a good estimate the carbon cost of the NZOC data ecosystem, which will be largely be driven by the compute costs. Where the compute is being undertaken onboard the ship or autonomous vehicle the carbon emissions due to the power consumption for that compute will be dictated by the carbon efficiency of the vessels themselves. If their power is net zero then so is that of the compute. However, much of the compute is likely to be in the cloud or elsewhere away from the ship infrastructure, in

which case the carbon cost becomes dictated by the power consumption at source. The big tech companies are promising net zero carbon (for their supply chain and product use) on the timescales of NZOC, but it is not clear how realistic this is. Understanding the balance of the location of the compute infrastructure between cloud, shore and ship is something that needs to be done to estimate the NZOC data ecosystem carbon costs.

Equality, Diversity and Inclusion

The ability for scientists and technologists to more completely engage with the data collection process irrespective of their location is expected to have a significant positive impact on those with disabilities, and those unable to travel due to family or caring responsibilities.

In addition, free and open data that is easily accessible to all, including without specialist IT equipment (for example on apps on mobile phones) will facilitate developing societies to access data.

A fully realised NZOC data ecosystem therefore has the potential for significant ED&I benefits.

Scientific Data Licensing and Intellectual Property Rights

Although the ambition is to move towards free and open data, scientific data is subject to a range of licensing and usage restrictions conditions, including embargoing for a period, and will continue to be so into the foreseeable future. Any data ecosystem will need to be sufficiently flexible to be able to maintain the data restriction information through the full data journey from collection to use. Where data are combined this is potentially non-trivial and requires the data management process to be well-implemented to reflect this requirement. The developers of the data ecosystem will need to reassure scientists and the data owners that their data is being managed properly and all elements of the IPR and fair usage requirements are being respected.

Other considerations relating to data and its legal status that will need to be considered include:

- Law of the sea and data implications,
- Personal data subject to the data protection act[41], particularly if Citizen Science data are received,
- Regulations surrounding the use of communications systems,
- Sharing of data in foreign EEZs,
- Environmental information regulations (2004)[42],
- CARE principles[43]

Implementing a data ecosystem that allows the propagation of varying ethical, licensing and IPR restrictions through the data journey is fundamental. This would have to include allowing variable access rights to data from the data ecosystem. Open access to data vs moratoriums on release until PIs have had the chance to publish will be an on-going discussion. Any data ecosystem should facilitate open and easy sharing but allow restricted sharing with clear and actionable changes to the data visibility depending upon the data licensing depending. The tools to facilitate this will need to be implemented, and underpinned by appropriate metadata.

[41] https://www.gov.uk/data-protection
[42] https://www.legislation.gov.uk/uksi/2004/3391/contents/made
[43] https://www.gida-global.org/care

Recommendation 26: NZOC develops a data policy that covers the range of likely IPR and use cases, and links the data management framework to the policy to allow data with a range of IPR conditions to be handled.

Review of Commercial Priorities and Opportunities for Collaboration

The sorts of digital resources described within this report are too broad in scope for NOC, or the ocean science community, to develop in isolation. Partnerships outside our traditional collaborations are likely to be needed for these ambitions to be realised. The engagement within the NZOC project of non-academic actors demonstrates there are organisations who are ready to engage with the ocean community for this sort of activity. The ocean community has domain expertise in ocean science and ocean observing, but will need to partner, most notably, with technology companies who are developing much of the technology needed and with the data sciences communities who are developing the tools needed to make the most of the ocean observations. There are also a wide range of industries that operate in the marine environment who are developing similar infrastructures that could be interesting to partner with. The increasingly visible sustainability agenda means industry and technology companies, including those with a historically poor carbon record, are keen on partnering with the science community in activities that are, or at least could be perceived to be, contributing to a more carbon neutral world. This provides an opportunity for the ocean sciences community to engage with a broader range of industrial partners than perhaps might otherwise have been possible; it is however incumbent upon our community to ensure that any partnering provides genuine value to the net zero oceanographic capability and any relationship must be ethically evaluated.

Recommendation 27: Ensure there is an appropriate ethical framework for partnering with technology and industrial partners.

