Marine Biodiversity Hub

National Environmental Science Programme

Data discoverability and accessibility

Report from workshops on marine imagery and biological specimen data

September 2018

Rachel Przeslawski, Inke Falkner, Scott Foster, Seb Mancini, Scott Bainbridge, Narissa Bax, Andrew Carroll, Emma Flukes, Manuel Gonzalez-Riviero, Tim Langlois, Kirrily Moore, Mark Rehbein, Katherine Tattersall, Dave Watts, Alan Williams, Mathew Wyatt

Project D2: Standard Operating Procedures (SOP) for survey design, condition assessment and trend detection

9 May 2019 *Milestone 27 – Research Plan v4 (2018)*

Australian Government
Geoscience Australia

CSIRO

UNIVERSITY of TASMANIA

THE UNIVERSITY OF WESTERN AUSTRALIA

IMOS
Integrated Marine Observing System

Enquiries should be addressed to:
Rachel Przeslawski
rachel.przeslawski@ga.gov.au

Preferred Citation

Przeslawski R, Falkner I, Foster S, Mancini S, Bainbridge S, Bax N, Carroll A, Flukes E, Gonzalez-Riviero M, Langlois T, Moore K, Rehbein M, Tattersall K, Watts D, Williams A, Wyatt M (2019). Data Discoverability and Accessibility: Report from Workshops on Marine Imagery and Biological Specimen Data. Report to the National Environmental Science Program, Marine Biodiversity Hub. *Geoscience Australia.*

Project Leader's Distribution List

Parks Australia	
IMOS	Tim Moltmann, Indi Hodgson-Johnston, Ana Lara-Lopez
All co-authors and workshop invitees	

Copyright

Acknowledgement

This work was undertaken for the Marine Biodiversity Hub, a collaborative partnership supported through funding from the Australian Government's National Environmental Science Program (NESP). NESP Marine Biodiversity Hub partners include the University of Tasmania; CSIRO, Geoscience Australia, Australian Institute of Marine Science, Museum Victoria, Charles Darwin University, the University of Western Australia, Integrated Marine Observing System, NSW Office of Environment and Heritage, NSW Department of Primary Industries. RP publishes with the permission of the CEO of Geoscience Australia.

We are extremely grateful to the participants of the Data Discoverability and Accessibility Workshops listed in Appendices A and B. Their insightful comments and discussions are the foundation of this report.

Important Disclaimer

The NESP Marine Biodiversity Hub advises that the information contained in this publication comprises general statements based on scientific research. The reader is advised and needs to be aware that such information may be incomplete or unable to be used in any specific situation. No reliance or actions must therefore be made on that information without seeking prior expert professional, scientific and technical advice. To the extent permitted by law, the NESP Marine Biodiversity Hub (including its host organisation, employees, partners and consultants) excludes all liability to any person for any consequences, including but not limited to all losses, damages, costs, expenses and any other compensation, arising directly or indirectly from using this publication (in part or in whole) and any information or material contained in it.

Contents

List of Figures

List of Tables

EXECUTIVE SUMMARY

As the rate of marine data acquisition increases, so too does the need for that data to abide by the FAIR (findable, accessible, interoperable, reusable) principles. From the nation's perspective, a coherent and assessable data source(s) enables smarter use and management of our marine estate. From a researcher's perspective, open data can be advantageous by increasing citations, media attention, collaborations, jobs and funding opportunities. It is therefore vital that researchers and research organisations strive to release all marine metadata and data so that it is discoverable and accessible by the public.

With the development of national standards (*Field Manuals for Marine Sampling to Monitor Australian Waters),* it became clear that we were unable to advocate a national standard for data release for many data types (bathymetry, marine imagery, biological specimen data) because we either do not yet have suitable digital infrastructure or clear links between existing infrastructure. To meet these challenges, workshops were held in the months following the release of the field manuals, focusing on issues with data discoverability and accessibility for two major data types:

- Marine imagery was the focus of a Data Discoverability and Accessibility Workshop hosted by the NESP Marine Hub and the Australian Ocean Data Network (AODN) on 6-7 September 2018 at Geoscience Australia in Canberra.
- Biological specimen data was the focus of a Data Discoverability and Accessibility Workshop hosted by the NESP Marine Hub and the AODN on -77 September 2018 at CSIRO in Hobart.

This report describes the findings of these two workshops.

In the context of these workshops and this report, *data discoverability* refers to whether a particular dataset or associated meta data is findable as defined by the FAIR principles, such as its inclusion in a known spatial portal that allows a user to search a region of interest. *Data accessibility* refers to the actual dataset itself (not just the meta data in isolation) being available to the public. This may include direct inclusion or links to spatial portals or a standalone collection. It does not include datasets available only upon request to an individual or agency.

For marine imagery, the major challenges for making data discoverable and accessible are related to digital platforms for data storage, annotation, and visualisation. Specific barriers include: i) poorly defined characteristics and linkages between existing online platforms for marine imagery which results in confusion over which platforms should be used, and ii) lack of optimised workflows for these platforms. Although some organisations within the Australian marine community are attempting to address these issues, the geographic focus of these organisations (i.e. tropical and temperate) mean that several groups are undertaking similar but independent initiatives. An opportunity therefore exists for these groups to collaborate and to develop a clear national standard and workflow for marine imagery with the end goal of an open national library of imagery and annotations that could be applied to a range of research questions and management needs. For biological specimens, the main

challenges associated with making data discoverable and accessible are: i) the limited expertise and associated funding support available to identify specimens and ii) the lack of immutable identifiers which percolates to other issues (e.g. inaccurate identifications, duplicate records, incorrect or incomplete meta data). Recommendations against each of these general challenges, as well as more specific barriers identified by workshop participants are listed in Sections 2.6 and 3.6.

Both workshops successfully brought together key players working in Australia with marine imagery or biological specimen data to identify the main challenges to making their data discoverable and accessible. More importantly, the workshop participants provided a way forward by developing clear lists of recommendations. Ultimately, we hope that this workshop report represents a foundation from which future programs can be developed, funded, and implemented to ensure the development of clear and consistent national workflows that are underpinned by stable, enduring and user-friendly digital infrastructure. Although there is an initial investment in time and resources required to appropriately develop national workflows for data sharing, marine researchers will ultimately spend less time on data management due to clear and efficient pipelines. To maximise national benefit, ongoing consultation and collaboration with key national agencies (e.g. AODN, NMSC) will be vital for future developments in this space.

1. INTRODUCTION

1.1 Importance of Data Discoverability and Accessibility

As the rate of marine data acquisition increases, so too does the need for that data to abide by the FAIR (findable, accessible, interoperable, reusable) principles (Wilkinson et al. 2016). Even historical data, which tends to be notoriously difficult to find and reuse, can and should be made publicly available (Easterday et al. 2018).

Scientists are showing an increased willingness to share data, although there is also an increasing perception of risk associated with this behaviour (Tenopir et al. 2015). The European Commission has published revised guidelines on FAIR data management as part of the Horizon 2020 work program to assist organisations in making their research data FAIR and managed properly (Commission 2016), and several institutions have developed workflows to accommodate this (e.g. Alfred Wegener Institute (Koppe et al. 2018)). The OECD also acknowledges the importance of data sharing to society and has been working on policy frameworks, cost and benefit analyses and case studies, that illustrate the benefits of opening government data since 2014 (OECD 2018). In 2017, Australia was rated in the top third of countries making an effort regarding open data, accessibility and government support for data reuse (Figure 1).

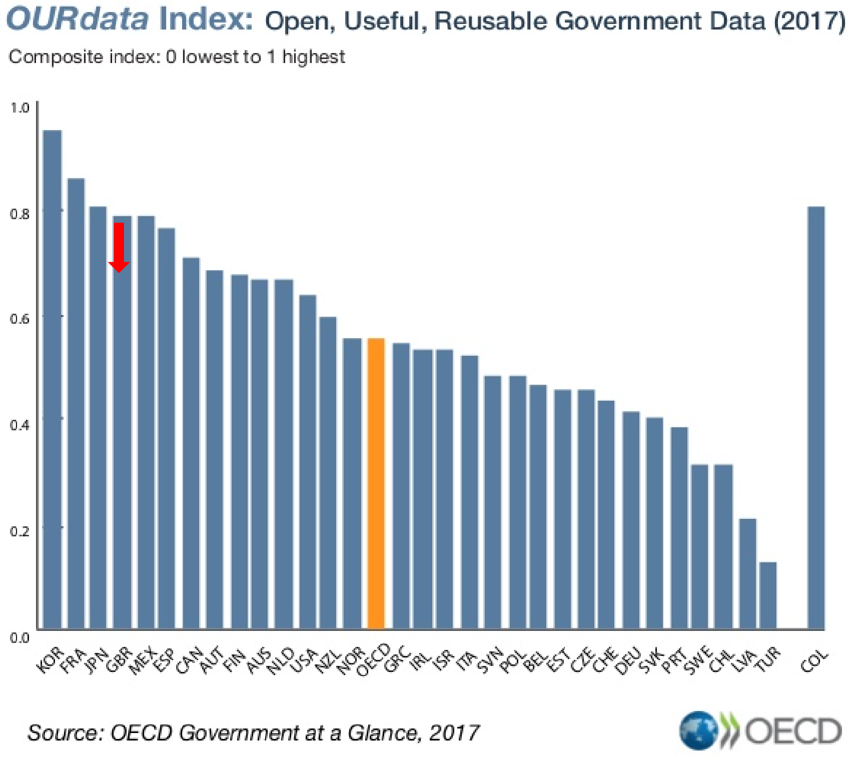


Figure 1 The OECD OURdata Index assesses governments' efforts to implement open data in the three critical areas - Openness, Usefulness and Re-usability of government data. Australia highlighted with the red arrow performs better than the OECD member country average.

Although an increasing number of organisations have embraced the concept of open data, many are still reluctant to do this due to security or privacy issues (Bearzi and Gimenez 2018; Pearce-Higgins et al 2018). Nevertheless, from a researcher's perspective, open data can be advantageous by increasing citations, media attention, collaborations, jobs and funding opportunities (McKiernan et al. 2016).

As such, all marine metadata and data should be publicly released so that it is discoverable and accessible, unless circumstances require otherwise (e.g. confidentiality clause or embargo for commercial work). Even in situations when data cannot be shared, the metadata should be made available so that future research effort can be better guided by existing sampling locations. Refer to (Stocks et al. 2016) for further details on appropriate information management including useful advice on data quality control and data sharing.

1.2 Field Manuals

A package of field manuals was released by the NESP Marine Biodiversity Hub in early 2018 with the aim of promoting national standard operating procedures (SOPs) for marine monitoring (www.nespmarine.edu.au/field-manuals). The field manuals targeted six key marine benthic sampling platforms that were identified based on frequency of use in marine sampling and monitoring programs: Multibeam sonar (MBES), Autonomous Underwater Vehicles (AUVs), benthic Baited Remote Underwater Video (BRUVs), towed video, grabs and box cores, and sleds and trawls. Each field manual focuses on data acquisition and post-processing including data management, particularly as applied to marine monitoring (Przeslawski and Foster 2018). In developing the 'Data Release' section of each field manual, it became clear that we were unable to advocate a national standard for data release for many data types (bathymetry, marine imagery, biological specimen data) because we either do not yet have suitable digital infrastructure or clear links between existing infrastructure.

