Oxygen data platform scoping Workshop
11-12 November 2019, Sopot, Poland

Preamble

The overarching goal of this initial 2 days scoping workshop is to develop a roadmap among the community interested in the issue of ocean deoxygenation towards an **open access oxygen data platform for the world ocean**. By this, it is meant a quality controlled (data quality flags assigned based on consensus reached by data contributors and users) data synthesis product, with underlying raw data available in one place or if impossible then distributed but available, with metadata clearly defined and available for each data and with a DOI assigned to each data set.

The oxygen data synthesis product should include ultimately all eulerian and lagrangian observations, i.e. Winkler titrations measurements, sensors data on CTDs and on fixed moorings/time series, sensors on BGC-ARGO floats and on gliders/wavegliders and on any remote vehicle/platform. A strategy could be to include first additional eulerian (sensors on CTD data) then tackle the lagrangian oxygen data. The philosophy to be followed should be inspired from the SOCAT initiative, a community-driven effort for the community users.

In an attempt to prioritize potential steps towards the creation of such an oxygen data synthesis product, several levels of improvements to current oxygen data management could be proposed.

First level improvements would be to gather data from the existing databases in which data is readily freely accessible in electronic format, without restriction, to remove duplicate data and to define consistent quality checks agreed by the community.

Second level improvements are to identify and correct information from additional datasets for OMZs, coastal hypoxic sites and other ocean depths and regions based on experts' recommendation.

Integration with the Argo, OceanGliders and coastal regional community is targeted through the efforts of GO_2NE, WESTPAC, VOICE, EBUS SCOR WG, EMODnet Chemistry, NOAA, etc.

Organization committee:

Denise Breitburg, Smithsonian Research Center, GO_2NE Co-Chair
Hernan Garcia, NOAA, World Ocean Database
Veronique Garçon, CNRS-Legos, IOCCP focal point for Oxygen
Marilaure Grégoire, University of Liege, GO_2NE Co-Chair
Andreas Oschlies, Geomar, SFB 754 Executive Board
Maciej Telszewski, IOCCP
Kirsten Isensee, IOC and GO_2NE Secretariat

Agenda

Day One: Monday November 11, 2019

Morning (Chair Marilaure Grégoire)

9:00-9:20 Welcome, introduction and local logistics (Kirsten Isensee (IOC UNESCO), Maciej Telszewski (IOCCP))
9:20-9:30 Scoping workshop objectives and expected outcomes (Véronique Garçon (CNRS))
9:30-9:40 Scoping workshop objectives and expected outcomes (cont'd) (Marilaure Grégoire (Univ. Liège)) & Denise Breitburg

The SOCAT approach and spirit

9:40-10:10 Benjamin Pfeil (University of Bergen) or /and Kevin O'Brien (NOAA)

What does exist presently in terms of open access data, data portals, data products and related quality controls of raw oxygen data?

10:10-10:40 Hernan Garcia (NOAA) WOD 2018 presentation
10:40-11:10 Toste Tanhua (GLODAP co-chair) GLODAPv2 presentation
11:10-11:30 Health break
11:30-12:00 Sunke Schmidtko (GEOMAR): Oxygen data atlas presentation
12:00-12:30 Catherine Schmechtig (INSU/CNRS) BGC-ARGO profiling float oxygen data set Coriolis data center
12:30-13:30 Lunch

Afternoon (Chair Véronique Garçon)

13:30-14:00 Alessandra Giorgetti (OGS- NODC) EMODNETdata base, oxygen data quality control
14:00-14:30 Bob Diaz (Virginia Institute) The oxygen coastal data base
14:30-15:00 Jacob Carstensen (Aarhus University) The Baltic sea oxygen data base. Integration and data quality control procedure

15:00-15:30	Henri Bittig (IOW), Oxygen optode Sensors: calibration, recommendations from the SCOR-142 WG
15:30-16:00	Emilio Garcia Robledo (Cadiz University), oxygen in OMZs. (How) can we correct?
16:00-16:20	*Health break*

16:20- 18:30 Groups splitting and identification of common questions:

- O_2 data provided by sensors on CTD,
- O_2 data provided by fixed moorings/time series,
- O_2 data provided by BGC–Argo, gliders, wavegliders, remote vehicles/platforms
- O_2 measured in estuaries and the coastal ocean

Each subgroup will work in parallel but participants should be free to rotate between groups. Each group should agree on a rapporteur reporting back on the second day of the workshop.
Based on each sensor/technique used for sampling/measuring, each group should discuss the following questions/topics:

1. How to implement standard and uniformed automatic quality checks on the whole data set and subjective check?

2. Definition of specific quality control/correction based on the technique/sensors used for sampling and what kind of metadata are needed to conduct this.
3. How to implement the recommended quality control and to build the reference oxygen data set?
4. Definition/update of existing quality flags.