Big Tech

Technology companies such as Microsoft and IBM, both of whom were represented at the Data Ecosystem Workshop, are developing data sciences frameworks that include the software and hardware required to deliver the data sciences enabled NZOC data ecosystem. They have serious ambitions around their carbon emissions, with Microsoft for example planning to be net zero for carbon (including offsetting) by 2030 and net negative beyond that.

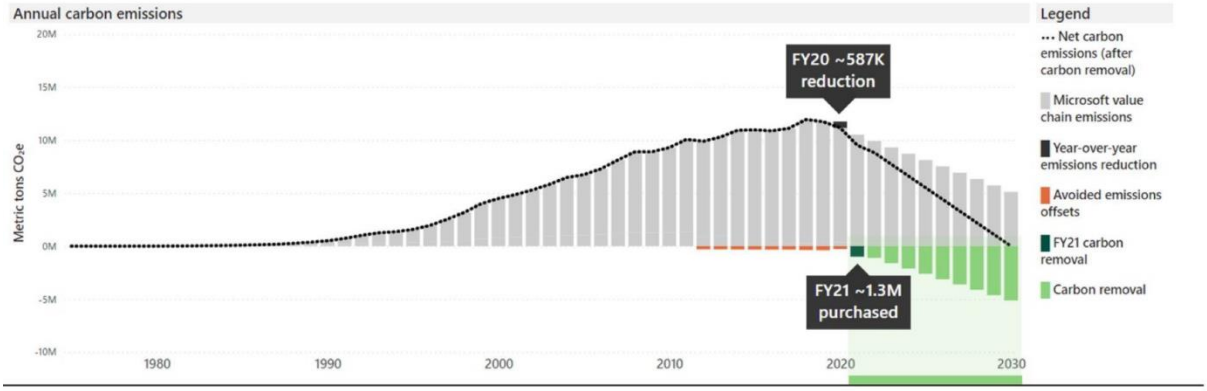


Figure 15: Microsoft's net zero ambitions to be net zero by 2030 (including offsetting) as presented at the NZOC Data Ecosystem workshop by Prof Arribas, Microsoft Sustainability Science Lead for Europe. Other organisations have similar ambitions.

It is therefore important to engage with technology providers to design and develop systems on the 2030 timescale that meet the NZOC data ecosystem requirements. It is only through partnering with these technology companies we will be able to develop the infrastructure required to build the NZOC data

ecosystem. To set the scene for the large-scale NZOC data infrastructure, smaller pilot projects will need to be developed to build the expertise in working in this way. Partnering with the big tech companies seems a quick win to make progress on the data ecosystem infrastructure build. We should be ensuring that we partner with those organisations with realistic ambitions for net zero on the 2030 timescales.

Recommendation 28: Build relationships with the big tech companies with realistic and ambitious net zero plans through pilot data ecosystem builds.

Academia

Within UKRI and the academic community there lies a solid foundation of data sciences skills. They are actively seeking real-world applications for their skills, and a close relationship between oceanographers with data sciences skills and leading developers of data science techniques in, for example EPSRC, will provide the most effective way for developing the AI/ML tools required for the NZOC Data Ecosystem. UKRI have significant ambitions in the Digital Twin arena, although these are not yet fully aligning with NERC Digital Environment and NZOC data ecosystem plans. Ensuring that the NERC / NZOC activities align with the broader UKRI ambitions is important to ensure NERC / NZOC benefit from the EPSRC and STFC capabilities in this area. As an aside, UKRI recognise it is important to ensure that the built environment is supported by suitable environmental information and so a mutually beneficial partnership seems realistic. The Centre for the Digital Built Britain has already been mentioned widely within this report; any engagement is likely to include some links to the CDBB, which is providing the overall framework for a lot of these activities.

Recommendation 29: Develop pilot projects involving data science specialists (in academia and the broader UKRI community) and ocean observers and scientists

Defence

The use of autonomy for generating the data for Defence Geointelligence has been an ambition of the Royal Navy for a number of years. The MOD's Digital Strategy[44] makes it clear that the Defence sector is taking the digital revolution, and particularly the use of data and data sciences as part of a modern way of working for the armed forces, very seriously. There are obvious and clear overlaps between the sort of data ecosystem that the Navy will require at sea, connecting their autonomy to their operational units and the command and control needed to do so, with data ecosystem described here for NZOC.

