OBIS Data Quality Control Project Team report

GitHub: https://github.com/iobis/quality-taskteam
OceanExpert: https://oceanexpert.org/group/441

Terms of Reference of QCPT

Motivation and scope

Data quality is important for any downstream users of the biodiversity data in the OBIS network. It can have significant impacts on the ease with which data from different sources can be re-used and the conclusions that can be drawn. Thus, ensuring the best quality data is a major concern for any OBIS node. To that aim, this task team was set up to further develop a common framework for the assessment and management of the quality of data submitted to OBIS.

The scope of this task team includes:

- Quality control (QC) criteria
- Align QC criteria with other data stakeholders/standards including:
    - IODE (use of IODE primary flags)
    - GBIF
    - TDWG (Darwin Core)
- Presentation of QC results to user (flags, interpreted values)

Goals and outcomes

The task team aims to:

- Establish a set of common QC criteria for OBIS data
- Provide solutions/recommendations to data quality related issues

The measures of success include:

- Reduced data quality flags across OBIS nodes
- Resolved data quality related issues in relevant OBIS GitHub repositories

Accomplishments

The project team carried out a number of actions in order to accomplish the aforementioned goals, implementing a collaboration framework for Quality Control Best Practices within OBIS. These actions are expanded on in the OBIS Steering Group reports in the Appendix of this document and can be summarized in the following sections:

Implementation of a QC-dedicated collaboration framework

The QCPT began the implementation of a community of practice in summer of 2021 under the enthusiastic coordination of Yi-Ming Gan, from AntOBIS, as chair of the project team with:

- The creation of a project team dedicated slack channel to ease communication
- The scheduling of online monthly meetings
- Setting up a dedicated github repository.

These three measures boosted collaboration between nodes but also between the project team members and the OBIS secretariat in the field of data quality control management.

The slack channel was used to share ad-hoc questions, doubts and specific use-cases about the formatting and needed quality checks of different types of OBIS data. Those questions and their answers would then be posted in the dedicated github repository as issues in order to keep track of their resolution, allowing to assign them to an expert or responsible person when necessary.

The online monthly meetings were used to align the quality procedures of all the OBIS nodes with each other and with the quality checks of the OBIS secretariat. These meetings served as a way to collaboratively solve common problems and to steer the direction and priorities of the project team. The running minutes of those monthly meetings can be accessed via the Appendix of this document.

Monthly presentation of OBIS nodes data quality procedures

From April 2022, the project team initiated node presentations during the online monthly meetings where each node manager would brief the rest of the project team on their data cleaning and transformation process. The goal of this initiative was to foster knowledge exchange, to identify areas that need assistance and to streamline best practices and data quality processes between nodes.

Six OBIS nodes (Yi-Ming Gan - AntOBIS, Abby Benson - OBIS-USA, John Nicholls - OBIS-OPI, Sachit Rajbhandari - OBIS-Australia, Georgia Sarafidou - MedOBIS and Braulio Fernandez - ESP OBIS) presented their data quality procedures and challenges encountered. Although highly appreciated by the project team members, the node presentations had to be discontinued due to lack of time available to invest in their preparation caused by the prioritization of other overwhelming project team tasks.

Coordination of data laundry events

Two data laundry events were organized by the project team with support from the OBIS secretariat: 8-12 November 2021 and 20-22 April 2022. The goal of the data laundry events were for OBIS nodes to resolve the quality issues of datasets in OBIS. Nine OBIS nodes investigated datasets from their nodes in each of these events.

Four sessions of data laundry meetings were organized in each event where node managers discussed data quality issues with the project team. In 2021, actions were taken on more than 20 datasets while 34 datasets were being investigated in 2022. During these events, the task team identified several needs from OBIS nodes: (i) Guidance for data with limited information on OBIS required fields and (ii) improvement of the documentation of the OBIS QC pipelines and other data validation tools within the OBIS network.

