Digital Antarctica

High-level Current State document

July 2021

Australian Antarctic Program Partnership
Institute for Marine and Antarctic Studies
20 Castray Esplanade, Battery Point
Tasmania, 7004, Australia
aappartnership.org.au

AAPP
Private Bag 80,
Hobart Tasmania,
7001

Contents

Version information

Version	Description	Author	Date
1.0	Initial version for stakeholder review	Rob Jennings	28/05/2021
2.0	Release version	Rob Jennings	19/07/2021

Executive Summary

A major role of data centres within the Australian Antarctic Program Partnership (AAPP) is to archive and deliver data to anyone who needs and has permission to access those data, including researchers, policy makers, and members of the public.

Researchers rely heavily on data that they either generate as part of their project, collect from existing instruments, or find from existing data sources, and must ensure that any data which support their publication are readily available. Data centres are central to much of this data processing.

Digital Antarctica is an initiative that has the goal of improving the processing of Australian Antarctic and Southern Ocean data by ensuring that the data used to create that research – and the data products created from that research – are easier to find, collate and integrate. The *Digital Antarctica High-level Scope document* outlines a high-level scope of what *Digital Antarctica* aims to achieve.

This document, the *High-level Current State*, gives a general overview of the current state of data capture and delivery within the AAPP, and includes an overview of the current state of the various data centres involved with the AAPP, describing their data remit, services and standards.

AAPP Data Centre Current State

Workflow of data from capture to publication

Each of the organisations within the AAPP perform distinct functions and capture data for their specific purposes, and each organisation's function, data capabilities and processes are listed in the sections below. However, the general processes of generating, archiving and publishing data are similar across those organisations:

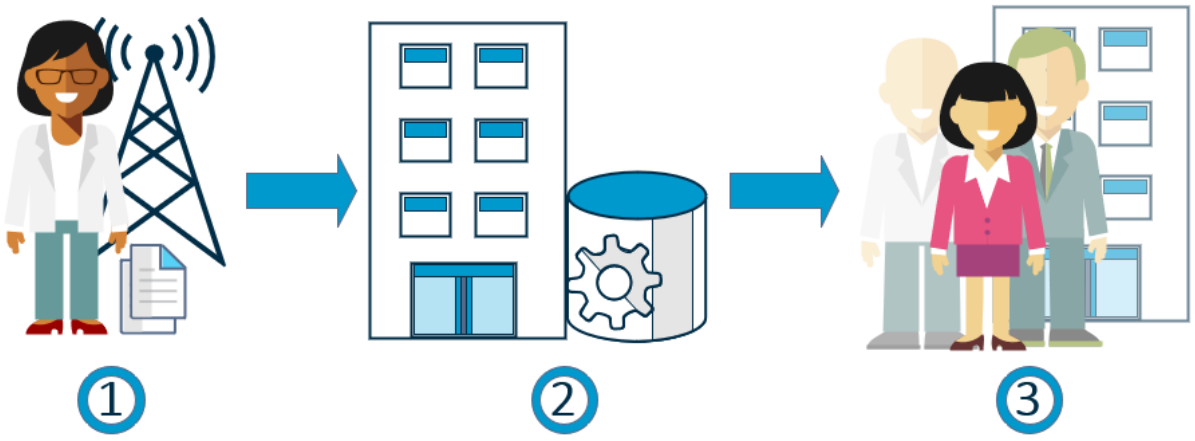

Figure 1 – An overview of the data collection process. A data collector collects or generates data which is submitted to the data centre. The data is processed and made available to the end users.

1. Broadly, raw data capture methods include observations, operational data (from sources such as installed instruments and satellites) and data routinely captured on research vessels. Data and data products are also generated as part of research projects. Additionally, data may be processed by personnel before being submitted (or made available) to a data centre for publication.
2. Within an organisation's data centre, further processing generally occurs to the data to ensure they are ready for publication. Processing, either before or after submission to the data centre, can include:
   - *Quality assurance* – Checking the data for scientific or data integrity quality. Data centres typically do not scrutinise scientific quality; however, they may check data for consistency and adherence to required data standards.
   - *Metadata processing* – Before data can be published and findable, they must have attached metadata. Each data centre has differing requirements for metadata. Most data centres will work with researchers to assist in metadata creation, and will check metadata for completeness and accuracy, updating the metadata with additional information where required.
3. Once data and metadata have been processed, the organisation's data centre will publish the data. Data is made available to the public via:
   - *Data portal/catalogue search* – Every data centre within the AAPP provides a search of its catalogue, available to the public. Some organisations provide multiple data portals catering to specific data types and user types.
   - *Web service* – Most data centres provide web services for direct machine-to-machine access to data. *This* enables regular and systemised harvesting of data, as well as direct access to data via analytical tools.
   - *Other means* – This includes direct access to the data (where required) and tools such as phone and web apps.

