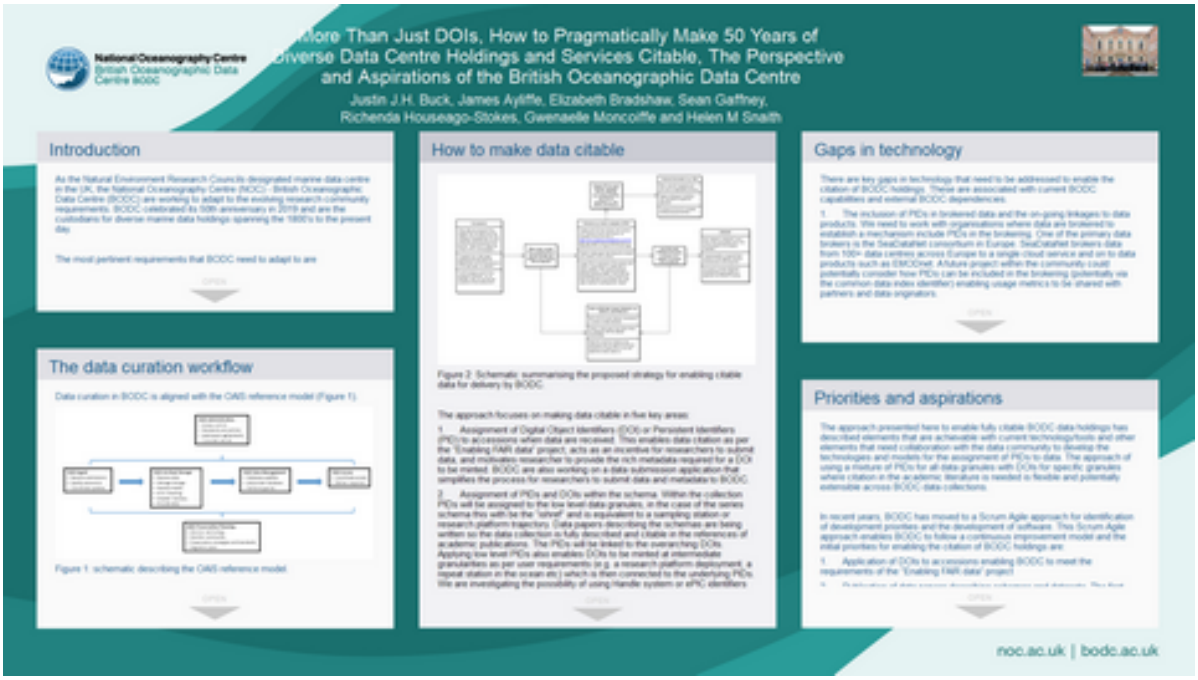
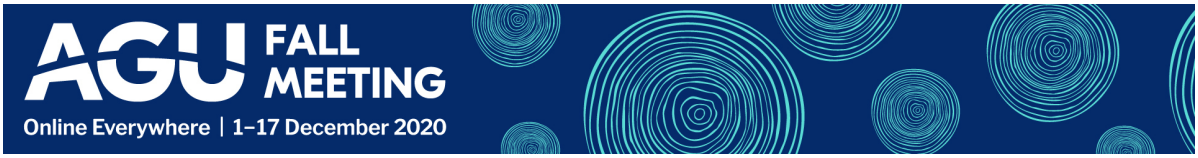
More Than Just DOIs, How to Pragmatically Make 50 Years of Diverse Data Centre Holdings and Services Citable, The Perspective and Aspirations of the British Oceanographic Data Centre



Justin J.H. Buck, James Ayliffe, Elizabeth Bradshaw, Sean Gaffney,
Richenda Houseago-Stokes, Gwenaelle Moncoiffe and Helen M Snaith

ENTER NAMES OF AFFILIATED INSTITUTIONS

PRESENTED AT:



INTRODUCTION

As the Natural Environment Research Councils designated marine data centre in the UK, the National Oceanography Centre (NOC) - British Oceanographic Data Centre (BODC) are working to adapt to the evolving research community requirements. BODC celebrated its 50th anniversary in 2019 and are the custodians for diverse marine data holdings spanning the 1800's to the present day.