These meetings proved to be an excellent forum for knowledge sharing and collaborative problem-solving. Special attention was given to issues related to:

- On-land QC flag
- ScientificNames (unknown taxa, vernacular names, temporary names, confidence of Identification
- Dates (invalid dates, historical dates)
- Depth values in relation to bathymetry

Collaboration with the OBIS Capacity Development and the Communication and Outreach Task Teams

The online monthly meetings of this project team were open to any other OBIS SG members. This facilitated that often members of other OBIS Task Teams would join these meetings to gain insights on the latest decisions, trends and issues faced by the OBIS quality control project team. The participation of the Capacity Development and Communication and Outreach Task Teams in the QCPT monthly meetings proved to be beneficial for the development of training materials and the OBIS manual.

Launch of an OBIS user survey

To close the gap between data users and OBIS, a survey was launched to collect user feedback on accessing and using OBIS data. The survey consisted of 28 questions that would help

improve the quality control currently implemented for datasets in OBIS and to guide the development of quality control measures and protocols including the fitness-for-use labels.

The full results of the survey and all textual responses with some annotations from the secretariat are available here: https://oceanexpert.org/document/32616 . The secretariat performed a first analysis of the free text responses and found that most issues and suggestions can be categorized as:

- Lacking or unclear documentation
- Data quality and completeness issues
- Insufficient quality control
- Data access user experience
- Training requirements

The results of this survey were shown to be beneficial to understand the challenges that the OBIS community is facing in terms of data accessibility and usage.

Launch of an internal OBIS QCPT members survey

As the project team was advancing, a reduction in participation from the OBIS nodes was observed. An internal OBIS survey was launched in order to assess the reasons behind this lack of participation. The results of said survey can be found here and have been summarized in the Appendix of this document.

The survey outcome showed that the lack or reduction of participation in the project team was caused primarily by the following items:

- Lack of dedicated funding / lack of time.
- Goals were too broad
- Low tangible return in the short term
- Meetings were focused on a specific time zone

Collaboration in the development of the GBIF Grand Unified Data Model

The project team did not only look into standardizing the data quality procedures within the OBIS network but also invested into aligning with the global community, actively collaborating with the GBIF and TDWG networks. These two communities have been collecting case studies since 2021 in order to expand the current data model, based on the Darwin Core Archive standard, to tackle its limitations when recording different types of biodiversity in situ detections.

The project team consulted various OBIS nodes as well as GBIF nodes (GBIF Norway) to collect marine data related use cases for this purpose. Four use cases were submitted:

- The Autonomous Reef Monitoring Structures (ARMS) MBON use case provided by EurOBIS,
- Two environmental and community measurements use cases provided by GBIF Norway (featuring the Nansen Legacy project) and AntOBIS and
- Animal tracking data use case from AntOBIS.

Alignment of the OBIS QC pipelines with the TDWG and GBIF quality checks

A joint online meeting between OBIS QCPT, OBIS Historical Data Project Team, OBIS Secretariat, GBIF Secretariat and TDWG Biodiversity Data Quality (BDQ) task group 2 (TG2) was held on 3rd February 2023. The outcomes of this meeting were:

- The alignment of the OBIS quality checks with the core tests and assertions developed by the TDWG Biodiversity Data Quality Tests and Assertions task group (BDQ TG2). The mapping is summarized on a wiki as part of the QCPT GitHub repository.
- The incorporation of a link to the LifeWatch & EMODnet Biology QC tool in a dialog box when users add their dataset to the OBIS network on IPT, available since IPT version 2.7.3.
- A summary of the OBIS and GBIF quality control tools, data flagging approaches and procedures pre- and post- publication with the aim of standardizing the quality control procedures across networks. This summary triggered an open Github discussion on merging the GBIF and OBIS validators:

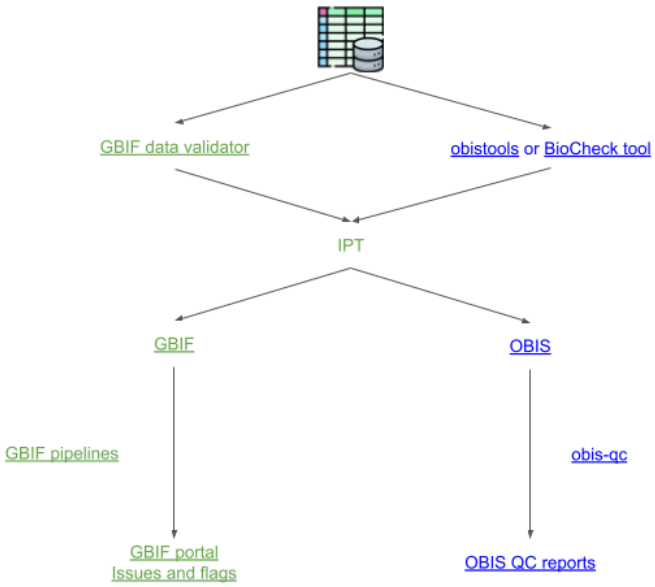


Figure 1. Diagram showing the different data validation tools and procedures in GBIF and OBIS.

Review and integration into OBIS of the marine related datasets from GBIF

Since the release of IPT 2.5.2, GBIF publishers can link their datasets to different networks in IPT. OBIS is one of the networks that can be selected for each IPT resource. GBIF and OBIS recommend that all marine publishers select the OBIS network where appropriate. The OBIS secretariat lists those marine datasets that are not yet in OBIS as issues to this GitHub repository and indicate which OBIS node(s) should endorse this dataset. Once endorsed, OBIS harvests the dataset directly from the source IPT and lists it on the OBIS node page.

By the end of this project team in 2023, 451 datasets were published as “OBIS” datasets within GBIF. The QCPT made sure to prompt the corresponding OBIS nodes to endorse and close these issues. Of those 451 Github issues, 287 were closed and 164 remained open. Several of the open issues deal with taxon checklist datasets, which OBIS cannot handle yet. The closing of the remaining issues should be addressed by the corresponding OBIS node at their earliest convenience.

Products of QCPT

In addition to the increase in quality and quantity in data quality and networking within OBIS and with the broader biodiversity informatics community, the project team accomplished tasks resulted into the creation of two more tangible products. These products can me summarized in:

TDWG abstract

Gan Y-M, Perez Perez R, Provoost P, Benson A, Peralta Brichtova AC, Lawrence E, Nicholls J, Konjarla J, Sarafidou G, Saeedi H, Lear D, Penzlin A, Wambiji N, Appeltans W (2023) Promoting High-Quality Data in OBIS: Insights from the OBIS Data Quality Assessment and Enhancement Project Team . Biodiversity Information Science and Standards 7: e112018. https://doi.org/10.3897/biss.7.112018

OBIS parameters for core tests and assertions

The project team completed the alignment of all obis-qc quality checks to the Core Tests and Assertions developed by TDWG Biodiversity Data Quality Tests and Assertions task group (BDQ TG2). obis-qc is a Python library developed by the OBIS Secretariat that powers the quality checks behind the OBIS portal. The mapping is summarized in the project team wiki although it could be more easily editable via this spreadsheet.

Challenges encountered

During the lifetime of this project team, a number of challenges have been identified and should be taken into account for the successful coordination and completion of future project and task teams within the OBIS network. These challenges can be summarized in:

- Lack of user feedback. The goals of the project team were to develop the fitness for use profiles for OBIS data and to improve the OBIS data quality pipeline. Since we lacked the information of which profiles were needed from the users and how the data was being used, it was difficult to steer the direction and the development. Hence, we carried out a user survey to close this gap.

- Bias in feedback and interactions from OBIS nodes. We noticed that there is a bias in the nodes that are actively participating in the project team activities. The reasons for inactivity of certain nodes were not known. Therefore, an internal OBIS survey was launched to better understand the internal challenges affecting the OBIS nodes' participation in the network activities.

- Technology restrictions in certain countries. We were notified that due to restrictions from certain countries, they could not participate in the events hosted due to the technology used to host the event (e.g., Google).

- Lack of funding for project team members, leading to a reduction in commitment levels of some members.

- Overwhelming tasks in the initial scope of the project, resulting in the team reducing the scope to focus achievable goals such as aligning the QC of obis-qc with TDWG core tests and GBIF pipelines.

- The QCPT did not have the capacity to compare QC from different existing tools to align with the TDWG core tests and GBIF pipelines, which could result in divergence of QC approaches between obis-qc and other tools.

