**GOOS Observations Coordination Group
Cross-Network Data Implementation Strategy**

**April 14, 2023 - Version 1.4**

**Introduction**

Data management and data flows are key components of both the Global Ocean Observing System (GOOS) strategy and its roadmap. The GOOS 2030 Strategy[[1]](#footnote-0) is a “call to action for ocean observation” and, in combination with the GOOS 2030 Strategy Roadmap[[2]](#footnote-1),, looks to define a path forward to meet increased societal needs for a fully integrated global ocean observing system by improving data interoperability among the global in situ networks.

Within GOOS, the Observations Coordination Group (OCG) works to coordinate the activities of the global ocean observing networks and programs to ensure an effective and integrated GOOS. One of the eight strategic FOCI for the OCG is Data Management and the OCG is committed to improving data interoperability both within the global networks as well as for external stakeholders, such as the World Meteorological Organization (WMO), the International Oceanographic and Data Exchange (IODE) communities and others.

The OCG data management activities aim to ensure that high-quality data are:

1. Freely available though findable and accessible web-based services,
2. Well documented,
3. Preserved for future generations and,
4. Citable.

In order to better understand the current state of the data management infrastructure that underpinned the GOOS, the OCG conducted a data interoperability analysis of the OCG global networks to map the data flows and identify opportunities for improvement. This process illustrated the complex and, at times, chaotic nature of these systems (see Appendix B for more information). The results of the data flow mapping effort, along with careful examination of existing data strategies from the WMO, UN Decade and IODE, was critical to the development of this OCG Cross-Network Data Implementation Strategy. The Strategy is based around the FAIR (Findable, Accessible, Interoperable and Reusable) data principles and provides specific implementation requirements for the global networks to meet the OCG data goals. This Strategy will be a living document, enabling agility in the face of technological and data management innovations.

As the complexity of the data landscape continues to grow, it is more important than ever to enable the trusted exchange of data and information through interoperable, accessible and timely data services. The ocean community has galvanized around the FAIR data principles for data exchange, but more specific requirements are needed in order to actually achieve interoperability. Data discovery issues continue to wreak havoc with users ability to find and use the data they want or need. Efforts abound in the community to develop data strategies to address these shortcomings, as mentioned previously. Rather than yet another data strategy, the GOOS OCG felt it was best to provide an implementation strategy which would align with existing strategies and point the way to achieving the stated goals of all of those strategies: improved interoperability, discoverability and usability of ocean data and information.

**The OCG Cross-Network Data Implementation Strategy**

**OCG Data Vision**

To provide integrated, interoperable, and FAIR-compliant access to the OCG Data System (encompassing data and metadata activities of the OCG networks, DACs, GDACs, OceanOPS and designated approved data access points) for international stakeholders, and to continue to improve access through innovation and stakeholder dialogue while engaging both with the OCG and international ocean data communities.

**OCG Data Goals**

* Enhance FAIR compliance of metadata, data and data services across all OCG networks
* Provide access to data endpoints (ie, global repositories) for each OCG network which can be federated through common middleware (such as ERDDAP)
* Ensure high quality data are available in near real time from the GTS and/or other data access services
* Ensure high quality data/metadata are discoverable and harvestable
* Ensure high quality data/metadata are available through identified global repositories
* Ensure data are fully documented and required metadata is available through OceanOPS
* Ensure data/metadata are properly archived and citable to maximize reuse

**OCG Data and Metadata Implementation Requirements**Networks are encouraged to meet all of the following data and metadata requirements and to continue maturing in these areas. The OCG will actively work to support and assist networks in achieving these data attributes.