The Royal Navy are already investing significant research effort in this area. An example is the QinetiQ led Maritime Autonomous Platform Exploitation (MAPLE)[45] project. Maple is developing the autonomy command and control systems required for the Royal Navy with a focus of to revisit the Information Architecture, and functionality, to generate a specification that will enable the MoD to procure a Command and Control (C2) system (Smith et al, 2020). These specifications could both inform and be informed by future NZOC data ecosystem developments. The defence interests go beyond the UK, with NATO also working on a Digital Ocean Project with a very similar concept to the NZOC Data Ecosystem. The opportunities to collaborate with, and contribute to, the UK and NATO Defence priorities will be significantly reinforced if NOC are leading the way in the digital space around ships, autonomy, and sensing of the environment.

[44] 20210421 - MOD Digital Strategy - Update - Final.pdf (publishing.service.gov.uk), accessed 11/06/2021
[45] QinetiQ to lead next phase of Maritime Autonomous Platform Exploitation (MAPLE) project (navyrecognition.com)

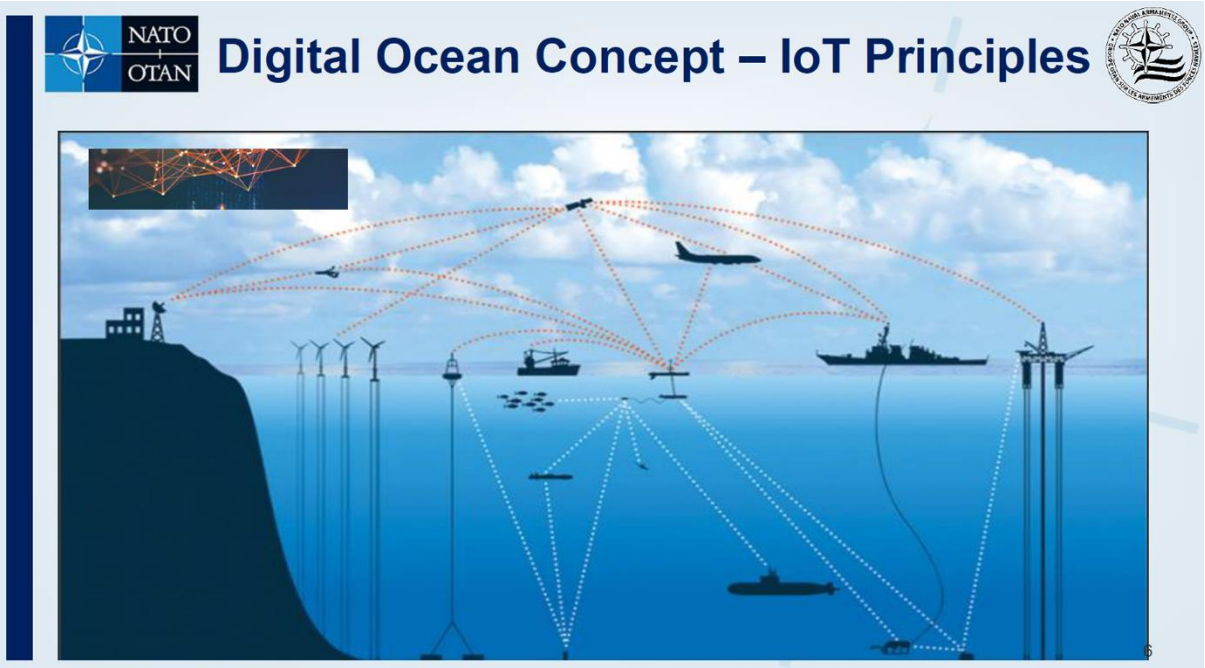


Figure 16: A schematic of the NATO vision of the digital ocean which very closely mirrors the NZOC Data Ecosystem vision presented in this report. Presented by Sean Trevethan, Director NATO MUS Innovation and Coordination Cell at the Unmanned Maritime Systems Technology Conference, May 2021.