1.3 Data Discoverability and Accessibility Initiatives

To meet these challenges workshops were held following the field manuals release, focusing on issues with data discoverability and accessibility for two major data types:

- Marine imagery was the focus of a Data Discoverability and Accessibility Workshop hosted by the NESP Marine Hub and the Australian Ocean Data Network (AODN) 6-7 September 2018 at Geoscience Australia in Canberra.

- Biological specimen data was the focus of a Data Discoverability and Accessibility Workshop hosted by the NESP Marine Hub and the AODN 26-77 September 2018 at CSIRO in Hobart.

The overarching aim of the workshops was to discuss current developments and identify key actions needed to establish national data workflows. Agendas for each workshop are included in Appendix A and Appendix B. For each workshop, a half day was devoted to presentations representing various agency capabilities and interests. The bulk of each workshop was spent in breakout groups to identify current and future workflows and associated barriers to making data accessible and discoverable. The final session in each workshop compiled a list of recommendations against each barrier.

We anticipate that the workshop recommendations listed in this report will generate actions such as collaborative project proposals, funding prioritisation, and governance establishment. Overall, such actions will increase the amount of marine biological data that is accessible and discoverable. Although there is an initial investment in time and resources required to appropriately develop national workflows for data sharing, marine researchers will ultimately spend less time on data management due to clear and efficient pipelines. In combination with national SOPs, the recommendations in this report allow for collatable and comparable data to be pooled over multiple spatial and temporal scales, thereby contributing to various monitoring objectives.

1.4 Report Format & Definitions

This report is divided into two main parts: Section 2 covers the marine imagery workshop and associated recommendations, and Section 3 covers the biological specimen data workshop.

Throughout this report, *data discoverability* refers to whether a particular dataset or associated meta data is findable as defined by the FAIR principles, such as its inclusion in a known spatial portal that allows a user to search a region of interest (e.g. Australian Ocean Data Network, Atlas of Living Australia). *Data accessibility* refers to the actual dataset itself (not just the meta data) being available to the public. This may include direct inclusion or links to spatial portals or a standalone collection. It does not include datasets available only upon request to an individual or agency.

2. MARINE IMAGERY

Data Discoverability and Accessibility Workshop I – Marine Imagery had 23 participants representing ecologists, programmers, engineers, and statisticians. The Agenda and Minutes for this workshop are included in Appendices A and C, respectively.

2.1 Objectives and Scope

The objectives for this workshop were to:

- Discuss current developments and workflows in marine imagery, including the widespread use of multiple imagery types and digital platforms;
- Prepare a list of challenges associated with making marine imagery and annotations publicly available.
- Develop a set of recommendations that address these challenges.

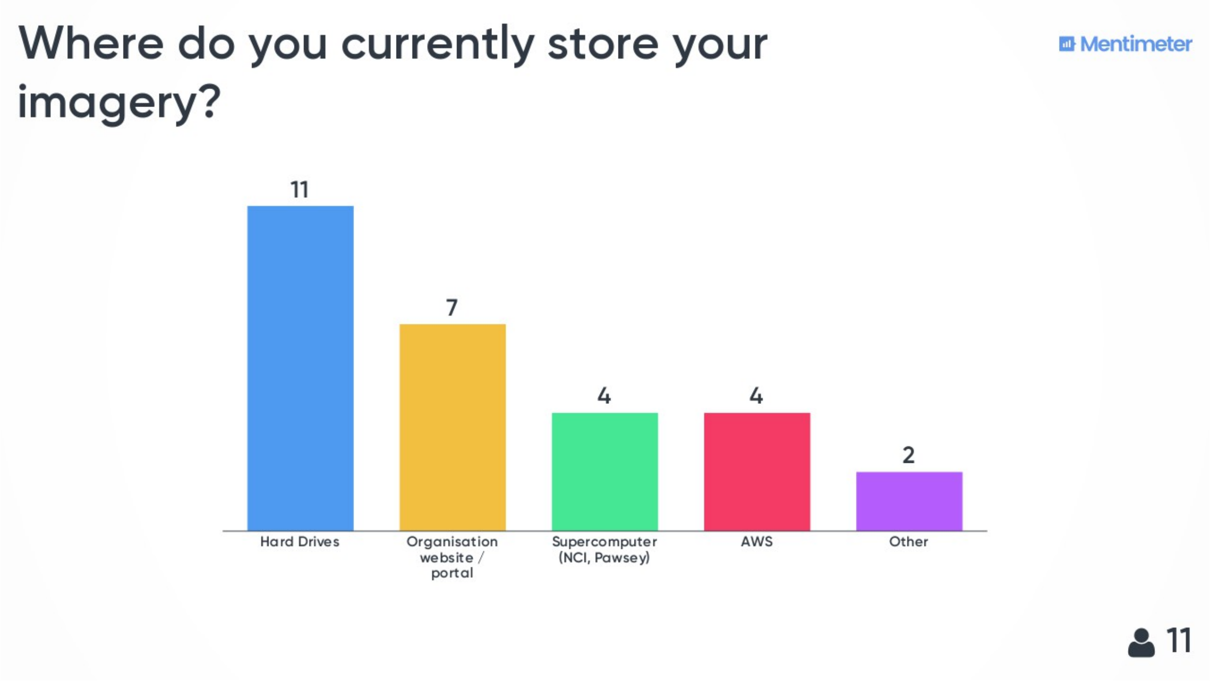
This workshop targeted benthic and demersal imagery, but the issues and recommendations raised are equally applicable to pelagic imagery. Both imagery and associated post-processing annotations were considered within the scope of this workshop. Out of scope were topics such as classification schemes, automated classifications, imagery analyses and interpretation and equipment specifications.

2.2 Institution Presentations

Workshop participants gave an update on current seafloor image collection developments from their agency's or platform's perspective. The following dot points highlight the most important aspects:

- The majority of agencies currently store their imagery on hard drives and organisational repositories, which means image collections are largely undiscoverable and inaccessible (Figure 2, Table 1).
- Annotations were stored in a wider range of locations depending on the type of platform (hard drives, organisational repositories, online platforms including Squidle+, GlobalArchive, BenthoBox), some of which are discoverable and/or accessible (Figure 3, Table 1).
- The IMOS AODN data portal has a new metadata catalogue, which holds 12,000 records from 209 datasets, harvested from many sources including CSIRO and the AIMS. This can be expanded to accommodate external datasets. These data are discoverable and accessible.
- Squidle+ (www.squidle.org) is an image annotation and data management platform developed by Greybits with support from Schmidt Ocean Institute, NeCTAR, and IMOS. It is currently used by many workshop participants to store AUV imagery and image annotations (Table 1). Images are publicly accessible, while annotations are discoverable only.
- GlobalArchive (www.globalarchive.org) is a marine imaging tool developed by Tim Langlois and his team at UWA. It is targeted at mono/stereo BRUV and DOV image annotations, which are publicly discoverable but not accessible, with the exception of project collaborators who can access the data and projects that have been made open access. A new workflow is currently being developed, which will allow for image storage and public annotation access.

- Geoscience Australia stores all imagery on NCI, where they are accessible if the catalogue is directly searched (http://dapds00.nci.org.au/thredds/catalog/fk1/catalog.html). They are not discoverable through spatial portals (e.g. AODN, SeaMap Australia), with the exception of AUV imagery acquired through the Australian Centre for Field Robotics.
- AIMS is currently in the process of developing a single, institution-wide image management repository and storage solution to accommodate its large, multi-platform image collection. BenthoBox, a machine-learning assisted image annotation tool, allows rapid image scoring by researchers. At an agency level, imagery is currently not publicly accessible or discoverable, although imagery from some discrete surveys are.
- CSIRO uses a range of image and annotation storage platforms, none of which are publicly accessible (Table 1). Imagery from discrete surveys may be publicly discoverable.
- Most of the IMAS imagery is stored in-house and is not accessible or discoverable. The exceptions to this are AUV imagery and annotations and BRUV annotations which are stored on Squidle+/AODN and GlobalArchive, respectively. (Table 1).
- Most of the NSW DPI imagery is also stored in-house and is not accessible. However, most AUV imagery and annotations and BRUV annotations are stored on Squidle+ and GlobalArchive respectively, but annotations are not publicly accessible (Table 1).



Where do you currently store your imagery?

Figure 2 Eleven workshop participants responded to this online question conducted during the workshop. According to the survey two thirds of marine imagery is currently stored on hard drives and institutional repositories. Respondents do not necessarily reflect the population of data-users or data-consumers.

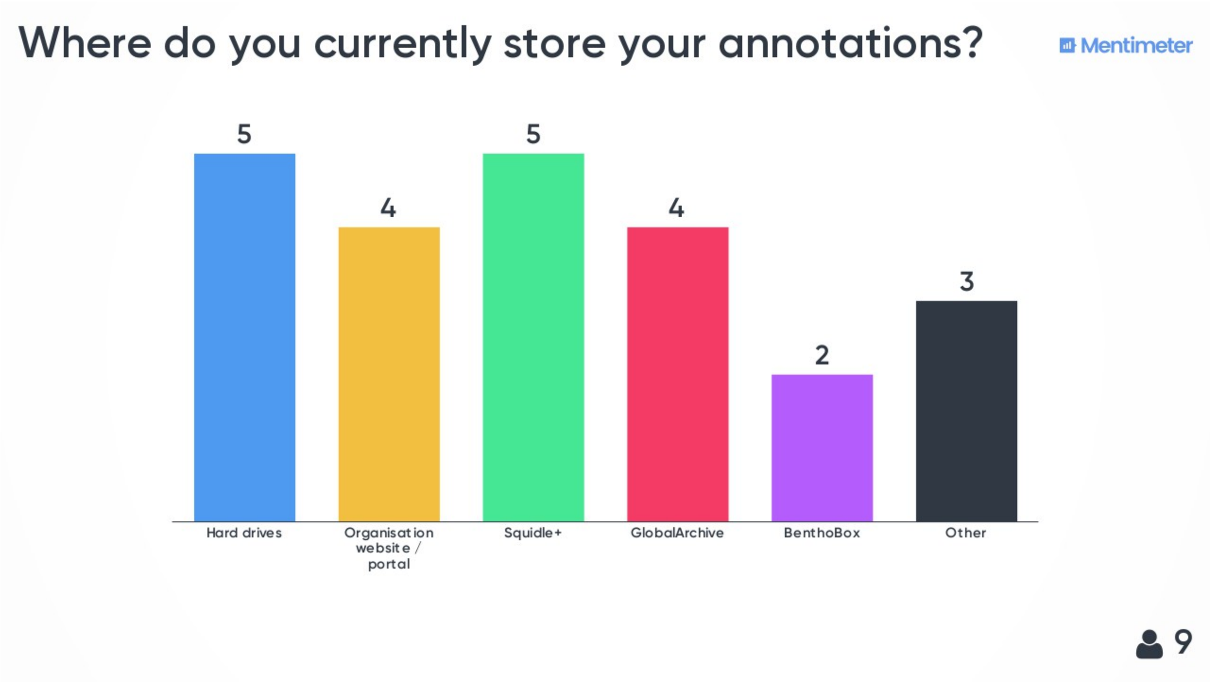


Figure 3 Nine workshop participants responded to this online question asked during the workshop. According to the survey, image annotations are stored in various places depending on the platform with which the image was taken. Respondents do not necessarily reflect the population of data-users or data-consumers.

Table 1 Overview of marine imagery as presented by course participants, and separately from NT DENR (Neil Smit) and Deakin University (Daniel Ierodiaconou).