Australian Antarctic Division

Organisation overview

The Australian Antarctic Division (AAD) is a division of the Department of Agriculture, Water and the Environment, and undertakes and administers research projects in and around Antarctica and the Southern Ocean.

The Australian Antarctic Data Centre (AADC) is the primary data centre for the AAD as well as most data recorded for Australian Antarctic Program science projects. As such, the AADC hosts a wide variety of Antarctic and Southern Ocean research and datatypes, encompassing any science data relating to the Australian Antarctic Program.

Data acquisition and processing

Researchers for the AAD collect data manually as well as operationally via a number of methods, including meteorological facilities; ionospheric activity monitoring; seismic, magnetic and GPS networks; and hydrographic and bathymetric mapping.

Datasets are submitted to the AADC via its portal[1] or, under special circumstances, may be directly submitted to the AADC via other methods (e.g. email, cloud storage or direct transfer into the AADC S3 file store). Scientists submitting the data are responsible for the quality of the data being submitted.

As AADC data submission is closely linked to Australian Antarctic Program science projects, the AADC manually tracks those science projects to ensure that they are meeting their data capture, metadata and publication obligations, and actively contacts researchers when project data are due or overdue.

Once submitted, the AADC will create DOIs for data and will make the data available, where applicable.

Metadata and data policy

All data submitted to the AADC must have associated metadata. Metadata must be created within the AADC website (if data are submitted manually, metadata must still be supplied via the website). The AADC uses the DIF metadata standard[2], and metadata are checked by the AADC to ensure general quality of the metadata as well as adherence to that standard and to the AADC's internal DIF profile (developed to ensure compatibility with other standards, including the Marine Community Profile (MCP) and ISO 19115-1).

The AADC publishes its data policy[3] outlining its responsibilities and the responsibilities of anyone publishing data via the AADC, as well as detailing requirements and processes around such matters as licensing, metadata and publication.

Data publication and availability

In keeping with Antarctic Treaty obligations, most data on the AADC portal are available to the general public. Some data require an email address to be delivered and some data are subject to an embargo period (during which data may not be made publicly available). As per the AADC data policy, a small number of datasets are not made public as they are sensitive or commercial-in-confidence (although commercial data may undergo generalisation to render the data publishable in some circumstances), and data may be embargoed from publication for a period.

The AADC website hosts a general data discovery portal for data consumers to search for research data, as well as specialised data portals such as the Automated Weather System, Map Catalogue and

[1] *https://data.aad.gov.au/eds/dataset/new*
[2] *For more information on DIF, see https://earthdata.nasa.gov/esdis/eso/standards-and-references/directory-interchange-format-dif-standard*
[3] *https://data.aad.gov.au/aadc/about/data_policy.cfm*

Gazetteer. Data are also available via OGC web services on GeoServer[4] and RESTful APIs, which have associated documentation on their use[5]. Files are stored in, and delivered via, Amazon S3 services, which can support other cloud services and applications.

Each metadata file in the AADC's catalogue is enriched with schema.org tags to ensure that the metadata records are discoverable via Google's Dataset Search. AADC data is also discoverable via the RDA catalogue from ARDC and other external portals.

The AADC uses Matomo and Google Analytics to track site and download usage, however they have not implemented a full dashboard or reporting from these tools due to staffing limitations. As such, there is limited visibility into the usage of AADC data. The AADC website, including the data portals and usage statistic tracking, is undergoing a redesign to improve its usability and modernity. The anticipated release date for the new website will be in the latter half of 2021. This redesign will assist data creators and data consumers, and will also improve usage statistics.

Challenges and opportunities

Staffing limitations have posed a challenge for the AADC. Due to recent staffing shortages, the AADC has not had the development resources required in regards to developing, maintaining and improving the data tools required to capture, process and serve data. However, these staffing shortages have now been addressed, and the AADC is looking to address these areas as well as increase operational support for the wider Australian Antarctic Division, lifting its focus from purely data into other aspects of technical support (e.g. coding for logistics).

Bureau of Meteorology

Organisation overview

The Bureau of Meteorology (the Bureau) is Australia's national weather, climate and water agency, providing a wide range of products and services to support informed decision-making by governments, emergency services, industry and the community. The Bureau's products and services include a range of observations, forecasts, warnings, analyses and advice covering Australia's atmosphere, water, ocean and space environments.