Recommendation 30: Explore opportunities for co-development of the NZOC data ecosystem with interested parties working within the Defence sector

International Activities

The digital environment is an increasingly active area internationally. DestinationEarth has already been mentioned and will hopefully provide opportunity for collaboration and co-development of the some of the infrastructure and best practice discussed throughout the report.

The UK is active in the UN Decade of Ocean Science for Sustainable Development, and in the Global Ocean Observing System (GOOS), which is framing a significant number of programmes and projects at the time of the writing of this report. The UKs G7 Presidency has given the opportunity to promote relevant activities in this area, and there will be a Digital Twin workshop in early 2022 which will allow the UK to show leadership in this area and foster international partnerships with leaders in the field. Longer term, the UN Decade Programme Digital Twins of The Ocean (DITTO) will provide thought leadership in this area. UK engagement in the programme is already secured and will be continued.

Marine Industry

According to the Inmarsat Report: Industrial IoT on land and at sea[46], it is the maritime industry that has adopted IoT solutions more than any other sector and will continue to accelerate in this area. As the marine sector moves to increasingly crewless vessels, with remote monitoring and control, the communications infrastructure required to support developments within the Maritime Sector and may accelerate some of the communications capabilities needed for the NZOC data ecosystem. The development of Digital Twins is

[46] Inmarsat_IIoT_on_land_and_at_sea_Maritime.PDF accessed 11/06/2021

already progressing in the maritime industry, for example at Kongsberg where Digital Twins and artificial intelligence are already well advanced[47].

Additionally, some of the data that the shipping industry will collect as part of their smart systems approaches will undoubtedly be of benefit to the ocean community providing opportunities for adding to the network of observations in support of the data ecosystem (see Citizen Science below).

Interestingly, according to the Inmarsat report the most important impediment to using IoT solutions in the maritime space is the lack of timeliness in data, with just over half of the report's contributors considering the time lag between data collection and availability for use being the primary reason not to take advantage of IoT approaches. This chimes with the NZOC requirement for increased real-time data and suggests there may be solutions within the industrial space for the telemetry that will support the NZOC data ecosystem in the future.

Having said the above, it is not obvious that a pro-active engagement with maritime industries will provide any added value above and beyond the progress expected independently within the digital space and so no recommendation on engagement with the Marine Industry is being put forward.

Citizen Science

Citizen Science, and the Internet of Things, brings a whole new dimension to the potential for data gathering. Vessels and infrastructure in the marine environment will increasingly be equipped with sensors, and themselves be working within a decision-making process that is informed by ever increasing data collection in the marine environment.

This provides an opportunity to engage with marine users to share data, an opportunity that should continue to be monitored and avenues explored to ensure that the most is made of the citizen science data available both to add value to the research cruise decision making but also to add to the data available for use by scientific users of NZOC data.

Recommendation 31: Support and engage with Citizen Science opportunities for gathering data in the marine environment.

Carbon Calculation

The greening government: ICT and digital services strategy 2020-2025[48] provides a bases from which to consider carbon implications and guide the recommendations on the carbon footprint for a Net-Zero digital ecosystem. Additionally, NOC has a published environment policy and on-going ISO14001 certification with the goal of reducing carbon costs of ICT. The NZOC digital ecosystem needs to be aligned with an incorporated into existing carbon monitoring and greening activity.

The Greening Government Commitment separates carbon emissions into 3 categories, Scopes 1 -3 which are the basis for mandatory GHG reporting in the UK.

[47] https://www.kongsberg.com/digital/solutions/kognitwin-energy/
https://www.kongsberg.com/digital/

[48] https://www.gov.uk/government/publications/greening-government-ict-and-digital-services-strategy-2020-2025

- Scope 1 —The Green House Gas (GHG) emissions that are made directly, for example while running its boilers and vehicles. Includes ships emissions. The data ecosystem may lead to Scope 1 emissions if it leads to increased ships emissions or running of locally generated power.
- Scope 2 — These are the indirect emissions, largely through the consumption of electricity or energy it consumed for heating and cooling buildings or running IT infrastructure. Where the NZOC data ecosystem has capital infrastructure for supporting the data ecosystem these will come under Scope 2 emissions.
- Scope 3 — This includes all emissions within the NZOC value chain, through bought in services or indirectly because of supply chain emissions. These would include any cloud-based compute server emissions, or those incurred through using data communications networks. It also includes any business travel. These are likely to be by far the biggest contributor to the NZOC data ecosystem carbon cost and the hardest to predict.