Institution	**Imagery Platforms**	**Data Size**	**Imagery**		**Annotation**	
			Repository	**Discoverable and accessible?**	**Repository**	**Discoverable and accessible?**
Geoscience Australia	Towed Video, AUV	<10 TB	NCI	Accessible only	None	No
CSIRO (Hobart)	Towed video, cage-mounted and long-line cameras, AOS (acoustic optical system), BOAGS, BRUV	<10 TB	CSIRO Cloud Repository	No	*Annotations* CSIRO Oracle VARS	No
					Metadata MarLIN	Discoverable only
AIMS	BRUV, Drones, Towed platforms, SeaSim cameras, web and time-lapse cameras, multi-files data sets like 360VR, hyper-spec and multi-beam data	~200 TB	National Archives (old video media from coral survey's for one long term project – recent years, data are stored on filesystem. Data no longer sent to NA) Central filesystem (new reef surveys, bits of everything) External HDD (BRUV, Multibeam, hyper-spec) Tape media (Old BRUV, Old Reef Surveys)	Mostly No	Oracle DB (annotation are generated via several applications, one is benthobox) GlobalArchive (limited) a small selection of BRUV video (if any)	No

Institution	Imagery Platforms	Data Size	Imagery		Annotation	
			Repository	Discoverable and accessible?	Repository	Discoverable and accessible?
			WAMSI (limited), copies of the WAMSI related project collections.			
NSW DPI	Towed video	~10 TB	In-house	No	No	No
	BRUV		In-house	No	GlobalArchive	Discoverable only
	AUV		Squidle+	Yes	Squidle+	No
	ROV		In-house	No	No	No
	Diver collected imagery		In-house	No	No	No
IMAS (UTAS)	Towed video, drop camera	~50 TB	In-house	No	No	No
	AUV		AODN and Squidle+	Yes	Partially on AODN	Yes, for scored image subset
	BRUV		In-house	No	GlobalArchive	Discoverable only
	ROV		In-house	No	No, potential to be linked to GlobalArchive for fish imagery and SQ+ for habitat	No
	Diver collected video		Reef Life Survey website	Yes	Reef Life Survey website	Yes
UWA	BRUV	500 TB	In-house	Yes	GlobalArchive	Yes

Institution	Imagery Platforms	Data Size	Imagery		Annotation	
			Repository	Discoverable and accessible?	Repository	Discoverable and accessible?
NT DENR	Towed video, stills	<3 TB	Local NAS drive	Not accessible Partially discoverable through Geoscience Australia data portal	Benthic database held by DENR	Yes, from Feb 2019 through NT spatial atlas and NT spatial library
	BRUV	<2 TB	On tapes	No	Local drive (text annotation to be entered into DENR Spatial database 2019)	Annotation Yes, once in spatial database Metadata, only discoverable
Deakin University	Towed video (video SD & Stereo HD)	<10 TB	Deakin server	No	AODN – Victorian data portal	Yes
	Towed video (down facing stills)	<1 TB	Deakin server	No	No	No
	BRUV	~23 TB	Deakin server	No	GlobalArchive	Discoverable only
	diver collected imagery	<1 TB	Deakin server + Reef Life Survey website	Yes	No	No

2.3 Current Workflows – Marine Imagery

Current imagery workflows vary among organisations and marine imagery acquisition platforms (BRUV, AUV, UVC, towed):

- BRUV imagery has the most developed workflow, resulting in meta data and annotations that are discoverable and accessible on Global Archive.
- AUV imagery has a workflow linked to Squidle+ and AODN, ensuring both imagery and annotations are discoverable and accessible if the user permits this. However, this workflow is currently only applicable to imagery collected by the *AUV-Sirius* and other platforms facilitated by the Australian Centre for Field Robotics (University of Sydney).
- UVC imagery has a workflow associated with IMOS and the AODN, ensuring both images and annotations are discoverable and accessible. However, this workflow is currently only used by the Reef Life Survey (https://reeflifesurvey.imas.utas.edu.au/portal/search).
- Towed imagery is associated with the least mature workflow, despite being the oldest imagery platform. Imagery is often housed on hard drives or internal networks, with workflows varying among organisations. Imagery and annotations are rarely accessible and even less often discoverable (e.g. Geoscience Australia's National Marine Imagery Collection on the NCI).

More information about current workflows can be found in the imagery workshop Minutes in Appendix B.

2.4 Ideal Workflows – Marine Imagery

Ideal imagery workflows should ultimately result in all meta data, imagery and annotations being publicly discoverable and accessible. Steps for a proposed ideal workflow are summarised in Figure 4 and described below:

1) Imagery is acquired using standard methods for a given platform (Przeslawski & Foster 2018).

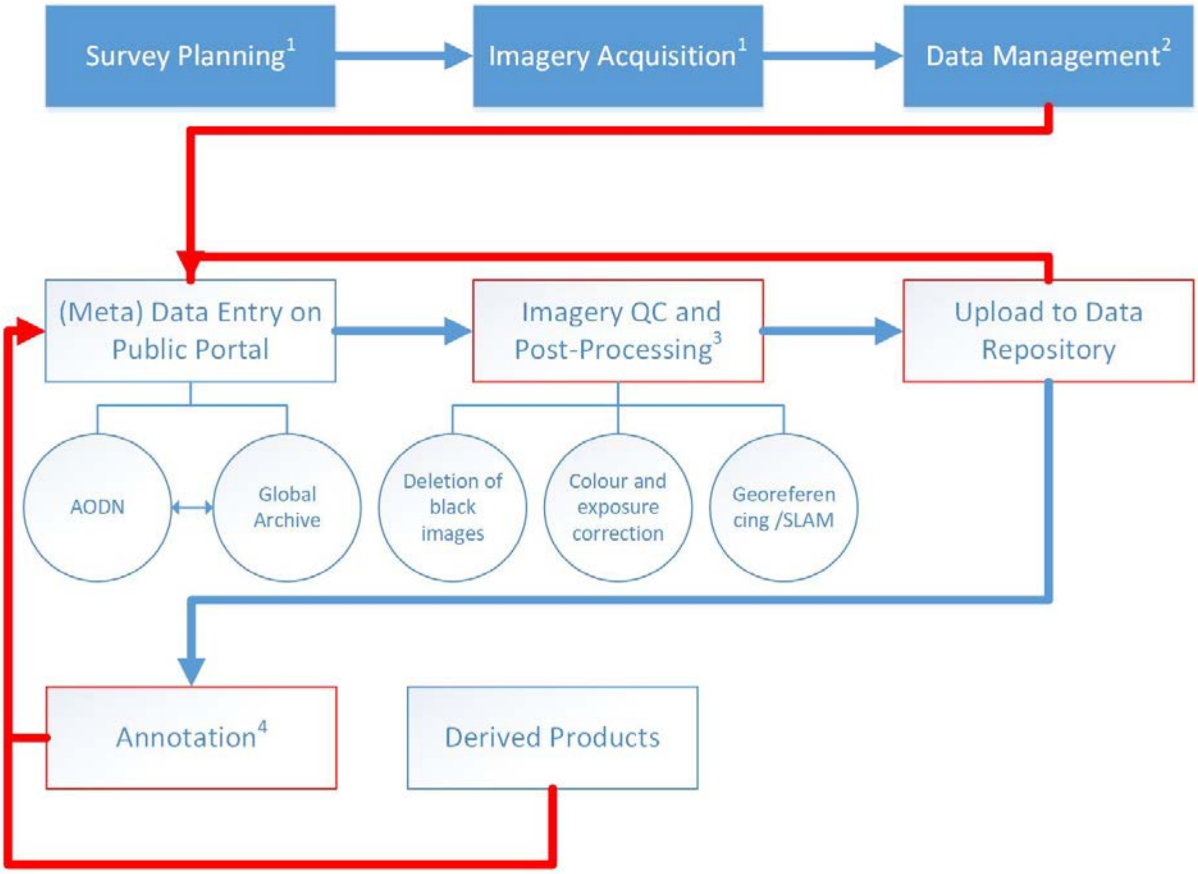
2) Meta data is entered into a standard template and linked to the AODN.

3) Imagery is uploaded into a permanent repository with a stable URL. This ensures accessibility of imagery.

4) The URLs of the imagery are linked to relevant meta data and AODN. This ensures discoverability of imagery.

5) Annotations are uploaded to an appropriate online system (e.g. Global Archive, Squidle+, Benthobox) and linked to AODN to ensure accessibility and discoverability of annotations.

More information about ideal workflows can be found in the imagery workshop Minutes in Appendix B.



[1] Follows national SOPs in Przeslawski and Foster 2018
[2] Includes back-ups
[3] Semi-automated or automated
[4] Includes manual (e.g. Event Measure), semi-automated (e.g. Squidle+, Benthobox), and in-house (e.g. CSIRO) annotation platforms

Figure 4 Idea workflow for marine imagery. Blue boxes represent activities undertaken during the survey, and white boxes represent post-survey activities. Red boxes and lines indicate activities and linkages that require further develop to be incorporated into a national workflow.

2.5 Barriers and Challenges

The two major challenges for making marine imagery data discoverable and accessible are related to digital platforms for data storage, annotation, and visualisation:

i) The characteristics and linkages between the existing online platforms for marine imagery are poorly defined, resulting in confusion over which platforms to use.

ii) Workflows for these platforms are not yet optimised.

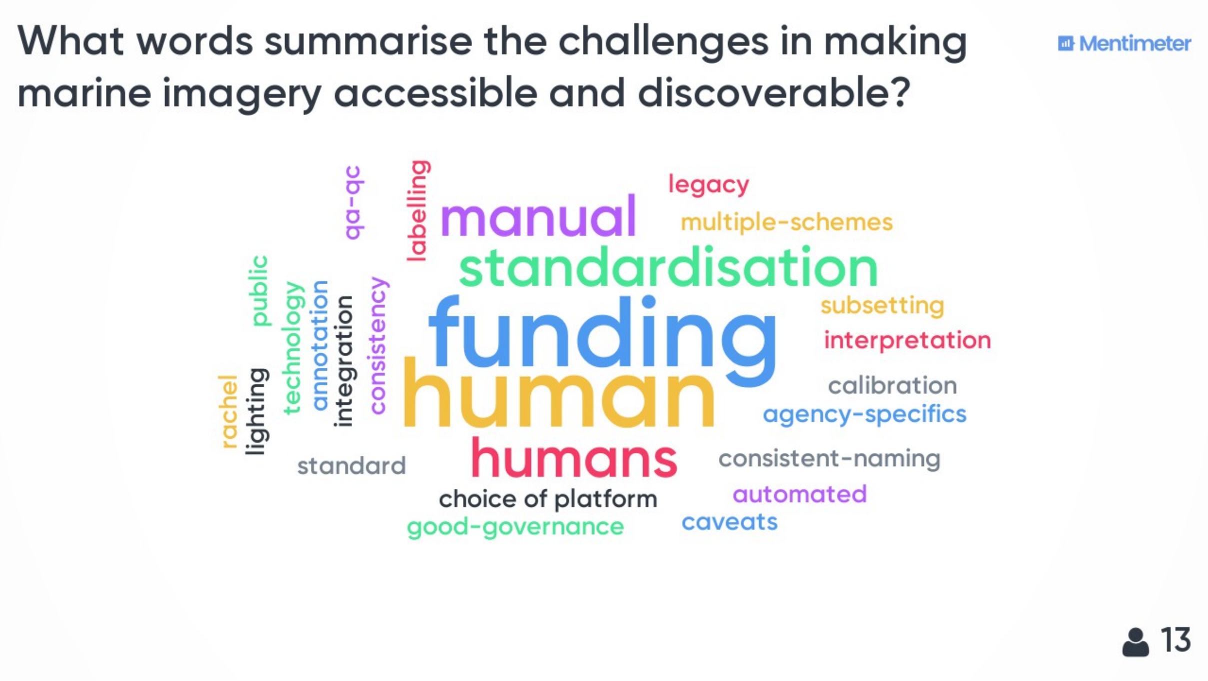
Several promising platforms have recently been developed to facilitate discoverability and accessibility of marine imagery data, but these have yet to be fully integrated into a national workflow. Squidle+ (www.squidle.org) and GlobalArchive (www.globalarchive.org) are currently recommended in the NESP field manuals as the recommended platforms to ensure discoverability and accessibility of marine imagery (Przeslawski & Foster 2018), but the platforms currently have no long-term support and alternative platforms are being used and further developed that have potential to achieve similar discoverability and accessibility goals (e.g. Benthobox at https://benthobox.com).