Outstanding tasks

Tasks that are deemed out of current scope were listed in the following GitHub issues: https://github.com/iobis/quality-taskteam/issues?q=label%3Awontfix+is%3Aclosed

Future directions and Recommendations

While the project team addressed most of the tasks set out at the beginning, they were not always fully completed by the end of the project. In order to avoid the challenges encountered during these three years and with the intention of bringing OBIS data and procedures up to best practices, this project team has acquired an overwhelming understanding on the following topics:

- Need to align data quality procedures

    - Within OBIS, the obis-qc pipelines, and the R packages “obistools” and “EMODnetBiocheck” need to be aligned for a standardization of quality control procedures within the network.

    - It is essential that OBIS also aligns its quality procedures with the broader biodiversity informatics communities such as GBIF and TDWG.

- Full time data manager in OBIS Secretariat needed
    - A permanent position for a full time data manager in the OBIS Secretariat is currently lacking but is vital in order to participate in working groups related to OBIS operations such as TDWG BDQ task group 2 in time. This would help to implement and maintain the outcome and vision of this project team.

- Optimisation of OBIS Task and Project Teams functioning needed
    - An assessment of how OBIS Teams are designed and carried out in order to be made sustainable is essential for the correct functioning of these teams.

Acknowledgements

The work of this OBIS project team wouldn't have been possible without the relentless motivation of its members and everyone that participated in its activities and contributed to the valuable discussions, especially the usual suspects below where a large proportion of the works was carried out on a voluntary basis:

Ward Appeltans, Pieter Provoost, Elizabeth Lawrence, Serita Van Der Wal (OBIS Secretariat), Yi-Ming Gan (AntOBIS), Abby Benson (OBIS-USA), Rubén Pérez Pérez (EurOBIS), John Nicholls (OBIS-OPI), Johnny Konjarla (IndOBIS), Georgia Sarafidou (MedOBIS), Ana Carolina Peralta (Caribbean OBIS)

The project team also benefited from the contributions of individuals outside the immediate OBIS data quality control project team. Although they were not formal members, they were always open to discussions and eager to share their expertise. Special thanks are extended to Lee Belbin, Arthur Chapman, John Wieczorek and Paul Morris for their invaluable support.

Appendix

2021 work plan

OBIS Data Quality Assessment and Enhancement Project Team	
The OBIS DQC team is directed to	Yi Ming Gan from AntOBIS has taken over as

collaborate with TDWG, GBIF, ALA to develop a framework for the assessment and management of data quality using a fitness for use approach (Veiga et al. 2017). This can be built upon a mapping of the available QC checks in OBIS against other biodiversity quality checks and validations In this case we will have a cross-mapping matrix, and we will be able to see what we are missing or could do differently, and probably there are additional checks that we should implement for any new data types (images, DNA, tracking, habitat...).	chairperson from Hanieh. There is a dedicated group in the OBIS slack now, and summary of the discussions of the monthly meetings is available 2021 monthly meetings minutes Data cleaning exercise here
Gather and classify user stories to form contextually themed use cases, such as species distribution modeling, invasive species, etc.	User survey is drafted to grasp the need and different use cases from users.
Define a core set of standardised tests and associated assertions based on Darwin Core terms (Wieczorek et al. 2012).	Will make use of tests and assertions created by TDWG Biodiversity Data Quality Task Group 2 (BDQ TG2) and take GBIF's implementation of this into consideration: https://github.com/tdwg/bdq/issues/192 Decided that all standardized tests will be aligned with the tests and assertions developed by BDQ TG2. Task team will only focus on marine specific QC.
Develop QC flags for missing metadata.	Will make use of tests and assertions created by TDWG Biodiversity Data Quality Task Group 2 (BDQ TG2) and take GBIF's implementation of this into consideration: https://github.com/tdwg/bdq/issues/192 Further discussions needed
Gather OBIS input for new data model by GBIF and TDWG	Google doc is shared with Data QC task team members (done, submitted 4 use cases)

2022 work plan

OBIS Data Quality Assessment and Enhancement Project Team - Chair Yi-Ming Gan & Ruben Perez Perez	