Data acquisition and processing

The bureau collects a large amount of operational data, via routine observations which contribute to its core service (i.e. forecasting). Examples of data collections include automatic weather stations, radiosonde launches, drifting buoys, ship observations, tide observations and satellites. Data are collected by the Bureau from all over Australia and the world, including Macquarie Island and Australia's Antarctic stations. Where possible, data are transmitted directly into systems; however, due to the size and location of the Antarctic data, full satellite datasets are not transmitted immediately back to mainland Australia. Rather, they are returned by ship on physical media once or twice a year.

Once operational data are collected they are, along with data from worldwide meteorological agencies, analysed to generate a forecasting data model. Data products, such as sea ice edge analyses and forecasts, are then derived from the model. These products form the majority of the public facing data that the bureau is responsible for.

Along with its operational data, the Bureau also performs ad hoc data collection for specific requests. Data collected on behalf of other projects are generally considered to be the responsibility of those projects. In this sense, the Bureau acts as a "data collection tool", rather than as the custodians of the data.

[4] *See https://data.aad.gov.au/geoserver/web/*
[5] *Fpr an example, see https://data.aad.gov.au/aws/api/docs/*

Metadata and data policy

The Bureau is responsible for creating, quality controlling and maintaining metadata of operational data and the derived data products. Metadata are created using internal standards based on ISO 19115.

The bureau publishes information about its data catalogue, including licensing and metadata, at its "About the Catalogue" page[6]

Data publication and availability

Data and data products are findable via the Bureau's metadata catalogue[7] which provides access to all of the Bureau's data that has associated metadata. Real-time weather observations and forecasts are available via the Bureau's main website[8] and their publicly available Android and iOS apps[9]. Archived observations are also available through the website's Climate Data Online portal[10], which provides a search of the Bureau's data recorded at stations using standard spatial data search facilities.

Data are also available to find and download directly via FTP[11], via their WIS catalogue,[12] and the Weather Station Directory[13]. Most data are available for free, but the Bureau also provides some services via a cost recovery model.

Like Antarctic data, all meteorological data, including the Bureau's, are shared worldwide for the purposes of contributing to all nations' forecasting and climate services. Data are distributed to interested parties on the Bureau's distribution networks and to the World Meteorological Organisation on its Direct Broadcast Network, DBNet[14]. Observation data and data products are also shared with equivalent international meteorological organisations, such as the National Weather Service in the US, and are also used for climate change modelling by organisations such as Europe's Copernicus Climate Change Service.[15]

Bureau data and data products are also available via the National Computational Infrastructure (NCI) for users within the NCI ecosystem.

Challenges and Opportunities

There is a challenge around serving and cataloguing the sheer volume of data that are generated daily by the Bureau. Through routine observations, forecasts and other analysis, petabytes of data and data products are generated which must be made available. Additionally, there is a requirement to ensure historical data is made available.

The Bureau is about to initiate an IT and data services uplift program, called ROBUST, with the aim to build more external facing services around the online platforms.

Commonwealth Scientific and Industrial Research Organisation

Organisation overview

CSIRO works with industry, government and the research community to find scientific solutions to problems within those sectors. The Marine National Facility (MNF), owned and operated by CSIRO, is

[6] *http://www.bom.gov.au/metadata/catalogue/about.shtml*
[7] *http://www.bom.gov.au/metadata/catalogue/index.shtml*
[8] *http://www.bom.gov.au/*
[9] *http://www.bom.gov.au/app/*
[10] *http://www.bom.gov.au/climate/data/*
[11] *http://www.bom.gov.au/catalogue/anon-ftp.shtml*
[12] *http://wis.bom.gov.au/openwis-user-portal/srv/en/main.home*
[13] *http://www.bom.gov.au/climate/data/stations/*
[14] *http://www.wmo.int/pages/prog/sat/dbnet-implementation_en.php*
[15] *https://climate.copernicus.eu/*

focussed specifically on research in and around the ocean (e.g. biological, oceanographic, hydrology, geological and atmospheric research) with data collected via its ship, *RV Investigator*.

Data acquisition and processing

All data collected on research voyages are archived in the CSIRO National Collections and Marine Infrastructure (NCMI) Information and Data Centre (IDC). Summaries of some data are streamed directly from the ship during voyages and are shared in near real-time with other organisations without going through significant quality control, although some automatic checks are performed to ensure data stability. Once a voyage is completed all data are removed from the ship on disk and uploaded to the CSIRO network on shore. Selected data are then quality controlled and, when ready, archived within the IDC. The unchecked summaries of data made available during the voyage are also archived and made available via FTP and the Near Real-time Underway Data (NRUD) database onshore. Due to the nature of the research, some project datasets may take time to process by the research project before being ready for archiving. In these cases, the IDC may follow those projects up to ensure that the data are published as required.