The most pertinent requirements that BODC need to adapt to are

- Findable, Accessible, Interoperable and Reusable (FAIR) principles (Wilkinson et al. 2018, Tanhua et al. 2019)
- Transparency, Responsibility, User focus, Sustainability and Technology (TRUST) principles for digital repositories (Lin et al. 2020)
- Collective benefit, Authority for Control, Responsibility and Ethics (CARE, Carroll et al. 2020)
- Open Archival Information Systems (OAIS) Reference Model, ISO 16363 (http://www.oais.info/)
- Transition from World Data System to CoreTrustSeal accreditation in 2021 (https://www.coretrustseal.org/).

With such a legacy of holdings, adapting to the rapidly evolving community requirements is a challenging situation with pragmatic approaches to meet requirements needed. A key element that spans these requirements is making data citable. This was identified as a priority in the outcomes of the "Enabling FAIR data" project (Stall et al. 2018).

This poster describes the focus and ambitions for how BODC aims to make all its holding citable, using the "series schema", one of BODCs primary data collections, as the case study. The poster will introduce the data curation workflow before presenting how relevant elements can be made citable. Gaps in current technology will be described before closing with a summary of the development priorities in achieving data capability.

THE DATA CURATION WORKFLOW

Data curation in BODC is aligned with the OAIS reference model (Figure 1).

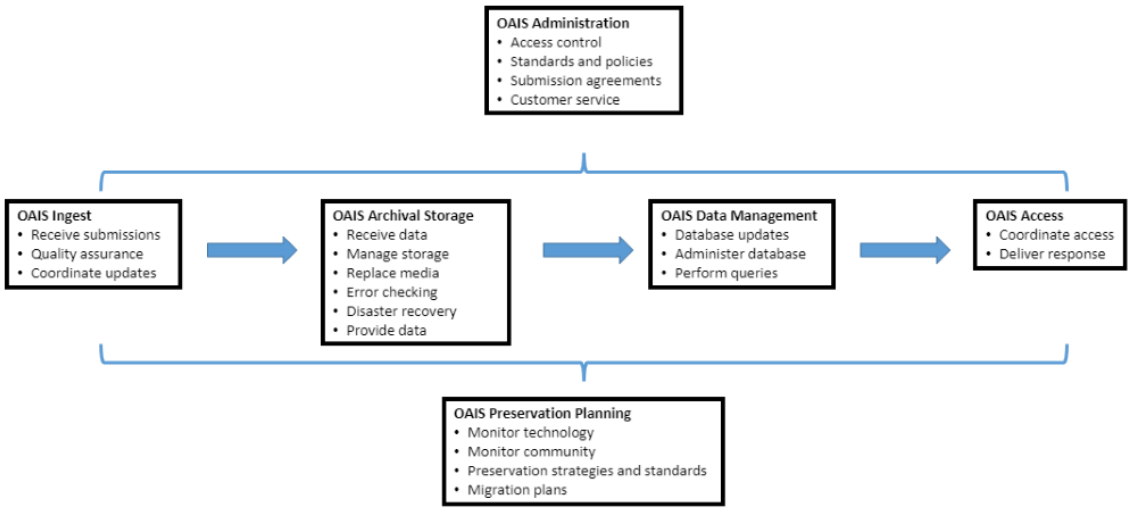


Figure 1: schematic describing the OAIS reference model.

Each component of the OAIS reference model will now be described.

Ingest – When data are submitted to BODC for curation, data managers quality assure submissions and assemble the metadata necessary for curation.

Archival storage - Digital data submissions (as received) are placed in a long-term accession and stored in triplicate across multiple sites.

Data management - Appropriate data are transferred into a standard internal format, documentation assembled and metadata loaded into BODC databases, with source variable names, units and essential metadata converted to controlled vocabularies from the NERC Vocabulary Server (NVS https://vocab.nerc.ac.uk/).

Access - delivered via the BODC website and brokered to partner repositories such as SeaDataNet. Discovery metadata is aligned with the SeaDataNet community standards. BODC are also progressively implementing data services such as the US ERDDAP tool (Snowden et al., 2019) to provide Application Program Interface (API) based data access.

Administration – Access control is attained by assigning a data policy to each data series (a series granule is an ocean profile, research vessel trajectory, etc) and this policy is used to administer access when data are requested from the BODC website or the requests team assemble data for manual requests. Data are converted to community formats including Ocean Data View ASCII and SeaDataNet NetCDF, with data described using terms from the NERC vocabulary server. BODC submission agreements are documented on the BODC website and customer service is assured with a dedicated requests team that serve data following local regulations including General Data Protection Regulation (GDPR) 2018 and Environmental Information Regulations (EIR) 2004.