Publish the OBIS user survey on various pages and send out to known OBIS users (e.g., publication authors).	Before end of June 2022		Survey published
Provide recommendations for issues mentioned in data laundry reports and monthly node presentations.	Mid 2023		Drop this task from the project team, no time and resources
Respond to GitHub issues that are related to issues raised in data laundry reports and monthly node presentations based on recommendations above.	Mid 2023		Drop this task from the project team, no time and resources
Review tests and assertions from TDWG and OBIS QC pipelines for the fields related to issues raised in data laundry reports and monthly node presentations.	DecMid 2023		Ongoing. Currently aligning obis-qc to tests and assertions from TDWG. This will be our only focus until the end of the year
Liaise with CDTT and COTT to assist in developing and identifying events, issues and training opportunities.	Mid 2023		Going well Taxonomic collecting questions will set up meeting with Leen We reviewed the training topics with Elizabeth. Support is limited to individual level as we lack time and resources.
Review quality issues of OBIS datasets from GBIF and explore solutions.	Mid 2023		paused Pieter to look into Italian datasets which are already in OBIS through EurOBIS Drop this task from the project team.

2023 work plan

OBIS Data Quality Project Team			
Align obis-qc with TDWG core tests and assertions and GBIF pipelines	All team members	Dec 2023	0
Report on activities to SG-OBIS	OBIS DQPT chair	at least one month prior to the SG meeting	0

OBIS Steering Group and Executive Committee reports

- SG-OBIS-10 report: https://oceanexpert.org/document/30481
- SG-OBIS-11 report: https://oceanexpert.org/document/32657
- EC-OBIS-5 report: https://oceanexpert.org/document/33552

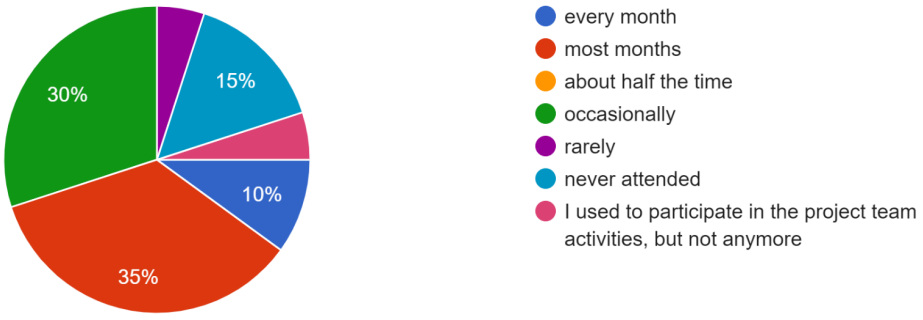
OBIS QCPT monthly meetings running minutes

- 2021 Running minutes - OBIS data QC task team
- 2022_OBIS-data-qc-running-minutes
- 2023_OBIS-data-qc-running-minutes

Results of the Internal OBIS QCPT member survey

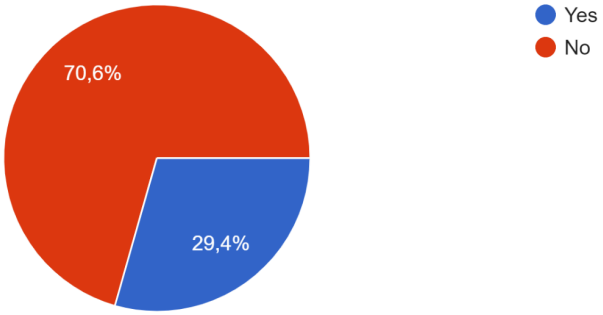
How frequent did you participate in Data QCPT's activities since 2021?

20 respuestas



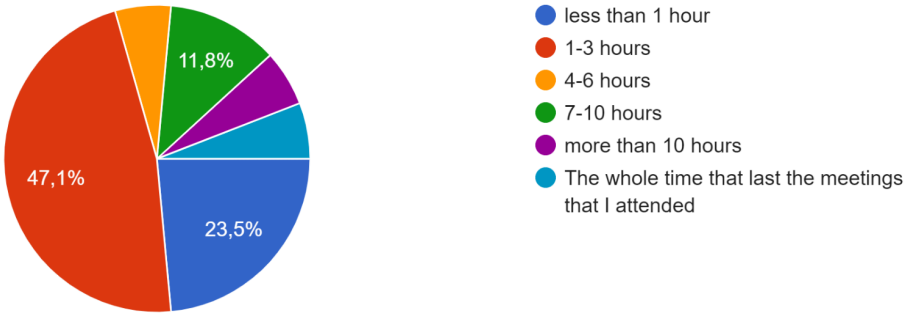
Is your involvement in the Data QCPT activities (partially) funded?