Quality control is performed by the data centre and domain experts, and includes cross-checks for data consistency across different datasets as well as checking the quality of the datasets themselves.

Metadata and data policy

While on the ship, researchers complete online forms to provide the basic metadata details for their data. As part of the archiving process, the data centre uses this information, as well as information drawn directly from the instruments and the voyage details, to create metadata records in the ISO 19115-3 standard, using the MCP where relevant. Standard MNF vocabularies are used, as well as vocabularies provided by other organisations such as BODC and AODN. Standard parameter names for onboard instruments are being developed.

The MNF has recently published its data policy[16] which outlines the policy for data and sample management for all projects using Marine National Facility (MNF) capabilities.

Data publication and availability

MNF data are available on a number of different portals within and outside of CSIRO. MARLIN3 is the MNF's primary metadata portal[17], providing access to most MNF data. Large datasets can be requested via the service desk. CSIRO also hosts other portals for MNF data, available at their data page[18] and MNF data are also available via external portals, including Geoscience Australia's AusSeabed, the AODN, and the Atlas of Living Australia. MNF data are also used by the Bureau of Meteorology in its operations.

The MNF also publishes a number of Web Services which enable the discovery of different data, including real-time tracking of the *RV Investigator*[19]. It provides links to the Oceans and Atmosphere GeoServer site[20] and OBIS Australia[21], the Ocean Biodiversity Information System which includes MNF data.

There are tracking facilities and logging available to record which pages people visit, and which data are accessed. While there is no funding imperative to report on usage, these statistics are used to generate reports which contribute to understanding the impact of the MNF. Research is used by,

[16] https://mnf.csiro.au/en/About/Policies/For-Chief-Scientists-and-Voyage-Applicants/Data-Management-Policy
[17] https://marlin.csiro.au/geonetwork/srv/eng/catalog.search#/search?resultType=csiro&sortBy=changeDate&from=1&to=20
[18] https://mnf.csiro.au/en/MNF-Data
[19] https://www.marine.csiro.au/data/services/
[20] http://www.cmar.csiro.au/geoserver/web/
[21] http://www.obis.org.au/

among others, universities, *Investigator* researchers looking for associated data and the general public (through the CSIRO portals and/or the external portals mentioned above).

Challenges and Opportunities

MNF has recognised the need to improve the workflow of metadata information from data capture to published metadata. In particular it is looking to improve the capture of core metadata information onboard *RV Investigator*.

There is a recognised requirement to improve the processing of underway data, and to ensure that data (as they are captured and after it is processed) better conforms to recognised standards. This will streamline processing and improve the interoperability and discovery of data. A particular challenge in this area is that physical and biological data standards differ greatly, and finding a standard way of delivering those data together can present difficulties.

Geoscience Australia

Organisation overview

Geoscience Australia[22] (GA) gathers and provides a wide range of geological and geographic data to meet the Australian Government's geoscience requirements. GA's remit includes all Australian jurisdictions, including marine and Antarctic regions. The type of data that it provides includes, but is not limited to, marine data, geology, geophysics, geochemistry, satellite imagery, and other imagery.

Data acquisition and processing

GA data are collected through a variety of methods, including scientific projects, operationally via instrumentation, and via purchase for specific purposes. Data are processed by the relevant science area in their own project IT infrastructure. Scientists creating data for GA are responsible for ensuring data are ready for release, which includes: creating and/or checking metadata; arranging appropriate licencing; and obtaining approvals. Peer review and quality checks are performed by GA domain experts, looking for consistency and accuracy in both data and metadata.

While significant amounts of source data and project data are stored on-premises, large datasets, including some imagery, may be stored at facilities such as NCI. Most GA datasets are stored in the cloud via Amazon Web Services (AWS) or other services for distribution and an increasing amount of data processing is occurring in the cloud. GA is in the process of implementing an on-premise archive and Corporate Data Store (CDS) to ensure the management and preservation of its most valuable data assets. Data types vary, reflecting the wide range of research gathered (e.g. images, database tables, netCDF, other file types).

Metadata and data policy

GA datasets have 2 levels of metadata, public and internal:

- Public metadata are applied to all datasets that are published and available via GA's enterprise catalogue (eCat[26]), and uses the ISO 19115-3 standard. GA uses their published profile for ISO 19115[23] to ensure that metadata are recorded consistently within the ISO standard. Additionally, GA is working on creating and using defined vocabularies based on NASA's GCMD[24].
- Domain specific internal metadata do not adhere to any individual standard and is variable depending on the data captured. It is internally discoverable in principle and is used to detail additional scientific information about the data in question. It commonly forms part of the dataset itself (i.e., within a data header), but may also occur as a separate file following a standard (e.g., las or miniSeed).