Breakout Tuesday November 12, 2019
Morning (Chair Hernan Garcia)

9:00-9:20	Short report back from afternoon discussion
9:20-12:00	Continued discussion, recommendations elaboration for each subgroup : *Session 1, Session 2, Session 3, and Session 4.*

Session 1: O_2 sensors on CTD and calibration with Winkler measurements
Session 2: O_2 data provided by fixed moorings/time series
Session 3: BGC–Argo- O_2 with *in situ* air calibrations and gliders/wavegliders/remote vehicles/platforms O_2 data
Session 4: O_2 measured in estuaries and the coastal ocean

Special attention should be paid to very low oxygen environments which require proper validation for O_2 Winkler titrations as well as for O_2 sensors.

12:00-13:00 Lunch break

Afternoon (Chair Maciej Telszewski)

13:00-15:00	Reports from each session
15:00-17:00	General discussion, elaboration of commonly agreed strategy to go forward, dream team on each task and future agenda?

17:00 Closure

List of participants:

Alessandra Giorgetti	EMODNET
Artur Palacz	IOCCP
Benjamin Pfeil	University of Bergen
Robert Diaz	Virginia institute of Marine Sciences
Catherine Schmechtig	CNRS
Daniel Conley	Lund University
Emilio Garcia Robledo	Cadiz University
Fei Chai	University of Maine/SIO
Hela Mehrtens	GEOMAR
Henry Bittig	IOW
Hernan Garcia	NOAA
Hiroshi Uchida	Jamstec
Jacob Carstensen	Aarhus University
Kim Currie	NIWA
Kirsten Isensee	IOC-UNESCO
Maciej Telszewski	IOCCP
Marilaure Gregoire	University of Liege
Masao Ishii	JMA-MRI
Shigeki Hosoda	JAMSTEC
Siv Lauvset	University of Bergen
Sunke Schmidtko	GEOMAR
Toste Tanhua	GEOMAR
Veronique Garçon	CNRS-LEGOS
Andreas Oschlies	GEOMAR (remotely)
Ivonne Montes	IGP (remotely)

Minutes of the Oxygen data platform scoping Workshop

Day One: Monday November 11, 2019

Morning (Chair Marilaure Grégoire)

A warm welcome and details of local logistics were first provided by our host in Sopot : **Maciej Telszewski** from IOCCP, and the GO_2NE project officer **Kirsten Isensee** from IOC UNESCO gave the audience a short introduction of the workshop motivations. A quick tour de table took place with 2 scientists (A. Oschlies and I. Montes) being connected remotely.

Then **Véronique Garçon** from CNRS developed the Scoping workshop objectives and expected outcomes. Although this meeting was delayed due to the US shutdown in fall 2018, it is finally taking place thanks to the funding and staff efforts from IOCCP, IOC, NOAA, SFB754, and GO_2NE. The rationale is the robust evidence of oxygen loss in the open and coastal ocean. This decrease is non uniform, it varies within basins and with depth. The ensemble of IPCC models simulates over the last 50 years a declining trend of a 0.05 µmol kg^{-1} yr^{-1} whereas the oxygen observations indicate a trend of 0.06 µmol $kg^{-1}yr^{-1}$, quite close but with a very different latitudinal distribution in the tropical thermocline.
There is thus an urgent need to define a roadmap towards the development of an open access oxygen data platform for the world ocean, quality controlled, together with data synthesis products, metadata clearly defined and DOIs, and also to focus on one unit for oxygen concentration.