***Real Time Data***

| **OCG-R1.** | Data shall be exchanged in real time (with minimum delay) via the WIS/GTS of the WMO in approved formats/templates. |
| --- | --- |
| **OCG-R2.** | Data shall be available in real time or near-real time on the Internet through interoperable services (preferably ERDDAP) freely and without any restriction. Community agreed quality control procedures shall be applied in real-time and adjusted values made available when possible |

***Delayed Mode Data***

| **OCG-R3.** | Each network shall have at least one identified Global Data Repository. This Global Data Repository may be one or multiple (mirrored) repositories, or they may be data endpoints that can be federated into a virtual global repository. |
| --- | --- |
| **OCG-R4.** | Data and data products shall be available through publicly accessible ERDDAP services. These distributed ERDDAP services will be federated under a single OCG ERDDAP focal point. |
| **OCG-R5.** | NetCDF is the preferred data file format, though ERDDAP services can act as a data format translator if needed. |
| **OCG-R6.** | Additional metadata should be available through the Global Data Repository and harvestable by machine-2-machine services. |

***Metadata***

| **OCG-R7.** | Networks shall have a defined uniform metadata content that includes at least the minimum OceanOPS requirements. |
| --- | --- |
| **OCG-R8.** | Metadata shall be based upon a well-documented community standard, including a persistent WMO/WIGOS identifier and use controlled vocabularies. |
| **OCG-R9.** | Metadata shall be exchanged with OceanOPS utilizing machine-2-machine services and avoiding multiple redundant manual transmissions. |

***Best Practices***

| **OCG-R10.** | Each network should have an active data team. |
| --- | --- |
| **OCG-R11.** | Each network should have identified best practices on data infrastructure and workflows and data Q.C. |
| **OCG-R12.** | Raw/real-time data, delayed mode data and data products should be archived and have unique identifiers created (i.e., Digital Object Identifier (DOI)) for citation and reuse. |

**Key Elements of the implementation Strategy**

**ERDDAP**

Through experimentation and various pilot efforts, OCG has identified the data platform ERDDAP[[3]](#footnote-2) as the key to improving interoperability and meeting FAIR compliance guidelines within OCG and the global ocean community. ERDDAP is an open source data broker that is widely used within the marine meteorological and oceanographic community to document, serve and archive data. One of the strengths of ERDDAP is its ability to interact with many different formats of data. This capability allows communities to work in the format with which they are most familiar while sharing the exact same data to other communities in the formats they are most familiar with. In a sense, it can be considered an on-the-fly data translator. This ability to abstract users and producers from data formats is a huge advantage.

In addition, ERDDAP is capable of machine-to-machine interactions, as well as manual interactions. This means that users can easily develop programmatic access for ERDDAP data and metadata. Automatic metadata translation is also done on the fly, so that it is possible to harvest metadata in the ISO-19115[[4]](#footnote-3) standard, making it very straightforward for harvesting engines. Included in this could be data and/or metadata harvests to support WIGOS, Discovery tools, and the future IOC Ocean Data and Information System (ODIS), as well as to integrate seamlessly with the future WIS 2.0 decentralized GTS evolution. ERDDAP also supports, natively, schema.org[[5]](#footnote-4) markup language, which creates structured data in support of the semantic web to make metadata more readable by search engines such as Google, etc. There is no extra effort required for this to happen.

For a further discussion of the usefulness of ERDDAP as part of the data framework for OCG networks and its relevance to the FAIR data principles, please see Appendix B.

**Tracking OCG Network Data Management Enhancements**

The recommendations provided by the OCG data strategy implementation plan are meant to be specific and actionable. In order to track compliance with the recommendations, they must also be measurable. OceanOPS currently produces an assessment of network capabilities in their annual report card[[6]](#footnote-5) and we will continue to leverage that report to track improvements in the overall OCG data system. Figure 2 illustrates the relevant data categories where OCG will track each network's progress towards improved data practices. These scores will be based upon a matrix that OCG is developing to track readiness levels among the networks. Data, metadata and best practices are key components in determining readiness levels. It is important to note that the OCG data team and OceanOPS stand ready to assist networks in the implementation of these data recommendations.



Figure 2. OceanOPS report card from 2022. The data, metadata and best practice scores have been highlighted.

**OCG Links to the Community**

The development of this data implementation strategy did not occur in a vacuum, and in fact was closely coordinated with complementary efforts in the global community, including: 1) WMO Unified Data Policy, 2) IOC data strategy, 3) Development of the IODE Ocean Data Information System and Ocean InfoHub. One goal of the OCG data implementation strategy was to **ensure** that when data complies with the OCG recommendations, **it will also comply** with the strategy requirements from these external data management strategies. This is important to ensure that OCG-generated data is available to support the WMO and IOC stakeholders without adding extra burden upon the OCG data producers.