[22] http://www.ga.gov.au/
[23] http://dx.doi.org/10.11636/Record.2018.026
[24] https://earthdata.nasa.gov/earth-observation-data/find-data/idn/gcmd-keywords

GA are in the process of obtaining a Core Trust Seal[25] certification for the CDS, their primary data repository.

Data publication and availability

GA provides a number of data portals and web services through which public consumers can find and retrieve data. Data may also be accessed directly via HTTP and other file protocols for certain users, as well as on request. Some of GA's public portals (such as the "eCat" search[26]) are general purpose, while others (such as the Australian Flood Risk Information Portal[27]) are tailored for a specific use case. There are over 140 services available, providing a combination of OGC and ESRI map services. A list of many of the public web services that GA provides can be found on their services website[28]. There are, however, several specific examples of where GA hosts, publishes and delivers data on behalf of and in agreement of other organisations.

GA metadata are also available via portals from other institutions, including the RDA catalogue from ARDC, DTA's portal at data.gov.au and IMOS's AODN portal. GA's data are, by default, shared with a CC:BY 4.0 Creative Commons license[29], and as such any entity that shares/uses GA data must correctly attribute it. GA is considered to the be source of GA data, and in general, does not supply data to be replicated and hosted from other sites.

Any user can access published data as GA follows the Open Data principle and so data are available to all members of the public. GA does not systematically track data retrieval however, and so no specific information is available on which data are used or where. However, Antarctic data are mostly used by the marine and scientific community, and as the basis for advice to the Antarctic Treaty System, while broader GA data are also used in industry, government and other applications.

Challenges and opportunities

A significant challenge in terms of GA data is data discoverability. This is a symptom of both the vastly broad array of data that GA administers, and of the manual nature of the data submission and metadata creation processes. In terms of *Digital Antarctica*, there may be a challenge in separating Antarctic and Southern Ocean data from larger datasets.

Institute for Marine and Antarctic Studies

Organisation overview

The Institute for Marine and Antarctic Studies (IMAS) is the University of Tasmania's research facility dedicated to Marine and Antarctic science and education. IMAS gathers research data from expeditions and via its laboratories and on-site facilities.

Data acquisition and processing

IMAS obtains its data from a number of different sources, including voyages, lab work, deployed instruments and ad-hoc surveys for specific purposes or projects.

Data and metadata are primarily submitted to IMAS via its data submission tool; however, data may be submitted (or referenced) independently of the tool via standard sharing channels (email, file share service, offline media etc.). Some projects keep their own data, in which case only the metadata are submitted via the data submission tool. IMAS also hosts data for other collaborating organisations who are unable to host their own data (including the National Environmental Science Program Marine Biodiversity Hub, of which the University of Tasmania (UTAS) is the administering organisation). Data

[25] https://www.coretrustseal.org/
[26] https://ecat.ga.gov.au/geonetwork/srv/eng/catalog.search#/home
[27] https://afrip.ga.gov.au/flood-study-web/#/search
[28] http://services.ga.gov.au/
[29] https://creativecommons.org/licenses/by/4.0/

in these cases undergo the same processing as IMAS data; however, original data ownership and intellectual property rights are preserved.

Datasets that are submitted to IMAS are stored in MS SQL Server databases managed by UTAS. Data files that are not suitable for database storage (e.g. non-tabular data including video, imagery etc) are stored on a Research Data Storage (RDS) allocation in the Tasmanian Partnership for Advanced Computing (TPAC). The data management software stack was provisioned via the National Research Cloud (NECTAR) but has been moved to UTAS servers, providing a higher level of redundancy.

While some data feed directly into the database via instruments, data are generally submitted by scientists (rather than gathered directly from instruments) and so are assumed to be checked for quality by the submitting scientists.

Once data are submitted, the IMAS Data Manager is notified. The Data Manager reviews the data and metadata before publication, checking data for consistency and format (e.g. column naming conventions etc.) and, where necessary, editing the metadata to assign vocabulary keywords (based on submitted free text) and links to the data as required. This process includes creation of OGC web services (for geospatial tabular data) and inserting web service links into the metadata. If there are issues with the data, the Data Manager will address these where appropriate, or will contact the data provider. In instances where data has only been sparsely described, IMAS prioritises publication within reason, but will attempt to improve the standard of the metadata through discussions with the data provider.