Then **Marilaure Grégoire** from University of Liège and **Denise Breitburg** from the Smithsonian Institution , GO_2NE co-chairs, further elaborated the rationale. They highlighted the fact that there is not one single entry for downloading oxygen data, there exist different units for oxygen, different quality control checks, and different vertical coordinates, so methodology should be better defined as well as a rigorous protocol on how to get rid of duplicates in the data bases. Presently, no gridded product from lagrangian platforms (e.g. Argo) exists so the merging of platforms would be useful (see H. Bittig's talk). The political issue arises on the delimitation of coastal versus open domains. Oxygen Minimum Zones (OMZs) require particular specifications. Is it possible to correct back? Accurate measurements are paramount for the N budget in these regions.

Benjamin Pfeil fom University of Bergen, with contribution from Kevin O'Brien (NOAA), presented the SOCAT approach and spirit. He recalled it was a community effort which started in 2007. The challenges were the following: data were organized differently depending on originators, many data were not even published, and there was no agreement on quality control. SOCAT is now delivering the 7th version, accessible via different platforms. A major effort on data adjustments has been performed with file and units being uniformed, expocodes assigned, metadata collected and organized, primary quality controls with WOCE flags performed, secondary quality controls with cruise flags, and data sets following the standard operating procedures (SOPs).
In SOCAT, there are both regional groups and a global group. Automation is crucial for carrying regular updates and for better usability, and importantly more data can be dealt with. Some automation slides produced by O'Brien can be looked at XXXX. The policy for DOI delivery is as such:
the entire collection receives a DOI, each individual quality controlled data file receives a DOI, as well as the gridded product, and the data archived. Benefits of automation are obvious: leverage standards,

framework flexible, this could be easily adopted for oxygen as well. The Surface Ocean CO2 Reference Network, SOCONET, was also presented. SOCAT still suffers from a lack of sustained funding.
Questions were asked whether we can have lessons learned from the carbon community so that we avoid the same mistakes when working towards the oxygen products. Benjamin answered that SOCAT would be more than happy to provide a synthesized overview of issues worth considering when approaching data synthesis holistically. The need for a proper business plan with strong emphasis on the economic value of impact of the products was expressed and agreed upon, suggesting to work with economists to evaluate the impact. Oxygen actually might be easier than carbon for this task since there is a lot of impacts in the coastal zone.

Hernan Garcia from NOAA presented all data being used in the World Ocean Data (WOD) base, there are different sources with 86% international sources, and 14 % being US. There is a continuous feedback between the WOD and WOA (World Ocean Atlas). They all are stored in the NOAA NCEI (National Center for Environmental Information). The WOA is a well cited product. There is a strict quality control/quality assurance cycle with common format, metadata, standard, units, and data integrity checks. The routine WOD quality control tests result in WOD flags, and the correction procedures are context specific with different methodologies for different depths.
Comparison with the GLODAP product shows that below 500 m, the difference is smaller than at higher depths between both products. Are the large discrepancies in the upper ocean (between 5 and 10 µmol kg^{-1}) due to the seasonal cycle? Difficulties in merging data from different resources (e.g. Argo and Winkler titration measurements) are highlighted. Hernan offers his view point for a merging approach : first dealing with Winkler titration measurements, then CTD/O_2 sensors data, then float and glider O_2 sensors data, and finally O_2 sensor onto moorings, and emergent sensors.
The whole ocean is covered since 1900 and the WOD contains 1.8 million of oxygen profiles. In addition to chemical variables, ancillary data are available. The base is centrally managed and needs sustainable resources. It benefits from a collaborative international effort (free and fair). The desired data quality is 1 µmol kg^{-1} in terms of reproducibility but the actual threshold is 3 µmol kg^{-1}. Quality tests are being carried out with objective metrics but also with subjective quality control, statistics and uncertainties are being provided. Variable specific quality control flags (it requires essential ocean variable -EOV- expertise, and to follow recommendation from IOC best practices for quality flag scheme QF) are in use. There is a tailored data access tool to select and to export data in the needed format. All data products have assigned DOIs. One highly desired product is a global and coastal oxygen content curve as a function of depth range. Hernan proposed to use the WOD as the cornerstone for building our global oxygen data platform. It was asked how many real coastal sites were present in the WOD. Everything that exists is in the WOD but the 1° resolution is not reflecting the real data, but it is possible to increase resolution down to 0.1°.