Scope 1 – direct fuel costs

Research vessels require fuel to operate – this includes the movement of the ship and generating electricity for all the internal workings, including on-board computing facilities. It is thought that the equivalent of a small server room, similar to the present NOC server room, it likely to be required on board each research vessel. The direct fuel cost of this server room is likely to be small in comparison to the ship's fuel requirements. The estimated annual total energy requirement of one research vessel is **35.6 Mill kWh**, assuming 2970 tonnes of Marine Gasoil (tMGO) pa fuel usage (see Table 3) and a conversion factor of 12,000 kWh per tMGO[49]. In contrast, the total power consumption of a server centre, based on present NOC server power consumption (see Table 4), is of the order **0.5 Mill kWh** per year.

	RRS MGO consumption (tonnes/day)	**Days per year**	**Fuel Consumption (tonnes / year)**	**Approx Carbon cost (tCO_2 / year)[50]**
Passage	11	70	770	2500
Science	9	225	2025	6500
Harbour	2.5	70	175	600
Total		**365**	**2970**	**9,600 tCO_2 / year**

Table 3: Approximate present day Marine Gasoil consumption by the NERC RRS Discovery under present day operating conditions. Passage is based on travel speed of 10 knots using two engines; faster steaming results in higher passage fuel consumption. Fuel consumption whilst doing science is lower than in passage as the ship spends more time at low speed or stationary.

Whether the present-day compute power consumption at NOC is a valid comparison is difficult to assess. The present trend is for reducing power associated with the compute, due to efficiency improvements, and this may continue. However, the power consumption of present technologies is not reducing apace with the increased processing power and is likely to become increasingly limiting. These ballpark estimates suggests that the power consumption of compute requirements on a future NZOC ship would be a small, but not insignificant, proportion of the ships total power consumption. The carbon efficiency of the future ship power source will obviously dictate whether this is a major concern; it seems unlikely but should be considered as a part of the net zero design considerations for the ship.

[49] Typical conversion factors for metered energy (vesma.com) accessed 03/08/2021
[50] Using 3.212 tonnes CO_2 per tonne fuel, https://www.ipcc-nggip.iges.or.jp/public/gp/bgp/2_4_Water-borne_Navigation.pdf accessed 03/08/2021

Scope 2 – indirect emissions

Indirect emissions will result from any infrastructure run as part of the NZOC function that requires non-locally generated power, such as mains electricity supplied computer resources. It is difficult to map the requirement for this area of emissions without first having a clear idea of the architecture of the data ecosystem, and how much of it will be on board a ship versus elsewhere, and if elsewhere whether the resource will be physically part of the NZOC system or a bought in (cloud) service. The (in prep) NERC Digital Strategy and its successors will set the NZOC agenda for compute strategies. The direction of travel is to a hybrid mix of local and cloud compute, consistent with the concepts of edge computing, so it is to be expected that an increasing proportion of the NERC data driven compute use will be in Scope 3 emissions.

ICT energy usage is monitored as part of NOCs on-going ISO14001 certification. The Carbon cost of the NOC Southampton is calculated as estimated tCO2e values based on the NOCS server room UPS meter for the last 5 years. Table 4 shows the year-on-year reductions in carbon cost of the server room. The drop in associated carbon emissions is primarily through decreasing Carbon Conversion Factors; every year the government publishes new figures for the amount of carbon associated with each unit of energy used that year, this is falling rapidly at the moment due to the increasing proportion of renewables in the national energy mix.

