Data Product Quality Best Practices

A white paper from the observatory best practices/lessons learned series

Tom Kearney
Leslie Smith
Chris Rutherford

June 30, 2019

Table of Contents

Executive Summary

Data Product Quality is broadly defined based on the fitness for use of data in a particular application. In this way the needs of the user dictate whether data can be considered of sufficient quality. With the increase in digital data and the separation between data generators and data users, it is important for observatories and aggregators to be clear about what their data represent and how the data have been processed. By clearly articulating these steps and utilizing community standards, data repositories can increase the trustworthiness of themselves as a resource and of their data. In this paper, we focus on this concept of trustworthiness, reliability, and user support. Specifically, this white paper examines the current trends and drivers for data quality by focusing on four key best practice topic areas: Data Quality Control Practices, Data Support Services, Metadata, and Interoperability.

In order to assess the state of the industry in these four topic areas, research was conducted including both a literature review and review of the websites of nine major observing systems and nine data aggregators. Research also included interviews with several selected observing system staff to refine and validate best practices and the best practice self-assessment tool.

Each of these best practices are discussed in detail, accompanied by context and literature references in the remainder of the white paper. Additionally, these best practices have been organized into a best practice Self-Assessment Tool that enables an existing or new organization to assess their current data product quality capabilities and maturity level. See Appendix.

Best practices described in this white paper are based on an extensive survey of existing observatory best practices. They represent an idealized world of achievable best practices, which are recognized to be challenging to implement. Each observatory has its own priorities and available resources, as such, the best practices described are aspirational. This best practice white paper objective is to provide a simplified, easy to understand and apply guide for self-assessment and planning. It does not represent a guide for technical assessments or implementation.

Data Quality Control Practices. Successful data quality control practices require the development of a data management plan (DMP). These plans should describe the lifecycle of a dataset from production to documentation to storage and then re-use as well as facilitate long-term preservation and access to the data. Once the DMP is set, individual quality control procedures can be documented. When possible, these procedures should be aligned with community recognized standards. As much as possible, automated flagging should be used in order to more quickly process larger amounts of real-time streaming data. Human-in-the-loop checks can then be used to investigate flagged issues and determine the root cause of larger issues. For smaller programs without streaming data, a cost-benefit analysis should be conducted to determine whether it is worthwhile to convert from completely human-in-the-loop quality control to automated flagging. In addition to flagging errors in the data, it is important that

sensors receive proper pre- and post-deployment calibration to check for sensor error, drift, or biofouling. Additionally, *in situ* data should be collected to ensure that the readings from the sensors accurately reflect the surrounding environment. Often data are made immediately accessible after automated flagging, as such multiple versions of the data will exist as the data goes through the remaining quality control procedures. Making versioning information available is critical.

Lastly, an organization should consider whether they would like to seek a formal certification of their repository. Repository certification generally entails an external assessment of the organization’s processes and documentation against a third-party defined best practice guideline. If the external assessors determine that the organization meets enough of the guideline requirements, the organization’s repository can be awarded a certification which generally increases a user’s confidence in the repository and its management practices.

Data Support Services. Data Support Services are an important element of user trust in data quality. Data support services is defined here as enabling users to effectively engage with observatory data, which requires user training, support and ancillary services related to data usage. If data are not accessible and understandable in a timely and complete manner, with comprehensive metadata and source information of the appropriate detail, they are not fit for use. As such a set of data support services can include an online repository of training materials, demo videos, FAQs, etc., as well as an interactive help desk, online forum and access to technical support. Additionally, the user experience can be enhanced within the registration portal by allowing the user to save settings, request recurring data product downloads, or provide access to community developed codes and API functions. The anonymity of online data downloads can be reduced through a user registration process. With this registration information, an observatory or data aggregator is able to get in touch with users more readily about errors or updates to the data. Data repositories who import data from other observatories can enhance data quality and usage by providing data import services including: Data curation and formatting.

Metadata. Metadata is critical for scientific research, as it enables discovery, analysis, reuse and sharing of scientific data. As such it is critical that sufficient detail accompany each dataset to enable its proper use and interpretation. Aligning metadata design, structure, and procedures with community recognized standards increases accessibility, usability, and interoperability of the metadata and the data itself. As there are many different standards, the most common standard used in a given field, or by the funding agency, should be employed. Once a standard and the minimal level of metadata needed are determined, metadata should be validated prior to final submission. Lastly, as the creation of metadata can be an onerous process, determining ways to automate metadata file creation is essential.

Interoperability. Interoperability is a complex topic, but at its root is the idea that one dataset is served in a format comparable enough with another that they can be integrated and manipulated together either by a computer or a human. As much of scientific research now requires data from multiple sources, interoperability between observatories and data aggregators is essential to answer your observatory’s thematic science questions. Given the breadth of the topic of interoperability, this white paper highlights interoperability best practices as well as supporting best practices outlined in other white papers and sections of this paper. Interoperability best practices focus around ways to make data comparable and accessible, for example through the use of community aligned standard vocabularies and data formats as well as utilizing similar frameworks for data download interfaces.

Scope

This white paper on Data Product Quality examines the current trends and drivers for data quality, identifies current industry best practices, and provides recommendations. Four critical areas are discussed with a focus on why they matter to Data Product Quality. The four Data Product Quality best practice topic areas are: Data QA/QC, Data Support Services, Metadata and Interoperability.

These best practices have been organized into four best practice Self-Assessment Tools that enables an existing or new organization to assess their current data product quality capabilities and maturity level. These tools can also be used to identify steps to achieve the next aspirational levels. See Appendix.

Background

Data quality is often defined based on the data's fitness for use in a particular application (US DOI 2008, Chapman 2005, QA4EO Task Team 2010, Juran 1964). That is to say, the same data may be deemed of high quality for one application, but too low for another. Ultimately, it is the end user that determines whether the data are of sufficient quality (QA4EO Task Team 2010). In some cases, the definition of data quality (fitness for purpose) is separated from information quality which indicates whether the data are meaningful and accurately reflect the real world (US DOI 2008). For the purposes of this paper, however, the term data quality encapsulates both data and information quality.

> *"Data are of high quality if they are fit for their intended uses in operations, decision making and planning." (Juran 1964)*

Scientific data and information has become increasingly digital. This new age of digital science opens the doors for greater collaboration, peer-review, and transparency, but also adds complexity to the issue of data product quality control. No longer are scientists publishing from their own datasets that they have taken from raw data to final datasets themselves, rather publications are a mix of data sources and in some cases the scientists may not have generated a single datum used in their analysis. Without the proper documentation of quality control processes, a user may be assuming the data are pristine when they have not been assessed properly, or a user may duplicate processes that have already occurred.

Increased access to digital data has led to the increased use of these data for real-time or long-term decision making (US IOOS 2017b). Ensuring that data available online are of high quality is of the utmost important when dealing with any data that may be used for decision making purposes.

With a separation between data generator and data user, it is imperative that data quality control practices be standardized, widely used, and properly communicated. With this in mind, a key driver identified for establishing best practices of data product quality is “Does this increase trustworthiness in the data?” If a user does not trust a data source, or loses trust in a data source after finding errors, they will not use that data. Thus, it is in an observatories best interest to develop processes and methods to increase trustworthiness in their data (SeaDataNet 2010). Reliability is an important element, as such many government agencies have specific rules and guidance regarding procedures for reviewing the quality of data before it is disseminated. For example, the United States Office of Management and Budget in 2002 released a set of guidelines known as the Information Quality Act which requires federal agencies to “develop procedures for ensuring the quality, objectivity, utility, and integrity of the information they disseminate” (US DOI 2008). Federally-funded scientific data disseminated through a government website is thus required to be quality controlled.

> *“Good quality research depends on good quality data and good quality data depends on good quality control methods.” (SeaDataNet 2010)*

> *“The foundation of ocean observing is each data point generated by ocean observations, and the quality of each data point depends on doing many specific things correctly, both before an observation is made and after it is received.” (QARTOD Project Plan)*

For the purposes of this paper, a data product is defined as any generated data beyond a raw data set. For example, data products can be data generated from calibrated instrument raw data streams that have been reviewed for quality control or manipulated to create a derived data set. Data Products are important in ensuring a broad user base as they reduce the price of entry for users by not forcing the user to process raw data themselves. However, as data products are manipulated prior to the user receiving them, it is important for the user to know and trust the quality control and processing methods to ensure that they are receiving data of sufficient quality for their analysis. One type of data product of special consideration are long-term, continuous observations served via a cyberinfrastructure that provides datasets for download on a made-to-order basis. In some cases, these datasets are referred to as “provisional” as they have only received a cursory automated quality control in order to allow the data to be displayed on the website in near-real-time. A data stream, or provisional dataset, could continue to grow for decades, but a user could easily pull a single week or day of data within that.

> *“Data quality is multidimensional, and involves data management, modelling and analysis, quality control and assurance, storage and presentation.” (Chapman 2005)*

This white paper presents a synthesis of industry best practices in data product quality, examines current methods used by observatories, and provides a framework across which an observatory could identify their current level of maturity. Specifically, this white paper will discuss data management plans, data quality control practices, metadata, interoperability, data support services, and data certification.

Methodology

This white paper is one of four in a series of best practice white papers. Other best practices white papers are: Data Identification, Citation and Tracking, Observatory Performance Metrics and Community Engagement. Similar methodology was used in each best practice white paper.

Best Practices Research and Synthesis

Data product quality best practices identification, research and synthesis was an iterative building process. As best practices were identified, they were researched, refined and validated using extensive literature reviews and website reviews of nine major observing systems and nine data aggregators. Sixteen of these serve data and were examined for this paper. Once this was completed, the best practices and best practice self-assessment tools were validated through interviews with staff from three relatively mature observatories. Due to the sensitive nature of research findings, the organizations examined during research are not identified. Literature review references are included.

Our authors focused on the following research objectives while conducting secondary research:

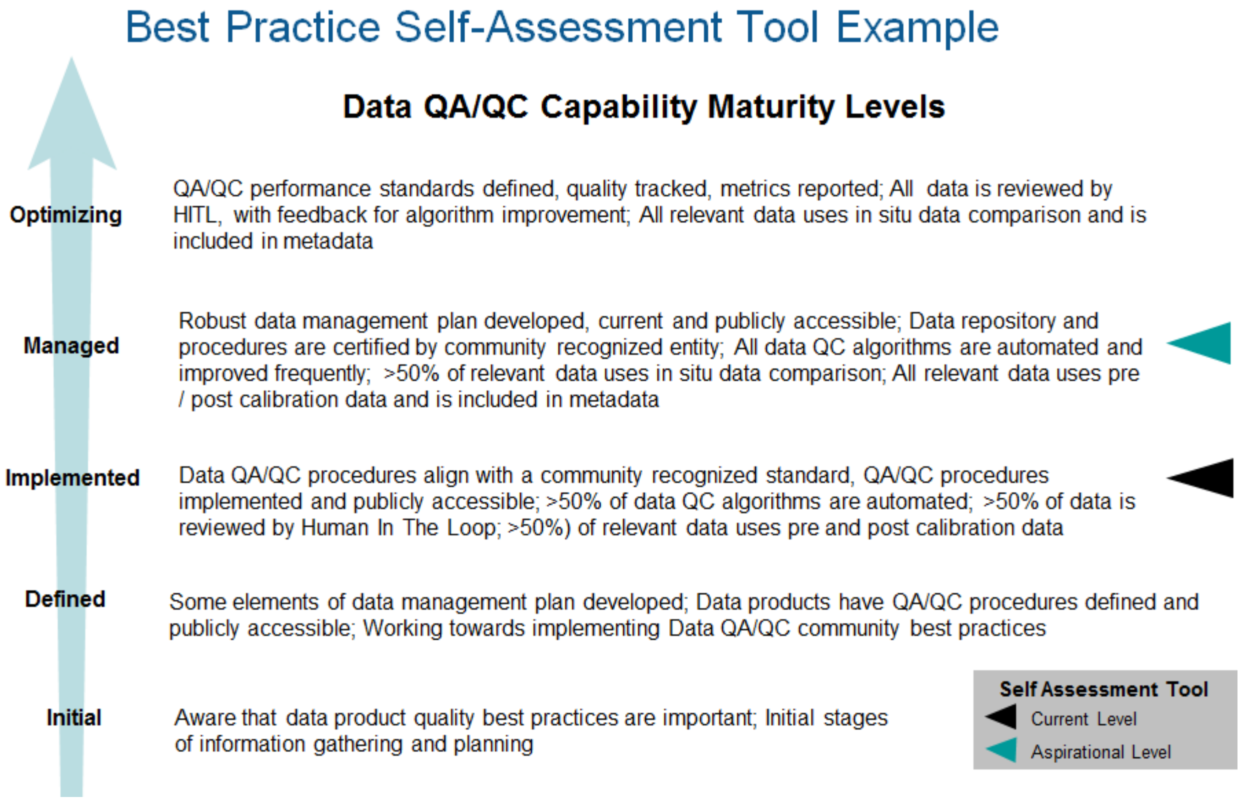
- Determine drivers for data product quality
- Determine high level requirements for data product quality
- Determine current state of industry capabilities to meet drivers

Best practice research information was synthesized from this research to identify and define best practices. As needed, secondary research was revisited to refine, test, and validate best practices. The goal of this research was to provide a high-level overview of the current state of the industry in implementing these best practices, this research is not meant to be a detailed technical assessment.

As best practices were identified and defined, a best practice self-assessment tool was developed. The best practice self-assessment tool was inspired by the Capability Maturity Model (CMM) developed by the Software Engineering Institute (SEI) at Carnegie Mellon University in 1986 (Paulk et al., 1993). The self-assessment tool creates a ranking of best practices (Figure 2), providing questions and scoring methodology. The tool ranking levels were validated through secondary and primary research. The scoring methodology provides flexibility for best practice variations across organizations. The self-assessment tool is intended to provide a structure for internal assessment and to identify aspirational improvements that can be implemented to increase maturity levels. It also provides context based on current industry wide best practice maturity levels.

Four Self-Assessment Tools are included in the Appendix: Data QA/QC, Metadata, Interoperability and Data Support Services. These best practice tools enable an existing or new organization to assess their current capabilities and maturity level. This tool can also be used to identify steps to achieve the next aspirational level. The best practice self-assessment tools and usage instructions are included in the Appendix.

Figure 2 displays one potential combination of capabilities, which results in maturity levels for a hypothetical observatory. Each observatory will have different combinations of capabilities, which aggregate to a certain maturity level. For example, one observatory may excel at data quality control, whereas another may excel at providing data support services. A simplified capability scoring method to determine levels are described in the Appendix.