Toste Tanhua from GEOMAR (GLODAP co-chair) presented the core Global Ocean Observing System data product version 2 (GLODAPv2) for biogeochemistry. It has been built so it is internally consistent and is moving into annual updates status. The primary quality controls are similar to those for the WOD (outliers, obvious errors, precision) with a published tool box. The secondary quality controls refer to accuracy, systematic bias, decisions on adjustments, and cross over analysis. For adjustments, many potential sources of uncertainty are complicating an otherwise straightforward assessment of cruise biases, such as : 1) temporal variability and long-time trends in parameter values on a particular location in the ocean, and 2) drifting or variable measurements imprecision and inaccuracy over the duration of a cruise. There exists a reference group for the different basins. The data suffer from a strong seasonal bias towards summer time. It is important to also consider the auxillary data, not only oxygen. In GLODAP oxygen data are also interpolated, mainly in the open ocean. It was asked if there are multiple submissions to different repositories and how to deal with these duplicates. Ideally one should submit only to one portal and there should be interoperability between portals. The question on how to go consistently about adjustments was raised. How adjustments influence the usability of the final product? Adjustments are based on deep data, but need to evaluate of course 'normal change'.

Sunke Schmidtko from GEOMAR gave a presentation of the oxygen data atlas he built for his Nature paper in 2017. He used many different data bases which has huge data overlap: Pangaea, Argo, Hydrobase3, NOAA WOD, CCHDO,.. There are 3-5 billion data points. He chose isopycnal mapping to compare data which are spatially apart and used a front finding algorithm. He did not change the way of mapping since the 90s, but optimal interpolation by downweighting of historical data could reduce the impact of old oxygen data. Absolute accuracy versus relative accuracy is important to be included. Fast marching algorithms may separate water masses such as shelf boundary currents from shelf interior waters. For trends and anomalies more effort is needed. Sufficient good data (more than 70%) can take care of false data (appropriate algorithm required). Fisheries agencies have very rich data sets. Good/modern statistics are needed. Even sparse data can be mapped with fast marching, front finding algorithms. Repository and products (which takes into account duplicates, errors, quality controls checks) should be both archived. Sunke advised not to trust any climatology since it only gives a broad idea of the oxygen field. It was asked whether we should build a premodern baseline, a specific mapping of historic data.

Catherine Schmechtig from INSU/CNRS gave an overview of the BGC-Argo profiling float oxygen data set in place at the Coriolis data center in Brest, France. At the GDAC, one can find all Core argo data files as well as the bgc-Argo data files. In the metadatafile, all specification on the sensors, parameters characteristics, factory calibration and calibration equation before the Argo float deployment are given. An Argo cookbook and quality control manual do exist. For each parameter 5 variables are provided: raw data, quality controlled data, adjusted data, adjusted quality controlled data, and the adjustment error. Data are available in 3 modes : R : real time quality controlled and flagged using an automated procedure, A : adjusted in real time with an automatic procedure, and D: delayed mode data which require the control and validation by a scientific expert. For scientific calibration, reports on post deployment calibration and adjustment information are available on the website. For the R mode, real time quality controlled, various tests are made: a global range test, a spike test which will be improved soon by a new spike test together with a biofouling test and a regional range test. Different methods exist for the *deoxy* variable adjustment. The Argo community is still working on near real time quality control, and testing a new definition for ongoing quality control since oxygen is important for nitrate and pH.

Alessandra Giorgetti from OGS- NODC presented the EMODNETdata base along with the oxygen data quality control. The base holds now more than 5 million data sets, with a focus on eutrophication, contaminants, and marine litter studies. It gives access to aggregated harmonized data sets building on seadatanet. It contains more than 600 000 oxygen data sets, different data platforms and many metadata sets. A quality control loop exists with a first quality control step on checking format and metadata, duplicates, and a second step on data aggregation with unit conversion (including depth pressure), high frequency data. Which kind of quality is being used at EMODNET?
They maintain a close collaboration with environmental users and customize metadata, internal validation like so that the quality of data sets is improved. It was asked whether NODC supplies real data? 85% are directly available, but some need specific permission for specific regions. The Black Sea

is a bit more difficult, metadata are fully available, there are more coastal data than open ocean data, they are processed within EMODNET. The EMODNET data base is not part of the WOD. It was mentioned the issue of versioning and stressed that it is important to know where are the 'real' data sets. It would be interesting to have automated flushing between data centres and to align the different quality controls between NODCs.