OceanOPS is a key component in this connection to external stakeholders, in particular the WMO and IODE. OceanOPS will be the metadata repository that is responsible for populating discovery and access systems though the WMO (OSCAR, WIGOS), the IOC (IODE ODIS and OIH) and the UN Ocean Decade (DiTTO, etc). This connection is why it is so critical that OCG network metadata is complete and up-to-date in the OceanOPS system.

**Working with the Networks**

The implementation of the OCG data requirements will depend heavily on the data teams of each network. The OCG Exec, through the activities of the Data Vice-chair, and OceanOPS, will do whatever they can to support enhancements to the data structures of OCG. This includes support for ERDDAP installation, configuration and usage, as well as support for the implementation of metadata services to support OceanOPS. This support will be accomplished through quarterly roundtables, direct interaction with network data teams when necessary, and attendance at relevant network workshops/meetings. The data flow mappings, which will be updated as required, have provided a benchmark for the current status of the OCG data infrastructure and the OCG annual meeting, along with the OceanOPS annual report card, will provide a way of assessing and reporting progress towards full implementation of the OCG Strategy.

**Conclusion**

The OCG Cross-Network Data Implementation Strategy is an ambitious effort to transform the data management framework of the GOOS global networks and also to provide a blueprint for how other projects/programs can improve the interoperability and usability of their own data. The Strategy is an attempt to add clarity to the vague notion of FAIR compliance by providing concrete and actionable methods to achieve such compliance. Achieving these data goals will have a significant impact on the availability and useability of the data generated by the OCG global networks and serve to maximize the value of that data by reaching as many users as possible in easy to use and understandable services and formats. The OCG will annually assess progress towards the overall data management goals, and adjust as necessary to reach them.

**APPENDIX A. The FAIR data Principles**

In 2016, Scientific Data published the ‘[**FAIR Guiding Principles for scientific data management and stewardship’**](http://www.nature.com/articles/sdata201618)[[7]](#footnote-6). The goal was to establish a set of guidelines to improve data sharing between and amongst communities.



Figure 1. Increasing data value through compliance with FAIR principles – from Soenen, K (2021). Reusing Open Data with ERDDAP and Python. <https://doi.org/10.5281/zenodo.5684719>

There are four guiding Principles:

1. Findable
	* The first step in using data is to find them. Metadata and data should be easy to find for both humans and computers. Machine-readable metadata are essential for automatic discovery of datasets and services, so this is an essential component
2. Accessible
	* Once the user finds the data or metadata, they must be able to have free and open access using standard protocols that support machine-to-machine interactions
3. Interoperable
	* To be interoperable, metadata must use standard vocabularies that follow the FAIR principles
4. Reusable
	* The ultimate goal of FAIR is to optimize the reuse of data. To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings.

The four principles do provide a guide for improving interoperability, but it is a rather vague guide, and it can be difficult to define actionable ways forward. The OCG data implementation strategy is an attempt to provide clear guidance on how to comply with the FAIR principles.

**APPENDIX B. OCG Data Flow Mapping**

**Introduction**

As part of the Observations Coordination Group (OCG) Data Strategy, and Metadata harmonization efforts, and in line with a request from the tenth GOOS Steering Committee meeting in September 2020, GOOS initiated a Data and Metadata mapping project to clearly map out the data and metadata flows for both real-time and delayed mode data, across all the recognized OCG global networks: Argo, OceanGliders, Data Buoy Cooperation Panel (DBCP), Ship Observations Team (Voluntary Observing Ships (VOS), Ships of Opportunity Program (SOOP), Automated Shipboard Aerological Programme (ASAP)), the Global Ocean Ship-Based Hydrographic Investigations Program (GO-SHIP), OceanSITES, Global Sea Level Observing System (GLOSS), Animal Borne Ocean Sensors (AniBOS), and High Frequency (HF) Radar.