Best Practice Self-Assessment Tool Example

Data QA/QC Capability Maturity Levels

Optimizing QA/QC performance standards defined, quality tracked, metrics reported; All data is reviewed by HITL, with feedback for algorithm improvement; All relevant data uses in situ data comparison and is included in metadata

Managed Robust data management plan developed, current and publicly accessible; Data repository and procedures are certified by community recognized entity; All data QC algorithms are automated and improved frequently; >50% of relevant data uses in situ data comparison; All relevant data uses pre / post calibration data and is included in metadata ◀

Implemented Data QA/QC procedures align with a community recognized standard, QA/QC procedures implemented and publicly accessible; >50% of data QC algorithms are automated; >50% of data is reviewed by Human In The Loop; >50%) of relevant data uses pre and post calibration data ◀

Defined Some elements of data management plan developed; Data products have QA/QC procedures defined and publicly accessible; Working towards implementing Data QA/QC community best practices

Initial Aware that data product quality best practices are important; Initial stages of information gathering and planning

Self Assessment Tool
◀ Current Level
◀ Aspirational Level

Best Practice Self-Assessment Tool Example

Results & Discussion

Development of a Data Management Plan

A data management plan (DMP) describes the lifecycle of a dataset from production to documentation to storage and then re-use. The primary goal of a data management plan is to facilitate long-term preservation and access to the data (GOMRI 2019, US NASA 2014, US NOAA 2015, US NSF 2018). As such, it must encompass the lifecycle of the data and how it will be produced, organized, stored, and shared (GOMRI 2019). In this way, a DMP also serves as a mechanism to facilitate the re-use of a dataset (GOMRI 2019, US NOAA 2015). Data Management Plans should be developed prior to any dataset production and should take into account stakeholder requirements for data availability and access.

Though there is no universal consensus with the observational community as to the format or components that must be included in a DMP, there are some common elements that can be seen across many organizations. Key elements of a data management plan include: the types of data collected, standards used for data and metadata format, policies for data access, sharing and re-use, and plans for archiving and preservation (US NSF 2018, US NASA 2014). Additionally, the DMP should clearly define roles and responsibilities within the organization for data management (GOMRI 2019). DMPs are not meant to describe sampling methods or data processing strategies (GOMRI 2019). Lastly, the DMP should stay current and up to date based on the needs of the program.

> *DQ BP1: Data Management Plan is developed, current, and publicly accessible.*

Data Management Plans are now required to be included in grant proposals from several federal agencies, e.g. NSF, NASA, and NOAA; NSF made data management plans a requirement for all proposals in 2011 (CODATA-ICSTI 2013). Similarly, in the United Kingdom, data management plans are mandates as part of the "Common Principles of Data Policy" (CODATA-ICSTI 2013). Having these as requirements from funding sources has created the necessary incentive to ensure wide community compliance with the practice.

> *"The goal of Data Management planning is to ensure that data are properly documented, made accessible, and preserved for future use" (US NOAA 2015)*

Of the 9 observatories examined, 5 describe DMPs on their website, 3 of these have direct links to data management plan PDFs, and one provides a "user manual" that provides similar information to a DMP. Two of the observatories that did not describe or post their DMP online are funded by the NSF, so it is assumed that they have one, it is just not publicly displayed.

> *"Improving data quality involves correcting defective data and implementing quality improvement procedures that ensure that the expected levels of data quality are achieved and maintained." (US DOI 2008)*

Data Quality Control Practices

Though there is no manual for data quality control in oceanographic research, several national and international organizations have sought to standardize QA/QC practices in order to provide a unified framework such that researchers can understand what QA/QC has been done to the data without needing to research too much into one specific organization. Utilizing the same QA/QC standards also promotes interoperability as it becomes easier to compare the quality of data between different data sources.

The Group on Earth Observations (GEO) is an international partnership of governments and organizations created in 2003 after the World Summit on Sustainable Development and the G8 summit (QA4EO Task Team 2010). The goal of this voluntary organization is to coordinate efforts and strategies across their partner organizations in order to standardize earth observations and facilitate their use in decision making. To this end, the group was tasked with establishing an international quality assurance framework to facilitate "harmonization and interoperability" of earth observation data (QA4EO Task Team 2010). The outcome of this was the Quality Assurance Framework for Earth Observation (QA4EO Task Team 2010) principles established in 2009.

> *"Data and derived products shall have associated with them an indicator of quality to enable users to assess their suitability for particular applications, i.e., their 'fitness for purpose'." (QA4EO Task Team 2010)*

The QA4EO outlines key guiding principles for data quality. These principles are then expanded into specific guideline documents to support their implementation. These principles have been paraphrased below (QA4EO Task Team 2010):

- All data and derived products must have an associated Quality Indicator based on documented quantitative assessment of its traceability to community agreed standards.

- Data product must be freely and readily available / accessible / useable in an unencumbered manner for the good of the GEOSS community, for both current and future users.
- Sound and effective harmonized documentation management is needed to facilitate and enhance interoperability and achieve the objectives of consistent and traceable quality information.

> *“The objective of data quality control of oceanographic data is…to ensure that the quality and errors of the data are apparent to the user, who has sufficient information to assess its suitability for a task.” (UNESCO 2013)*

In 2008, a joint forum on Oceanographic Data Management and Exchange Standards composed of individuals from the International Oceanographic Data and Information Exchange (IODE) and the Joint WMO-IOC Technical Commission for Oceanography and Marine Meteorology (JCOMM) established the Ocean Data Standards Pilot Project (UNESCO 2013). The purpose of the project is twofold 1) to create an achievable set of standards for ocean data management that can be broadly adopted and facilitate the exchange of data between data centers and 2) to establish a process for individual programs to submit their QA/QC standards in such a way that they can be recognized and accepted by the ocean community (UNESCO 2013).

What came out of this project was a set of IODE Quality Flag Standards that “define a common set of quality flags that can be used by data centers and projects.” (UNESCO 2013) Further details of these standards can be found in the automated flagging section of this paper.

The US Integrated Ocean Observing System (IOOS) established the Quality Assurance Quality Control of Real Time Oceanographic Data (QARTOD) project in 2012 in order to create standard processes to “identify the quality of oceanographic data in real time” (US IOOS 2017a). The true merit of the QARTOD lies in the grassroots nature of the effort, its community-driven process of creating common standards to gain community acceptance. The idea for QARTOD began much earlier, however, at a workshop in 2003 hosted by the National Data Buoy Center (US IOOS 2017b). The goal of this workshop was to “develop minimum standards for calibration, QA/QC methods and metadata” (US IOOS 2017b). From this initial meeting came five more meetings, the last of which served as the official kick-off meeting for the US IOOS QARTOD project.

IODE Quality Flag Standards were issued shortly after the first QARTOD manual was published and reviewing these together, it became clear that the early QARTOD standards shared a lot of similarities. As such, in an effort to create a common set of global principles, instead of maintaining two nearly identical standards, the QARTOD adopted the IODE scheme and adjusted their standards accordingly (US IOOS 2017a).

The QARTOD is a work in progress as more data products are defined, but the core principles that set the parameters of the QARTOD, it's Seven Data Management Laws remain unchanged. These laws are listed in their entirety below (US IOOS 2017b):

1. Every real-time observation distributed to the ocean community must be accompanied by a quality descriptor.
2. All observations should be subject to some level of automated real-time quality test.
3. Quality flags and quality test descriptions must be sufficiently described in the accompanying metadata.
4. Observers should independently verify or calibrate a sensor before deployment.
5. Observers should describe their method / calibration in the real-time metadata.
6. Observers should quantify the level of calibration accuracy and the associated expected error bounds.
7. Manual checks on the automated procedures, the real-time data collected, and the status of the observing system must be provided by the observer on a time-scale appropriate to ensure the integrity of the observing system.

From the principles, standards, and laws of these three organizations, a series of data product quality best practices for this white paper were generated:

DQ BP 1: Data Management Plan is developed, current, and publicly accessible.
DQ BP 2: Data QA/QC procedures are documented, maintained, and aligned with community recognized standards.
DQ BP 3: Data repository and procedures are certified by community recognized entity.
DQ BP 4: Data QC algorithms are automated, frequently reviewed, and improved.
DQ BP 5: Data QC procedures use humans in the loop with relevant subject matter experience.
DQ BP 6: Pre and post deployment calibrations are used to modify/annotate data and included in metadata.
DQ BP 7: In situ data are collected, used to modify/annotate data, and included in metadata.
DQ BP 8: Versioning of modified datasets are available and accessible.

Data quality control procedures utilized are dependent on the structure of the data provider, in terms of the relationship between the data provider and the data generator. There are three main models of this relationship within the 16 organizations examined for this paper:

1. **Centralized Quality Control**. This model is a closed system wherein data go straight from sensor collection into a universal cyberinfrastructure that both quality controls the data and serves it to the community. All data QA/QC occurs or is moderated at one central location. This model often occurs at large infrastructure projects.
2. **Decentralized Quality Control.** Under this model, individual groups are responsible for data collection and quality control separate from the overarching observatory that serves the data. The observatory can impose guidelines and requirements for quality control, but as quality control is decentralized it is unlikely that the exact protocols will be

conducted within each group and quality control processes necessarily have a time lag and cannot be done in near-real time. Additionally, different groups collect different data that require different testing so best practices may apply to some of the data but not others. Examples of this model include organizations that pull together data collected by researchers within a given region (e.g., GOMRI, MARACOOS, NANOOS).

3. **Data Generators are Separate from Data Providers.** This model encompasses Data Aggregators who receive data submitted from other organizations (e.g., NCEI, IRIS, AmeriFlux).

Eight of the sixteen organizations examined for this project generated their own data, of these, three followed the model of Type 1 and five followed the model of Type 2.

Quality Control refers to steps taken after data collection to determine the quality of the data collected and if possible perform corrections on the data. Quality Assurance refers to process in place during the collection of the data. Validation of data typically involves flagging and documenting suspected data, and then checking on those suspect records (Chapman 2005). In the following subsections, the elements of QA/QC are described in more detail, including Quality Assurance steps including proper procedural documentation and pre-deployment calibration, as well as Quality Control steps such as flagging, annotating, assessing, and correcting data.

Documentation of QA/QC Procedures

> *DQ BP 2: Data QA/QC procedures are documented, maintained, and aligned with community recognized standards.*

In order to fully communicate to the users what QA/QC has been done to the data, it is important to properly document QA/QC procedures keeping them up to date and version controlled in terms of protocols actually being used within the program and community accepted standards of what can be considered “good” data. Examples of this can vary from overarching discussion of the testing conducted (both automated and human-in-the-loop) and flagging/annotation provided to detailed, parameter-specific documentation of QA/QC procedures. In all cases, it is a best practice for these procedures to be vetted by the community, updated with some frequency, version controlled, and executed by trained staff or subject matter experts.

By utilizing a community accepted standard (e.g., QARTOD, QA4EO, IOC), the observatory can circumvent much of the vetting process that would be needed for an in-house procedure in order to fulfill the needs of this best practice. In this way, using a universal standard for quality control increases the overall reliability of the dataset (SeaDataNet 2010). Selecting a community standard can be challenging for an observatory that services several different communities. Though challenging it is an aspirational goal that datasets receive quality control that adheres to that communities selected standard.

Note that in cases where the same dataset serves multiple communities, it is recommended that consensus be reached regarding how to best meet the requirements of each community.

Of the 16 organizations examined for this white paper, 8 directly generated data and were analyzed for their QA/QC processes. Of those, 7 provided detailed information about quality control procedures on their website. The level of detail varied by organization. In some cases, procedures were broadly defined for the program as a whole, whereas in others detailed manuals were written and posted on the website as downloadable PDFs. Three of these organizations specifically cited that they utilized a community accepted standard for their quality control. In all cases, they utilized the QARTOD system regardless of whether they were based in the United States or abroad.

In addition to providing procedures in a publicly accessible format based on community standards, it is important to determine how successful the observatory is at implementing these procedures. For example, assessing how well the observatory followed DMP procedures, defining what constitutes successful implementation, and providing metrics of success to the community. This level of assessment leads to continuous process improvement.

Automated Data QC

With large, streaming datasets, it is not possible for a person to look at every datum that streams through to check it for errors and do so in a timely manner. Using automating flagging, real-time data can be inspected as they are collected and flagged for errors (SeaDataNet 2010). These flags can later be adjudicated by a human-in-the-loop and large sources of error can be investigated. But in the short term a user is able to download data in near-real-time that has received at least minimal quality control to it with flags highlighting good versus bad data. In this way, flagging alleviates some of the issues with making unchecked data publicly accessible and the downstream impacts of data being used in analysis prior to correction (Chapman 2005).

> *“One of the problems facing a real-time oceanographic observatory is the ability to provide a fast and accurate assessment of the data quality assurance and quality control (QAQC).” (Ocean Networks Canada)*

Automated data quality control flagging involves putting data through a series of algorithms to test for a set of potential errors in the data. For example, sensor drift, the sensor zeroing out, or large spikes in the dataset. If an error is found, the corresponding flag is associated with the data stream. In this way the data flag becomes a mechanism to communicate the results of the automated QC test to a user downloading the data (US IOOS 2017a). A summary flag can be attributed to data, which would then take into account the results from all of the tests; data are scored based on their worse overall score (US IOOS 2017a). Similarly, a derived data product inherits the worse score from the data used in its calculation (UNESCO 2013).

DQ BP 4: Data QC algorithms are automated, frequently reviewed, and improved.

Another benefit of automated flagging, is that quality flags enable users to filter data based upon known quality criteria (UNESCO 2013), allowing for data of poor or questionable quality to be filtered out by a user if they choose or all data kept. In this way, the use of automated flagging allows the user to download the full data set and then decide how to best use the data for their own purposes (US IOOS 2017a, Bermuda Biological Station for Research, Inc 1997).

One of the issues with automated flagging is that there is not one standardized global flagging system. As such, in the absence of good documentation a user may not know the meaning of the flagging notations. As articulated in the definition of data quality, it is critical for data users to know and understand what quality control practices have been done to the data prior to download. In order to remove these ambiguities, the International Oceanography Data and Information Exchange (IODE) has proposed a potential university quality control flagging system (UNESCO 2013). The system involves two levels, the first level is fixed and contains a list of 5 qualifiers - 1 = Good to 4 = Bad, 9 = missing.