17 respuestas



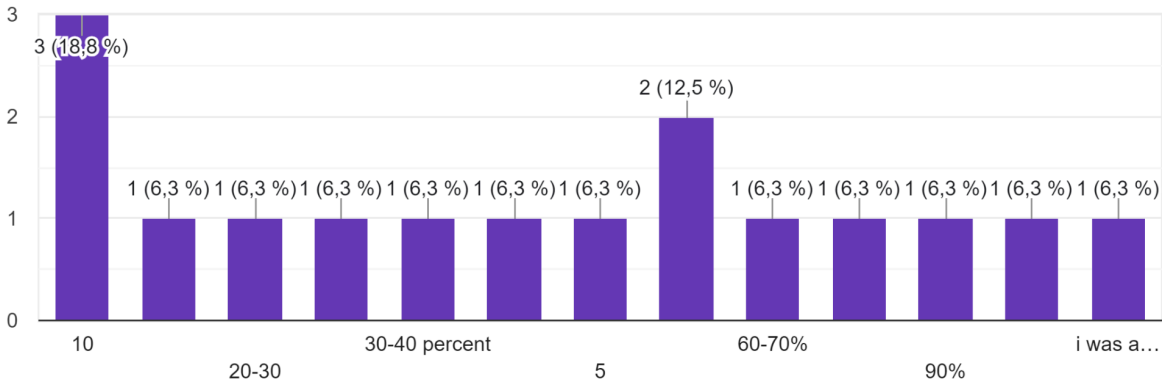
How many hours per month did you spend in the activities of the Data QCPT?

17 respuestas

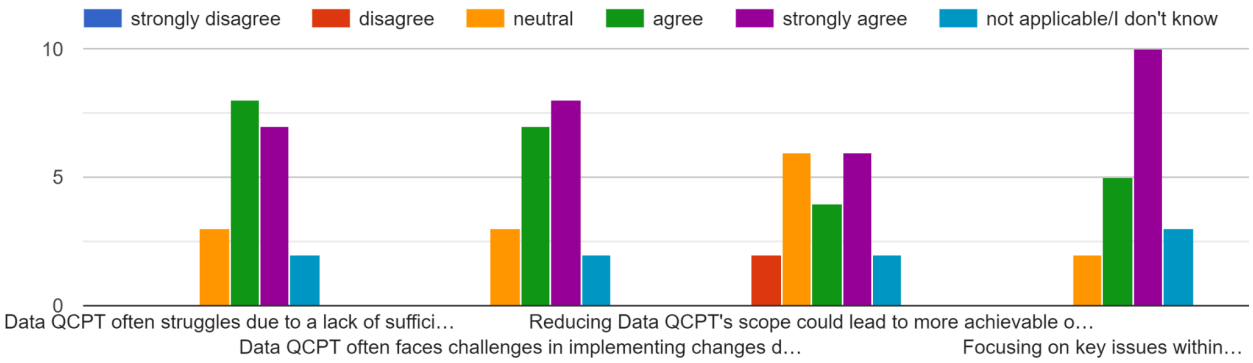


What is/was the percentage of your time spent in Data QCPT relative to other OBIS activities?