Metadata and data policy

The IMAS data submission tool guides users to provide data attributes, using vocabularies (sourced from AODN and international standard sources) where appropriate, which culminates in the creation of an ISO 19115-3 metadata record. IMAS has transitioned its metadata to ISO 19115-3 from the MCP implementation and are in the process of updating the data submission tool in partnership with AODN.

Data publication and availability

IMAS has a data discovery portal[30] which is an IMAS-hosted and branded installation of the AODN Open Geospatial portal (See IMOS Data publication and availability below) and metadata are also available on the IMAS GeoNetwork[31]. IMAS also hosts a large number of OGC web services on GeoServer[32] and has a small collection on THREDDS[33].

IMAS developed and maintains *Seamap Australia*, an Australian seabed habitat classification scheme and spatial database, which brings together benthic habitat mapping data from a number of sources and is used by anyone looking to study marine habitat, as well as divers and recreational fishers. IMAS is also working on the *Tasmanian Marine Atlas* concentrating on spatial decision-making outcomes for Tasmania.

IMAS data are shared with a number of other data aggregation services:

- Australian Ocean Data Network (AODN) – selected datasets are tagged for harvest into the AODN by the IMAS data manager, and are harvested nightly by AODN via a GeoNetwork CSW endpoint.

[30] https://data.imas.utas.edu.au/portal/search
[31] https://metadata.imas.utas.edu.au/geonetwork/srv/eng/main.home
[32] https://geoserver.imas.utas.edu.au/geoserver/web/
[33] https://thredds.imas.utas.edu.au/thredds/catalog.html

- Research Data Australia (RDA) – periodically harvests all IMAS records via a GeoNetwork OAI-PMH34 endpoint, transforming the data to their own schema (RIF-CS) before adding them to their datasets.
- Ocean Biodiversity Information System (OBIS) – interacts directly with IMAS GeoServer Web Services through the OBIS Australia IPT.

IMAS data are mainly used by scientists, but are also used by anyone interested in Tasmanian and Southern Ocean data or in data visualisation (e.g. artists, education etc.).

Challenges and Opportunities

Challenges are posed by the wide variety of types, formats, and scale of data. This makes it difficult to automate data ingestion and necessitates a significant amount of manual intervention by the Data Manager, which can be inefficient. A flexible approach to data management and publication has been necessary.

Seamap Australia was initially conceived to predominantly serve benthic habitat data, and to support the Seamap Australia habitat classification scheme. The portal received considerable uptake by the marine management sector and has since developed further into this space. A significant and ongoing challenge has been to identify and actively with engage Seamap Australia stakeholders to ensure that it meets its original goals while also addressing the needs of the growing user base to support data-driven management decisions.

Integrated Marine Observing System

Organisation overview

Australia's Integrated Marine Observing System (IMOS) is a national system of observing equipment covering physical, chemical and biological observations of marine environments. IMOS uses various data capture techniques (including static observing platforms, satellites and autonomous underwater vehicles) to collect data across all of Australia's ocean environments, including the Southern Ocean.

All IMOS data are made discoverable by its data facility, the Australian Ocean Data Network (AODN). The AODN portal also serves data from a number of organisations and research facilities across Australia, including a number of AAPP partners.

Data acquisition and processing

The AODN hosts all publishable IMOS data, as well as data from some AODN partners, individuals and other groups (e.g. organisations with limited data management capabilities). It also federates data hosted by other AODN partners via web services to be made available via the AODN portal. AODN partners include Government agencies, educational institutions and other organisations such as the CSIRO.

Data that are hosted by the AODN will pass through its data processing pipeline. Once in the pipeline, data is pre-processed, validated, processed and published into AWS, with metadata being stored on GeoNetwork. Validation includes manual and automated checks for compliance, as well as checks against metadata standards (e.g. Climate and Forecast convention and IMOS netCDF convention). The AODN does not quality control the science of the data, and assumes the data is publishable before submission.

Data and associated metadata collected by the following two IMOS facilities, National Reef Monitoring Network and Animal Tracking, do not go through the standard pipeline, instead being ingested via custom-purpose APIs into their own separate databases.

[34] *Open Archives Initiative Protocol for Metadata Harvesting* https://www.openarchives.org/pmh/

Metadata and data policy

For data to be discoverable via its portal, the AODN requires an ISO 19115-1 compliant metadata record. The AODN also uses published controlled vocabularies[35] within the ISO standard to assist in data discovery. The AODN will work with IMOS submitters to generate metadata records. AODN partners are expected to provide ISO 19115-1 compliant metadata with their data and/or publication web services.