"All data can have value to some users" (US IOOS 2017)

Value	Primary-Level Flag Short Name	Definition
1	Good	Passed documented required QC tests
2	Not evaluated, not available, or unknown	Used for data when no QC test performed or the information on quality is not available
3	Questionable/suspect	Failed non-critical documented metric or subjective test(s)
4	Bad	Failed critical documented QC test(s) or as assigned by the data provider
9	Missing data	Used as placeholder when data are missing

The goal is that these flags and their definitions would be used consistently across all global programs. For example, as previously discussed, the NOAA QARTOD has adopted the IOC 54: V3 flagging standard into their own standard protocol. The exceptions being that QARTOD discourages the use of Flag 2 “Not evaluated” and expands Flag 3 to include “Suspect or High Interest” to describe how a human-in-the-loop may want to focus on those data (US IOOS 2017a).

The second level of flagging provides the details to justify the level one flags and can be specific to a given organization. As second level flags can be individualized, IODE is requesting that adopters secondary level flags to the ocean data standards repository to be curated and maintained.

With so much faith being put into these automated algorithms to monitor data streams, it is also critical to review and update algorithms on a routine basis. Additionally, as technology and QA/QC algorithms are improved and refined, the quality and scope of the testing can be improved and the need for effort intensive human-in-the-loop interactions decreases.

Lastly, it is important for QC Flags to be archived with the data and metadata (US IOOS 2017a). To ensure the retention of this flagged information, it is important that when human-in-the loop testing is conducted, these new tests are added to, but do not overwrite the original QC flags. In this way versioning of quality control metrics can be maintained and a user is able to go back and access the exact information they did at the original download.

Of the 8 organizations that generated data, 6 used automated flagging of some kind on at least a portion of their data. It should be noted that for “Type 2” observatories there may be instances where some of the data quality control groups conduct automated testing and others did not. This is due to the fact that the type of data collected by the different groups can sufficiently differ that automated testing may be appropriate in some cases but not in others. For example, glider or mooring data may receive automated flagging, but species count data may exclusively receive human-in-the-loop review.

Human-in-the-Loop QC

Human-in-the-loop quality control can come in two main forms: 1) as a mechanism to review flags generated by automated algorithms and investigation of larger issues as they arise, and 2) as the primary, and in some cases only, quality control a dataset receives.

DP BP 5: Data QC procedures use humans in the loop with relevant subject matter experience.

After data has been processed through automated quality control algorithms, data are assigned flags that assign the data qualifiers ranging from good to questionable to bad. This information is available to the user. However, this only provides a portion of the puzzle needed for the user. Key information is missing in terms of why those data were flagged as bad, and more importantly whether the issue is something that can be addressed either in correcting existing data or ensuring that future data collected by that sensor or platform are good. For example, there could have been an error in the data entry of a calibration coefficient. Those data could then appear anomalous and would be flagged. A user downloading the data would see that they were flagged and likely remove them from their analysis. If someone within the observatory, however, was checking the flagged data, they could have found the error, updated the calibration coefficient, and provided good data to the user for use.

Human-in-the-loop testing is also important in terms of examining subtleties in the dataset. For example, an atmospheric pressure anomaly could be flagged within a dataset, but upon further examination by a quality control evaluator, it was determined that that anomaly was due to a very intense hurricane that passed over the array. Now, instead of those data being deemed as questionable, they have become an important asset in describe a unique, anomalous event in the area.

"Outlier detection can be a valuable validation method, but not all outliers are errors." (Chapman 2005)

Human-in-the-Loop QC is not just important in terms of further examining, catching, or correcting errors, it is also critical in determining the root of an error so that it can be corrected and prevented in future datasets. Data cleaning is a short-term tool, prevention of errors is the long-term goal (Chapman 2005).

As discussed in the automated QA/QC section, given the size of many continuous, streaming datasets, having a person look at every piece of data is not sustainable. However, not all observatories examined have these types of datasets. Thus, for some smaller projects with discrete datasets, it is still possible to have humans QA/QC the full dataset without making the process so onerous that it cannot be done in a timely fashion. Ultimately utilizing automated QA/QC with humans checking flagged data is the most sustainable and preferred method, but programs need to do a cost-benefit analysis to assess whether, given the types and quantity of data they collect, it is worthwhile to transition current manual methods to automated methods.

Two of the Type 1 observatories surveyed utilized human-in-the-loop testing as a means to review and cross check the results of the automated QC analysis. The third Type 1 observatory was a much smaller project that captured discrete datasets based on monthly surveys and as such all of the data were reviewed directly by the human analysts working with the samples. Within the five of the Type 2 observatories, much of the QA/QC work depended on the datasets being created. In some instances, data sets were sufficiently small that they were solely checked by humans in the loop, in other cases, datasets were bigger, and humans were used to check flags.

Calibration and In Situ Validation Data

> *DQ BP 6: Pre and post deployment calibrations are used to modify/annotate data and included in metadata.*

Sensor calibration is a critical element of data quality control. If an uncalibrated sensor is deployed, there is no way to know if the data being collected accurately reflect the environmental conditions. It is so critical that three of the QARTOD Seven Data Management Laws refer to it, namely that observers should independently verify or calibrate a sensor before deployment, describe their method / calibration in the real-time metadata, and quantify the level of calibration accuracy and the associated expected error bounds (US IOOS 2017a).

Sensor calibrations may be conducted by the vendor or "in house" at the observatory. Post-deployment calibrations may either be applied as a correction to the dataset or exist in separate datasets - e.g. streamed real-time data and recovered data with post corrections. In the case of post-deployment corrections, it is important that information about data versioning is also provided as the calibrated data may replace or be served alongside the original data from the sensor.

Making modifications and annotations publicly available and including calibration information in metadata presents a challenge to an observatory. Calibration procedures for different data products vary in terms of their level of complexity and in turn their level of documentation may also vary. The challenge for the observatory is in determining how much of the detail and complexity needs to be presented to the user in data product annotations and metadata files, and how much should be kept within internal documentation for record keeping.

DQ BP 7: In situ data are collected, used to modify/annotate data, and included in metadata.

All of the groups examined for this paper described some form of calibration process either on their website or in their data management or data quality control plan. In the case of the Type 2 observatories, this may have been described with some of the datasets but not necessarily all. In the case of one of the organizations examined, they have their own calibration facility focused on calibrating the observatories sensors.

In situ data collected either at the time of deployment/recovery or collected concurrently during all or part of a deployment can provide key metrics of sensor performance and data quality. Similar to calibration information, information about *in situ* data used to modify or annotate data should be included in the metadata, and the level of detail and complexity to be included in this documentation must be determined.

Examples of *in situ* data include (but are not limited to):

1. Overlapping deployments between new and old sensors or new and old platforms.
2. Co-located sampling with a similar sensor, or a sensor that measures a similar parameter.
3. Deployment near another observatory's infrastructure with similar sensors. For example, deployment of an Argo float near a surface mooring.
4. Co-located sensors (i.e. duplicate sensors) put on a platform for a full deployment.
5. Ship hydrographic cast, or other ship-based sensor, that measures conditions at the time of deployment/recovery.
6. Sample collection and analysis during deployment/recovery cruises. This could include water sample collection, sediment cores, trawls, or ROV collected samples.
7. Fly-by of a glider, or other autonomous vehicle near the deployed sensor.
8. Comparison with satellite data.

The goal of each of these is to provide data to compare to the deployed sensor to ensure the measurements from the sensor are representing the state of the environment. Pre-deployment checks can provide key information to ensure trust in the dataset, including, the instrument was properly calibrated, damage was not done to the sensor during transit causing it to fall out of calibration, and proper coefficients were entered into the system to calculate accurate data to name a few. Similarly, post-deployment checks can help to determine if the sensor lost calibration during its deployment. In some cases, post-deployment checks can be used to correct drift or apply a biofouling correction. In other cases, however, the instrument may have failed in an uncorrectable manner.

Note that if physical samples are used to provide calibration and *in situ* validation data, an accepted community standard or SOP should be followed in a certified laboratory and sample documentation included in the metadata.

Six of the eight organizations examined described doing some form of *in situ* validation to their data. In the case of the Type 2 observatories, this may have been described with some of the datasets but not necessarily all. In most cases the *in-situ* validation data described was collecting water or specimen samples.

Versioning

In this new era of large, online datasets with data streaming in real-time, versioning of data is a necessity (CODATA-ICSTI Task Group on Data Citation Standards and Practices 2013). With data constantly served and modified online, it is important for a user to understand corrections or updates made to the dataset and when they occurred. For example, if they downloaded the dataset and went back a year later to the same site they should be able to immediately see what changes were made to the data, e.g., a new, more sophisticated processing algorithm applied, an error in a calibration coefficient. It is for this reason that information on data versioning and provenance must be included with each dataset, preferably both on the dataset landing page and within the metadata.

> *DQ BP 8: Versioning of modified datasets are available and accessible.*

While immediately accessible data may include automated QC flagging, it does not include many of the later quality control steps, for example, any of the cross checks by humans-in-the-loop, corrections for post-deployment calibrations, or bio-fouling corrections. As such, by necessity there will be multiple versions of the data available online in order to make the data available quickly, but also perform QC tests of adequate rigor.

In terms of this paper, versioning is defined as manipulations of a dataset by a data collector. This could include an update to a calibration coefficient, drift correction, an updated processing algorithm, etc. Versioning can be handled in a couple ways. For example, each new version of data could be archived separately with the datasets linked together. Alternatively, the same data landing page could be used across versions with the history of data versions described on the page. In this second method, older files may or may not be available, though it is recommended that original data are maintained in conjunction with newer versions of the data (Chapman 2005). Not only is it critical to provide versioning information about the data, but it is also important that quality information about the data be carried across versions of the data, i.e., “propagated from end to end” (QA4EO Task Team 2010).

Of the organizations researched online, two provided information related to versioning of the datasets they served. In one case, version information was provided on the landing page as new versions of the discrete dataset were uploaded. The other instance involved continuous data and, in this case, a “change log” was included in the metadata file with dates when the data were modified. Another provided information on “date of last update” but no information of what happened surrounding the update.

Measuring Success

For example, Data Quality Control performance metrics, see Observatory Performance Metrics Best Practice white paper (DOI: http://dx.doi.org/10.25607/OBP-504) which provides a comprehensive performance metrics discussion.

Data Support Services

Data Support Services are an important element of user trust in data quality. Data support services is defined here as enabling users to effectively engage with observatory data, which requires user training, support and ancillary services related to data usage. If data are not accessible and understandable in a timely and complete manner, with comprehensive metadata and source information of the appropriate detail, they are not fit for use (Chapman 2005).

As streaming datasets increase in size and sophistication, so too do the data download platforms that support them. Though there are some standardized methods for data download within the oceanographic community (e.g., THREDDS, OPENDAP, ERDAPP) there is not universal download mechanism and often observatories use their own custom interface. As this is an area of rapid technological innovation, and there is an expansion of potential capabilities enabled by user registration, it is critical that users are assisted in how to effectively use the interface and the data.

An observatory best practice would be to provide support services for the best-informed use of data and metadata provided, including, but not limited to: data storage and repository services, data product catalogues, self-help tutorials and training, SME-led webinars and demonstrations, unique identification of original downloaded data sets and the QA/QC level at the time accessed, a helpdesk, and enhanced support and community interaction features available only to registered users. Data aggregators should ensure the components of this best practice are met by the data provider, or the data aggregator must fulfill that role to ensure this best practice is met. Data Support Services Best Practices are as follows:

> *“Effective feedback channels with users and suppliers is an easy and productive mechanism of improving data quality.” (Chapman 2005)*

DS BP 1: Data product user training and reference materials are publicly accessible (user self service).
DS BP 2: Enhanced user technical support for data products and system is available (Enhanced user support).
DS BP 3: Enhanced data user services are enabled by unique user registration.
DS BP 4: Users are provided with the ability to request recurring data product downloads.

DS BP 5: Users are provided with the ability to access and execute saved community code and API functions.
DS BP 6: User performance metrics are defined, tracked, and reported.
DS BP 7: Data Centers and Aggregators provide data management services.

A good example of a program that exemplifies these user serves is EUMETSAT[1], an intergovernmental organization that supplies weather and climate-related satellite data, images and products. This program provides several of the functions outlined in the above best practices list, including: a Help Desk, user notifications, product navigator, EO portal service that includes user registration, online and in person user training, interactive tools and software programs, technical documentation and training materials, and data licensing support (EUMETSAT 2018). Additionally, this organization focuses on reporting the performance of these operational services via biannual Central Operations Reports (EUMETSAT 2018). This reporting[2] includes performance metrics such as number of registered users, online data delivery, orders pending vs filled, delivery-time trends, and number of help desk queries and category of question asked.

User Training Materials

> *DS BP 1: Data product user training and reference materials are publicly accessible (user self service).*

An efficient way to reach a broad number of users is to provide them with an online repository of reference materials that includes the information needed to utilize data access interfaces as well as to understand what the data represent and what quality control has been done on them. For example, these reference materials could include quick start guides, tutorials, FAQs, demo videos, and recorded webinars. Basic training and reference materials would be those that cover broad aspects of data download and access. Advanced training and reference materials would cover specific technical aspects and data products including a data products catalog, data product specific QA/QC manuals, API guides, data processing algorithms.

Of the 16 organizations examined, 13 of them provided some type of online training or reference materials. Of those eight provided broad overview materials such as “how to” tutorials for using the online interface, an overview manual, or a FAQ page. Five organizations provided highly detailed training materials focusing on specific data products, data download tools, and processing algorithms. Of the three that did not include training materials, two did provide rudimentary tutorials of which buttons on their online interface did what.

[1] https://www.eumetsat.int
[2] https://www.eumetsat.int/website/home/Data/ServiceStatus/CentralOperationsReports/index.html?lang=EN

Though this is a very efficient system in that the reference material library can be built once and then used by an infinite number of users that visit the site, it is a passive engagement mechanism. As such there is no user feedback provided about the resources and there is no guarantee that the materials will answer all of the user’s questions or that the users will be able to understand all of the materials. As such, though these reference materials are very helpful and can help a new user get a better idea of what a system entails, or perhaps clarify a sticking point for a seasoned user, it should not be viewed as the only information needed to provide to a user.

User Technical Support

Effective communication with data users is a critical aspect of data quality (Chapman 2005). There are several facets to this: 1) active communication with users to ensure they are able to effectively and efficiently use the system access data to meet their scientific objectives and 2) two-way communication between the observatory and the user regarding data quality issues.

As discussed in the above section, while online user training materials can be a helpful resource for users and may answer some of their questions. Because this is a passive support mechanism it does not ensure that all of a user’s needs are met. As there is no way to predict every question a user may have or every stumbling block on an FAQ page, the only way to ensure all user needs are met and questions are answered is to provide an active support mechanism. For example, a user may request support via an email help desk ticket request, online forum, live chat, or phone contact from trained staff with access to specific data product expertise.