Our aims for this mapping exercise are to:  1) enable those outside the networks to better understand how network data moves through the global and national data management systems;  2) enable us in identifying gaps and areas where we can potentially improve or better support data and metadata access; 3) ensure that the required metadata is accessible and flowing into the OceanOPS monitoring system, 4) use as a base for a cross-network data strategy, ensuring that the data from global networks reaches existing and future global access points for both operational real-time and quality controlled delayed mode data.  This is a focused effort to increase value and visibility of OCG network data by improving its interoperability and ensuring that OCG data meets its critical role as an integral part of the global ocean information digital ecosystem.

The information and insight from this data mapping will inform and guide the development of an OCG Data Implementation Strategy, to better support observational network development in these areas, and recommend best practices to the community.  Understanding these data and metadata flows and identifying areas of enhancement are crucial in order to increase FAIR compliance of OCG network data and compliance with the WMO Unified Data Policy.

The data mapping contains data structures, QC elements, and key performance indicators such as data availability, timeliness, and completeness of metadata, to identify the current state of data and metadata flows for the networks. With this picture of the data flows within the different networks, we can also work with WMO and IODE to integrate this information into their data mapping efforts to extend mapping beyond the scope of just OCG and its networks.

Beyond the work of the GOOS OCG the data maps provided here are useful for global, national and regional observing and data management systems, GOOS partners and others that want to understand the existing data pathways.

**Description of the mapping**

For each OCG global network, the data mappings consist of three parts: real-time data, delayed mode data and metadata. The mappings illustrate the pathway of 1) real-time data to availability on the WMO GTS and/or through non-GTS services (e.g. GDACs), 2) delayed mode data from network specific repositories, and 3) metadata flow from networks to be available in the OceanOPS monitoring system, and from endpoint locations to OceanOPS to verify for example metadata quality, data arrival and timeliness.

The mapping effort is an on-going work as the OCG networks evolve in partners, EOVs/ECVs, and data pathways. While the initial effort concentrated on the delivery of individual observations or primary data, it will continue to map the products (e.g. gridded) made available, where possible.

The complete set of data mappings is available at <https://goosocean.org/index.php?option=com_oe&task=viewDocumentRecord&docID=31176>

**Appendix B. Further Discussion about ERDDAP and FAIR principles**

The RDA Working Group “FAIR data maturity model” (established in January 2019) aims to develop a common set of core assessment criteria for FAIRness, as an RDA Recommendation. As a result of the work done over the past year, they have released a first set of guidelines to evaluate FAIRness within the needs of a community[[8]](#footnote-7). These guidelines can provide an actionable framework, as OCG moves towards improved FAIR compliance, by identifying the elements of FAIRness to focus upon.

The maturity model uses indicators derived from the FAIR principles and categorizes them in three priority levels: Essential, Important, and Useful.

| *Priority* | Findable | Accessible | Interoperable | Reusable | Grand Total |
| --- | --- | --- | --- | --- | --- |
| Essential | 7 | 8 | 0 | 5 | 20 |
| Important | 0 | 3 | 7 | 4 | 14 |
| Useful | 0 | 1 | 5 | 1 | 7 |
| **Grand Total** | 7 | **12** | **12** | **10** | **41** |

*Table 1. The breakdown of FAIR principles by Priority level. (from https://doi.org/10.15497/RDA00045)*

As indicated in the table above, there are 20 indicators that are considered “Essential”, and a further 14 that are considered “Useful”. A more detailed discussion of what exactly these indicators are will follow below when we discuss how data systems can improve FAIRness.

To further illustrate the usefulness of ERDDAP as part of the data framework for OCG networks, let’s return to the RDA FAIR data maturity matrix. Table 2 illustrates the “Essential” indicators for data FAIRness. In the table, the rows marked green in the last column indicate capabilities that are provided by default through ERDDAP. In this first table, it is important to note that the cells that are red, have to do with ensuring the metadata content is complete. If the metadata content is complete, the cells would be green.