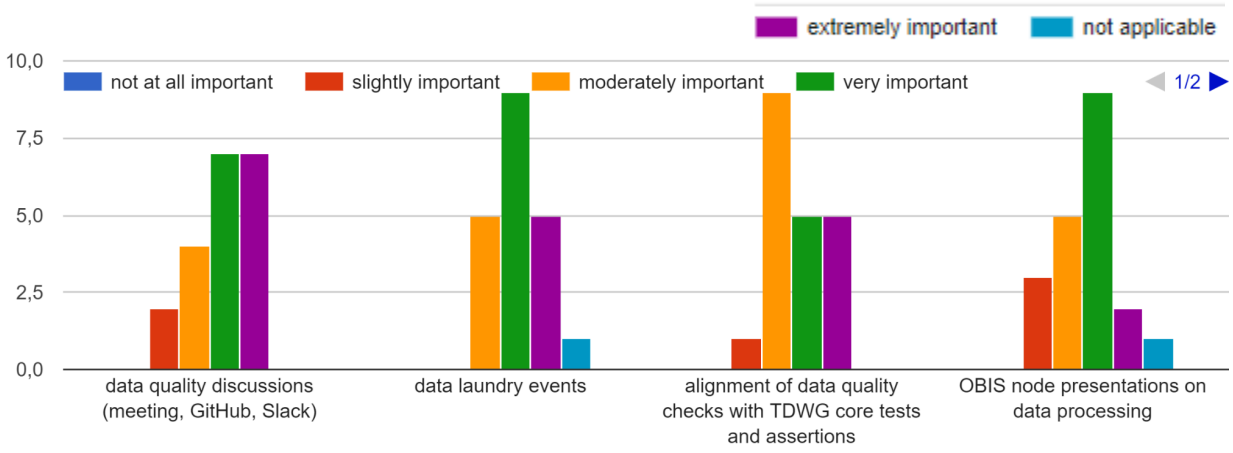
16 respuestas



Please rate the following statements

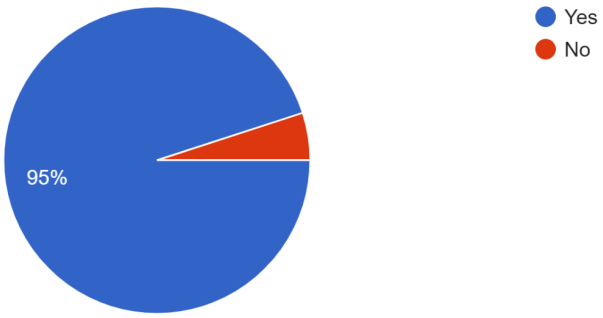


Which of the tasks from Data QCPT is important to you?



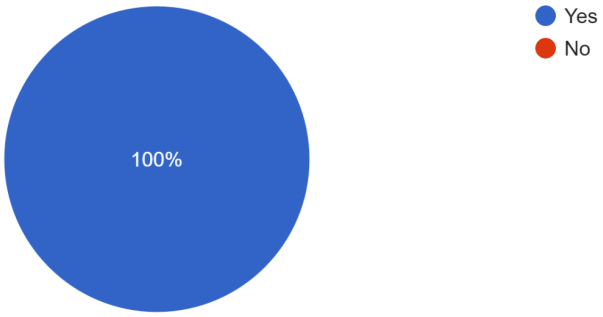
Do you think the outcomes of the Data QCPT (e.g. aligning the data quality checks with TDWG Core Tests and Assertions and GBIF pipelines) are import...BIS products? (e.g. OBIS portal, R packages etc.)

20 respuestas



Do you think having at least a full time data manager in OBIS Secretariat is essential to implement and maintain the outcome of Data QCPT?

20 respuestas



What will encourage you to participate in the activities of Data QCPT?	How much time per month can you allocate to the Data QCPT if we are to continue over the next year?	Is there anything else you'd like to share regarding Data QCPT? Your input is valuable!
	10 hours	
Earlier time schedule due to timezone	At least 1 to 2 hrs	Not at this moment