At a very basic direct user support level, a contact email address can be provided, but this is only an effective mechanism if there is a specific person within the organization assigned to prioritizing and addressing these requests. Depending on available resources, a preferred baseline method would be for the observatory to provide basic data product help desk services, including the ability for a user to initiate a trouble ticket and request ticket status. The help desk tracking system provides a key benefit as it enables issue tracking by the user and the observatory. Additionally, it facilitates communication within the program to help as issues may need to be addressed by several different individuals/departments. An additional level of service is to provide an option for the user to, in real-time, contact trained observatory staff with access to specific data product expertise. This could either be done using an online live chat feature or phone contact. Advanced data product help desk services would include a user’s ability to view ticket status online and receive automated notification of trouble ticket status. It should be noted that real-time chat support and advanced help desk tracking require a high level of resources.

DS BP 2: Enhanced user technical support for data products and system is available (Enhanced user support).

There are some resources, such as GitHub, that can provide a free, out of the box tool for issue tracking, help desk support, and development of a Knowledge Base.

The majority of the 16 organizations examined (10) only provided contact email addresses. Some of these were generic “info” or “contact” email addresses while others were specific data issue contact addresses or comment forms. It was unclear whether any of these submissions went into a tracked and ticketed help desk system or whether they just went to an individual(s) inbox. Four organizations provided online forums or message boards for users to post their comments and questions. With one of these, the interface was only available to registered users. The remaining two organizations did not provide any avenue through which a user could contact the organization.

Two-way communication of data quality issues is also important and can occur in several ways. First, is the communication between the user and the observatory if the user is having an issue with the data interface and cannot download data, or if the observatory is going to have a known outage. Second, is communicating to users who have downloaded data if an issue was found to have an error during a retrospective QC check. Lastly, is user’s communicator to the observatory when they have found an error in a dataset. With so much data being collected it is not feasible for an observatory to check every datum by hand and data users provide another critical set of eyes on the data. Additionally, users are typically doing a deep dive into the data and combining or comparing the data to other data sources, parameters, etc., and as such have a better chance of spotting any errors in the data (Chapman 2005).

Enhanced Services via User Registration

Given online security concerns, minimal functionality can be provided openly on an observatory’s cyberinfrastructure. In order to enhance the user experience and add further functionality, observatories can request that users register or use some type of common login to verify the authenticity of the user (Gries et al. 2018). Using a common login, such as CILogon, or a trusted third-party service like Google removes the need to locally manage registration as well as provides greater interoperability where the same login can be used across multiple repositories (Gries et al. 2018).

In addition to enhanced security, a side benefit of user registration is that it provides the opportunity for observatories to optimize a user’s experience through their account. One of the struggles of open online data is the anonymity of users such that the observatory does not know who exactly downloaded what data and therefore is unable to get in touch regarding errors and updates. User registration removes enough of that anonymity that the observatory, with permission, can connect with users to overall enhance the services they are able to provide.

DS BP 3: Enhanced data user services are enabled by unique user registration.

For example, a designated account for each user facilitates the ability to save data queries, receive notifications on previous downloaded data, participate in an online forum, and submit help tickets through the online portal. Additionally, if an email address was provided during registration, and proper permissions were obtained, this could facilitate the observatory directly contacting the user through a newsletter, or as needed observatory or software updates.

Advanced capabilities could also include the ability for a user to flag data and add annotations, in essence allowing a user to provide feedback to the observatory as a subject matter expert.

DS BP4: Users are provided with the ability to request recurring data product downloads.

As these are ‘ad hoc’ assessments, it is critical to have a human-in-the-loop review of the data prior to making any annotations or flag public. Even though a staff member would need to review any submissions it would still be an asset to the observatory to have more individuals examining the data.

Even with the benefits of user registration, its implementation can provide a challenge to an observatory depending on the level of sophistication of the observatory’s cyberinfrastructure. For smaller projects, it may not be time or cost effective to devote resources to the creation of such a sophisticated interface. Additionally, some users prefer not to register and just to have open access to the data. Observatories with registration may consider having some or all of their data available without the need for a login and then additional functionality and any embargoed data available with login.

Little more than half (9) of the organizations examined offered registration. Of these, three only provided registration for those that needed to submit or manage data in the system, not for data users. Of the six observatories that provided users with the ability to register for the system, half of these (3) did not describe any benefits associated with registration, one described registration as providing a mechanism through which to receive alerts and save settings, and two used registration as a way for users to collaborate with each other online.

Once within a user registration portal, it may be possible to design further tools to assist a user in finding and using the data. Specifically, the user’s ability to save and edit data export queries based on user defined criteria and data delivery format within the online data portal or set extract time periods in future and request delivery timing. An advanced service could include a machine-to-machine interface connection to the data wherein the user could extract time periods in the future and request delivery timing. While offering recurring downloads requests and saved settings is not a requirement, it is an aspirational addition to a user interface. As noted, one of the observatories did mention that a service provided to registered users was the ability to save settings.

DS BP 5: Users are provided with the ability to access and execute saved community code and API functions.

Additionally, once within the user portal, the user could be provided access to and collaborate on algorithms (e.g., Python, Matlab) or API functions generated by the observatory or other community members that can be used to download or process the observatory data. A challenge for the observatory is to determine where to code is executed, whether on the observatory's server, a third-party cloud server, or downloaded with the data and executed on the user's server. A less resource intensive alternative to presenting this code within the user portal would be to utilize a github repository where code can be down loaded or added and utilizing Jupyter[3] notebooks to provide online executable code. An added benefit of utilizing GitHub is that it allows for the code to be open access, supporting the FAIR principles referred to later in the paper.

DS BP 6: User performance metrics are defined, tracked, and reported.

While none of the organizations examined utilized this type of interface, several did have their own variations on making API code accessible to the user community. In some cases, they were example API code written by the observatory, while in others they were tools created by individuals either involved with the observatory or invited to contribute. In several cases, GitHub was used as the interface to share these code.

None of the observatories examined described tracking performance metrics in data support services on their websites. As these would be metrics used internally they may be keeping track of them but not displaying them on the website.

Data Aggregator Services

Data aggregators host data that their organization did not collect. The separation between collection and serving can lead to issues of data quality. As such, it is critical that data suppliers use sound procedures in data collection, documentation, and quality control (Chapman 2005) when supplying data to an aggregator. Data aggregators must require a certain amount of QA/QC applied to the data before they can be served online and documentation must be included. For example, a data aggregation site might become a trusted resource as most of the data hosted on its site are properly quality controlled and are of high quality. Without documentation and acceptance criteria, this could then put false trust on a dataset from a different supplier that has not received similar treatment.

[3] https://jupyter.org/

It is important to note that though aggregators can make recommendations, suggestions, and some requirements for data hosted on their site, they cannot control steps that the principal investigators collecting the data take in terms of data processing and quality control (BCO-DMO 2018). Additionally, it is important for data aggregators to request information from data submitters as to what processes were conducted on the data and then to share that information with the data aggregator users, so the information is not lost (UNESCO 2013).

DS BP 7: Data Centers and Aggregators provide data management services.

Once the data have been accepted by the data aggregator, the aggregator, as the custodian, or curator, of the data has responsibilities to maintain and improve the data throughout its lifetime (Chapman 2005). These responsibilities may include conducting quality control tests, creating proper documentation, archiving the data, providing information on and access to previous versions of the data (DQ BP 8), and providing data services to users. Data support services provided by a data aggregator can include data curation services, data formatting, data distribution to other repositories, and data access services. It should be noted that this best practice only refers to data aggregators that are actively curating data, not those that maintain references to data to broaden its discoverability.

Eleven of the 16 organizations whose websites were reviewed for this paper serve data generated from external sources, making them data aggregators. Of these, six provided data support services to data submitted to the aggregator and three provided strict guidance for submitters regarding requirements for how data must be processed prior to submission. Interestingly, two of the organizations that provide extensive data curation and management assistance limit their free assistance to organizations with funding associated with the data aggregator. This finding highlights just how labor-intensive proper curation in an aggregator can be and likely why others do not offer these services.

Measuring Success

For example, Data Support Services performance metrics, see Observatory Performance Metrics Best Practice white paper (DOI: http://dx.doi.org/10.25607/OBP-504) which provides a comprehensive performance metrics discussion.

Metadata

Metadata is critical for scientific research as it enables discovering, analyzing, reusing and sharing of scientific data[4]. For the purposes of this paper, metadata is broadly defined as descriptive data about data that can be processed by humans and computers. Other

[4] https://bigdata.ieee.org/images/files/pdf/Standards_for_Big_Datasets.pdf

organizations provide much more specific definitions, e.g., ISO/IEC 11179[5] and Dublin Core Metadata[6] describing data structures and resources.

Metadata best practices can be summarized as:

MD BP 1: Metadata aligns with community recognized standards.
MD BP 2: Sufficient metadata always accompanies data products.
MD BP 3: Validation of the metadata has been performed.
MD BP 4: Metadata file creation processes are automated.

Metadata - Current Needs

In many cases, the biggest issue with large, online datasets is not the data itself, but the metadata, the human and machine-readable descriptions that provide data about the data. It is not enough to simply put data online, "data are not usable until they can be 'explained' in a manner that both humans and computers can process" (Musen et al., 2015).

Metadata provides a required mechanism to find data online, yet studies find that 80% of data repositories are not applying metadata effectively. Metadata layers are ways for less technical users to interact with data mining systems. (NBD-PWG NIST Big Data Public Working Group 2018d)

Metadata Standard Alignment

Aligning metadata design, structure, and procedures with community recognized standards increases accessibility, usability, and interoperability of the metadata and the data itself. Example metadata standards include: Dublin Core Metadata Initiative (DCMI), ISO 19115-1[7]/19139[8], ISO/IEC 11179, ISO 2709[9] (Implemented as MARC21)/ISO 25964[10], and US Federal Geographic Data Committee (FGDC)[11].

An often-referenced metadata standard is the Dublin Core developed by the DCMI[12]. The original Dublin Core was first published in 1995. In 1998, discussions began regarding making it a standard within the US National Information Standards Organization (NISO). This led to the

[5] http://metadata-standards.org/11179/
[6] http://dublincore.org/
[7] https://www.iso.org/standard/53798.html
[8] https://www.iso.org/standard/32557.html
[9] https://www.iso.org/standard/41319.html
[10] https://www.iso.org/standard/53657.html
[11] https://www.fgdc.gov/
[12] http://dublincore.org/

publication of ANSI/NISO Z39.85-2001[13] and the ISO 15836-2003[14]. Additionally, publication of a “Part 2” to the ISO standard, covering several dozen new properties and classes is expected in 2019. Starting in 2002, DCMI grew into the role of "de facto" standards agency by maintaining its own, updated documentation for DCMI Metadata Terms. The DCMI Usage Board currently serves as the maintenance agency for ISO 15836[15].

Though the Dublin Core standard may be the most well-known, there is no universal standard for metadata. According to NIST, there are currently approximately 30 Metadata standards listed on the Digital Curation Centre (DCC) website[16]. (NBD-PWG NIST Big Data Public Working Group 2018e). This presents a problem that even if an organization seeks to adopt a community standard, which of the 30 standards should be used.

Similarly, though the Dublin Core is captured in an ISO standard, several other ISO standards also address metadata. ISO/IEC 11179 provides instruction for naming of the following items, concept, data element concept, conceptual domain, data element, and value domain. ISO/IEC 11179-5:2015[17] describes naming in a metadata registries (MDR); includes principles and rules by which naming conventions can be developed; and provides examples of naming conventions. Metadata quality can also be met using ISO 2709 (Implemented as MARC21) and thesaurus or ontology quality can be met by using ISO 25964. (NBD-PWG NIST Big Data Public Working Group 2018d) Additionally, a common ISO standard used by NOAA is the ISO 19115 standard.

> *MD BP 1: Metadata aligns with community recognized standards.*

Aligning with community standards improves metadata quality. Understanding the current status of metadata community standards is important in selecting which metadata standard to implement. This is a challenge as much work needs to be done within the community to solidify metadata standards. In the meantime, selecting the standard preferred by the organizations funding agency, or most commonly used by similar organizations with which data may be shared may be the best path forward.

Once a standard has been selected there is still the challenge of implementation. For a new observatory data repository, it is easier to implement a metadata standard from the beginning of data generation. Organizations that have existing metadata management systems will need to match any new metadata systems to the existing system, paying special attention to federation and integration issues (NBD-PWG NIST Big Data Public Working Group 2018d).

[13] https://groups.niso.org/apps/group_public/document.php?document_id=6577

[14] https://www.iso.org/standard/37629.html

[15] https://www.iso.org/standard/52142.html

[16] http://www.dcc.ac.uk/

[17] https://www.iso.org/standard/60341.html

Sufficient Metadata

As noted in the previous section, there is no real agreement in the scientific community on what should be a metadata standard. Similarly, there is no concurrence on what should be considered sufficient or minimal metadata. It is beyond the scope of this white paper to define what is considered sufficient metadata due the variability and uniqueness of the underlying subject matter requirements. Instead, we define sufficient metadata as, sufficient detail to enable proper use and interpretation of the data, including supporting interoperability. (NBD-PWG NIST Big Data Public Working Group 2018b)

> *MD BP 2: Sufficient metadata always accompanies data products.*

The data provider, as the source of data, is responsible to maintain information about the data, including origins, history and processing transformations of the metadata. It is critical that sufficient metadata accompany each data product served by an observatory or data aggregator (NBD-PWG NIST Big Data Public Working Group 2018b).

Alignment with one of the community standards provides necessary metadata guidelines. This best practice assumes metadata alignment with community standards and goes further to indicate that sufficient metadata always accompanies data product.

Guidance towards sufficient metadata may be found in prominent metadata community standards. For example, the 15 fields defined in the Dublin Core Simple Metadata[18] or the ISO 19115:2003 Geographic information that defines the minimum set of metadata required to serve the full range of metadata applications (data discovery, determining data fitness for use, data access, data transfer, and use of digital data).[19] Additionally, feedback from engaged users may help to define sufficiency for a programs specific user community.

Metadata Validation

Metadata Validation refers to quality control processes conducted by organizations creating, transforming, transferring and/or ingesting metadata. Observatory research information systems regularly import, cleanse, transform and prepare research information from internal and external data sources. An important factor for successful data integration is the removal of data errors (such as duplicates and harmonization of the data structure, inconsistent data and outdated data, etc.). These are essential tasks of data integration using extract, transform, and load (ETL) processes (Azeroual et al. 2019).