	EM93 (kWh)	CCF (kgCO2e/kwh)	kgCO2e	**tCO2e**
2016	518,275	0.41205	213555.2	**213.5552**
2017	365,128	0.35156	128364.4	**128.3644**
2018	321,495	0.28307	91005.59	**91.00559**
2019	324,602	0.2556	82968.27	**82.96827**
2020	341,043	0.23314	79510.77	**79.51077**

Table 4: Carbon costs per annum as estimated tCO2e values based on the NOC Southampton site server room UPS meter for the last 5 years

It is expected that the move to carbon neutral power at the NOC sites will reduce the carbon cost of computing over the coming years, in line with GGC commitments. This suggests that despite a likely future increase in the compute capacity required for the NZOC data ecosystem, the combination of the move to bought in / cloud services and improved carbon efficiency of the NOC estate will lead to a reduction in the Scope 2 data ecosystem carbon emissions.

Scope 3 – supply chain and business travel carbon costs

The largest contributor to NZOC carbon emissions associated with the data ecosystem is likely to be cloud-based compute emissions.

Cloud providers are increasingly conscious of their carbon emissions, with large tech companies such as Microsoft committing to ambitious net zero ambitions. As described in previous sections a key consideration when developing infrastructure and identifying a cloud services provider is to include carbon cost implications in the tendering process. An example of a carbon calculator for cloud infrastructure has been provided by Microsoft Azure[51]. One consideration when considering cloud versus local compute is the Power Utilisation Efficiency (PUE). PUE is a concept developed by the Green Grid, a non-profit consortium collaborating to

[51] https://azure.microsoft.com/en-us/blog/microsoft-sustainability-calculator-helps-enterprises-analyze-the-carbon-emissions-of-their-it-infrastructure/?cdn=disable

improve the resource efficiency of data centres and regularly shows cloud-based data centres run by big tech to have significantly better PUE stats than local solutions, which is unsurprising given the scalability of their solutions. Any solution developed by the NZOC data ecosystem should explore concepts such as PUE in the procurement process.

One area that the data ecosystem will support reduced Scope 3 emissions is in the reduced travel and transport related to the travelling to join the research expedition. An effective data ecosystem will increase the engagement of scientists in the Research Expedition remotely, so will help reduce travel by facilitating fewer flights with less science presence on board.

Recommendation 32: Use the greening government: ICT and digital services strategy 2020-2025, NOC environmental policy and on-going ISO14001 certification as the basis for the NZOC digital ecosystem carbon strategy

Recommendations Summarised

For ease of reference, the full list of recommendations is presented below, sorted by broad category of activity or theme.

Pilot Studies

9: User requirements for the NZOC data ecosystem applications should be gathered, and iteratively enhanced following experience in developing the user interfaces.

17: Commission a study to design and develop a seamless infrastructure between ships and on-shore facilities, allowing the deployment and running of models on ships, and data transfer between nodes, providing consistency between different parts of the system.

20: A pilot use case be developed for the data ecosystem that embodies a range of use cases to allow the scoping of the software required to fulfil the processing needs of a future data ecosystem

28: Build relationships with the big tech companies with realistic and ambitious net zero plans through pilot data ecosystem builds.

29: Develop pilot projects involving data science specialists (in academia and the broader UKRI community) and ocean observers and scientists

30: Explore opportunities for co-development of the NZOC data ecosystem with interested parties working within the Defence sector

Digital Skills

1: Do a skills audit to assess present gaps in critical digital skillsets, and plan for training the future generation of the NZOC workforce

14: Invest in Research Software Engineer careers to deliver UKRI/NERC world class science.

Best Practice

2: Activities in the NZOC data ecosystem are developed following national and international best practice, and particularly following the guidelines laid down with the National Digital Twin Programme (NDTp)

6: Maintain engagement with ocean Digital Twinning activities (such as, but not exclusively, DestinationEarth).

7: Engage with Information Management Frameworks to ensure the NZOC Data Management framework meets nationally and international interoperability requirements.

19: Follow industry standard best practices on cybersecurity, and in particular align with the UK Digital Twins Programme on cybersecurity

26: NZOC develops a data policy that covers the range of likely IPR and use cases, and links the data management framework to the policy to allow data with a range of IPR conditions to be handled.