Bob Diaz from the Virginia Institute presented his oxygen coastal data base, which goes back to 1995 where 44 sites were monitored at that time. Now the base includes up to 660 hypoxic sites. The availability of information is medium, meaning through the World Research Institute (WRI) and google ocean. There are no raw data per se but information on the impacts on fisheries and benthic organisms. To complement the data base, one should include as well area affected by hypoxia, the data on oxygen concentrations, causes, trends, links to cited references, integration with other global data bases, e.g. HABs, human footprint; in other words linking qualitative and quantitative information. DOIs is a good way to give proper credit to data providers and following the SOCAT data use policy should be recommended.

Jacob Carstensen from Aarhus University gave an overview of the Baltic sea oxygen data base together with integration and data quality control procedure. He first presented who are the end users of the data base. The Baltic sea data is rich : many long term data sets, 120 years of oxygen data, it shows a 10 fold expansion, and regular monitoring cruises. The data sets cover the whole Baltic Sea with in total 30 000 profiles. The Baltic sea has a difficult topography, and is very sensitive to eutrophication. The Baltic environmental database (BED) is jointly managed by DK (AU), DE (IOW), FI (SYKE), SE (SMHI) which are hooked up and data portal directs to national data bases. For other countries there is not a direct link so data are stored in Stockholm. The data quality assurance is done by the data provider using standards' procedures. There exist different levels of quality control, the last scientific check being done within the framework of the annual reporting. Characteristic features are : seasonal hypoxia, regional basis, bathymetry (not deep) has an impact, and wind events make a difference. Oxygen data are associated to salinity, temperature, nutrients and chlorophyll *a,* the quality control is based on national procedures, using the data is the best means to ensure a high quality. If data are not being used the monitoring will not be sustained. There is globally a problem of data sharing. The BED data base was developed for the Baltic Sea action plan to set nutrient targets to be delivered to the system. Industrial data are not included, and neither fish assessments but it may be worth to consider. There is a strong demand from policy makers who want to know the BS evolution. Up to now, there is no early warning system.

Henri Bittig from IOW presented the calibration recommendations of oxygen optode sensors from the SCOR-142 WG. He explained the different measurement options for oxygen in seawater, and the different time responses of sensors. All sensor data is out of calibration unless adjusted with *in situ* reference, storage drift and *in situ* drift. Sensors need a good multipoint calibration. One may refer to Bittig et al. (2018) for the best choice of sensors according to purpose of use. Understanding the characteristics of each sensor is key for sensor calibration, and it is important to have appropriate units (see SCOR WG 142 recommendations on conversions). It is crucial to make adjustments for consistency where needed and to make them in a transparent way. It is paramount to make easily transparent and identifiable where data come from for every single data point. Concerning Argo adjustments, they are automated with only a gain factor correction in BGC-Argo. The calibration in air is recent so it is delicate to decide what to do with the older sets, probably surface saturation and nearby profiles could help. It was asked whether an error factor can be provided for each data point, and this work is in progress.

Emilio Garcia Robledo from Cadiz University provided some insights on corrections for ultra low concentrations of oxygen in the OMZs. The focus is on the very low oxygen zones, there is some relevance of trace oxygen concentrations since under anoxic conditions anaerobic processes thrive.
Denitrification and anammox processes are inhibited at oxygen concentrations larger than 1 µmol l^{-1}, nitrification is highly efficient at nanomolar oxygen concentrations and maximum rates are reached at around 2 µmol l^{-1}. Different methodologies exist:
1) Winkler titrations: the precision is around 0.5 µmol l^{-1} but sampling is introducing oxygen. As oxygen is dissolved in polymers, water collection in Niskin bottles contaminates the sample by 0.5-1 µmol l^{-1}, but also from Niskin to glass bottle within 1 min 2.5-3.9 µmol l^{-1}, in anoxic waters it can be 5-7 µmol being introduced in 2 minutes. Winkler titrations are not appropriate for 15-20 µmol l^{-1} calibration (error 10%),
2) Seabird CTD SBE 43 : high temporal and spatial resolution, the unit is made of metal, but water flows to the sensor through polymer tubing, different profiles are usually acquired down and up,
3) STOX sensor : it allows an internal zero calibration, it is wise to run parallel CTD and STOX and use STOX for for CTD calibration,
4) Profiling buoys : some floats park at 1000 m within low oxygen conditions, there is no polymers so the oxygen contamination is reduced.
The take home messages are the following : 1) Winkler not reliable below 10-15 µmol l^{-1}, not to be used for calibration below 15-20 µmol l^{-1}, CTD does not reach anoxia but can be corrected with STOX or reference sites, 2) linear corrections to be done for offset, and 3) probably most Winkler measurements are overestimates.