Preservation planning – This is attained through engagement with the research data and marine research community including engaging with the Challenger Society for Marine Science, Research Data Alliance, World Data System, CODATA, European Geophysical Union, and American Geophysical Union. BODC has a well-established preservation strategy and mitigation plans which have proven to be robust.

The ability to cite data is required at several stages in the data curation workflow:

- The version of data submitted by the originator, as this is the version of data that is often associated with scientific publications by the originator. This is currently facilitated through manual assembly of a dataset in the BODC published data library which is based on the approach documented in the Ocean Data Publication Cookbook (IOC)

- The aggregated data product in each of the BODC schemas. This poster will use the series schema as a case study.

- Data that have been brokered to BODC partners or included in aggregated collections of data products. This requires data to be persistently identifiable at the same granule size that data are delivered by BODC.

In addition to the requirement for data to be citable BODC need to be able to track when data are cited in academic literature as data originators are increasingly interested in metrics on how the data they have submitted to BODC are reused.

HOW TO MAKE DATA CITABLE

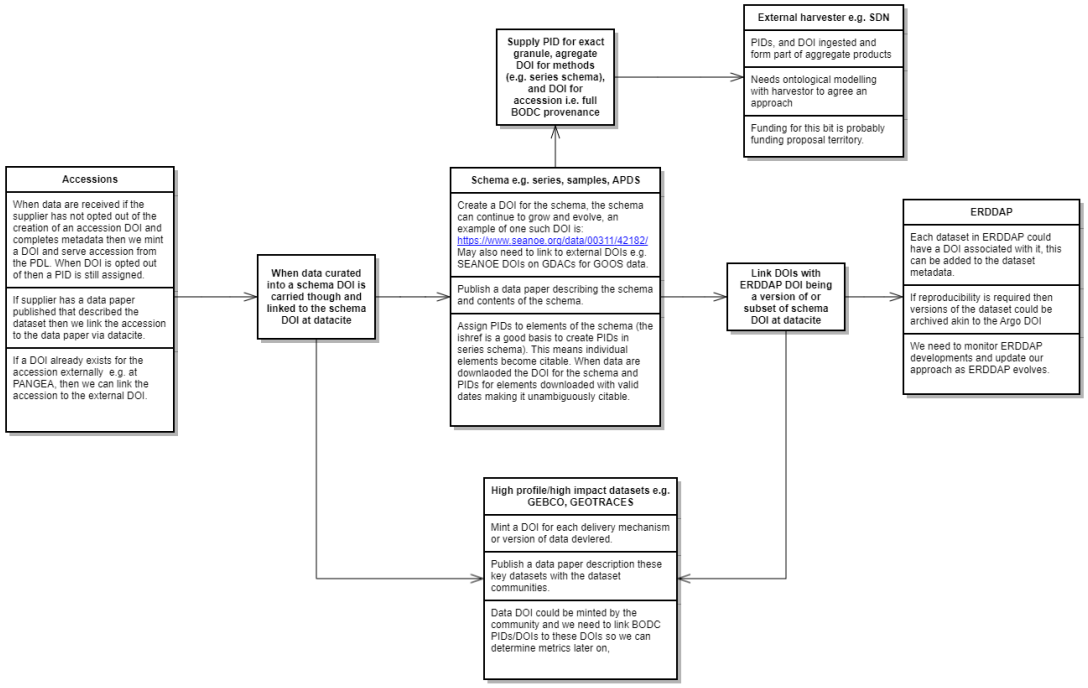


Figure 2: Schematic summarising the proposed strategy for enabling citable data for delivery by BODC.

The approach focuses on making data citable in five key areas:

1. Assignment of Digital Object Identifiers (DOI) or Persistent Identifiers (PID) to accessions when data are received. This enables data citation as per the “Enabling FAIR data” project, acts as an incentive for researchers to submit data, and motivates researcher to provide the rich metadata required for a DOI to be minted. BODC are also working on a data submission application that simplifies the process for researchers to submit data and metadata to BODC.

2. Assignment of PIDs and DOIs within the schema. Within the collection PIDs will be assigned to the low level data granules, in the case of the series schema this with be the “ishref” and is equivalent to a sampling station or research platform trajectory. Data papers describing the schemas are being written so the data collection is fully described and citable in the references of academic publications. The PIDs will be linked to the overarching DOIs. Applying low level PIDs also enables DOIs to be minted at intermediate granularities as per user requirements (e.g. a research platform deployment, a repeat station in the ocean etc) which is then connected to the underlying PIDs. We are investigating the possibility of using Handle system or ePIC identifiers (https://www.pidconsortium.net/) for the underlying PIDs; these have a lower maintenance overhead than DOIs and can be minted automatically once a mirror to the PID provider is setup. When data are supplied from the schema to users they would be provided with the DOI for the data paper and the PIDs included in their request so enable unambiguous citation of the data.