Workshop participants rated the lack of sustainable funding and automation as the biggest challenges for optimised imagery workflows (Figure 5). Participants were most concerned about the longevity of image analysis platforms and the lack of metadata standards as an impediment to effective marine image discoverability and accessibility (Figure 6). Consistent classification and a link to the AODN data platform were also rated important and participants acknowledged that the development of automated workflows will most likely be time-consuming and costly (Figure 6).

More specific barriers to accessible and discoverable marine imagery were also discussed in breakout groups, including:

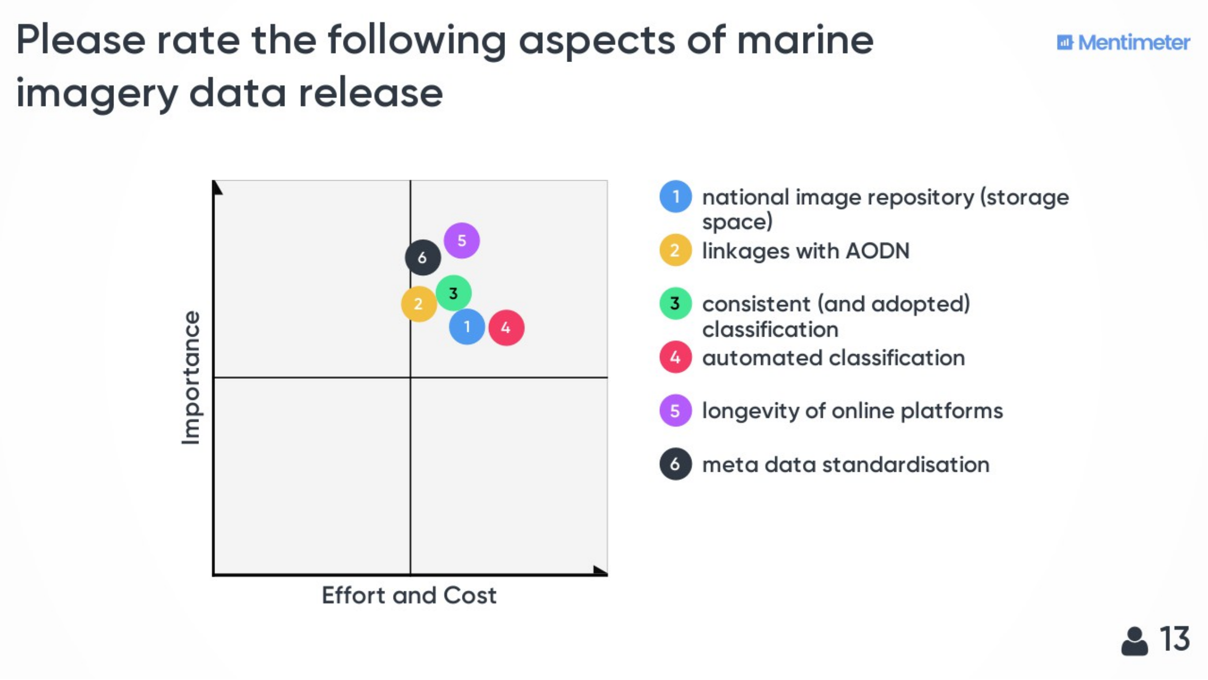
- There is a lack of communication and integration between the main Australian groups working on marine imagery data management and analysis;
- There are poorly defined characteristics and linkages between the existing online platforms related to marine imagery;
- There is a lack of a centralised marine imagery repository and tracking system; meaning some data may not be able to be harvested by data aggregators, whereas other data may have multiple copies;
- Bottlenecks exist during processing, imagery upload, and annotation on platforms
- Challenges exist with mapping between annotation methods (e.g. still point-based vs time window) and schema (e.g. CATAMI vs CBICS);
- Some imagery data platforms are maintained by individuals, with lack of long-term support or institutional backing;
- There is limited automation or clear workflows to accurately capture meta data for many imagery platforms;
- There are no champions for some of the imagery platforms;
- There is a lack of governance and focused working group(s); and
- There are few (or unarticulated) incentives to change the current paradigm.

During the course of the workshop and in the following weeks, it became obvious that there are at least two main groups working to address issues related to data management and analysis of marine imagery. The Australian Institute of Marine Science is streamlining its internal imagery process using an in-house version of BenthoBox, while organisations working in temperate Australia (UTAS, USYD, UWA) are developing their own process and linkages between platforms namely AODN, Squidle+, and Global Archive. These different geographic sectors of influence in the Australian marine community (i.e. tropical and temperate) mean that several groups are undertaking similar independent initiatives. An opportunity therefore exists for these groups to collaborate to develop a clear national standard and workflow for marine imagery.



What words summarise the challenges in making marine imagery accessible and discoverable?

Figure 5 Word cloud generated by 13 workshop participants describing the challenges they experience regarding marine imagery workflows.



Please rate the following aspects of marine imagery data release

1 national image repository (storage space)
2 linkages with AODN
3 consistent (and adopted) classification
4 automated classification
5 longevity of online platforms
6 meta data standardisation

Figure 6 Thirteen workshop participants rated the importance and burden in effort and cost involved in developing various aspects associated with making marine imagery discoverable and accessible.

2.6 Recommendations

The workshop group compiled several recommendations to address the barriers listed above.

Barrier 1: There is a lack of communication and integration between the main Australian groups working on marine imagery data management and analysis.

Recommendations:

- Hold a follow-up workshop on marine imagery data in 2019 with key staff from institutions with major marine imagery collections.
- Document the workflows from each group, as well as their bottlenecks and internal challenges.
- Identify differences in these workflows and assess whether these would affect marine imagery as nationally collatable and comparable data.
- Depending on the point above, incorporate one, both, or amalgamated workflows into the NESP SOPs through the next version of the towed imagery, AUV, BRUV, and ROV field manuals.
- Promote data-sharing best practice (FAIR).
- Consider AusSeabed and similar initiatives as models for partnering between institutions to integrate data.

Barrier 2: There are poorly defined characteristics and linkages between the existing online platforms related to marine imagery.

Recommendations:

- As a marine imagery community, we need to focus on improving consistency in annotation data and metadata rather than the platforms themselves There is therefore a need to:
    i. Define minimum and recommended standards for each imagery data type, including QA/QC (e.g. quantifying observer bias with annotations); and
    ii. Require that these minimum standards be followed. Use platforms that incorporate these standards as examples - these can change over time as new tools are developed to address niche research issues.
- Develop an infographic to articulate current digital platforms for marine data, including purpose, data type, and linkages. The NESP Marine Hub has this as a milestone for Project D2 and will work with the AODN to progress this in 2019.

Barrier 3: There is a lack of a centralised marine imagery repository and tracking system, meaning some data may not be able to be harvested by data aggregators, whereas other data may have multiple copies.

Recommendations:

- Provide a framework within which meta data, including version history, can be formally compiled, characterised, and visualised.
- Apply this framework to characterise marine imagery holdings for major institutions/platforms.
- Explore the possibility of a permanent marine imagery repository (including backups and security/sharing) with ARDC and other major agencies.

- Apply a data citation system (e.g. DOI) to facilitate tracking of data usage in any such image repository, as this would increase uptake by recognising contributors for their input to the repository.

Barrier 4: Bottlenecks exist during processing, imagery upload, and annotation on platforms

Recommendations:

- See points above regarding storage.
- Scope global solutions for large file size sharing, streaming, viewing, and access as related to Australian marine imagery (e.g. YouTube).
- Since the speed-of-access problem transcends marine imagery and likely applies to other data types (e.g. satellite imagery, bathymetry), NCRIS should be approached to see if they can develop a solution.

Barrier 5: There are challenges with mapping between annotation methods (e.g. still point-based vs time window) and schema (e.g. CATAMI vs CBICS)

Recommendations:

- Conduct a census of current annotation methods and schemes in relation to their purposes, including an online survey to gauge level of data quality, QA/QC methods, extent (spatial/temporal), biological resolution needed and applied
- Identify international initiatives in this space.
- Scope the adoption of a framework (e.g. software system) that allows the marine imagery community to cross-walk between schemes. It is important to facilitate mapping between CATAMI and other annotation schemes, as it seems unlikely that a single annotation system will be applied by everyone.
- Revisit and update the CATAMI national classification scheme, including morphospecies catalogue, including the development of a shared morphospecies library for national standardisation, all to be managed by a technical working group.
- Propose a national standard for QA/QC of marine imagery, including quantification of observer bias in annotations;
- Encourage scoring of imagery at the finest level possible so it can map up to all schemes (refer to NESP SOPs).

Barrier 6: Some of the imagery data platforms are maintained by individuals, with lack of long-term support or institutional backing

Recommendations:

- Scope a marine imagery collective (e.g. IMOS marine imagery node) and links to high level committees (NMSC) through to researchers and end-users to inform funding priorities. Potential funders are IMOS/AODN, government and universities (GA, CSIRO, UTAS, AIMS), ARC LIEF, SOI, Industry Partners (e.g. APPEA)

Barrier 7: There is limited automation or clear workflows to accurately capture meta data for many imagery platforms

Recommendations:

- Standardise metadata (adopt automated software to reduce human error and to increase efficiency)

- Develop a semi-automated process to reduce scoping time and human errors (ideas from Robotic Process Automation may apply). For example, semi- automated in-fill process to populate metadata.
- Enforce meta data standards (e.g. via permits or through vessel systems).
- In next version of relevant NESP field manuals, do the following:
  - Define minimum requirements for metadata (reduce prescriptiveness)
  - Ensure consistent formats and vocabularies (define)
  - Establish working groups by platform to develop standards and ensure uptake and compliance. Communication between groups is essential to ensure national standards are applied across platforms

Barrier 8: There are no champions for some of the imagery platforms, and we need someone to drive national synthesis

Recommendations:

- Compile a list of common requirements across these platforms to inform the design of tools that will support marine imaging around the country (data upload, storage, annotation, etc.).
- Identify champion(s) to focus on the national data products that should be delivered to inform state of the environment, marine parks monitoring, etc. This should start in a platform-agnostic manner by considering what needs to be delivered. Platforms, and survey patterns can then be selected and/or designed based on these requirements.

  At the time of writing this report, there are champions for the following platforms:

  I) AUV: Neville Barrett (UTAS) has been leading the IMOS AUV scientific working group through which data is delivered to AODN and accessible to the community, but there is still no consensus on the downstream processing and delivery of data collected outside the IMOS AUV facility.

  II) BRUVs: Tim Langlois (UWA) has been championing the need for a national repository through his work with Global Archive, and a national BRUV working group chaired by Euan Harvey was established in 2017 to coordinate national efforts.