Afternoon (Chair Véronique Garçon)

Groups splitting and identification of common questions:

- O_2 data provided by sensors on CTD,
- O_2 data provided by fixed moorings/time series,
- O_2 data provided by BGC–Argo, gliders, wavegliders, remote vehicles/platforms
- O_2 measured in estuaries and the coastal ocean

After a quick discussion we reorganized the groups discussion and decided it would be best to remain as a whole group to tackle the list of questions below :

How to implement standard and uniformed automatic quality checks on the whole data set and subjective check?
Definition of specific quality control/correction based on the technique/sensors used for sampling and what kind of metadata are needed to conduct this,
How to implement the recommended quality control and to build the reference oxygen data set?
Definition/update of existing quality flags,
Special attention should be paid to very low oxygen environments which require proper validation for O_2 Winkler titrations as well as for O_2 sensors.

What do we want a data portal for?

It was encouraged at the very start to “forget” about our own data set but rather focus on the way forward for the global oxygen data portal. What is the main objective? Scientific questions of course might differ (validation of models, deoxygenation occurrence and trends, net community production estimates, etc..) so there is no need to specify upfront what is the uncertainty. In the case of SOCAT for instance, it is not always possible to prescribe accuracy since some data providers may be excluded and some data may be used for specific applications. Some thought we should state upfront the desired applications since this will assign the required accuracy. Then it was mentioned that there exist several categories of stakeholders, those interested by the coastal and regional ocean and those focusing on the global ocean. For the coastal ocean one issue is resolution of the products wich is critical to detect impacts. In coastal regions high variability does not allow for interpolation. Metadata requirements are different for the coastal and open ocean. So ideally one should aim towards separate data bases for the global and coastal ocean using the same principles and interoperability will take care of availability of all data. The amount of data along the coasts is humongous. There is probably no way to dream about a global data base for coastal oxygen data. Rather it might be wiser to form regional hubs for coastal data since in regions people have similar interests and therefore are keener to provide the efforts (data, metadata development, co-funding etc.). All data does not need to sit in one place but data needs to be accessible and to have uniform metadata, sets of standards and operating procedures (SOPs) and agreed uncertainty/quality control estimates assigned to data points.

For the coastal ocean, it was stressed that data archeology might be really difficult providing that often there are numerous detached data points that cannot be connected to any longer-term trend or process or any specific element of variability. We will see all sorts of trends going into all directions. The key aspect will be to develop filtering mechanism to sieve out not useful data from data that can be used meaningfully. The efficient and wise way to go for coastal ocean data bases is to work jointly with National Oceanic Data Centers (NODCs) as this is probably the only existing semi-organized network of regional data collection efforts.

For the open ocean, we need to look how to harmonize data from different platforms such as GOA-ON hubs, GOOS GRA's, or NODCs. We also need to look a bit wider outside oxygen and connect to other oceanic parameters such as temperature and pH. Mention is made that much more coastal data exist than we can even start to imagine. For example, all aquaculture companies in Norway are obliged to monitor oxygen but these data is not submitted to any data center. One should be aware that metadata portals are sometimes blockers to the development of data sets as they make an impression that data is available in an organized form but in fact it is only metadata.

Discussion groups should focus on **(i) accuracy, quality, protocols, and on (ii) what sort of data base we need in open ocean and coastal systems, (iii) how to integrate data from various platforms, and (iv) What metadata is needed**. With the coastal zone there are issues with national interests so it is crucial to start building trust in regions by showing the positive impact of data sharing and showing examples of countries that offer oxygen data freely and hope that others will follow the suite. One way forward could be to start with discussing what is missing from the most developed places and how can we move forward with that by implementing lessons learned, protocols from others.