3. Working with communities on data papers, where BODC is a partner. BODC are part of several international data infrastructures or collaborations such as Argo and GEBCO. BODC could mint a DOI for each collection of data and this can be cited by the infrastructure. We are also working collaboratively with these communities to produce data papers. An example of one such paper is Wong et al. (2020).

4. Working with organisations where data are brokered to establish a mechanism to include PIDs in the brokering. BODC are a partner in the SeaDataNet (https://www.seadatanet.org/) consortium in Europe. SeaDataNet brokers data from 100+ data centres across Europe to a single cloud service and on to data products such as EMODNet (Martin Miguez et al., 2019). A future project in the community could potentially consider how PIDs can be included in the brokering enabling usage metrics to be shared with partners.

5. Assignment of PIDs or DOIs to data services on top of BODC holdings. BODC are progressively working to add more data from its schemas into the ERDDAP tool (Snowden et al. 2019). Assigning DOIs to datasets in ERDDAP which are connected to the underlying PIDs would enable requests from such services to be unambiguously cited by users.

In addition to enabling the citation of each element of data delivery, the accession DOIs will be linked to the ingested data PIDs, data papers and DOIs associated with products. These connections will potentially enable metrics on data usage to be derived at the accession level and data originators to know the impact of data they submitted to BODC.

GAPS IN TECHNOLOGY

There are key gaps in technology that need to be addressed to enable the citation of BODC holdings. These are associated with current BODC capabilities and external BODC dependencies.

1. The inclusion of PIDs in brokered data and the on-going linkages to data products. We need to work with organisations where data are brokered to establish a mechanism include PIDs in the brokering. One of the primary data brokers is the SeaDataNet consortium in Europe. SeaDataNet brokers data from 100+ data centres across Europe to a single cloud service and on to data products such as EMODnet. A future project within the community could potentially consider how PIDs can be included in the brokering (potentially via the common data index identifier) enabling usage metrics to be shared with partners and data originators.

2. The schemas are constantly evolving with data added daily and occasionally updated when we receive feedback on data. Data users will need to be able to cite the data they use unambiguously even when data versions are updated. This is not a challenge unique to BODC and Argo (2020) is an example of a solution to this situation. This does have the weakness that the data version granules are only locally resolvable. Thus, an approach where the PID for a granule is updated when a new version is published may be more appropriate, as this has the potential to be fully resolvable on international PID infrastructure.

3. This dependency on PIDs connected at multiple levels (accession DOI to internal PID to delivery DOI within BODC, and potentially additional layers when data are brokered beyond BODC) is, to our knowledge, untested. We are not aware if the PID infrastructure for Handles, ePIC, or DOI has been used this way. As such, currently technology may not enable the cited reference resolution we are aiming for to provide data usage metrics to data originators without development within the PID community.

4. The assignment of DOIs to data services is subject to the limitations describe above on evolving data and we are also dependent on the community development of tools such as ERDDAP for the method of implementation. We are keen to explore the possibilities of how PIDs can be applied to data services with these communities.

PRIORITIES AND ASPIRATIONS

The approach presented here to enable fully citable BODC data holdings has described elements that are achievable with current technology/tools and other elements that need collaboration with the data community to develop the technologies and models for the assignment of PIDs to data. The approach of using a mixture of PIDs for all data granules with DOIs for specific granules where citation in the academic literature is needed is flexible and potentially extensible across BODC data collections.

In recent years, BODC has moved to a Scrum Agile approach for identification of development priorities and the development of software. This Scrum Agile approach enables BODC to follow a continuous improvement model and the initial priorities for enabling the citation of BODC holdings are:

1. Application of DOIs to accessions enabling BODC to meet the requirements of the “Enabling FAIR data” project

2. Publication of data papers describing schemas and datasets. The first paper on the Series Schema is being drafted and collaborative papers are on-going with GEBCO being drafted and Argo published (other papers will be needed as communities mature).

3. Liaison with PID providers on the proposed approach to look at feasibility of the approach, detailed modelling and estimating costs. This engagement is on-going with ePIC.