  III) Diver-collected imagery has potential champions through Graham Edgar and Rick Stuart-Smith (UTAS). Their program Reef Life Survey has lots of data available (including time series), with national and international coverage. These workflows are not yet standardised with others, particularly the AIMS long-term monitoring program (see Przeslawski et al 2019).

To date, there have been no clear champions to date for Towed Imagery to standardise workflows, data products, and deliverables. Towed imagery systems show high variability in platform design, sensor suites (video, stills, water column parameters, etc.), survey objectives and deployment practice compared to the other platforms. Towed Imagery may therefore require several champions representing different types of systems and approaches (e.g. lightweight drop cameras, deep-sea platforms). Alan Williams (CSIRO) has agreed to be the nominal coordinator for towed imagery in the near future, as recommendations from this report are progressed.

Barrier 9: There is a lack of governance and focused working group

Recommendation:

- Identify existing groups (e.g. IMOS Benthic Monitoring Group, NMSC Baselines WG) to support funding proposals, revisit Terms of Reference, and develop a strategy document for moving forward as a united community (vision, communicate value, risk and mitigation, funding).

Barrier 10: There are few (or unarticulated) incentives to change the current paradigm

Recommendations:

- Describe why a researcher should make his/her data accessible/discoverable and abide by standards (and what happens when you don't).
- Promote this information.
- Develop automated high-level reporting that researchers can use.
- Liaise with funding agencies and regulators so they insist on best practices, including meta data standards and data accessibility.
- Avoid insistence on one-size-fits-all approach for all platforms and agencies; instead focus on bringing platform-specific and agency-specific workflows together so that data is, at the very least, accessible and discoverable and ideally comparable and collatable.
- Invest in user-friendly platforms that make it easy for researchers to submit appropriate meta data and data.

3. BIOLOGICAL SPECIMEN DATA

The Data Discoverability and Accessibility Workshop II – Biological Specimen Data had 22 participants including ecologists, taxonomists, curators, data managers and statisticians. The Agenda and Minutes for this workshop are included in Appendices B and D, respectively.

3.1 Objectives and Scope

The objectives for this workshop were to:

- Discuss current developments and workflows related to data associated with biological specimen identifications;
- Prepare a list of challenges associated with making accurate biological specimen data publicly available; and
- Develop a set of recommendations that address these challenges.

This workshop targets data from benthic and demersal macro-organisms, but many of the issues and recommendations raised are applicable to pelagic organisms, meiofauna, and microbes. Data from taxonomic identifications (e.g. presence only, presence/absence, species abundance/biomass matrix) and genetic sequencing were within the scope of this workshop. Out of scope for this workshop were topics such as sample curation, equipment specifications, ethics approvals and other permitting, and project-specific data (e.g. biochemical, ecotoxicological).

3.2 Institution Presentations

Workshop participants gave an update on the protocols associated with the curation of both biological specimens and biological specimen data from their agency’s perspective. The following dot points highlight the most important workshop outcomes are summarised in Table 2:

- According to the online survey conducted during the workshop the majority of biological specimens are currently stored in museums and/or at the researchers’ or third-party organisation (Figure 7).
- This online survey also showed that biological specimen data are stored in various places including personal hard drives, internal networks, organisational websites and/or several online databases (Figure 8).
- Several well-established online repositories exist for biological specimen data, each with different protocols for data standards, publication and revisions. This poses challenges involving inconsistencies and duplicates or different versions of data being published in several databases.
- Museums are custodians for both biological specimens and biological specimen data and therefore face additional challenges related to specimen curation, storage and coding.
- The AODN data portal currently holds the following biological data: plankton survey data, acoustic tracking and animal tagging data (which will not be discussed in this report), RLS data and fish/squid occurrence data from bottom trawls (NIWA). An integration of

ALA data into the AODN data portal is currently under development as is a workflow for data going to OBIS.

- The Atlas of Living Australia (ALA) (www.ala.org.au) is an NCRIS facility hosted by CSIRO, which hosts over 80 million species records from Australia, including from marine environments. These data are discoverable and accessible, but are presence only (i.e. species occurrence). ALA harvests from the Australian node of OBIS (OBISAU).
- The Ocean Biogeographic Information System (OBIS) (www.obis.org.au) is a network of country-specific nodes and taxa-specific groups (e.g. FishBase). OBIS Australia (OBISAU) currently holds more than 7 million discoverable and accessible records from numerous data providers and sources within Australia. OBISAU regularly checks for new data at ALA and, if useful, harvests it for subsequent publication to OBIS. OBISAU has some datasets containing absence records which are not yet handled by OBIS or ALA.
- The Australian Faunal Directory, managed by the Department of the Environment and Energy, is an online taxonomic catalogue of Australian animals, which provides the most up-to-date information on taxonomic names and classifications. Species lists are discoverable and accessible, but there are no georeferenced data.
- CSIRO sends most biological samples to museums for identification and storage, where many are registered over time. The original data are published on the CSIRO Data Trawler, often with the field identification (e.g. family, operational taxonomic unit). The data from registered samples at the museums are also submitted to ALA and OBIS via the OBISAU, often with the final identification. Only those biological specimens which are registered are discoverable through museums, ALA, and OBIS. Unregistered samples are generally not accessible.
- Geoscience Australia sends most biological samples to museums for identification, although they sometimes identify infauna to operational taxonomic unit in-house. Museum specimen data are followed up according to project-specific needs (i.e. sponge biodiversity study) and published on the GA website, where they are accessible but not discoverable.
- The Australian Antarctic Division has an extensive collection of Antarctic specimens, which are currently stored at the Division. The specimens and resulting data are currently not systematically curated. Biological data entry to the Australian Antarctic Data Centre is done on a project basis, where it is discoverable and accessible. There has been progress integrating a portion of the east Antarctic specimen data into ALA, but completion of this task is dependent on funding resources and available expert staff.
- The Tasmanian Museum & Art Gallery receives biological specimens from many Institutions and individuals. Due to time and funding constraints only a portion of these samples are curated and identified. Identified specimens are recorded in the museum database and uploaded to ALA, where they are discoverable and accessible. There are a large amount of unidentified (dark) samples with museums, research organisations, individual researchers that are not recorded and thus not discoverable or accessible.

Table 2 Overview of biological specimens and specimen data as presented by course participants.

Organisation	Data type	Specimens		Specimen data				
		Repository	Discoverable and accessible?	Data size	Repository	Discoverable and accessible?	Occurrence/Absence	Discoverable and accessible?
AODN	Plankton survey data Acoustic tracking Animal tagging RLS data Fish/squid occurrence data	N/A	No		AODN Data Portal	Yes		
ALA	Australian terrestrial and marine fauna (record only)	N/A	No	76 million records	ALA database	Yes	Occurrence only	Yes
OBIS AU	All taxa but dominated by fish and shallow water/coastal organisms	N/A	No	7 million records	OBISAU IPT and OBISAU database	Yes	Yes	Yes
Australian Faunal Directory	Australian terrestrial and marine fauna (record only)	N/A	No	>120,000 species/ sub-species	AFD database	Yes	No	Yes
CSIRO	Catch data Biological samples Specimen photographs	CSIRO storage and museums	No (CSIRO) Yes (Museums)		Museum Database/ CSIRO Data Trawler	Yes	Yes	Limited

Organisation	Data type	Specimens		Specimen data				
Geoscience Australia	Biological samples (marine fauna)	GA storage and museums	No (GA) Yes (Museums)		GA website	Accessible only	Yes	Accessible only
Australian Antarctic Division	Antarctic specimens	AAD storage and museums	No (AAD) Yes (Museums)		Australian Antarctic Data Centre	Discoverable and limited access	Occurrence data only	Discoverable and limited access
Tasmanian Museum & Art Gallery	Biological samples	Museum storage	Yes, via ALA or on request	>100 000 records (~100 000 dark records)	OZCAM Natural Values Atlas Atlas of Living Australia Museum databases for specific taxonomic groups	Yes, via ALA or on request	Occurrence data only	Yes, via ALA or on request
Western Australian Museum	Biological samples	Museum storage	Yes, via ALA or on request		Museum databases, with upload of data for select groups to OZCAM, ALA	Yes, via ALA or on request	Occurrence data only	Yes, via ALA or on request

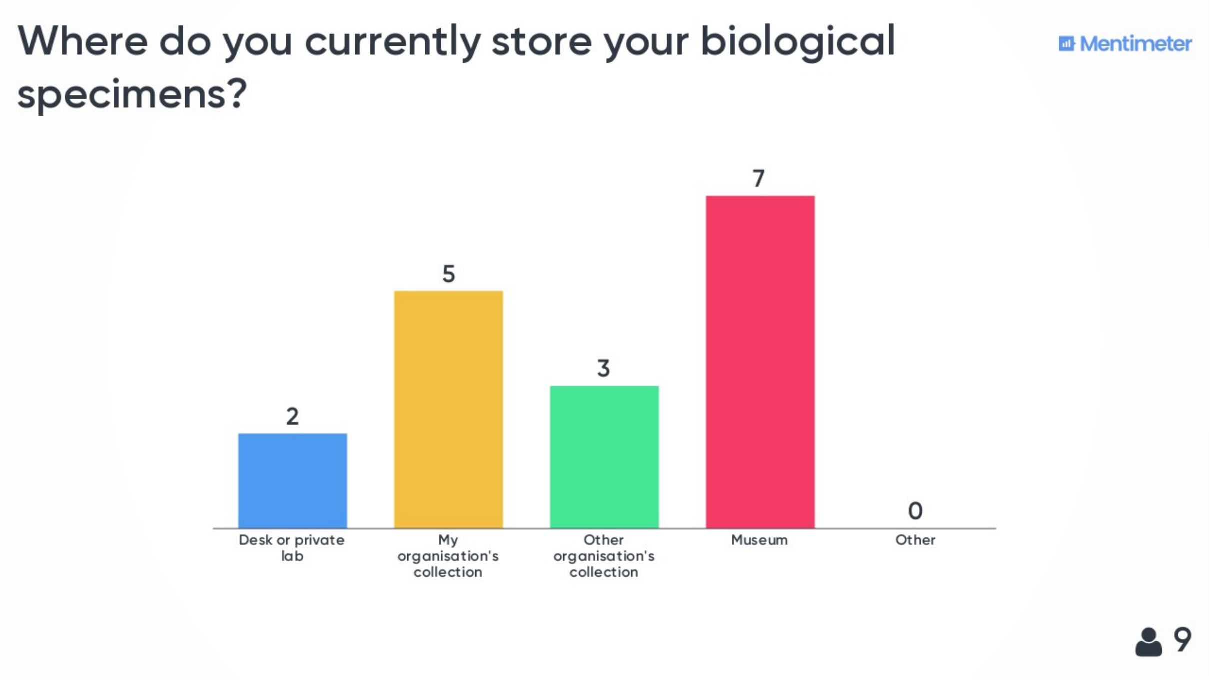


Figure 7 The majority of biological specimens are currently stored in museums and/or at the researchers' or third-party organisation.