Breakout groups could be : *Protocols, SOPs for ocean oxygen data management for open ocean and coastal ocean, SWOT analysis of WOD: what is missing and what is wanted (delayed mode and original raw data sets)? What is done with oxygen data? Accuracy and quality flags separate therefore importance of metadata, how to integrate data from various platforms in one product (data archeology and seamless integration of data coming from various observing systems)? how to give credit with data providers?*

Day Two: Tuesday November 12, 2019

Morning (Chair Hernan Garcia)

The chair made a short report from the previous day afternoon discussion and invited to develop further the discussion along the topics which were approved.

1) What does a new data set need to include as metadata (open and coastal ocean)?

The list of all key information in a machine readable form to be given as metadata should include: Method (winkler or sensor (model serial number)), Nation PI, Full resolution mode or downsampled version, Period, How many profiles, Does the profile goes down to the bottom, Indication of the tidal stage for coastal sites, Sensor identification, Calibration equation and calibration method reference, post-deployment calibration, Accuracy, precision, Reference to best practice, Postmodification - adjustment afterwards, Uncertainty, precision, reference material, DOIs, data adjustment, Auxillary data information, Detailed Sensor information, Version information, Secondary quality control conducted already, Software used for calibration and for analyzing the data sets, Collecting organization - analytical laboratory (especially in coastal areas), Platform information, Level of Quality Controls - automated, live, Which quality control flags scheme was used, Quality control recommendation of best practices to follow, Manufacturer of reagents, How the sensor is mounted on the platform in air, Upcast or downcast, Storage conditions of sensors prior to deployment, Reference to best practice document, Data version.
It was recommended to compare with Goa-on metadata file, Argo and BGC Argo metadata file, to share vocabularies with EMODNET, seadata net and climate forecasts, and to avoid free text.

2) What are the existent Quality Controls?

The First level of QC is automated but maybe it would be wise to limit automated QC. Is the measurement within the observable range, spike test, check for units**,** perform range tests in the region based on historical data to provide a smart check assigning a flag. Quality of T, S and P should be also taken into account, if they are provided. A new tool was developed within Atlantos for primary control that could be used. Checks should be made on density and depth inversion, and on excessive gradients. Standard Operating Procedures (SOPs) should be followed for flagging. The Argo cookbook is a nice reference to be inspired from. Individual flags for the different checks on/ off – fail/ not fail should be implemented. QCs need to be platform specific but as a start Argo QCs can be largely adopted by the Glider community for their QCs. Biofouling flagging which is extremely important for the coastal ocean should be done at this stage.
The Second level of QC should detect outliers and the Third level should involve Glodap QC – the text book - but there is some concern about the deep ocean. How to take into account the warming trend? Glodap methods for specific measurements should thus be examined with great care. It would be beneficial to standardize BGC Argo QC (and also include biofouling check). Get inspired by the ICOS data life cycle flow chart for oxygen data.

Refinements of what is a secondary QC were provided: assessment of systematic bias (drift, time response), keep temperature and salinity. It should also be specific to platforms. The Argo automated adjustment protocols can be consulted (see Thierry et al., 2018, Argo quality control manual for dissolved oxygen concentration at https://archimer.ifremer.fr/doc/00354/46542/). Argo is preparing a decision tree impact of S and T on dissolved oxygen. For Glider QCs, one can refer to Argo as a starting point, always the possibility of post calibration and cross reference to Argo but there is a certain need to establish a proper post-calibration. Wavegliders should be treated more like moorings (surface automonous vehicles, e.g. saildrones). For moorings, there is no specific document, especially for coastal moorings, a recommendation is made to consult at IOOS (real time data, integrated ocean observing system) or the FIXO3, Jerico Next EU projects. HOT and BATS moorings protocols documents can be consulted as well as the Monterey Bay MBARI (coastal ocean) ones. For CTD sensors, on GO-Ship cruises it is important to state how far above the bottom to stop. For discrete sampling with Winkler on Niskin bottles, the GO-Ship procedure for the open ocean maybe not that good for the coastal cruises, please refer to the International Quality controlled Ocean Data base IQUOD. In the Baltic sea, the procedure is bottom dripping a Niskin bottle to measure bottom. For surface underway instruments or ferry boxes, there is no known manual. For sea bottom observatories the question was raised of the existence of proper recommendations for oxygen measurements. There should be specific work to be done on the coastal measurements.

3) What does a data center need to serve?