In addition to these priorities, we are keen to get feedback from the community including AGU Earth Space and science informatics and this poster forms part of that process.

AUTHOR INFORMATION

Contact details for BODC staff are available at https://www.bodc.ac.uk/about/contact_us/staff_list/. This section will include a short biography for each co-author of this poster.

Dr Justin J.H. Buck (https://orcid.org/0000-0002-3587-4889) – Justin is a senior data manager, product owner and member of the Series schema team in BODC. Justin specialises in the management of data from autonomous platforms and has worked across many elements of BODC systems and architecture over the last 11 years. Justin is also no stranger to the complexities and challenges of enabling data citation having previously collaborated with Ifremer on the Argo DOI for dynamic data.

James Ayliffe (https://orcid.org/0000-0002-7363-5029) is a marine data manager at the British Oceanographic Data Centre, with over 8 years' experience in marine data management. His background is in physical oceanography and marine policy. James has in interest in how data flows work across multi-disciplinary, multi domain networks. James is a member of the GOA-ON North-East Atlantic hub and MEDIN DAC working group.

Elizabeth Bradshaw is a sea level data manager at the British Oceanographic Data Centre, quality controlling national, European and international sea level data and ensuring its availability for future reuse. She is head of the Permanent Service for Mean Sea Level and also a member of the GLOSS Group of Experts. She has a particular interest in data archaeology and rescuing historical sea level data.

Sean Gaffney is the Database Administrator at the British Oceanographic Data Centre, with 15 years of marine data management expertise, concentrated primarily on coastal shelf datasets, His background is in ocean optics and he has an interest in good practice archiving and management of ocean model outputs. He is a core team member of the Marine Environmental Data and Information Network, specialising in geospatial discovery metadata.

Richenda Houseago-Stokes is a senior data manager at the British Oceanographic Centre. Within BODC Richenda is responsible for the Sample Schema data and managing the Natural Environment Research Council (NERC) portfolio of grants.

Gwenaëlle Moncoiffé is a biological oceanographer and a senior data manager with 30 years of research and data management experience. She leads the Vocabulary Management Group in BODC, a team responsible for maintaining and further developing the NERC Vocabulary Server and its associated services. She is a member of the data citation team in BODC. Gwenaëlle is also a member of the Research Data Alliance and co-chair a working group on semantic interoperability within and across scientific domains. She has an interest in implementing FAIR data principles in BODC and promoting them within the scientific community.

Helen M Snaith is a senior data manager at the BODC, Head of the BODC's Southampton Site and member of the senior management team who develop and maintain the BODC strategy and delivery plans. Helen has more than 30 years research experience in Oceanography, having specialised for many years in the development of satellite altimetry. She has been involved in the development and quality control of new products for a range of satellite missions, including the European Space Agency's marine Explorer missions (CryoSAT, GOCE and SMOS) and the recent Sentinel series of satellite altimeters. She has also led in developing services to access these data, in combination with observational data, for coastal and Storm Surge research. Within BODC, she has a responsibility for delivery of 'High Volume' data such as ocean model output, swath-based marine geoscience data and other non-numerical data sources including passive acoustics and imagery.

ABSTRACT

The British Oceanographic Data Centre (BODC) celebrated its 50th anniversary in 2019. It holds data collected from 1773 to the present day. Holdings are multidisciplinary, heterogeneous data reflecting the full range of disciplines, platforms, temporal and spatial fieldwork scales typically encountered in oceanographic research and monitoring. These collections vary in granularity and contain data which are at different stages of curation ranging from raw data to standardised data products.

BODC need to improve data services to meet the developing the expectations of the research community. These include the FAIR data principles, TRUSTed repository guidelines and CoreTrustSeal accreditation. This is a significant challenge within the constraints of resource available (both financial and human). The initial focus for BODC is making holdings citable with the following aspirations:

1. Application of DOIs to data at the point of receipt by BODC.
2. Publication of data papers and publication of DOIs for data products. Application of persistent identifiers to low level data granules where DOIs are not feasible.
3. Application of persistent identifiers to datasets included in BODC API services and versioning of these data.
4. Work with organisations or groups who include data curated by BODC in their products to enable the provenance of data to be unambiguous.
5. Work with communities on joint data papers where BODC are a partner organisation.

This will enable each type of data served by BODC to be unambiguously citable. The initial effort is being directed towards the application of DOIs to data submissions and publication of data papers for BODC curated data products.