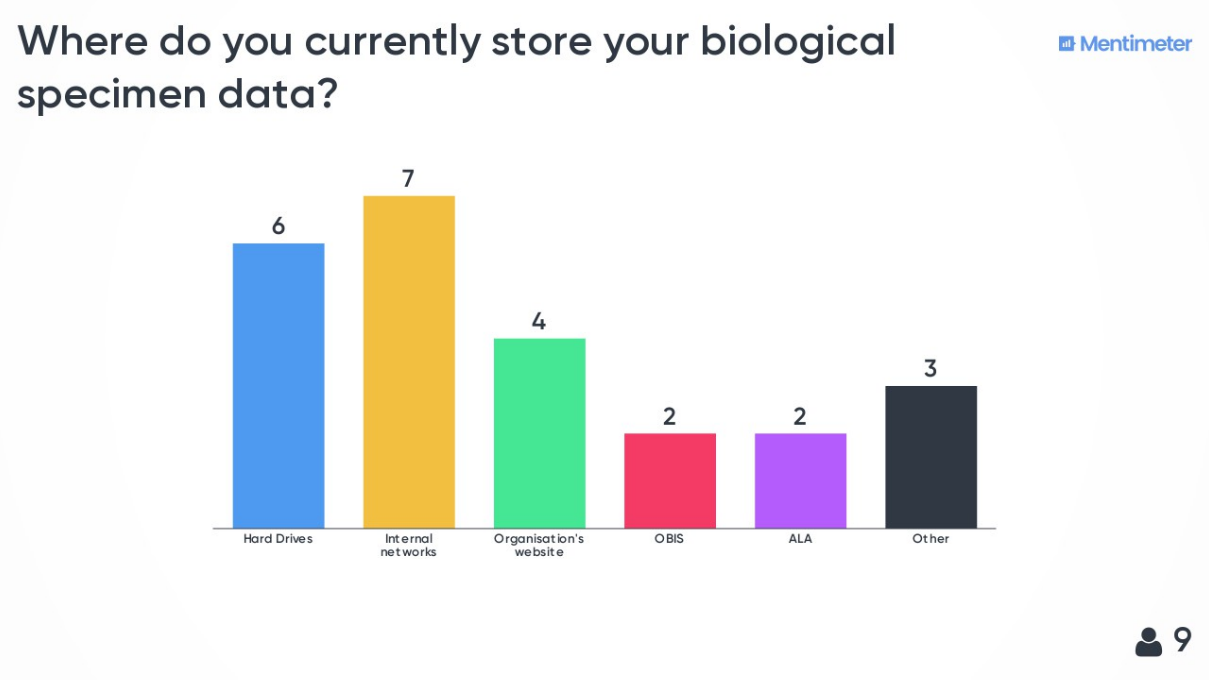


Figure 8 Biological specimen data are stored in various places including personal hard drives, internal networks, organisational websites and/or several online databases

3.3 Current Workflows – Biological Specimen Data

Current workflows for biological specimen data vary dramatically among and even within institutions. Workflows are often *ad hoc* and depend on resources allocated to taxonomic identifications, survey objectives, and available taxonomic expertise.

Broadly speaking, current workflows involve the following steps:

1. *Survey planning and statistical considerations*. In some cases, a broad team of taxonomists is engaged to ensure all specimens are appropriately collected, sorted, preserved, and curated. Unfortunately, it is still far too common that specimens are collected with little plan for non-target organisms. A plan for open data related to specimens is rarely developed at this stage.

2. *Specimen collection*. This often involves multiple researchers and institutions, each of which may be interested in a discrete taxonomic group. Most data are transcribed onto paper and then digitally transcribed in personal spreadsheets during spare time onboard, although there have been efforts onboard the *R.V. Investigator* to establish web-based and ship-based data management systems and workflows (K. Moore, D. Watts, personal communication).

3. *Meta data dissemination*. Immediately after a marine survey is completed, there is often limited effort to ensure that appropriate meta data (i.e. why, where, when and how samples were collected) is made publicly available. The AODN meta data portal can facilitate this, but it is rarely used for benthic and demersal biological sampling.

4. *Onshore Specimen Processing and Curation*. Specimens from target taxa are lodged at museums decided during the survey planning phase (ideal) or post-survey (not-so-ideal). Non-target taxa are often lodged at the same museum(s) (where they are unlikely to be identified) or the institution leading the survey (where they may indefinitely remain in storage).

5. *Data acquisition*. Specimens are identified and sometimes associated with abundance or biomass to develop a species matrix. This step can span several years because specimens are identified by various taxonomists, often with limited time and funding. There is an inconsistent and poorly linked backflow of these updated identifications back to the original collector.

6. *Data harvest and dissemination*. Currently, museum databases are harvested automatically and regularly by the ALA which then incorporate the species occurrence into their platforms where it becomes discoverable and accessible. There are no automated processes for other institutions to integrate their data, although this can be directly submitted to individuals at ALA and OBISAU. There are circumstances where the same data is harvested more than once, and these can appear as undetectable duplicates, chiefly due to lack of unique and immutable identifiers.

7. *Data other than species occurrence*. Absences, abundances, biomass, and genetic analyses are not usually included in museum data (or do not accurately reflect sampling, as sometimes only a voucher specimen is sent to the museum), and they are not generally considered by ALA or OBIS. Some institutions make this information available via supplementary material in publications (Przeslawski et al. 2014), in primary publications (e.g. *Scientific Data*), or institutional repositories, but the associated data are not discoverable in the major portals (e.g. AODN, ALA, OBIS).

3.4 Ideal Workflows – Biological Specimen Data

Ideal workflows for biological specimens and associated data ultimately result in meta data and data that are discoverable and accessible. Steps for a proposed ideal workflow are summarised in Figure 9 and described below:

1) The survey is planned and biological specimens are acquired using standard methods for a given platform (Przeslawski & Foster 2018). A unique and immutable identifier is assigned for each grouping of specimens made at the finest-resolution (e.g. phylum, class, OTU). Each specimen or group of specimens is photographed with this unique identifier. Ideally, this identifier would be a globally unique number and registered at an international organisation such as has been developed for geological samples with the International Geo Sample Number.

2) Meta data including a unique and immutable identifier are entered into a standard template and linked to the AODN. This ensures the sample locations (via a bounding box) are publicly available as soon as possible after the completion of a survey, even if the associated taxonomic identifications and species occurrence data take much longer.

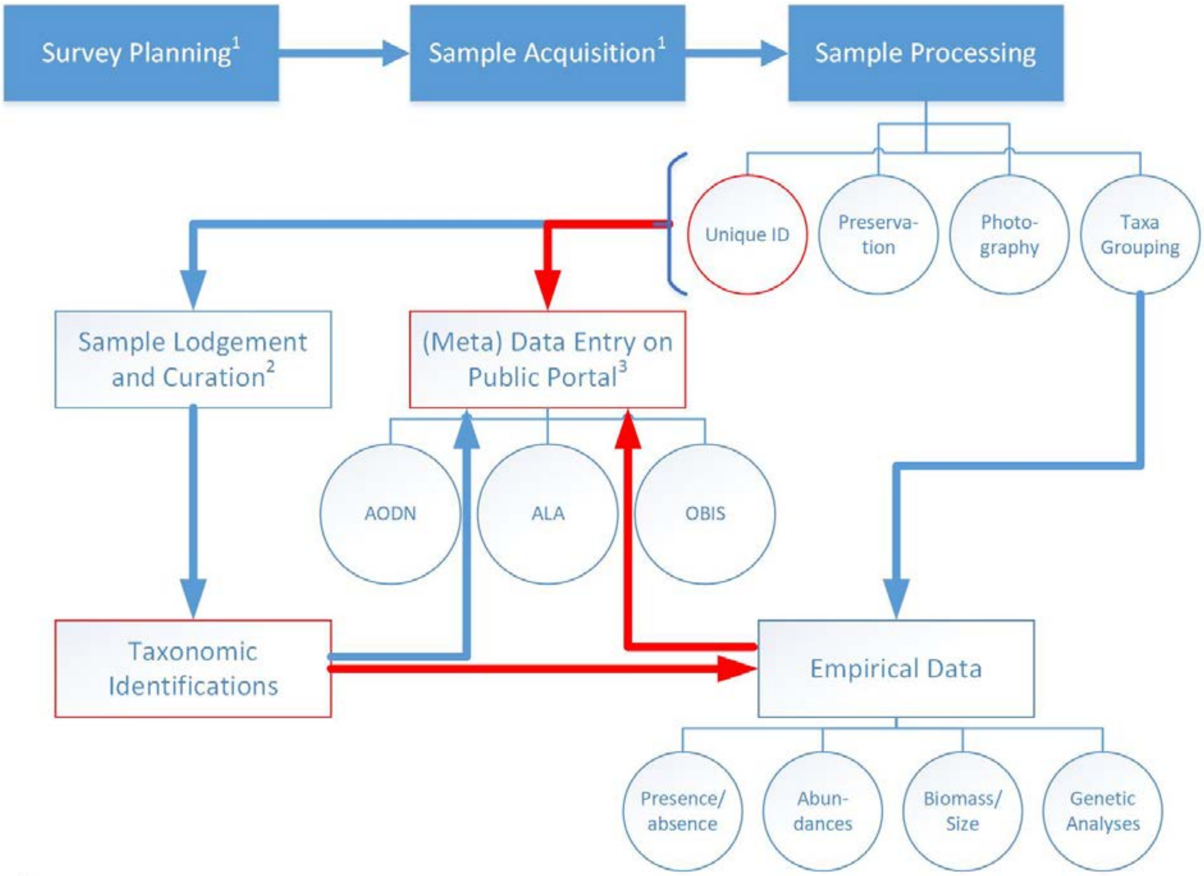
3) Specimens are lodged at appropriate museums or research organisations (e.g. CSIRO National Fish Collection), with an agreement to share taxonomic identifications with the survey leader as they progress. The unique identifier is included in the museum database, along with a newly assigned museum registration number if required, noting multiple registration numbers can eventuate from one unique identifier (e.g. if the museum separates the initial onboard grouping into multiple species or individuals). The unique identifier now becomes XXX-YYY where XXX is the original onboard identifier, and YYY is the museum registration number.

4) As they progress, taxonomic identifications are entered into the relevant museum database, along with survey identifier and the unique identifier. This data will be harvested by the ALA and OBIS using an automated system which should link to the survey identifier and unique identifier. Over time, all taxonomic identifications from a given survey should therefore be available via the ALA, even if taxonomic identifications take years to complete among various individuals or institutions.

5) If available, associated data on absences, abundances, biomass and genetic information should be made publicly accessible, as this is important crucial ecological information.

a. OBISAU provides capability for researchers to include absences, abundances and biomass, as well as environmental data, if this information is available.

b. Alternatively, researchers should link the relevant information (e.g. species composition matrix) to the meta data submitted to the AODN immediately after the survey (Step 2 above).



[1] Follows national SOPs in Przeslawski and Foster 2018
[2] Can include multiple institutions, may include temporary lodgement at collecting institution while preliminary identifications (e.g. OTUs) are undertaken
[3] Includes associated environmental data collected in the sampling platform

Figure 9 Ideal workflow for biological specimen data. Blue boxes represent activities undertaken onboard, and white boxes represent post-survey activities. Red boxes and lines indicate activities and linkages that require further development to be incorporated into a national workflow.