1. Having a more focused goals with short term return or implementation. Working on 1 small and specific goal at a time, with the possibility for all the OBIS members to implement, change or enhance our QC procedures in a short term. 2. Have some more attractive compensation for the data laundry events efforts from the data managers side. The nodes dedicate time to check and fix datasets but this is very time consuming and most of the time our institutions don't allow us to dedicate time for that. Maybe this could be a special event paid by OBIS, like an in person workshop were we work hard fixing dataset flags and at the same time we have the opportunity to enjoy sharing among colleagues during e. g. 1 whole week. Like a kind of data camp event, were we work but also enjoy having dinner together or just visiting the city hosting the workshop.	2-3 hours/month	We all are engaged in many voluntary tasks, as many organizations use to work in this way. At some point this is not sustainable, unfortunately I don't have a business model at the top of my head to share as a suggestion. Maybe in another opportunity I will share ideas.
More specific goals for the team, and how we can help further these goals as individuals	5-8 hours	You're doing a great job leading this team! Shorter term, achievable goals should help take the weight off your plate
I don't need encouragement - these activities are vital!	2	This activity lies at the root of OBIS and should be a core function to ensure the continued survival of OBIS.
Setting up achievable and specific goals in the short term. Dedicated funding.	8h	Just thanks to a highly committed and enthusiastic chair

More focused activities, especially ones that directly improve the work of my node or make it easier to operate an OBIS node.	1.5 hours per month	
Sustainable approach - that I do not have to do this at my personal time because when we did not get any funding, all of the sudden I have 5 project teams from OBIS that go to my personal time. Good return of investment - if what I contribute to Data QCPT will help me to gain time in my work as a data manager Shorter duration, focused goal - the only reason I show up now is because I want this to end	3-4 hours max	I have no idea how the work will be maintained. What we are doing right now is a snapshot, however GBIF pipelines and TDWG core tests continue to evolve. It is not sustainable to have the project team completing this but no maintenance planned. Cost of implementation should perhaps be considered prior to forming any working groups. I am sorry that OBIS receive this little budget ... I also do not think that it is reasonable for Pieter to work on this at his personal time. It is also a little discouraging when we work so hard on the decision but it cannot be implemented. To me, we missed the testing and feedback part of the cycle when implementation is not funded. We decided on something (QC parameters) which could have been refined based on testing results, but we could not do that due to implementation delays (though this is nobody's fault). I am frustrated at the system. Please also take into consideration the funding of OBIS' nodes in nodes health assessment, not just based on the data publication. Many members seem to be participating in this project team at their personal time. When nodes are struggling with funding, strategies can be developed to obtain funding through regional collaboration etc.

I just don't have enough time to participate right now. Currently, our node has limited staff, although we are prioritizing an effort to hire more staff for the next year. Unfortunately, participation on all project and task teams is secondary to the node management and non-OBIS responsibilities of node staff.	0 hours until additional staff is hired.	I think the Data QCPT is a critical project for OBIS, and I'm looking for ways to participate in the future.
To share knowledge, updating myself on latest developments and implementing those at our node, discussing challenges that we face with data and finding solutions.	4 hours	It is essential to have this team, as data quality should be our first priority when are providing data as a service. It is important to train all node managers and data managers on the quality flags that we developing.
the team spirit and the willingness of all members to remain engaged to the project's scope	1-3 hours/month	:)
Best data publishing	Annually is prefer	
	More than 5 hours	
	1/4 day	Seek synergy with DT Bioflow
Fewer demands on my time from non-OBIS activities	2-3 hours	
Address general issues		
Acquire a better understanding of how this and other task teams work, as well as, to have a good ability in quality control topics in order to collaborate appropiately instead of delay the process.	I don't know the quantity of time invested by the people who works in this team but I could say that I could allocate an average time of this. But how I mention before I would like to acquire more knowledge in order to give an appropriate collaboration.	I would like to share my thoughts about the importance of this task team. Data quality controls are essential to publish appropriately the datasets in OBIS, it make the stored data in OBIS reliable to use them and then publish it in important journals. No one question the datasets because OBIS has a good quality control in comparison with other databases.

When my needs align with the activities of QCPT i benefit well from their activities, even asynchronously. If I have specific insight I will endeavour to be of use in turn.	2hrs/month if called upon	I appreciate being polled via survey as it's sometimes difficult to see the true reach of an initiative like this. Many folks like myself quietly reviewing and benefitting from the time and care taken by the data quality team and the materials and methods being worked through. I hope we can convince anyone who needs to hear it that this work is vital and should continue.
Starting from the data preparation before uploading the dataset allows the data manager to keep track of what's happening. But the main issue that hinders participation is the time difference; usually the online session happens in the evening for the Asian region.	1-2 hour	
motivated team and single achievements	1-4h	