[18] http://www.dublincore.org/specifications/dublin-core/dces/

[19] https://www.iso.org/standard/26020.html

Preparation activity is where the transformation portion of the ETL/ELT cycle is likely performed, although analytics activity will also likely perform advanced parts of the transformation. Tasks performed by this activity could include data validation (e.g., checksums/hashes, format checks), cleansing (e.g., eliminating bad records/fields), outlier removal, standardization, reformatting, or encapsulating. Extract, transform, and load (ETL) and Extract, load, transform (ELT) are variations of similar processes. The difference is where the transformation occurs, at source location or target location. (NBD-PWG NIST Big Data Public Working Group 2018d)

> *MD BP 3: Validation of the metadata has been performed*

For example, NOAA Environmental Data Management has an ISO Metadata Validation tool, CEdit. A metadata record can be validated using the ISO XML Schema Definition (XSD) schema[20].

Amidst most of the use cases for data integration is an absolute need to maximize data quality, which helps to ensure accuracy. Data must be cleaned to provide quality and accurate analytic outputs. This is especially true in cases where automated integration systems are in play. Data preparation has been cited as consuming the majority of time and expense to process data. While quality is not mandatory for integration, it is commonly the most important element (NBD-PWG NIST Big Data Public Working Group 2018d).

Automated Metadata Creation

Metadata has many potential data sources, for example, pre/post calibration coefficient data, co-located sample data, geographic data, and instrument service history to name a few. As discussed above, observatory research information systems regularly import, cleanse, transform and prepare metadata from internal and external data sources (Azeroual et al. 2019).

> *MD BP 4: Metadata file creation processes are automated.*

Metadata processing is often a blend of manual and automated processing steps. Manual metadata processing can negatively impact both quality and resources required. Understanding and improving the metadata processing workflow by adding automation programming will increase metadata quality and reduce the labor required.

Automated metadata extraction can help improve efficiency for time and resource management as well as alleviate the problems associated to the “metadata resource bottleneck”. The successful application of automated metadata extraction requires informed solutions that are based on a broad understanding and integration of existing methods and tools. In particular, solutions should include the identification of weak links in the metadata

[20] https://geo-ide.noaa.gov/wiki/index.php?title=About_Collection_Metadata_Editing_Tool#Validating_Metadata

collection workflow and be firmly grounded in strict quality control at each stage of extraction (Dobreva et al. 2013).

Measuring Success

For example, Metadata performance metrics, see Observatory Performance Metrics Best Practice white paper (DOI: http://dx.doi.org/10.25607/OBP-504) which provides a comprehensive performance metrics discussion.

Interoperability

Interoperability in Data Product Quality

Interoperability is defined as the ability of information systems and stakeholders to exchange and make use of information. Further, it is the ability of different information systems, devices or applications to connect, in a coordinated manner, within and across organizational boundaries to access, exchange and cooperatively use data amongst stakeholders[21].

There is a significant increase in digital data sharing, which results in data being available for the end user from multiple repositories, including the original data generator, external supporting data repositories and data archive repositories. Environmental research data repositories provide much needed services for data preservation and data dissemination to diverse communities (Gries et al. 2018).

Trust that data from the original data generator used in research will be same as the data represented in any repository is essential to data quality credibility. Interoperability best practices address methods to increase effective data transmission and accurate replication.

Big Data and Interoperability - Current Needs

There is broad agreement among commercial, academic, and government leaders about the remarkable potential of open access online data to spark innovation, fuel commerce, and drive progress. ‘Big Data’ is the common term used to describe the deluge of data in today’s networked, digitized, sensor-laden, and information-driven world (NBD-PWG NIST Big Data Public Working Group 2018a).

Huge amounts of research data stored in a multitude of research data repository platforms can often only be used by a small fraction of their potential scientific user community. This can be caused by differences in data semantics, underlying data models and metadata schemas, whose complexity and number prevents scientists from accessing and/or fully utilizing the data. In some cases, the issue is not necessarily the need for more information about the data, it is

[21] https://www.himss.org/library/interoperability-standards/what-is-interoperability

that irrelevant data need to be distinguished and filtered away from the necessary data to achieve interoperability. The lack of interoperability between research data repository platforms causes research data not to be used to their full potential (RDA Research Data Repository Interoperability Working Group 2017).

There is an urgent need to improve the infrastructure supporting the reuse of scholarly data. To address this, a diverse set of stakeholders have come together to design and jointly endorse a set of FAIR Data Principles. These four foundational principles - Findability, Accessibility, Interoperability, and Reusability - are meant to serve as a guide for data producers. The FAIR Principles put specific emphasis on enhancing the ability of machines to automatically find and use data, in addition to supporting its reuse by individuals. It is important to note that these principles apply not only to data, but also to the algorithms, tools, and workflows that led to that data (Wilkinson et al. 2016).

The FAIR Principles provide conceptual interoperability guidelines. One source of Interoperability implementation research and guidelines is provided by the National Institute of Standards and Technology (NIST). NIST is currently stimulating collaboration among data industry professionals to provide effective interoperability through the NIST Big Data Interoperability Framework. NIST is also working to define and prioritize Big Data requirements, including interoperability, portability, reusability, extensibility, data usage, analytics, and technology infrastructure (NBD-PWG NIST Big Data Public Working Group 2018a).

Interoperability is a complex topic, for the purposes of this white paper, interoperability best practices will be described at a high level, providing key best practices that enable interoperability. Reference citations are provided for further information.

Interoperability Best Practices

Environmental research data repositories provide much needed services for data preservation and data dissemination to diverse communities. Due to independent development these repositories serve their communities well, but were developed with different technologies, data models, and using different ontologies[14]. Data repositories can address these interoperability challenges by working together to align on best practices to facilitate interoperability between repositories and observatories.

Interoperability best practices incorporate supporting best practices previously defined in this white paper, as well as the Data Identification, Citation and Tracking white paper. It is necessary to incorporate these supporting best practices as they also address aspects of interoperability. The Self-Assessment Tool in the Appendix also incorporates the supporting best practices. They are listed together here for ease of reference.

Key for Interoperability Best Practice List:

IO is Interoperability
DI is Data Identification, Citation and Tracking
DQ is Data Quality Assurance / Quality Control
MD is Metadata
DS is Data Support Services

Interoperability Best Practices:

IO BP 1: Controlled vocabularies & defined ontologies are used that adhere to community standards.
IO BP 2: Community recognized and supported data format standards are used.
IO BP 3: Dataset provenance is clearly documented and available to users.
IO BP 4: Third-party systems are synchronized with evolving resources.
IO BP 5: Clear data and community software reuse statement (license) is provided.
IO BP 6: Community aligned usability design concepts are employed for user interfaces and tools.
IO BP 7: Actively participate in interoperability best practices communities

Interoperability Supporting Best Practices Previously Defined:

DI BP 1: Persistent data identifiers are associated with all data
DI BP 3: Data identifiers are maintained throughout the life cycle of the data, including when observatory data are transferred to data aggregators
DI BP 4: Information about data versioning and provenance is available and accessible
DQ BP 8: Versioning of modified datasets are available and accessible
MD BP 1: Metadata aligns with community recognized standards.
MD BP 2: Sufficient metadata always accompanies data products.
MD BP 3: Validation of the metadata has been performed.
MD BP 4: Metadata file creation processes are automated.
DS BP 3: Enhanced data user services are enabled by unique user registration.
DS BP 5: Users are provided with the ability to access and execute saved community code and API functions.

Controlled Vocabulary & Defined Ontology

Semantic uniformity is an important component of interoperability as it ensures that the vocabulary used both within and between data repositories is aligned such that systems can exchange data (technical alignment) and users can easily understand definitions (usability alignment). Two interoperability mechanisms to achieve semantic uniformity are controlled vocabularies and mapping to ontologies.

Controlled vocabularies provide a way to organize knowledge for subsequent retrieval. In library and information science, a controlled vocabulary is a carefully selected list of words and phrases, which are used to tag units of information (document or work) so that they may be more easily retrieved by a search[22]. Ontologies, on the other hand, are shared vocabularies that are used to describe components of a particular discipline and the relationships among these components. By mapping a vocabulary to an ontology, you make it easier for others (or even the future you) to understand your data[23]. This contrasts with controlled vocabularies that are merely lists of predefined, authorized terms.

IO BP 1: Controlled vocabularies & defined ontologies are used that adhere to community standards.

Standards for describing relationships between different data sources, and standards for maintaining metadata context relationships will have substantial impact on interoperability. Semantic platforms to enhance information discovery and data integration applications may provide solutions in this area, the Resource Description Framework (RDF) and ontology mapping seem to be the front runners in the race to provide semantic uniformity. RDF graphs are leading the way for visualization and ontologies have become accepted methods for descriptions of elements. (NBD-PWG NIST Big Data Public Working Group. 2018d)

RDF graphical models represents the relationship between data elements. The data elements are nodes, and the relationship is represented as a link between nodes. Graph storage models essentially represent each data element as a series of subject, predicate, and object triples. Often, the available types of objects and relationships are described via controlled vocabularies or ontologies (NBD-PWG NIST Big Data Public Working Group. 2018b).

Community standards for controlled vocabularies and ontologies help enable interoperability. They are intended to standardize data semantics across data repositories. ISO 25964[24] provides guidelines for vocabulary interoperability and can be used when defining a vocabulary or ontology for a program (NBD-PWG NIST Big Data Public Working Group. 2018d).

Community Data Format Standards

Data is the central currency of science, but the nature of scientific data has changed dramatically with the rapid pace of technology. This change has led to the development of a wide variety of data formats, dataset sizes, data complexity, data use cases, and data sharing practices (Hart et al. 2016).

[22] https://semwebtec.wordpress.com/2010/11/23/contolled-vocabulary-vs-ontology/

[23] https://guides.library.yale.edu/rdm_healthsci/metadata_schema

[24] https://www.niso.org/schemas/iso25964

A broad overview of data repositories show that they accept a wide range of data types in a wide variety of formats, and generally do not attempt to integrate or harmonize the deposited data and place few restrictions (or requirements) on the descriptors of the data deposition. The resulting data ecosystem is becoming more diverse, and less integrated, thereby exacerbating the discovery and re-usability problem for both human and computational stakeholders (Wilkinson et al. 2016). Interestingly, it appears the industry is responding to this challenge as is demonstrated by our examination of data aggregators that found that nine of the eleven organizations surveyed did some form of data curation or had strict data guidelines for submittal. This may highlight that the organizations were selected may be uniquely further along in terms of this best practice.

IO BP 2: Community recognized and supported data format standards.

Community recognized and supported non-proprietary data format standards include:
Open-source Project for a Network Data Access Protocol (OPeNDAP), Sensor Observation Service (SOS), Sensor Web Enablement (SWE), Observations and Measurements (O&M), Web Map Service (WMS), Web Coverage Service (WCS), and Environmental Research Division Data Access Program (ERDDAP)[25].

Common data formats used by stakeholders that align with community standards increase interoperability.

Data Provenance

IO BP 3: Dataset provenance is clearly documented and available to users.

Data provenance is an important aspect of interoperability as it provides the history of the data, including a record of data source and all transformations, which especially important when data is exchanged between repositories. Transformations to raw data to data products can lead to non-trivial differences in results, which are impossible to explain without sufficient data provenance reporting. Particularly when data is exchanged between repositories (Pasquier et al. 2017).

Technical tools for tracing data lineage are currently under development. In a data repository environment, the data lineage problem is that of tracing repository data items back to the original source items from which they were derived. Cui and Widom (2003) formally define lineage tracing challenges in data repository transformations, and present algorithms for lineage tracing.

[25] https://ioos.noaa.gov/data/contribute-data/data-access-services/

If data provenance becomes a well-established convention, eventually the provenance metadata associated with each dataset will provide the complete data record. Such a record enables data users to give credit to both the authors of a referenced dataset as well as all the contributors of datasets and software from which the data were derived. As a result, this provides incentives for researchers to share resources (data, code, and process) as it will increase visibility and recognition of their work (Pasquier et al. 2017).

Data Synchronization

Data synchronization is defined as reliable real-time, asynchronous, streaming, and batch processing to extract and or update data from/to data repositories using automated programs. Examples sources and repositories include: Centralized, distributed, cloud data sources, sensors or instruments (NBD-PWG NIST Big Data Public Working Group 2018d).

> *IO BP 4: Third-party systems are synchronized with evolving resources.*

The exponential growth of data is already resulting in the development of new theories and methods addressing topics from synchronization of data across large distributed computing environments to provide consistency in high-volume and high-velocity environments (NBD-PWG NIST Big Data Public Working Group 2018d).

Providing data repository functionality that supports automated data exchange supports interoperability. Indicators that a data repository is supporting data synchronization for external users are application programming interfaces for automated data exchange (APIs), data packages, and metadata (Wiggins et al. 2018).

Data Sharing Rights

> *IO BP 5: Clear data and community software reuse statement (license) is provided.*

To encourage the sharing of data, data must be as open as possible and allow for reuse without restriction. This can be communicated through a data license that describes how the data can be accessed, used, and shared.

Clearly articulating data sharing rights will allow science to advance by encouraging the community to collaborate and build on existing work. Similar requirements are needed for all digital research products, especially software[26]. Open data sharing is echoed in the FAIR data principles that require data be released with a clear and accessible data usage license

[26] https://rd-alliance.org/sites/default/files/attachment/2018-03-23_RDA_P11_Legal-Interoperability_Stall.pdf

(RDA Research Data Repository Interoperability Working Group 2017).

To speed adoption time, accessibility, quality, and usability must be broadened, and proprietary barriers to sharing data must be overcome (NBD-PWG NIST Big Data Public Working Group 2018c).

Common User Interface and Tools

From a user perspective of accessing data from different data sources that use community aligned usability design concepts for user interfaces can increase usability, accessibility and user adoption. Usability and accessibility are interrelated concepts.

> *IO BP 6: Community aligned usability design concepts are employed for user interfaces and tools.*

According to NIST, user adoption and utilization can be increased by focusing on community-based standards for usability and accessibility, which removes barriers to effectively using data from different data repositories (NBD-PWG NIST Big Data Public Working Group 2018c). As such, observatories and data aggregators should align development of the user interface with common community practices.

Guidance for applying usability and accessibility best practice concepts to improve user's interaction with an online interface can be found in ISO 9241-11:2018[27].

Other sources for guidance on usability and accessibility include:

Web Content Accessibility Guidelines 2.0[28]
User Agent Accessibility Guidelines 1.0[29]
Authoring Tool Accessibility Guidelines 1.0[30]
WAI Education and Outreach Working Group[31]
WAI Interest Group[32]

[27] https://www.iso.org/standard/63500.html
[28] http://www.w3.org/TR/WCAG20/
[29] http://www.w3.org/TR/UAAG10/
[30] http://www.w3.org/TR/ATAG10/
[31] http://www.w3.org/WAI/EO/
[32] http://www.w3.org/WAI/IG/

Interoperability Best Practice Communities

> *IO BP 7: Actively participate in interoperability best practices communities*

Actively participating in interoperability best practice communities is one way to learn about and contribute to the development of interoperability best practices. There are many groups, a sample includes:

- Research Data Repository Interoperability WG[33]
- Open Archives Initiative[34]
- DataONE Users Group[35]
- Coalition for Publishing Data in the Earth and Space Sciences (COPDESS)[36]

It is important to note that involvement in these communities can require resources beyond the scope of smaller observatories. In lieu of participation in these organizations, smaller observatories can seek out opportunities at conferences to connect with these groups.