3.5 Barriers and Challenges

The main challenges associated with making biological specimen data discoverable and accessible are i) the limited resources available to identify specimens and ii) the lack of immutable identifiers which then percolates to other specific barriers (listed below) (Figure 10, Figure 11). For the latter, an example is that one benthic sled deployment may be associated with the following identifiers, all different and potentially assigned by different institutions at various times from specimen acquisition through to curation: i) a sample

number representing the entire catch, ii) a lot number representing a coarse grouping (e.g. phylum, family), iii) an identifier representing operational taxonomic unit, iv) a registration number during lodgement at a museum, and v) a refined identifier once taxonomic identification to species is complete. Similar challenges exist in other disciplines strongly linked to specimens including botany (Nelson et al. 2018) and geology, and in the latter case seem to have been somewhat resolved by the creation of an International Geo Sample Number (IGSN) Minting Service (https://www.auscope.org.au/igsn-info).
Workshop participants rated controlled meta data standards and correct taxonomic identifications as most important followed by linkages between platforms and the correction of errors including duplicates. According to workshop participants, identifying specimens accurately and correcting errors manually involves the most resources and time (Figure 11).

More specific barriers to accessible and discoverable biological specimen data were also discussed in breakout groups, including:

- Taxonomic experts generally are not involved in pre-survey planning to identify expertise, resources and taxonomic resolution needed.

- The workflow from specimen collection to data upload is complex, unclear and involves multiple institutions and often a very long duration. (e.g. identifiers change, metadata is not linked, taxonomic changes and updates are not consistently integrated).

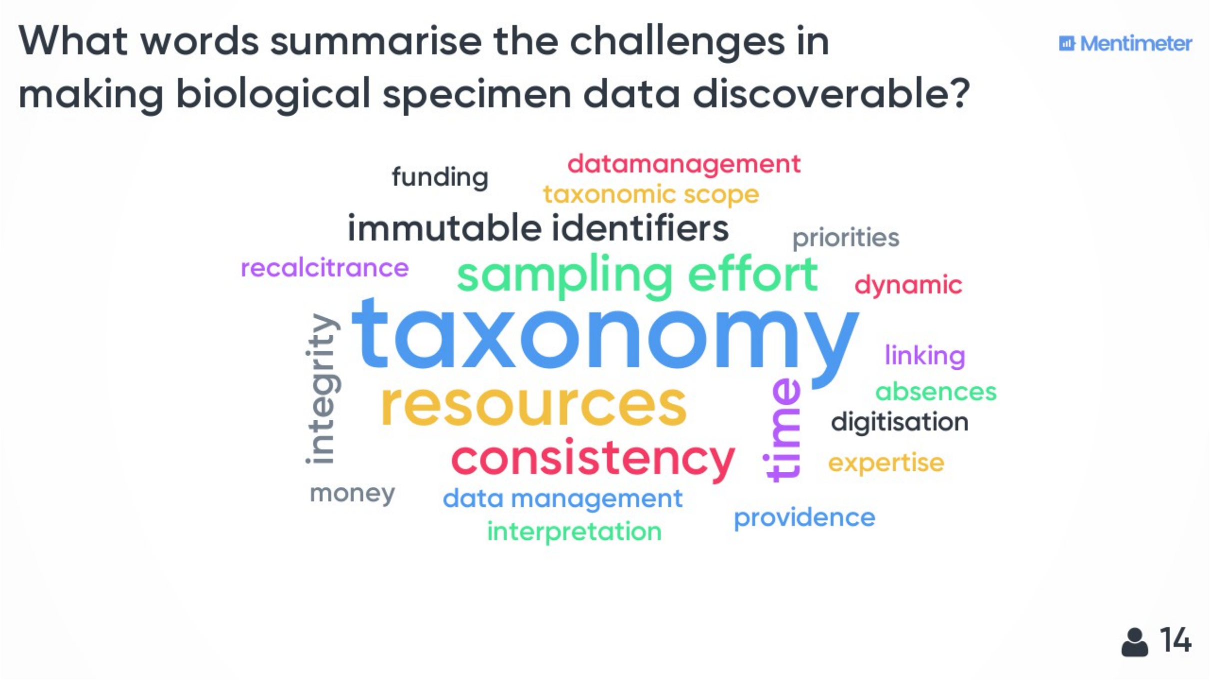
- It is unclear how to best deal with the large volume of legacy and/or unidentified (dark) specimens, which are currently undocumented. There are limited resources and time to work them up and make them discoverable and accessible.

- Species occurrence data from sled-collected specimens eventually appears on the ALA or OBIS, but this is an often circuitous route through museums and very rarely includes species absences or sampling effort information crucial for ecological and monitoring purposes.

- There seems to be little traction establishing and getting uptake of metadata and data standards (e.g. DarwinCore).

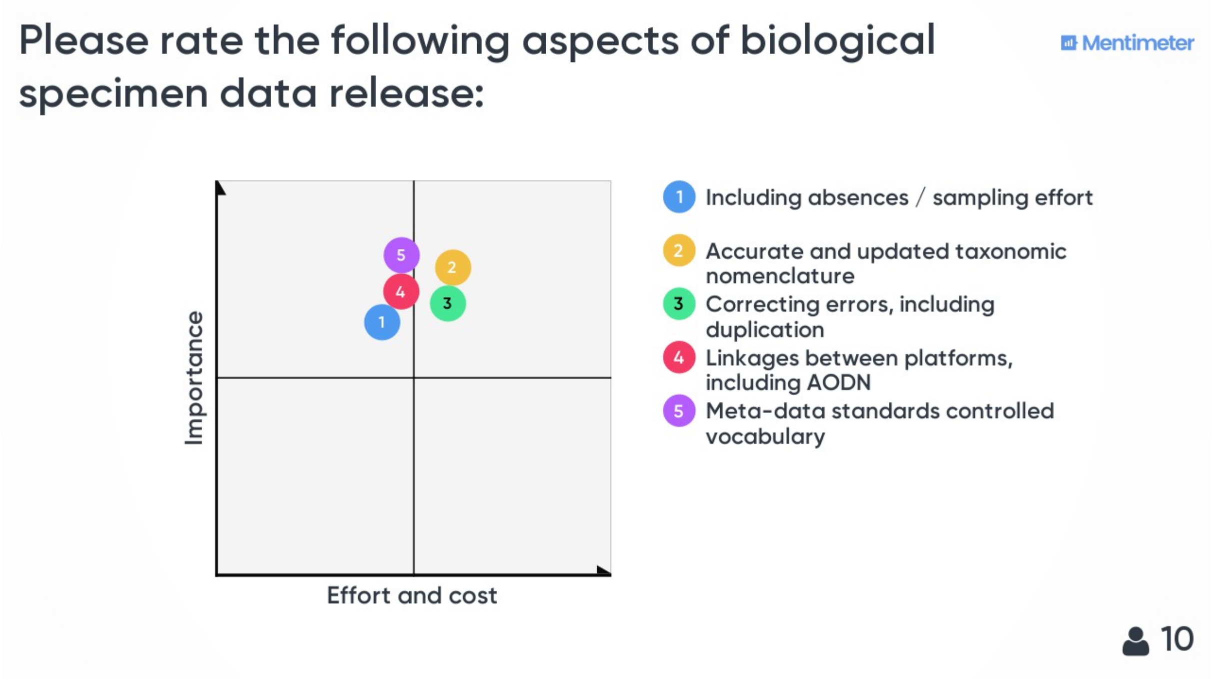
- There are few incentives to follow standards and release data (or deterrents not to). These are time-consuming tasks, and the current system disincentives these by prioritising funding for data collection and analysis, rather than data and specimen management.

- There are no clear and consistent links and feedback loops between platforms (e.g. AODN, ALA, OBIS), and the ones that do exist are often not fully automated or are dependent on a single person.



What words summarise the challenges in making biological specimen data discoverable?

Figure 10 Word cloud generated by 14 workshop participants describing the challenges they experience regarding biological specimen workflows.



Please rate the following aspects of biological specimen data release:

1. Including absences / sampling effort
2. Accurate and updated taxonomic nomenclature
3. Correcting errors, including duplication
4. Linkages between platforms, including AODN
5. Meta-data standards controlled vocabulary

Figure 11 Ten workshop participants rated controlled meta data standards and correct taxonomic identifications as most important followed by the correction of errors including duplicates.

3.6 Recommendations

The workshop group compiled several recommendations to address the barriers listed above.

Barrier 1: Taxonomic experts are not always suitably involved in pre-survey planning

Recommendations:

- Researchers should follow national SOPs during survey planning (Przeslawski and Foster 2018), specifically by contacting taxonomic expert(s) during pre-planning survey period to secure commitment, logistics and required taxonomic resolution.
- Version 2 of the NESP field manuals should elucidate incentives for uptake of appropriate pre-survey planning.
- Researchers should consider the available time and funding in tandem with the survey needs, as taxonomists may not be needed in the first place (e.g. species inventory exists and OTU-level required)
- Review current user-friendly identification keys and scope the development of needed keys, in conjunction with expert taxonomists and field ecologists (e.g. CSIRO's FishIDer, NIWA's Marine Identification Guides).

Barrier 2: The workflow from specimen collection to data upload may be complex, unclear and involves multiple institutions and often a very long duration.

Recommendations:

- Databases and data formats need clear standards (WoRMS), and this must be clearly communicated to field scientists and museum curators.
- A central platform with immutable identifiers must be identified or established from which mutable identifiers (e.g. refined taxonomies) can be linked (e.g. identifier R2R program). The critical elements in an identifier must be identified to help standardise any immutable and linked identifiers.
- Explore harnessing workflows related to specimens of other disciplines (e.g. herbarium specimens (Nelson et al. 2018) and geological specimens (International Geo Sample Network))
- Digitise the collection process where possible. Consider a single ship-based database (i.e. such that RV Investigator, RV Nuyina, RV Solander have the same system)
- Scope funding and other resources needed to automate or streamline the error correction process in the ALA-museum feedback loop.
- Scope national QC process for biological specimen data, possibly via ALA or AODN
- Both collecting institutes and museums should apply standardised (and preferable automated) tests to detect duplicate records.

Barrier 3: It is unclear how best to deal with the large volume of legacy and/or unidentified (dark) specimens, which are currently undocumented.

Recommendations:

- Generate a 'gumtree' type service for collected-but-yet-to-be-identified samples. Explore similar initiatives (e.g. Otlet) to adapt or harness existing infrastructure or architecture
- Determine if there's broad or national interest in doing this. Consider key taxa or bite-size project-based approach.

Barrier 4: Species occurrence data is a circuitous route through museums and very rarely includes species absences or sampling effort information crucial for ecological and monitoring purposes.

Recommendations:

- Engage more broadly with OBIS International to gauge their interest in supporting this work and/or communicate this capability of OBISAU more broadly to researchers in Australia. Refer to De Pooter et al. (2017) for specific suggestions related to expanding OBIS beyond species occurrence data.
- Ensure that matrices containing species abundances, absences, and other data can be submitted to AODN and linked to relevant meta data.
- Include a new step in V2 national SOP to encourage researchers to link their absences, abundance, biomass, or genetic analyses to the meta data already submitted to AODN.

Barrier 5: There seems to be little traction establishing and getting uptake of metadata and data standards (e.g. DarwinCore).

Recommendations:

- Consider establishing a working group to clearly advocate for data standards, including a communication plan to maximise uptake and scope incentives. This group should link with the Faunal Collections Informatics Group (FCIG), a museums data working group tasked with implementing DarwinCore across museums.
- Consider building use-case for an ideal provenance chain.

Barrier 6: There are few incentives to follow standards and release data (or deterrents not to).

Recommendations:

- Researchers, managers, curators, and data managers must articulate incentives to support open data infrastructure and abide by standards. This can be included in the V2 of the national SOPs as well as agency-specific SOPs.
- The MNF, IMOS/AODN and major research institutions should promote their open data policies and support digital infrastructure that enables this.

Barrier 7: There are no clear links and feedback loops between platforms (AODN, ALA, OBIS).