The AODN publishes its data policy[36] and guidelines for contributing data[37] (either to be hosted by the AODN or to have metadata made available via the AODN portal) to assist anyone wishing to have their data discoverable through the AODN portal.

Data publication and availability

The AODN portal is an online tool for searching all data submitted to the AODN (including all published IMOS data). It guides its users through a progressive system of data refinement (including facet searching, keywords and location) to provide them with the metadata and data (or links to the data) they require.

The AODN hosts a large number of netCDF files from a number of different organisations on its THREDDS catalogue[38]. These files are available through the AODN portal, but can be accessed through an interface in the browser or can be integrated directly into GIS software.

The AODN publishes a list of web services[39] to enable integration of AODN data and AODN services. The portal software (AODN Open Geospatial Portal) is open source, and can be installed locally with source code available via github[40]. They also publish guidance on how to integrate AODN netCDF files directly into development environments such at MATLAB, Python and R, and include code samples in its code library[41].

The primary purpose of the AODN portal is to search, discover and download research data. Therefore, its audience is anyone interested in that data: primarily researchers, but also government agencies as well as commercial and private concerns.

The AODN can track usage and report on elements such as number of searches or downloads of a particular dataset and how data are accessed (e.g. via the Portal, GeoServer and THREDDS), however these statistics are not used to drive development of the portal.

Challenges and Opportunities

The AODN portal provides access to a broad range of data, and this breadth of data is the source of many of its challenges.

- With the large amount of data involved, there is an imperative to avoid manual intervention of data on its workflow from submission to publication. Currently this manual work comes in the form of assistance with metadata, dealing with exceptions and setting up bespoke services and pipelines. This is exacerbated by the dual services (hosting and federation) that the AODN provides.
- As one of the portal's goals is to bring together data in a single interaction, the wide variety of data types and feature types (e.g. point, profile, trajectory, timeseries and gridded) creates challenges in best presenting that data.

[35] https://vocabs.ardc.edu.au/#!/?q=AODN
[36] https://imos.org.au/fileadmin/user_upload/shared/IMOS_General/documents/internal/IMOS_Policy_documents/4.3_AODN_data_policy_May16_Final.pdf
[37] https://help.aodn.org.au/contributing-data/
[38] http://thredds.aodn.org.au/thredds/catalog.html
[39] https://help.aodn.org.au/web-services/
[40] https://github.com/aodn/aodn-portal
[41] https://help.aodn.org.au/aodn-data-tools/user-code-library/

- Because of this the portal must remain somewhat generic. While users and contributors would like it to meet their specific needs, this can be difficult to achieve due to of the large amount of work those features would require to be applied across data features. Additionally, greater features and more streamlined ingestion are only possible with stricter metadata and data standards, which then provide a barrier to contributing partners.

The AODN partnerships provide linkages to other data related services (e.g. AusSeabed), and there is an opportunity to develop these linkages.

Appendix 1 – Summary tables

Data collection

This summary shows how each organisation collects data. For the purpose of the summary, data collection has been divided into three main activities:

- *Science project* –Data is collected or generated as part of a research project with specific project outcomes
- *Operational* – Data is collected via standing instruments and established ongoing processes.
- *Research vessel* – similar to operational collection, data is collected from a research vessel regardless of the projects undertaken on the vessel

	Science project	Operational	Research vessel
AAD	Y		Y
BoM		Y	
CSIRO	Y		Y
GA	Y	Y	
IMAS	Y	Y	Y
IMOS		Y	

Data acquisition

The processes that each organisation uses to ingest data into its data centre are outlined below. These processes do not assume or preclude processing of the data (e.g. quality assurance) as part of the data submission process.

- *Portal* – Data is submitted manually via an online portal
- *Direct contact* – Data is submitted manually by contacting the data centre in question. Data is provided on portable media, email, cloud sharing or other manual service
- *Project process* – Data is submitted or made available as part of standard project procedure, e.g. to “publish” data, a project may change a status on a dataset or move data to a public space
- *Data push/pull* – Data is gathered automatically from operational instruments and vessels, or from other data sources. External sources may also push data through pipelines.

	Portal	Direct contact	Project process	Data pull/push
AAD	Y	Y		
BoM				Y
CSIRO			Y	Y
GA			Y	Y
IMAS	Y	Y		Y
IMOS				Y

Metadata

This summary shows the metadata standard employed by the data centre, including any relevant profile that may be in place.