This session tried to list what are the prime functions a data center needs to serve. The following list is far from being exhaustive: Best estimate for measured oxygen concentration value together with raw data, Uncertainty estimate, Results of the QC coming from the data provider, e.g. in case of sensor failure, Auxillary information, Data and metadata, Searchable for metadata, data and best value, Standardized format to download and auxillary data, Facilitate meta and data submissions, Services to data provider, cross updating, DOIs assignments and updates, Outputs tailored in one format, Storage for eternity and being available, Labelling of stations if they are part of a bigger sampling scheme, Feedback function, User statistics, capacity of downloading statistics, recognition of PIs and give credit to funding agencies, credit throughout the processs: important for data provider and funding for data centre, Download it in something not under matlab format, No limitation of size download, interoperability between webservices, Help desk, Release dates of data versions, Clear statement on methodology, Desire to have community blessing, statement on how you can use it, data use license, Citation guidance, For modeler needs: Gridded data product with interpolation as well as fixed lines, repeat hydrography data, and time series data.

4) Which data products?

Visualisation products should be detached from data centres. Raw data should not contain any products, they can be linked but not necessary be part of data centres. Annual release of gridded products (and without interpolation such as SOCAT like) should be ensured. What kind of products should be feasible with the data centre services provided: 5 products for the different measurement and one huge integrated product so we can define the requirements on quality control and meta data taken into account while working at the same time on integration between the different products. Do we provide a product only for moorings or we include them with Argo? So rather start with 3 products and short term goals, a climatology for Winkler and CTD-O2 measurements then adopt a staggered approach, later including float sensors, glider oxygen sensors, and moorings. What would be the ideal approach for the coastal ocean: Merging of gliders, CTD and Winkler oxygen measurement?

5) Scientific outcomes and services

A product is a community reference integrated internally consistent database. For the coastal ocean, the need is on high resolution product in special areas with huge coverage fit for purpose. These include: deoxygenation trend – gridded product plus map of error, objective mapping, net production, hypoxic/area/severity – interpolated data, effects of oxygen decline on marine life – habitat compression costs of low oxygen. For the open ocean, outcomes can be: deoxygenation trend area/volume of low oxygen waters – integrated internally consistent database in highly vulnerable areas, oxycline variability, ocean oxygen content graph, and regional budget assessments. It is of paramount importance to come up with a business plan where we state the economic value of oxygen decrease (Euroseas project). Other ecosystem services include : model validation for improved predictions (e.g. for aquaculture) – products needed near real time observation (QC), air sea interface (underway systems), and bottom water fluxes to document the interstitial oxygen content.

6) What data sets do we need to consider?

Task teams for the different measured data will be defined (see section 7).

7) Roadmap products of the workshop

News item will be posted at IOCCP and IOC websites, along with an item in the IOC newsletter.

All activities mentioned here are leading to products and outcomes, this should be achieved to be in 3-5 years. Winkler measurements data base maybe in a year and the others to follow up. For each task team, metadata are required and desired.

Tentative schedule is as follows: Merging Hernan Garcia's white paper with minutes from Sopot meeting by July 2020, in a year from now, another group meeting is to be planned. Kirsten and Maciej will help with coordination of tasks, 6-months to a year workplans to be achieved. It is crucial to communicate with regional stakeholders and scientists: Gil Jacinto, Wajih Naqvi, and Minhan Dai. Ocean Science meeting could be an avenue to meet. Minutes (handled by Veronique) and shortened version of Hernan's manuscript (handled by Marilaure) to be sent in February to the group and put in a google doc for further editings.

6 workplans that need to be reflected in the white paper:

1) CTD O_2 integrating - Hernan, Veronique, Emilio, Marilaure
2) Winkler O_2 – Sunke, Emilio, Hernan, Jacob, Bob
3) Argo – Henry, Catherine, Siv, Fei, Xing, Virginie?, Ivonne
4) Glider – Henry, Rudnick? Ask him? Fei to contact him
5) Moorings – Henry, Hiroshi, Bjorn to be asked, Francisco?, Nico Lange, Denise Breitburg (her laboratory), Nancy Rabalais?
6) Underway – Toste, Siv, orway is doing something along those lines, maybe somebody from Ifremer? Loic Petit DL, Copernicus IMR

Structure of data centre interoperability taskteam: Benjamin, Kirsten, Hernan, Alessandra?
SCOR WG proposal in April 2020 to be submitted?