Recommendations:

- Continue the work begun as part of the Marine Research Data Cloud project to harmonise web services and capability for queries and outputs, and establish consistent service across infrastructures.
- Streamline the feedback loop between individual museums and ALA/OBIS.
- Scope Europe’s controlled vocabularies regarding sampling methodologies.

4. CONCLUSION

Data sharing is increasingly recognised as important and expected among scientists, but is still hampered by major barriers (Tenopir et al. 2015). Global initiatives are underway to develop national repositories for environmental data, automated analyses and annotation, and visualisation platforms to aide evidence-based decision (e.g. Coalition for Publishing Data in the Earth and Space Sciences, https://copdess.org). However, we have yet to establish a national-scale systematic end-to-end workflow and associated infrastructure for publishing collated marine data included in this report (imagery, biological specimen data).

The two Data Discoverability and Accessibility workshops described in this report successfully brought together key players working with marine imagery and biological specimen data to identify the main challenges to making their data discoverable and accessible. More importantly, the participants in these workshops provided a way forward through the establishment of clear lists of recommendations to address these challenges (Section 2.6 and 3.6).

In these recommendations, we have deliberately avoided naming individuals or agencies that could implement or fund these solutions, except where already embedded in a funded project plan (e.g. NESP milestone to develop infographic). This was outside the scope of the workshops and may also unnecessarily narrow options to address discrete recommendations.

Ultimately, we hope that this workshop report represents a foundation from which future programs can be developed, funded, and implemented to ensure clear and consistent national workflows underpinned by stable and user-friendly digital infrastructure. A follow-up workshop will be held in late 2019 to progress recommendations in this report related to marine imagery. To maximise national benefit, ongoing consultation and collaboration with key national agencies (e.g. AODN, NMSC) will be vital for future developments in this space.

REFERENCES

De Pooter D, Appeltans W, Bailly N, Bristol S, Deneudt K, Eliezer M, Fujioka E, Giorgetti A, Goldstein P, Lewis M, Lipizer M, Mackay K, Marin M, Moncoiffé G, Nikolopoulou S, Provoost P, Rauch S, Roubicek A, Torres C, van de Putte A, Vandepitte L, Vanhoorne B, Vinci M, Wambiji N, Watts D, Klein Salas E, Hernandez F. 2017. Toward a new data standard for combined marine biological and environmental datasets - expanding OBIS beyond species occurrences. Biodiversity Data Journal 5: e10989. https://doi.org/10.3897/BDJ.5.e10989

Easterday, K., T. Paulson, P. DasMohapatra, P. Alagona, S. Feirer, and M. Kelly. 2018. From the Field to the Cloud: A Review of Three Approaches to Sharing Historical Data From Field Stations Using Principles From Data Science. Frontiers in Environmental Science 6.

European Commission. 2016. H2020 Programme - Guidelines on FAIR Data Management in Horizon 2020.

Koppe, R., P. Gerchow, A. Macario, A. Haas, C. Schäfer-Neth, S. Rehmcke, A. Walter, T. Düde, P. Weidinger, A. Schäfer, and H. Pfeiffenberger. 2018. SENSOR.awi.de: Management of heterogeneous platforms and sensors.in RDA 11th Plenary, Berlin.

McKiernan, E. C., P. E. Bourne, C. T. Brown, S. Buck, A. Kenall, J. Lin, D. McDougall, B. A. Nosek, K. Ram, C. K. Soderberg, J. R. Spies, K. Thaney, A. Updegrove, K. H. Woo, and T. Yarkoni. 2016. How open science helps researchers succeed. eLife 5:e16800.

Nelson, G., P. Sweeney, and E. Gilbert. 2018. Use of globally unique identifiers (GUIDs) to link herbarium specimen records to physical specimens. Applications in Plant Sciences 6:e1027.

OECD. 2018. Open Government Data Report: Enhancing Policy Maturity for Sustainable Impact. Organisation for Economic Cooperation and Development, Paris.

Przeslawski, R., B. Alvarez, C. Battershill, and T. Smith. 2014. Sponge biodiversity and ecology of the Van Diemen Rise and eastern Joseph Bonaparte Gulf, northern Australia. Hydrobiologia 730:1-16.

Przeslawski, R., and S. Foster. 2018. Field Manuals for Marine Sampling to Monitor Australian Waters. National Environmental Science Programme, Marine Biodiversity Hub.

Przeslawski R, Bodrossy L, Carroll A, Cheal A, Depczynski M, Foster S, Hardesty BD, Hedge P, Langlois T, Lara-Lopez A, Lepastrier A, Mancini S, Miller K, Monk J, Navarro M, Nichol S, Sagar S, Stuart-Smith R, van de Kamp J, Williams J. 2019. Scoping of new field manuals for marine sampling in Australian waters. Report to the National Environmental Science Programme, Marine Biodiversity Hub. Geoscience Australia.

Stocks, K. I., N. J. Stout, and T. M. Shank. 2016. Information management strategies for deep-sea biology. Pages 368-385. Biological Sampling in the Deep Sea. Wiley Blackwell, West Sussex.

Tenopir, C., E. D. Dalton, S. Allard, M. Frame, I. Pjesivac, B. Birch, D. Pollock, and K. Dorsett. 2015. Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide. PLOS ONE 10:e0134826.

Wilkinson, M. D., M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. G. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. C. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, and B. Mons. 2016. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3:160018.

APPENDIX A – AGENDA FOR DATA DISCOVERABILITY AND ACCESSIBILITY WORKSHOP I – MARINE IMAGERY

6-7 September 2018
Scrivener Room, Geoscience Australia, Canberra

Objective: The aim of the workshop is to discuss current developments and identify key actions needed to establish a national workflow related to marine benthic and demersal imagery and annotations.

Attendees: Rachel Przeslawski (GA)*, Sebastian Mancini (AODN)*, Andrew Carroll (GA), Ari Friedman (Greybits), Tim Langlois (UWA), Oscar Pizarro (USYD), Alan Jordan (NSW DPI), Scott Foster (CSIRO), Mat Wyatt (AIMS), Manuel Gonzalez-Riviero (AIMS), Jac Monk (IMAS), Nev Barrett (IMAS), Maggie Tran (GA), Alix Post (GA), Scott Nichol (GA) Mark Rehbein (AIMS), Peter Walsh (UTAS), Steph Bagala (Macquarie), Inke Falkner (GA), Julia Martin (ARDC), Melanie Barlow (ARDC), Stefan Williams (USYD), Franzis Althaus (CSIRO)
* Chair

Day 1

9:00	Workshop opening and introductions		Andrew Carroll, Seb Mancini
9:30	Purpose and scope of workshop		
		NESP objectives	Andrew Carroll
		AODN objectives	Seb Mancini
10:15	Morning tea		
10:45	Current developments		
	10:45	Squidle+	Ari Friedman
	11:20	Global Archive	Tim Langlois
	11:55	GA imagery collection	Andrew Carroll
	12:05	AIMS imagery collection	Mat Wyatt, Manuel Gonzalez-Riviero
	12:15	CSIRO imagery collection	Franzis Althaus
	12:25	IMAS imagery collection	Peter Walsh, Nev Barrett
	12:35	Automated image analysis (ARC-LIEF proposal)	Oscar Pizarro
	12:45	Other agency (e.g. state) perspectives	Alan Jordan, all
13:00	Lunch		
13:45	Discussion: Linkages and gaps		All
14:00	Activity: What is the current workflow(s) for marine imagery?		Chairs: Tim Langlois, Franzis Althaus, Jac Monk
14:45	Arvo tea		
15:15	Presentation of current workflows, including major issues		Chairs

15:45	Activity: What is the ideal workflow?	Chairs: Scott Foster, Andrew Carroll, Alan Jordan
16:15	Presentation of ideal workflows	Chairs
17:00	Day 1 close	
18:00	Dinner at Kingston Foreshore	

Day 2

9:00	Recap	Rachel Przeslawski, Seb Mancini
9:10	Discussion: Identify barriers to ideal workflow	All
10:15	Morning tea	
10:45	Discussion: Action needed for each barrier to ideal workflow	All
13:00	Lunch	
14:00	Discussion: Where to from here?	All
14:45	Workshop summary	Rachel Przeslawski, Seb Mancini
15:00	Workshop close	

APPENDIX B – AGENDA FOR DATA DISCOVERABILITY AND ACCESSIBILITY WORKSHOP II – BIOLOGICAL SPECIMEN DATA

26-27 September 2018
Freycinet Room, CSIRO, Hobart

Objective: The aim of the workshop is to discuss current developments and identify key actions needed to establish a national workflow related to data associated with biological specimen identifications.

Attendees: Rachel Przeslawski (GA)*, Sebastian Mancini (AODN)*, Scott Foster (CSIRO), Katherine Tattersall (CSIRO), Dave Watts (CSIRO), Pamela Brodie (CSIRO), Narissa Bax (IMAS), Dave Connell (AAD), Jonny Stark (AAD), Johnathan Kool (AAD), Emma Flukes (NESP/UTAS), Peggy Newman (ALA), Miles Nicholls (ALA), Felicity McEnnulty (CSIRO), Glenn Johnstone (AAD), Haylee Weaver (Aust Faunal Directory), Anthony Whalen (Aust Faunal Directory), Kirrily Moore (Tasmanian Museum and Art Gallery TMAG), Karen Gowlett-Holmes (CSIRO), Lisa Kirkendale (Western Australian Museum WAM), Inke Falkner (GA), Peter Walsh (UTAS)
* Chair

Day 1

9:00	Workshop opening and introductions		Rachel Przeslawski, Seb Mancini
9:30	Purpose and scope of workshop		Rachel Przeslawski, Seb Mancini
		NESP objectives	Rachel Przeslawski
		AODN objectives	Seb Mancini
10:15	Morning tea		
10:45	Current developments and workflows		
	10:45	Atlas of Living Australia	Peggy Newman
	11:15	OBIS	Dave Watts
	11:45	National Species List / Australian Faunal Directory	Anthony Whalen Haylee Weaver
	12:00	CSIRO biological specimen data	Karen Gowlett-Holmes
	12:15	GA biological specimen data	Rachel Przeslawski
	12:30	AAD biological specimen data	Jonny Stark, Glenn Johnstone
	12:45	Museum perspective	Kirrily Moore
13:00	Lunch		
13:45	Absences and sampling effort		Scott Foster
14:15	Activity: What is the current workflow(s) for biological specimen data release?		Chairs: Scott, Glenn, Kirrily
15:15	Afternoon tea		
15:45	Presentation of workflows, including major issues		Chairs / all
17:00	Day 1 close		
17:30	Drinks with IMOS Task Team meeting attendees, Salamanca Place, TBC		

Day 2

9:00	Recap	Rachel Przeslawski, Seb Mancini
9:10	Activity: What is the ideal workflow(s)?	Chairs: Seb Mancini, Haylee Weaver
9:45	Presentation of ideal workflows	Chairs/all
10:15	Morning tea	
10:45	Discussion: Identify the barriers to ideal workflow(s)	All
11:30	Discussion: Recommendations to address each barrier	All
13:00	Lunch	
14:00	Discussion: Where to from here?	All
14:30	Workshop summary	Rachel Przeslawski, Seb Mancini
15:00	Workshop close	

www.nespmarine.edu.au

Contact:

Rachel Przeslawski
Geoscience Australia

Address | GPO Box 378, Canberra ACT, 2601
email | rachel.przeslawski@ga.gov.au
tel | +61 2 6249 9101