National Oceanography Centre
British Oceanographic Data Centre BODC

(https://agu.confex.com/data/abstract/agu/fm20/7/0/Paper_749807_abstract_716912_0.png)

REFERENCES

Argo (2020). Argo float data and metadata from Global Data Assembly Centre (Argo GDAC). SEANOE. https://doi.org/10.17882/42182

Carroll, SR, et al. 2020. The CARE Principles for Indigenous Data Governance. Data Science Journal, 19: 43, pp. 1–12. DOI: https://doi.org/10.5334/dsj-2020-043

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

Martín Míguez B, Novellino A, Vinci M, Claus S, Calewaert J-B, Vallius H, Schmitt T, Pititto A, Giorgetti A, Askew N, Iona S, Schaap D, Pinardi N, Harpham Q, Kater BJ, Populus J, She J, Palazov AV, McMeel O, Oset P, Lear D, Manzella GMR, Gorringe P, Simoncelli S, Larkin K, Holdsworth N, Arvanitidis CD, Molina Jack ME, Chaves Montero MM, Herman PMJ and Hernandez F (2019) The European Marine Observation and Data Network (EMODnet): Visions and Roles of the Gateway to Marine Data in Europe. Front. Mar. Sci. 6:313. doi: 10.3389/fmars.2019.00313

Ocean Data Publication Cookbook. Paris: UNESCO. (Manuals and Guides. Intergovernmental Oceanographic Commission, 64), (IOC/MG/64).

Snowden D, Tsontos VM, Handegard NO, Zarate M, O' Brien K, Casey KS, Smith N, Sagen H, Bailey K, Lewis MN and Arms SC (2019) Data Interoperability Between Elements of the Global Ocean Observing System. Front. Mar. Sci. 6:442. doi: 10.3389/fmars.2019.00442

Stall S, Cruse P, Cousijn H, Cutcher-Gershenfeld J, de Waard A, Hanson B, Heber J, Lehnert K, Parsons M, Robinson E, Witt M, Wyborn L, Yarmey L (2018). Data Sharing and Citations: New Author Guidelines Promoting Open and FAIR Data in the Earth, Space, and Environmental Sciences. Science Editor. 2018;41(3):83-87.

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

Wilkinson, M. D., Sansone, S.-A., Schultes, E., Doorn, P., Bonino da Silva Santos, L. O., and Dumontier, M. (2018). A design framework and exemplar metrics for FAIRness. Sci. Data 5:180118. doi: 10.1038/sdata.2018.118

Wong APS, Wijffels SE, Riser SC, Pouliquen S, Hosoda S, Roemmich D, Gilson J, Johnson GC, Martini K, Murphy DJ, Scanderbeg M, Bhaskar TVSU, Buck JJH, Merceur F, Carval T, Maze G, Cabanes C, André X, Poffa N, Yashayaev I, Barker PM, Guinehut S, Belbéoch M, Ignaszewski M, Baringer MO, Schmid C, Lyman JM, McTaggart KE, Purkey SG, Zilberman N, Alkire MB, Swift D, Owens WB, Jayne SR, Hersh C, Robbins P, West-Mack D, Bahr F, Yoshida S, Sutton PJH, Cancouët R, Coatanoan C, Dobbler D, Juan AG, Gourrion J, Kolodziejczyk N, Bernard V, Bourlès B, Claustre H, D'Ortenzio F, Le Reste S, Le Traon P-Y, Rannou J-P, Saout-Grit C, Speich S, Thierry V, Verbrugge N, Angel-Benavides IM, Klein B, Notarstefano G, Poulain P-M, Vélez-Belchí P, Suga T, Ando K, Iwasaska N, Kobayashi T, Masuda S, Oka E, Sato K, Nakamura T, Sato K, Takatsuki Y, Yoshida T, Cowley R, Lovell JL, Oke PR, van Wijk EM, Carse F, Donnelly M, Gould WJ, Gowers K, King BA, Loch SG, Mowat M, Turton J, Rama Rao EP, Ravichandran M, Freeland HJ, Gaboury I, Gilbert D, Greenan BJW, Ouellet M, Ross T, Tran A, Dong M, Liu Z, Xu J, Kang K, Jo H, Kim S-D and Park H-M (2020) Argo Data 1999–2019: Two Million Temperature-Salinity Profiles and Subsurface Velocity Observations From a Global Array of Profiling Floats. Front. Mar. Sci. 7:700. doi: 10.3389/fmars.2020.00700