Initiating the Global Data Assembly Centres for Marine Biogeochemistry

Benjamin Pfeil, Bjerknes Climate Data Centre, University of Bergen and Bjerknes Centre for Climate Research, Norway, Benjamin.Pfeil@uib.no
Maciej Telszewski, International Ocean Carbon Coordination Project of the Scientific Committee on Oceanic Research (SCOR), and Intergovernmental Oceanographic Commission (IOC) of UNESCO, Institute of Oceanology of Polish Academy of Sciences, m.telszewski@ioccp.org
Kevin O'Brien, University of Washington/JISAO, USA, kob@uw.edu
Toste Tanhua, Geomar, Helmholtz Centre for Ocean Research, ttanhua@geomar.de
Are Olsen, University of Bergen and Bjerknes Centre for Climate Research, are.olsen@uib.no
Eugene Burger, NOAA, USA, Eugene.Burger@noaa.gov
Alex Kozyr, Carbon Dioxide Information Analysis Center, Environmental Sciences Division Oak Ridge National Laboratory, kozyra@ornl.gov

For the past decades the international ocean biogeochemistry community has mainly used and depended upon one global data center, the Carbon Dioxide Information Analysis Center ocean trace gases section (CDIAC-Oceans) at the U.S. Department of Energy's Oak Ridge National Laboratory, USA. CDIAC-Oceans provides data management support for ocean carbon measurements from Repeat Section cruises, VOS/SOOP lines, time series and moorings data, has accommodated most community requests for data archival and data access and has also actively engaged with the science community, supporting large synthesis projects like SOCAT, the LDEO Database, GLODAP, CARINA, PACIFICA and GLODAPv2. The funding support for the ocean trace gases section of CDIAC has been endangered several times in the past and puts in jeopardy the uninterrupted data management that the ocean biogeochemical data community has come to rely upon as well as the trust and recognition from the scientific community that CDIAC-Oceans has built though decades of interactions. The loss of CDIAC-Oceans will likely have a negative impact on ocean carbon data submissions and reduction in value added products.

The uncertainty of funding for CDIAC highlights the vulnerability of a system that relies too heavily on individual data managers or institutions. At the same time, it provides an opportunity to review the requirements for modern data access and data management systems that have evolved significantly during the last decades and which currently are not being met through the CDIAC infrastructure. Operational data management systems that (a) provide automated data ingestion, (b) conform to modern standards for data and metadata, (c) utilize standard vocabularies, (d) have easy-to-use data access tools, and (e) provide stable data citations are driven not only by user requirements, but also by funding and government agencies as they promote open access to data. In the discussion of CDIAC funding and the vulnerability of ocean biogeochemistry data, we see a strong opportunity to implement a data management infrastructure that can thrive in the modern world of integrated science data.

A modern data management infrastructure needs to be established in which existing data centers (e.g. CDIAC, CCHDO, BCO-DMO, PANGAEA) and data from various other networks (e.g. OceanSites, Argo) can be integrated through interoperable discovery and access services. This is essential for providing access

to data, while at the same time ensuring that credit for data creation and data synthesis products is appropriately assigned. We propose to mimic the successful data management approach implemented for the Argo profiling float network (http://www.argodatamgt.org). The Argo network addresses national funding agency requirements of having data housed in specific locales by setting up two Global Data Assembly Centers (GDACs), one in the US and one in Europe. Data holdings are mirrored between the data centers and can be accessed through either one. This redundancy makes access to the data collection, by nature, more resilient.

We suggest establishing a system of Global Data Assembly Centers for ocean biogeochemistry (e.g. GDAC-OBGC) where two initial GDACs are established, each with specific roles and responsibilities. The two GDACS will be complementary systems that will leverage the unique capabilities of each, to provide a complete solution for data ingestion, data quality control, data access, data citation and data archival.

A strong focus will be on interoperable access of standards compliant carbonate system data and metadata, irrespective of where they are archived. In addition, it is paramount that support for automated data ingestion, both for real time and delayed mode sources, be integrated into the data management workflow. This is crucial to being able to keep pace with the higher volume of data now being generated by autonomous platforms. First order quality control checks built into the automated ingestion streams can further reduce the quality control burden. By providing interoperable access, and adhering to standards and conventions, this framework will make future data synthesis products and activities much more efficient than with the current non-centralized data management system.

Another important emphasis of the GDAC will be an external review process by ensuring that (a) data are being quality assured and controlled according to community agreed standards, (b) direct feedback is given to the data source, (c) duplicates are being identified and resulting issues are resolved, (d) metadata are complete according to community agreed best practices or existing standards, (e) data and metadata are available through interoperable services, (f) reports are made to IODE and JCOMM Committees on data management status and activities, (g) data citation practices as outlined by the Research Data Alliance (RDA) and DataCite are incorporated, (h) data requests and searches from users can be reproduced and (i) there is clear tracking of the complete data lifecycle for each dataset. The last three bullet points are often overlooked but are increasingly becoming more important to ensure that PIs get credit for data they create and that users/reviewers can reproduce the exact data requests for data that is referenced in scientific publications.

The implementation of the above framework will facilitate continuation of the data synthesis and assessment products such as GLODAP, SOCAT and create a foundation for additional data products, including the integration of data such as time series data and coastal data. In addition, the implementation of such a framework will support compatible efforts internationally, providing a cohesive process toward more uniform data management strategies within the ocean biogeochemistry community. In the long term, such efforts will provide a significant cost savings by reducing data management overhead as well reducing the data management burden on individual scientists.