Measuring Success

For example, Interoperability performance metrics, see Observatory Performance Metrics Best Practice white paper (DOI: http://dx.doi.org/10.25607/OBP-504) which provides a comprehensive performance metrics discussion.

Data Certification

The concept of “certification” is used in many ways and can have many meanings depending on the context of the term’s use. To clarify this term and its use for purposes of this paper we define the term “certification” as follows:

“[a] Formal procedure by which an accredited or authorized person or agency assesses and verifies (and attests in writing by issuing a certificate) the attributes, characteristics, quality, qualification, or status of individuals or organizations, goods or services, procedures or processes, or events or situations, in accordance with established requirements or standards.”[37]

Based on this definition, there are several certifications that are relevant for this paper.

1) **Repository Certification.** The certification of a repository in terms of their processes of data storage and ensuring the long-term integrity of the data. Note that this is not a certification of the data itself within the repository. Repository certification is considered valuable as it gives individuals a higher level of confidence in the long-term storage and

[33] https://www.rd-alliance.org/groups/research-data-repository-interoperability-wg.html

[34] https://www.openarchives.org/

[35] https://www.dataone.org/dataone-users-group

[36] http://www.copdess.org/enabling-fair-data-project/

[37] http://www.businessdictionary.com/definition/certification.html

access of the data, to ensure the data is not intentionally or unintentionally modified or corrupted over time.

2) **Data Certification.** The certification of a particular dataset by a third party using predefined validation/authentication criteria to ensure the usefulness of the data in a court of law. These data are referred to as "certified data". Though this is not common, one example is data certification conducted by the National Oceanic and Atmospheric Administration (NOAA) National Centers for Environmental Information (NCEI)[38].

For the purposes of this document, only the repository certification use case is discussed in detail. The data certification use case is a very specific example and typically only applicable to data provided in support of legal proceedings. Because the topic of QA/QC is still relevant for observatories even though there is no formal, third-party designed standard for activities specifically related to scientific data to be certified against, we will discuss the concept of self-attestations and their value at the end of this section as well.

Repository Certification

> *DQ BP 3: Data repository and procedures are certified by community recognized entity.*

The concept of repository certification is based on a set of prescribed, best-practice-based standards to justify that a repository meets a minimum level of rigor. In this case a repository could be an observatory or a data aggregator, repository simply refers to organizations that host data online. Having a certification from a third-party standards and validation body gives the repository a positive reputation in the eyes of its users. In this particular case, there have been several certifying bodies who provided data repository certification levels for vendors to utilize.

Until recently, the major players in this space were the DANS (Data Archiving and Networked Services), an institute of the Royal Netherlands Academy of Arts and Sciences (KNAW) and the World Data System of the International Science Council (WDS-ISC). Initially, each of these organizations released their own requirements for standards for repository certification. Observatories and data aggregators desiring to be certified by one or both of these accrediting bodies could conduct a self-assessment against specific certification requirements for a fee. Once completed, the self-assessment results and supporting documentation were sent to the certifying bodies review board for peer review. If the certification board agreed that the observatory/aggregator met the certification requirements, they would be awarded the certification seal (logo) to display on their website. DANS issued the Data Seal of Approval (DSA) as its certification and WDS-IDC issued the WDS Regular Member certification.

[38] https://www.ncdc.noaa.gov/customer-support/certification-data

In September 2017, DANS and WDS-ISC joined forces to create a new certification standard called Core Trust Seal (CTS). This new certification does not replace previous certifications from DSA and WDS but harmonizes the requirements of both certifying bodies into a new set of certification requirements. Repositories with existing WDS or DSA certifications are not automatically "grandfathered" into the CTS certification in part because the CTS certification is designed to be both a harmonization of these standards and an update to the level of required rigor beyond what was previously required for either the WDS or DSA certification. With the advent of the new CTS certification, both the WDS and DSA organizations stopped taking new or renewal requests for their previous NDA and WDS certifications in 2018 and now point all groups to the CTS site for repository certification instructions.

The CTS certification is composed of 16 requirement guidelines[39] which are reviewed and updated every 3 years with an application process similar to the DSA and WDS certifications. Eventually, the TrustSeal organization plans to collaborate with DANS and WDS-ISC to develop a global framework for repository certification levels that move from the existing "core" level certification, to the "extended" level certification (Nestor-Seal DIN 31644), to the "formal" certification (ISO 16363 - Trusted Digital Repository) level[40].

The following provides a distribution of existing certified repositories around the world[41]:

- WDS Certified Repositories [55]
- DSA Certified Repositories [34]
- DSA & WDS Certified Repositories [1]
- CTS Certified Repositories [53]

Data Attestation and the Need for Data Certification Standards

Though achieving certification by an independent reviewing body is an aspirational goal for the observatories, it is not achievable for many organizations. The financial and time resource commitments needed to achieve and maintain certification create a barrier of entry for smaller organizations. The community should continue to strive for this best practice, making third party certification more accessible and common place, but in the meantime, it is important to examine existing strategies used by observatories to justify their methodology, attestation regarding the quality of data provided.

An attestation regarding the data provided by an organization speaks to the QA/QC processes used on the data before it is made available to the public and provides some, higher-level of confidence in the quality of the data being provided (beyond generally making data available to the public). An attestation regarding methods and processes used by an observatory to QA/QC their data does not necessarily indicate that the data provided by the organization are good or

[39] https://easy.dans.knaw.nl/ui/datasets/id/easy-dataset:72520/tab/2

[40] https://www.coretrustseal.org/wp-content/uploads/2017/01/Intro_To_Core_Trustworthy_Data_Repositories_Requirements_2016-11.pdf

[41] https://www.coretrustseal.org/why-certification/certified-repositories/

bad, just that the processes used on the data meet some necessary level of rigor as defined and implemented by the attesting organization. Data entry errors, calibration errors, and other user errors can still affect the accuracy of even the best processed data.

An attestation has little value without publicly available documentation to both define what QA/QC processes are being used, how they are being implemented, and documentation to prove that what is documented in the processes are actually being done consistently over time. Some examples of these QA/QC processes may involve the calibration and testing of instrumentation that collect data readings to ensure these readings are correctly and consistently being captured or activities to ensure vendor-provided algorithms have been correctly implemented on these instruments. Changes to algorithms or calibrations also need to be sufficiently documented and the user community should be notified to understand any corrections.

Although an organization attestation regarding the QA/QC of data they provide is a good first step in building the scientific value of the data to the greater community, the certification of data, as discussed above, is still the "gold standard" by which all scientific data is measured. Developing one or more global standards specific to the QA/QC needs of scientific data should be a focus of activity for the scientific community moving forward. Since a great deal of standard development work has already been done related to the topics of QA and QC by the International Standards Organization (ISO), perhaps working with this group to develop specialized standards in the existing ISO 9000 series could be a reasonable next step. ISO has already created specialized ISO documents for sector specific applications of the ISO 9001 standard[42] so some level of precedence for this type of activity has already been set. Providing high quality scientific data to the public is the goal of all observatories, providing globally recognized standards to help organizations do this more effectively in a provable way should be a priority for the global scientific community for the near future.

As discussed in this section, the concept of data certification typically falls into one of two main categories: repository certification and data certification.

Repository certification typically involves a third-party group that reviews existing repository documentation against a pre-defined, best practices standard managed and updated by members of the third-party group. For repository certification activities, success can be directly measured based on whether or not the repository can successfully meet the certification requirement or which level of certification the repository meets when validated by the third-party group.

Data certification does not have a common set of standards or third-party validation organizations to issue certifications. This is partly due to the fact that there are so many types of data that a one-size-fits-all for data quality analysis and certification would be very difficult to support. Since no specific standard is currently used to certify data, best practices like those

[42] https://www.iso.org/iso-9001-quality-management.html

provided in this paper should be considered as a means to measure success (i.e. providing high quality data to the scientific/user community). These measures can be used as a means to provide an organization’s attestation related to the quality of the data provided.

Conclusion/Recommendations

Data Product Quality is broadly defined based on the fitness for use of data in a particular application. In this way the needs of the user dictates whether data can be considered of sufficient quality. With the increase in digital data and the separation between data generator and data user, it is important for observatories and aggregators to be clear about what their data represent and how the data have been processed. By clearly articulating these steps and utilizing community standards, data repositories can increase the trustworthiness of themselves as a resource and of their data. The following are recommendations for observatories and data aggregators for how to develop processes and procedures to ensure their data are of high 8quality and are viewed that way by the community.

Best practices described in this white paper are recognized to be challenging to implement. Each observatory has its own priorities and available resources, as such, the best practices described are aspirational. This best practice white paper objective is to provide a simplified, easy to understand and apply guide for self-assessment and planning. It does not represent a guide for technical assessments or implementation.

Recommendation #1: Work with an existing standards-making organization (Ex: ISO) to develop one or more community recognized standards for Quality Assurance and Quality Control of scientific data. This will provide all data providers a common standard or set of standards to utilize for their data. It will also provide the basis for third-party organizations to offer certification services against these recognized standard(s). **DQ BP 3.**

Recommendation #2: Develop and make publicly accessible a data management plan and data QA/QC procedures based around community recognized standards. The procedures should include pre- and post-deployment calibrations as well as in situ validation. **DQ BP 1,2,6,7.**

Recommendation #3: As much as possible, automated flagging should be used in order to process the large amount of real time streaming data. Human-in-the-loop checks can then be used to investigate flagged issues and determine the root cause of larger issues. For smaller programs without streaming data, a cost-benefit analysis should be conducted to determine whether it is worthwhile to convert from completely human-in-the-loop quality control to automated flagging. **DQ BP 4,5.**

Recommendation #4: Versioning of modified data sets need to be made available and accessible. **DQ BP 8.**

Recommendation #5: Use a multi-layer approach for user technical support is key. Providing an online repository of videos and written materials allows for self-service support and the ability to reach a very large audience with minimal effort. This repository, however, may not answer every question a user has, as such an interactive medium such as a help desk or online forum is also needed to provide active support. **DS BP 1,2.**

Recommendation #6: Request that users register for your online data portal and utilize the user portal as a mechanism to enhance their experience. Doing this can solve many logistical issues, for example, it provides an open means of communication with the observatory (i.e. trouble ticket) and fellow users (e.g., online forum), it allows for more potential human-in-the-loop/SME checking of the data, and it allows for more efficient use of the system by the users as they can set up their own data downloads and can leverage publicly accessible codes to analyze their data. Registration also enables the organization to better track user performance metrics. **DS BP 2,3,4,5,6.**

Recommendation #7: Data aggregators not only need to provide support to users downloading data, but also data providers to ensure proper curation, formatting, distribution, and access. **DS BP 7.**

#8: Metadata should be provided with all data that provides sufficient information to enable the data’s proper use and interpretation. Where possible these metadata should also align with a community recognized standard and be validated prior to finalization. **MD BP 1,2,3.**

Recommendation #9: As much as possible, metadata should be created using automated processes. **MD BP 4.**

Recommendation #10: In order to ensure interoperability, the following should be done when creating a serving data online: community aligned standard vocabularies and data formats, similar frameworks for data download interfaces, persistent data identifiers, provide information on data versioning and provenance, and provide metadata sufficient information to enable the data’s proper use and interpretation that has been validated and aligns with a community standard. **IO BP 1,2,3,6; DI BP 1,4; DQ BP 8; MD BP 1,2,3.**

References

Azeroual, O., G. Saake, and M. Abuosba. 2019. ETL Best Practices for Data Quality Checks in RIS Databases. Informatics 6, 10.

Bermuda Biological Station for Research, Inc. 1997. Bermuda Atlantic Time-series Study Methods. 136pp

Biological and Chemical Oceanography Data Management Office. 2018/ BCO-DMO Data Management Guidelines Manual: a collection of best practice recommendations for collecting and sharing biogeochemical and ecological oceanographic data and metadata. Woods Hole, MA, Biological and Chemical Oceanography Data Management Office (BCO-DMO), 14pp. http://hdl.handle.net/11329/434

Chapman, A. D. 2005. Principles of Data Quality, Version 1.0. Report for the Global Biodiversity Information Facility, Copenhagen.

Cui, Y. and J. Widom. 2003. Lineage Tracing for General Data Warehouse Transformations. The VLDB Journal 12:41-58. https://doi.org/10.1007/s00778-002-0083-8

Dobreva, M., Y. Kim, and S. Ross. 2013. Automated Metadata Generation. DCC Digital Curation Manual, J. Davidson, S. Ross, M. Day (eds), Retrieved May 8, 2019, from http://www.dcc.ac.uk/resources/curation-reference-manual/completed-chapters/automated-metadata-extraction

EUMETSAT. 2018. Operational Services Specification, v2H. EUM/OPS/SPE/09/0810. 128 pp.

Gries, C., A. Budden, C. Laney, M. O'Brien, M. Servilla, W. Sheldon, et al. 2018. Facilitating and Improving Environmental Research Data Repository Interoperability. Data Science Journal, 17: 22, pp. 1–8. DOI: https://doi.org/10.5334/dsj-2018-022

Gulf of Mexico Research Initiative. 2019. How to Develop a Data Management Plan for the Gulf of Mexico Research Initiative Information & Data Cooperative. Accessed April 2019. https://data.gulfresearchinitiative.org/data-management-planning

Hart, E.M., P. Barmby, D. LeBauer, F. Michonnea, S. Moun, P. Mulrooney, et al. 2016. Ten Simple Rules for Digital Data Storage. PLoS Comput Biol 12(10): e1005097. doi:10.1371/ journal. pcbi.1005097

Intergovernmental Oceanographic Commission of UNESCO. 2013.Ocean Data Standards, Vol.3: Recommendation for a Quality Flag Scheme for the Exchange of Oceanographic and Marine Meteorological Data. (IOC Manuals and Guides, 54, Vol. 3.) 12 pp. (English.) (IOC/2013/MG/54-3)

Juran, J.M. and A.B. Godfrey. 1999. Juran's Quality Handbook. Fifth Edition, McGraw-Hill.