32: Use the greening government: ICT and digital services strategy 2020-2025, NOC environmental policy and on-going ISO14001 certification as the basis for the NZOC digital ecosystem carbon strategy

Hardware and Telecommunications

3: Map the computational architecture on the research ships, autonomous platforms and potential 3rd party communication architecture required for an NZOC Digital Twin

4: Scope a scalable data lake architecture to receive NZOC data, developing pilot projects to develop expertise in managing the data architecture across cloud, ship and shore-based infrastructure.

12: Scope the data communication required to support communications at sea and integration within the NZOC Digital Twin

13: Evaluate the compute infrastructure needed for the data ecosystem, including the relative benefits of on-premise, JASMIN (or other common NERC/UKRI facility) and cloud compute.

16: Invest in computing infrastructure to deploy and run components of the data ecosystem on the ship, including HPC and data lake facilities

18: Consider the case of integration of edge computing into the NZOC data ecosystem.

Data Management

8: Design an open planning ecosystem with well-defined standards and open APIs to allow shared planning between different actors of the system

10: Invest in the design and develop open standards to enable coordinated command and control of autonomous vehicles and ships between different institutions (NOC, BAS …)

21: Develop a data management infrastructure that allows effective deposit, curation and access to NZOC datasets from TRUST accredited facilities, with FAIR interfaces and carrying sufficient provenance and quality information.

22: Sensor / platform development should include the end-to-end data management as an intrinsic, and where needed funded, part of the design from the outset.

23: The NZOC digital ecosystem enables unambiguous data provenance including data versioning.

Software

15: Ensure there is a software licensing strategy that fosters collaboration and innovation, boosting the environmental sciences software ecosystem, whilst maintaining secure code

Data Sciences and Modelling

5: Develop a modelling capability that can dynamically assimilate observations in a moving frame (ship-following) at a resolution relevant for observation collection decision making, using modelling tools (e.g. NEMO, ERSEM, NEMOVAR) already well-established in the community.

24: Create a community of support around AI/ML in the NZOC data ecosystem

25: Develop a skills base in the NZOC community in AI/ML to support developing the data ecosystem workflows

Collaboration

2: Activities in the NZOC data ecosystem are developed following national and international best practice, and particularly following the guidelines laid down with the National Digital Twin Programme (NDTp)

11: Develop a multidisciplinary/multi-council strategy to increase collaborations on the development of novel autonomous planning and optimization systems to maximize the usage of autonomous assets at sea (NERC-EPSRC crossovers)

27: Ensure there is an appropriate ethical framework for partnering with technology and industrial partners

31: Support and engage with Citizen Science opportunities for gathering data in the marine environment.

References

Al-Ali, A. R., Gupta, R., Batool, T. Z., Landolsi, T., Aloul, F., & Nabulsi, A. Al. (2020). Digital twin conceptual model within the context of internet of things. Future Internet, 12(10), 1–15. https://doi.org/10.3390/fi12100163

Lin, D., Crabtree, J., Dillo, I. *et al.* The TRUST Principles for digital repositories. *Sci Data* **7,** 144 (2020). https://doi.org/10.1038/s41597-020-0486-7

Smith, P, Astle, J, & Biggs, W. (2020). Towards deployment, how the UK MAPLE architecture is being developed ready for exploitation and its role at the centre of international experimentation involving maritime unmanned systems. Presented at the International Naval Engineering Conference and Exhibition (iNEC 2020).

Tanhua T, Pouliquen S, Hausman J, O'Brien K, Bricher P, de Bruin T, Buck JJH, Burger EF, Carval T, Casey KS, Diggs S, Giorgetti A, Glaves H, Harscoat V, Kinkade D, Muelbert JH, Novellino A, Pfeil B, Pulsifer PL, Van de Putte A, Robinson E, Schaap D, Smirnov A, Smith N, Snowden D, Spears T, Stall S, Tacoma M, Thijsse P, Tronstad S, Vandenberghe T, Wengren M, Wyborn L and Zhao Z (2019) Ocean FAIR Data Services. *Front. Mar. Sci.* 6:440. doi: 10.3389/fmars.2019.00440

Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3,** 160018 (2016). https://doi.org/10.1038/sdata.2016.18

Annex: Cybersecurity Report