- P – Primary standard used for public data
- S – Secondary standard. This standard may be used as a reference in creating the primary standard, or may be used alongside the primary standard for the purposes of cross compatibility or for internal use

	DIF	ISO 19115-1	ISO 19115-3	ISO Profile?	Internal
AAD	P	S		MCP	DIF profile
BoM		S			P
CSIRO			P	MCP	
GA			P	Internal	S (internal data)
IMAS			P		
IMOS		P			

Data publication and availability

This summary table shows the methods that data is made available to the data centre uses. For the purposes of the summary, the following methods are defined:

- *Data portal* – A public portal to search for the majority of data available
- *Specialised sub-portal/s* – Portals that serve a subset of data, or are developed with specific user types in mind.
- *Direct access* – Access to data files directly via cloud share/FTP etc
- *Web service* – Systematic access to data
- *Other* – Other access (specified)
- *External portals* – Data from this portal is available on other portals external to the organisation (either by data duplication or by federated catalogue search/access)

	Data portal/ catalogue	Specialised sub-portal/s	Direct access	Web service	Other	External portals
AAD	Y	Y	Y	Y		Y
BoM	Y	Y	Y		Y[42]	Y
CSIRO	Y	Y	Y	Y		Y
GA	Y	Y	Y	Y	Y[43]	Y
IMAS	Y	Y	Y	Y	Y[44]	Y
IMOS	Y			Y	Y[45]	

[42] *E.g. Android and iOS apps*
[43] *THREDDS*
[44] *THREDDS*
[45] *THREDDS*

Appendix 2 – Glossary

Term	Description
AADC	Australian Antarctic Data Centre
AAPP	Australian Antarctic Program Partnership
AODN	Australian Ocean Data Network
API	Application Programming Interface – an interface for machine to machine communication.
ARDC	Australian Research Data Council
AWS	Amazon Web Service
BODC	British Oceanographic Data Centre
BoM	Bureau of Meteorology
Bureau	Bureau of Meteorology
CC:BY	See Creative Commons Attribution License
Creative Commons Attribution License	Creative Commons licence which enables the sharing and altering of data under the restriction that correct attribution must be provided.
CSIRO	Commonwealth Scientific and Industrial Research Organisation
CSW	Catalogue Services for the Web
DTA	Digital Transformation Agency, who hosts https://data.gov.au/
DIF	Directory Interchange Format, a metadata standard developed by NASA for describing Earth science data.
ESRI	Environmental Systems Research Institute
FTP	File Transfer Protocol – a protocol for transferring files over a network
GA	Geoscience Australia
GCMD	Global Change Master Directory. NASA's international data collection resource. Now available via https://idn.ceos.org/index.html
GeoNetwork	Software for cataloguing geospatial data
GeoServer	A web enabled server that allows users to connect to geospatial data. http://geoserver.org/
GIS	Geographical Information System
Google Analytics	A service provided by Google that tracks and reports traffic over web and web-based pages
IDC	Information and Data Centre
IMAS	Institute for Marine and Antarctic Studies
IMOS	Integrated Marine Observing System
ISO	International Organisation for Standardisation
ISO 19115	Standard for recording geographic information. ISO 19115-1 is used to record fundamental geographical information ISO 19115-2 is used to record extensions for acquisition and processing of geographic information

	ISO 19115-3 defines an integrated XML implementation of ISO 19115-1, ISO 19115-2 and some elements of IS 19139 ISO 19115 standards may be implemented in their raw state, or may have a profile (such as the MCP) applied to them.
Marine Community Profile	A profile applied to the ISO 19115 standards to suit the needs of the marine community
Matomo	A service that tracks and reports traffic over web and web-based pages
MCP	See Marine Community Profile
MNF	Marine National Facility
NCI	National Computer Infrastructure – a high performance computing and storage facility
NCMI	National Collections and Marine Infrastructure
netCDF	Network Common Data Form – a form of array-oriented scientific data.
NRUD	Near Real-time Underway Data
OBIS	Ocean Biodiversity Information System
OGC Web services	Web services built to the standards defined by the Open Geospatial Consortium including: OGC: WCS: Web Coverage Service for geospatial information (space/time coverage) OGC: WFS: Web Feature Service for geographical features OGC: WMS: Web Map Service for map information OGC: WMTS: Web Map Tile Service for map tile information
RESTful API	An API built using the Representational State Transfer web service architecture.
RDA	Research Data Australia
RDS	Research Data Storage
S3	Amazon's "Simple Storage Service" which provides storage and other services via an online web service interface.
SQL	Structured Query Language
THREDDS	Thematic Real-time Environmental Distributed Data Service – a service that provides human and machine access to data files, including netCDF files.
TPAC	Tasmanian Partnership for Advanced Computing
UTAS	University of Tasmania
Web Service	A web enabled API that facilitates machine to machine communication across the internet.
WIS	WMO Information System
WMO	World Meteorological Organisation