Musen M.A., C.A. Bean, K.-H. Cheung, M. Dumontier, K.A. Durante, O. Gevaert, A. Gonzalez-Beltran, et al. 2015. The Center for Expanded Data Annotation and Retrieval. Journal of the American Medical Informatics Association, JAMIA. 22 (6) 1148-1152

NBD-PWG NIST Big Data Public Working Group. 2018a. NIST Big Data Interoperability Framework: Volume 1, Definitions. Version 2. National Institute of Standards and Technology, Gaithersburg, MD. https://doi.org/10.6028/NIST.SP.1500-1r1

NBD-PWG NIST Big Data Public Working Group. 2018b. NIST Big Data Interoperability Framework: Volume 2, Big Data Taxonomies. Version 2. National Institute of Standards and Technology, Gaithersburg, MD. https://doi.org/10.6028/NIST.SP.1500-2r1

NBD-PWG NIST Big Data Public Working Group. 2018c. NIST Big Data Interoperability Framework: Volume 3, Big Data Use Cases and General Requirements. Version 2. National Institute of Standards and Technology, Gaithersburg, MD. https://doi.org/10.6028/NIST.SP.1500-3r1

NBD-PWG NIST Big Data Public Working Group. 2018d. NIST Big Data Interoperability Framework: Volume 7, Big Data Standards Roadmap. Version 2. National Institute of Standards and Technology, Gaithersburg, MD. https://doi.org/10.6028/NIST.SP.1500-7r1

NBD-PWG NIST Big Data Public Working Group. 2018e. NIST Big Data Interoperability Framework: Volume 9, Adoption and Modernization. National Institute of Standards and Technology, Gaithersburg, MD. https://doi.org/10.6028/NIST.SP.1500-10

Pasquier T., M.K. Lau, A. Trisovic, E.R. Boose, B. Couturier, M. Crosas, et al. 2017. If these data could talk. Scientific Data 4:170114. doi: 10.1038/sdata.2017.114.

QA4EO Task Team. 2010. A Quality Assurance Framework for Earth Observation: Principles, Version 4.0. Group on Earth Observations.

RDA Research Data Repository Interoperability Working Group. 2017. Research Data Repository Interoperability Primer, Version 1.0. http://doi.org/10.15497/RDA00020

SeaDataNet. 2010. Data quality control procedures, Version 2.0, 6th Framework of EC DG Research.

Wiggins, A., R. Bonney, G. LeBuhn, J.K. Parrish, and J.F. Weltzin. 2018. A Science Products Inventory for Citizen-Science Planning and Evaluation. BioScience 68(6):436–444. https://doi.org/10.1093/biosci/biy028

Wilkinson, M.D., M. Dumontier, I.J. Aalbersberg, G. Appleton, M. Axton, A. Baak, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3:160018. doi: 10.1038/sdata.2016.18.

U.S. Department of the Interior. 2008. Data Quality Management Guide, Version 1.0. 70pp

U.S. Integrated Ocean Observing System. 2017a. Manual for the Use of Real-Time Oceanographic Data Quality Control Flags, Version 1.1. 43 pp.

U.S. Integrated Ocean Observing System. 2017b. QARTOD Project Plan Accomplishments for 2012-2016 and Update for 2017-2021, Version 2.0. 48 pp.

US National Science Foundation. 2018. Proposal & Award Policies and Procedures Guide. NSF 1801. OMB Control Number 3145-0058. Effective January 29, 2018. 181 pp.

US National Oceanic and Atmospheric Administration. 2015. Data Management Planning Procedural Directive, Version 2.0.1. Effective January 1, 2015. 9 pp.

US National Aeronautics and Space Administration. 2014. NASA Plan for Increasing Access to the Results of Scientific Research. NP-2015-05-1796-HQ. 28pp.

Appendix

Best Practice Self-Assessment Tools

Four Self-Assessment Tools are included below:

Data QA/QC
Data Support Services
Metadata
Interoperability

Each best practice self-assessment tool enables an existing or new organization to assess their current capabilities and maturity level for each topic. This tool can also be used to identify steps to achieve the next aspirational level. This white paper is intended to provide a Self-Assessment Tool methodology for an organization to identify and plan for improvements in people, process, and technology.

Each observatory has its own priorities and available resources, as such, the best practices scoring descriptions and point values described will likely need to be modified. The Self-Assessment Tool methodology provides a structure which can be modified to adapt to each unique observatory.

Steps for Using the Self-Assessment Tool

1. Review Best Practices List
2. Review Example of a completed best practice self-assessment
3. Determine Self-Assessment Capability Scoring
4. Determine Maturity Levels

Data QA/QC Self-Assessment Tool

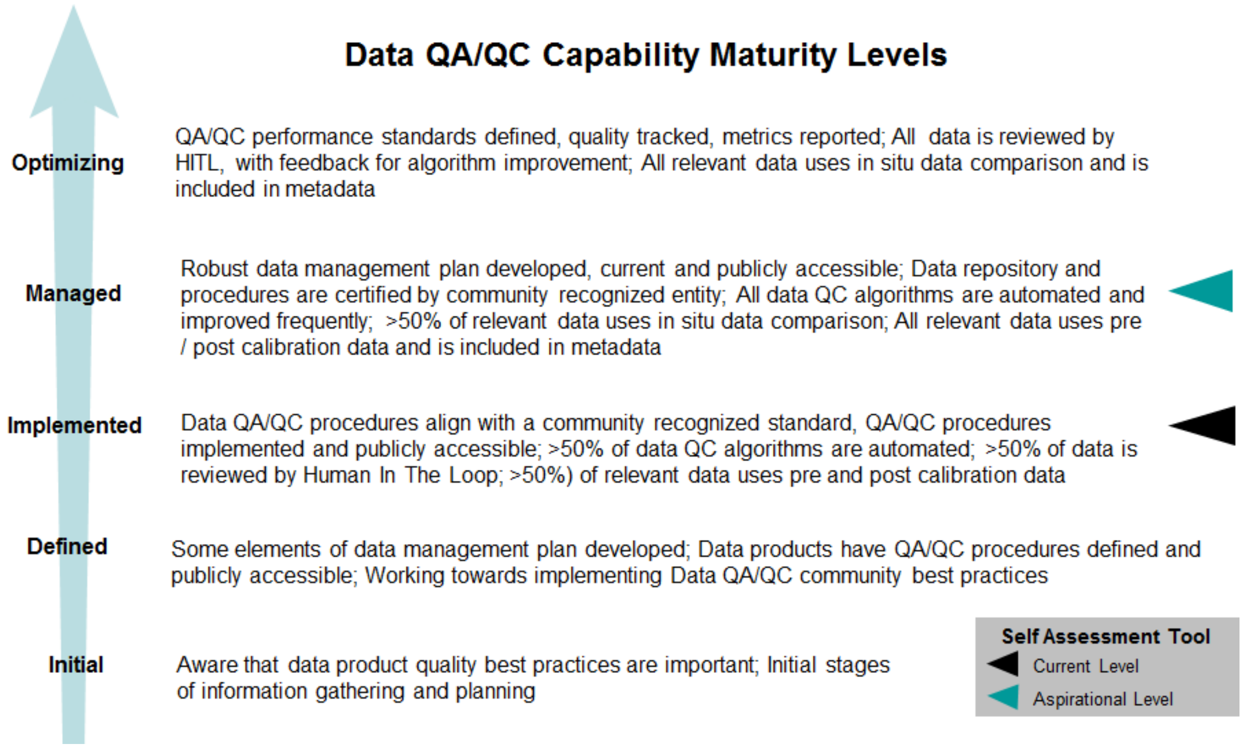
1. Best Practices List

DQ BP 1: Data Management Plan is developed, current, and publicly accessible.
DQ BP 2: Data QA/QC procedures are documented, maintained, and aligned with community recognized standards.
DQ BP 3: Data repository and procedures are certified by community recognized entity.
DQ BP 4: Data QC algorithms are automated, frequently reviewed, and improved.
DQ BP 5: Data QC procedures use humans in the loop with relevant subject matter experience.
DQ BP 6: Pre and post deployment calibrations are used to modify/annotate data and included in metadata.
DQ BP 7: In situ data are collected, used to modify/annotate data, and included in metadata.

DQ BP 8: Versioning of modified datasets are available and accessible.

2. Example Of Data Product Quality Best Practice Self-Assessment

The example below displays one potential combination of capabilities, which results in maturity levels for a hypothetical observatory. Each observatory will have different combinations of capabilities, which aggregate to a certain maturity level. For example, one observatory may excel at tracking and reporting data citations, whereas another may excel at providing data citation guidance. A simplified capability scoring method is described in the next step.



Data QA/QC Capability Maturity Levels

Optimizing QA/QC performance standards defined, quality tracked, metrics reported; All data is reviewed by HITL, with feedback for algorithm improvement; All relevant data uses in situ data comparison and is included in metadata

Managed Robust data management plan developed, current and publicly accessible; Data repository and procedures are certified by community recognized entity; All data QC algorithms are automated and improved frequently; >50% of relevant data uses in situ data comparison; All relevant data uses pre / post calibration data and is included in metadata ◀

Implemented Data QA/QC procedures align with a community recognized standard, QA/QC procedures implemented and publicly accessible; >50% of data QC algorithms are automated; >50% of data is reviewed by Human In The Loop; >50%) of relevant data uses pre and post calibration data ◀

Defined Some elements of data management plan developed; Data products have QA/QC procedures defined and publicly accessible; Working towards implementing Data QA/QC community best practices

Initial Aware that data product quality best practices are important; Initial stages of information gathering and planning

Self Assessment Tool
◀ Current Level
◀ Aspirational Level

3. Data QA/QC - Self Assessment Capability Scoring

For each best practice, determine the capability maturity score for your observatory. Only select one capability score per best practice. It is assumed each capability score is inclusive of prior score. Note: Score assumes if capability maturity not present, score is 0.

DQ BP 1: <u>Data Management Plan is developed, current, and publicly accessible.</u>
Examples include: A data management plan that describes the lifecycle of a dataset from production to documentation to storage and then re-use. The primary goal of a data management plan is to facilitate the long-term preservation and access to the data (NASA, NOAA, NSF, GOMRI). Key elements of a data management plan include: the types of data collected, standards used for data and metadata format, policies for data access, sharing, and re-use, and plans for archiving and preservation (NSF, NASA).

DQ BP 1 Self-Assessment Capability Scoring:

- Some elements of data management plan developed – 1 point
- Robust data management plan developed, current and publicly accessible – 2 points

DQ BP 2: **Data QA/QC procedures are documented, maintained, and aligned with community recognized standards.** Examples include parameter-specific data QA/QC procedures for each type of data/product provided, (both automated and human-in-the-loop), updated with some frequency, and executed by trained staff or subject matter experts. Data QA/QC SOPs align with a community recognized standard that defines how the data is QA/QCed (see data certification #1). Standards include: QARTOD, QA4EO, IOC. Though many standards exist, there is a global effort by the IOC to create a unified set of standards (IOC). These principles were then adopted by the IOOS system who have created manuals for the application of these standards for various common oceanographic sensors (QARTOD). Uniform set of QA/QC performance standards are defined, quality-tracked and metrics reported.

DQ BP 2: Self-Assessment Capability Scoring:

- Data products have QA/QC procedures developed and publicly accessible – 1 point
- Data QA/QC SOPs align with a community recognized standard – 2 points
- QA/QC performance standards defined, quality tracked, metrics reported – 3 points

DQ BP 3: Data repository and procedures are certified by community recognized entity. Examples include: The data repository and procedures are based on a set of prescribed, best-practice-based standards that align with community recognized entity such as: TrustSeal Organization, Data Seal of Approval (DSA) and The World Data System of the International Science Council (WDS).

DQ BP 3: Self-Assessment Capability Scoring:

- Data repository and procedures are certified by community recognized entity – 2 points

DQ BP 4: Data QC algorithms are automated, frequently reviewed and improved. Assumes all data goes through QC processing. Examples include: Automated QC Flagging and annotation of QC flags, post-calibration corrections, and field verification corrections. At a minimum data QC automation includes QC flagging.

DQ BP 4: Self-Assessment Capability Scoring:

- Significant (>50%) of data QC algorithms are automated – 1 point
- All data QC algorithms are automated, reviewed and improved frequently – 2 points

DQ BP 5: Data QC procedures use humans in the loop with relevant subject matter experience. Assumes humans in the loop (HITL) Data QC follows automated procedures.

Examples include: annotation of QC flags, post-calibration corrections, and field verification corrections. HITL provides necessary information for data QC algorithm improvements.

DQ BP 5: Self-Assessment Capability Scoring:

- Significant (>50%) portion of data is reviewed by HITL – 1 point
- All data is reviewed by HITL, with feedback for algorithm improvement – 2 points

DQ BP 6: Pre and post deployment calibrations are used to modify/annotate data and included in metadata. Examples include: Instrument calibrations from the refurbishment vendor or "in house" calibration services from the observatory. Post-deployment calibrations may either be provided or applied as a correction to the dataset or exist in separate datasets - e.g. streamed real-time data and recovered data with post corrections.

DQ BP 6: Self-Assessment Capability Scoring:

- Significant (>50%) portion of relevant data uses pre and post calibration data – 1 point
- All relevant data uses pre / post calibration data and is included in metadata – 2 points

DQ BP 7: In situ data are collected, used to modify/annotate data, and included in metadata. Examples include: 1) co-located sampling with a similar sensor, 2) co-located duplicate sensors, 3) ship-based sensor, 4) water sample collection and analysis, 5) autonomous vehicle, 6) satellite data. Physical samples are processed using accepted community standards. In situ data and associated corrections are included in metadata.

DQ BP 7: Self-Assessment Capability Scoring:

- Significant (>50%) portion of relevant data uses in situ data comparison – 1 point
- All relevant data uses in situ data comparison and is included in metadata – 2 points

DQ BP 8: Versioning of modified datasets are available and accessible. Examples include: Data products updated for corrections after initial release (post release processing). Versioning of QC flags so users know if/when something has changed. Store/make accessible versions of data. Datasets may change over time for many reasons. The ability to access/replicate data used in research or transferred to another data repository is important for data verification.

DQ BP 8: Self-Assessment Capability Scoring:

- Versioning of modified datasets are available and accessible – 2 points

4. Determine Maturity Levels

1. Add up your Data QA/QC capability score points from above to determine your current capability maturity level.

2. Identify your aspirational Data QA/QC maturity level by selecting a desired best practice capability score. Add up your desired capability score points to determine your aspirational capability maturity level.

Initial Level	0-1 points
Defined Level	2-5 points
Implemented Level	6-8 points
Managing Level	9-12 points
Optimizing Level	12+ points

Data Support Services Self-Assessment Tool

Steps for Using the Self-Assessment Tool

1. Review Best Practices List
2. Review Example of a completed best practice self-assessment
3. Determine Self-Assessment Capability Scoring
4. Determine Maturity Levels

1. Best Practices List

DS BP 1: Data product user training and reference materials are publicly accessible (user self service).