| **ID** | **Essential indicators** | ERDDAP |
| --- | --- | --- |
| RDA-F1-01M/D | Metadata/Data is identified by a persistent identifier |  |
| RDA-F1-02M/D | Metadata/Data is identified by a globally unique identifier |  |
| RDA-F2-01M | Rich metadata is provided to allow discovery |  |
| RDA-F3-01M | Metadata includes the identifier for the data |  |
| RDA-F4-01M | Metadata is offered in such a way that it can be harvested and indexed |  |
| RDA-A1-02M/D | Metadata/Data can be accessed manually (i.e. with human intervention) |  |
| RDA-A1-03M/D | Metadata/Data identifier resolves to a digital object |  |
| RDA-A1-04M | Metadata is accessed through standardized protocol |  |
| RDA-A1-04D | Data is accessible through standardized protocol |  |
| RDA-A1.1-01M | Metadata is accessible through a free access protocol |  |
| RDA-A2-01M | Metadata is guaranteed to remain available after data is no longer available |  |
| RDA-R1-01M | Plurality of accurate and relevant attributes are provided to allow reuse |  |
| RDA-R1.1-01M | Metadata includes information about the data reuse license  |  |
| RDA-R1.3-01M | Metadata complies with a community standard |  |
| RDA-R1.3-01D | Data complies with a community standard |  |
| RDA-R1.3-02M | Metadata is expressed in compliance with a machine-understandable community standard |  |

*Table 2. Essential indicators from RDA maturity models. Green or red indicates if having the data in ERDDAP automatically means support of that indicator - Green for yes, Red for No, or not applicable.*

If we delve further into the RDA FAIR maturity matrix to examine the “Important” indicators (table 3), we find, once again, that ERDDAP provides, by default, support for all of the indicators that aren’t reliant on metadata content. Again, if the content exists in the metadata, the column would be green.

| **ID** | **Important indicators** | ERDDAP |
| --- | --- | --- |
| RDA-A1-01M | Metadata contains information to enable the user to get access to the data |  |
| RDA-A1-05D | Data can be accessed automatically (i.e. by a computer program) |  |
| RDA-A1.1-01D | Data is accessible through a free access protocol |  |
| RDA-I1-01M | Metadata uses knowledge representation expressed in standardized format |  |
| RDA-I1-01D | Data uses knowledge representation expressed in standardized format |  |
| RDA-I1-02M | Metadata uses machine-understandable knowledge representation |  |
| RDA-I1-02D | Data uses machine-understandable knowledge representation |  |
| RDA-I2-01M | Metadata uses FAIR-compliant vocabularies |  |
| RDA-I3-01M | Metadata includes references to other metadata |  |
| RDA-I3-03M | Metadata includes qualified references to other metadata |  |
| RDA-R1.1-02M | Metadata refers to a standard reuse license |  |
| RDA-R1.1-03M | Metadata refers to a machine-understandable reuse license |  |
| RDA-R1.2-01M | Metadata includes provenance information according to community- specific standards |  |
| RDA-R1.3-02D | Data is expressed in compliance with a machine-understandable community standard |  |

*Table 3. Important indicators from RDA maturity models. Green or red indicates if having the data in ERDDAP automatically means support of that indicator - Green for yes, Red for No, or not applicable.*

Integrating ERDDAP into the OCG data framework, in combination with ensuring that data contains complete metadata documentation, will significantly improve the interoperability levels of OCG network data, and ensure that it is findable, accessible and usable across both the ocean community, but also many others.



Figure 3. ERDDAP connects users to data easily by eliminating the complexity of supporting differing sources and formats (from Soenen, K (2021). Reusing Open Data with ERDDAP and Python. <https://doi.org/10.5281/zenodo.5684719>)

1. https://www.goosocean.org/index.php?option=com\_oe&task=viewDocumentRecord&docID=24590 [↑](#footnote-ref-0)
2. https://www.goosocean.org/index.php?option=com\_oe&task=viewDocumentRecord&docID=26687 [↑](#footnote-ref-1)
3. https://coastwatch.pfeg.noaa.gov/erddap/ [↑](#footnote-ref-2)
4. https://www.iso.org/standard/53798.html [↑](#footnote-ref-3)
5. https://schema.org/ [↑](#footnote-ref-4)
6. https://www.ocean-ops.org/reportcard2022/ [↑](#footnote-ref-5)
7. Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18 [↑](#footnote-ref-6)
8. **RDA** FAIR Data Maturity Model Working Group (2020). FAIR Data Maturity Model: specification and guidelines. *Research Data Alliance*. DOI: [10.15497/RDA00045](https://doi.org/10.15497/RDA00045) [↑](#footnote-ref-7)