DS BP 2: Enhanced user technical support for data products and system is available (Enhanced user support).

DS BP 3: Enhanced data user services are enabled by unique user registration.

DS BP 4: Users are provided with the ability to request recurring data product downloads.

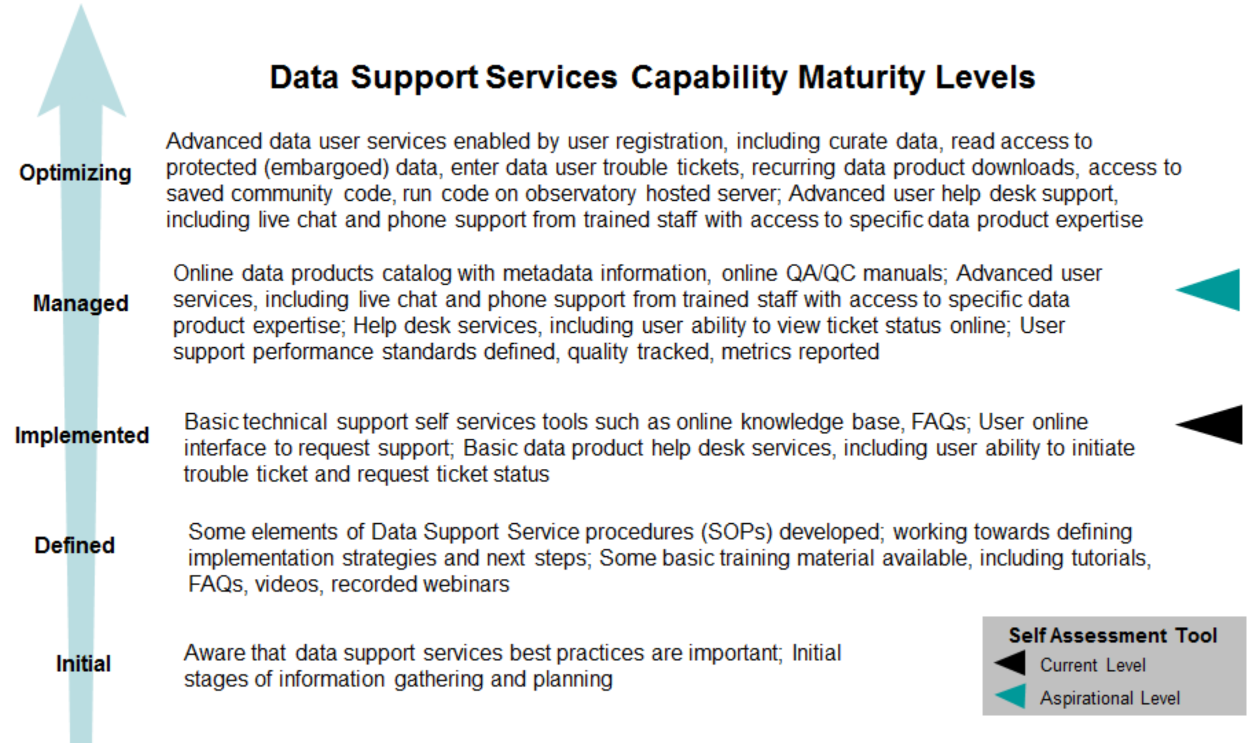
DS BP 5: Users are provided with the ability to access and execute saved community code and API functions.

DS BP 6: User performance metrics are defined, tracked, and reported.

DS BP 7: Data Centers and Aggregators provide data management services.

2. Example Of Data Support Services Best Practice Self-Assessment

The example below displays one potential combination of capabilities, which results in maturity levels for a hypothetical observatory. Each observatory will have different combinations of capabilities, which aggregate to a certain maturity level. For example, one observatory may excel at tracking and reporting data citations, whereas another may excel at providing data citation guidance. A simplified capability scoring method is described in the next step.



Data Support Services Capability Maturity Levels

Optimizing: Advanced data user services enabled by user registration, including curate data, read access to protected (embargoed) data, enter data user trouble tickets, recurring data product downloads, access to saved community code, run code on observatory hosted server; Advanced user help desk support, including live chat and phone support from trained staff with access to specific data product expertise

Managed: Online data products catalog with metadata information, online QA/QC manuals; Advanced user services, including live chat and phone support from trained staff with access to specific data product expertise; Help desk services, including user ability to view ticket status online; User support performance standards defined, quality tracked, metrics reported

Implemented: Basic technical support self services tools such as online knowledge base, FAQs; User online interface to request support; Basic data product help desk services, including user ability to initiate trouble ticket and request ticket status

Defined: Some elements of Data Support Service procedures (SOPs) developed; working towards defining implementation strategies and next steps; Some basic training material available, including tutorials, FAQs, videos, recorded webinars

Initial: Aware that data support services best practices are important; Initial stages of information gathering and planning

Self Assessment Tool
◀ Current Level
◀ Aspirational Level

3. Data Support Services - Self Assessment Capability Scoring

For each best practice, determine the capability maturity score for your observatory. Only select one capability score per best practice. It is assumed each capability score is inclusive of prior score. Note: Score assumes if capability maturity not present, score is 0.

DS BP 1: Data product user training and reference materials are publicly accessible (user self service). Examples include: Data Product training materials to access the data and use the data such as quick start guides, tutorials, FAQs, demo videos, recorded webinars. Advanced training material examples include: Online data products catalog; Data product specific QA manuals; Metadata information; API guides; data processing algorithms, moderated user question blog.

DS BP 1: Self-Assessment Capability Scoring:

- Basic training and reference materials cover broad aspects of data download and access. For example, tutorials, FAQs, “how-to” videos, recorded webinars – 1 point
- Advanced training and reference materials cover technical aspects and specific data products including data products catalog; data product specific QA/QC manuals; API guides; data processing algorithms – 2 points

DS BP 2: Enhanced user technical support for data products and system is available (Enhanced user support). Examples include: User is able to request live support via email requests, live chat and phone contact from trained staff with access to specific data product expertise. User is able to initiate trouble tickets and receive timely trouble ticket status.

DS BP 2: Self-Assessment Capability Scoring:

- Basic data product help desk services, including user ability to initiate trouble ticket and request ticket status – 1 point
- User able to access technical support via email requests, live chat and phone contact from trained staff with access to specific data product expertise – 2 points
- Advanced data product help desk services, including user ability to view ticket status online, receive automated notification of trouble ticket status – 3 points

DS BP 3: Enhanced data user services are enabled by unique user registration. Examples include: Ability to save data queries, receive notifications on previous downloaded data, receive newsletters, participate in online forums, and observatory updates. Advanced services include: Ability to curate data (flag data, add annotations), read access to protected (embargoed) data, enter data user trouble tickets, provide data product input.

DS BP 3: Self-Assessment Capability Scoring:

- Data user services enabled by user registration, including save data queries, receive notifications on previous downloaded data, receive newsletters and observatory data product updates – 1 point
- Advanced data user services enabled by user registration, including curate data (flag data, add annotations), read access to restricted access data, enter data user trouble tickets, provide data product input – 2 points

DS BP 4: Users are provided with the ability to request recurring data product downloads. Assumes user registration. Examples include: Ability for user to save and edit data export query based on user defined criteria and data delivery format, set extract time periods in future and request delivery timing. Advanced service includes machine-to-machine interface connection to the data.

DS BP 4: Self-Assessment Capability Scoring:

- Recurring data product downloads enabled by user registration, including ability for user to save and edit data export query based on user defined criteria and data delivery format – 1 point
- Advanced recurring data product downloads enabled by user registration, set extract time periods in future and request delivery timing. Advanced service includes machine-to-machine interface connection to the data – 2 points

DS BP 5: Users are provided with the ability to access and execute saved community code and API functions. Assumes user registration. Examples include: User access to saved executable code and API functions in github repository, access to hosted Matlab/Python code. Advanced service includes ability to: Run code on observatory hosted server, add personal executable code and API functions, Copy and edit community executable code.

DS BP 5: Self-Assessment Capability Scoring:

- Access to saved community code enabled by user registration, including the ability for user access to saved executable code and API functions in github repository, access to hosted Matlab/Python code – 1 point
- Advanced access to saved community code enabled by user registration, including ability to run code on observatory hosted server, add personal executable code and API functions, copy and edit community executable code – 2 points

DS BP 6: User performance metrics are defined, tracked, and reported. Examples include: Uniform set of User support and Customer experience performance standards are defined, responsiveness and resolution quality-tracked and metrics reported.

DS BP 6: Self-Assessment Capability Scoring:

- Some user support performance standards defined, in process of implementing – 1 point
- User support performance standards defined, quality tracked, metrics reported – 2 points

DS BP 7: Data Centers and Aggregators provide data management services. Examples include: Data curation services, data formatting, data distribution to other repositories, data access and technical support services.

DS BP 7: Self-Assessment Capability Scoring:

- Data curation services, data formatting, data distribution to other repositories, data access services – 1 point

4. Determine Maturity Levels

1. Add up your Data Support Services capability score points from above to determine your current capability maturity level.
2. Identify your aspirational Data Support Services maturity level by selecting a desired best practice capability score. Add up your desired capability score points to determine your aspirational capability maturity level

Initial Level	0 points
Defined Level	1-2 points
Implemented Level	3-4 points
Managing Level	5-7 points
Optimizing Level	7+ points

Metadata Self-Assessment Tool

Steps for Using the Self-Assessment Tool

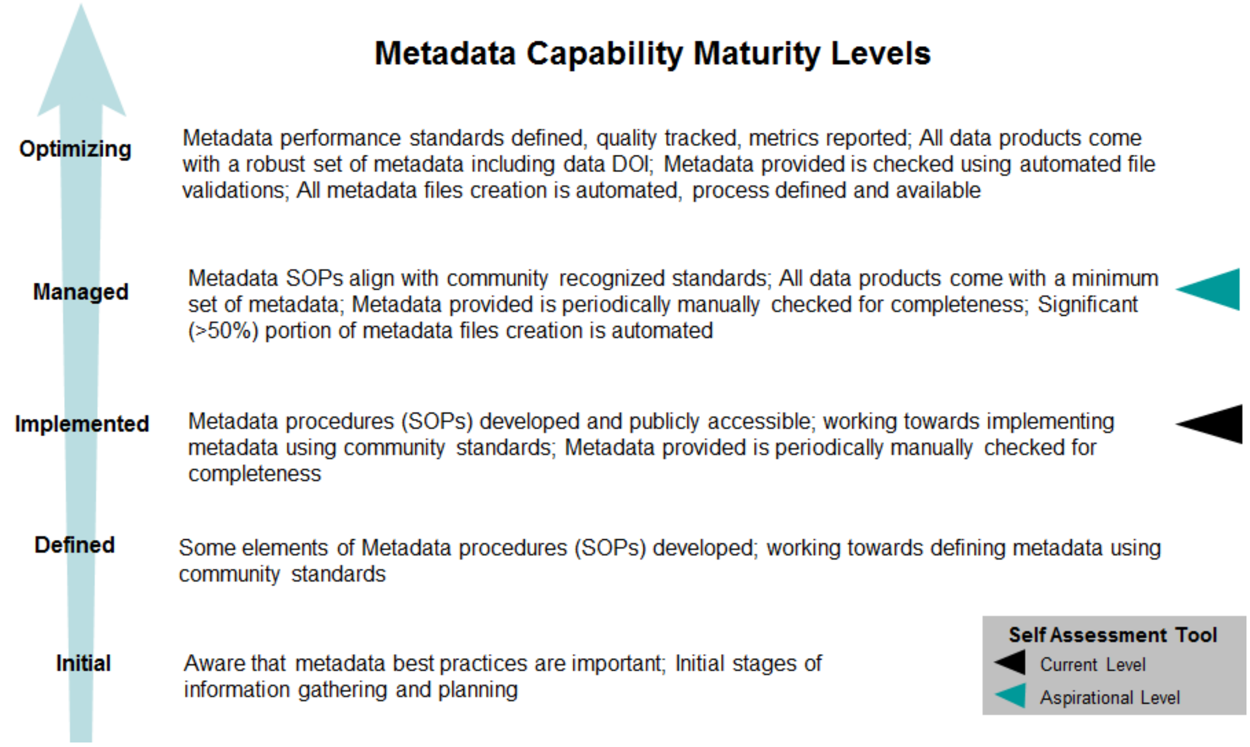
1. Review Best Practices List
2. Review Example of a completed best practice self-assessment
3. Determine Self-Assessment Capability Scoring
4. Determine Maturity Levels

1. Best Practices List

MD BP 1: Metadata aligns with community recognized standards.
MD BP 2: Sufficient metadata always accompanies data products.
MD BP 3: Validation of the metadata has been performed.
MD BP 4: Metadata file creation processes are automated.

2. Example of Metadata Best Practice Self-Assessment

The example below displays one potential combination of capabilities, which results in maturity levels for a hypothetical observatory. Each observatory will have different combinations of capabilities, which aggregate to a certain maturity level. For example, one observatory may excel at tracking and reporting data citations, whereas another may excel at providing data citation guidance. A simplified capability scoring method is described in the next step.



Metadata Capability Maturity Levels

Optimizing Metadata performance standards defined, quality tracked, metrics reported; All data products come with a robust set of metadata including data DOI; Metadata provided is checked using automated file validations; All metadata files creation is automated, process defined and available

Managed Metadata SOPs align with community recognized standards; All data products come with a minimum set of metadata; Metadata provided is periodically manually checked for completeness; Significant (>50%) portion of metadata files creation is automated ◀

Implemented Metadata procedures (SOPs) developed and publicly accessible; working towards implementing metadata using community standards; Metadata provided is periodically manually checked for completeness ◀

Defined Some elements of Metadata procedures (SOPs) developed; working towards defining metadata using community standards

Initial Aware that metadata best practices are important; Initial stages of information gathering and planning

3. Data Support Services - Self Assessment Capability Scoring

For each best practice, determine the capability maturity score for your observatory. Only select one capability score per best practice. It is assumed each capability score is inclusive of prior score. Note: Score assumes if capability maturity not present, score is 0.

MD BP 1: Metadata aligns with community recognized standards. Assumes for available data products. Examples include: Publicly accessible metadata procedures, element definitions, file formats, example file format content and metadata tutorials. Example metadata standards include: Dublin Core Metadata Initiative (DCMI), ISO 19115-1/19139, US Federal Geographic Data Committee (FGDC). Metadata quality performance standards are defined, quality-tracked and metrics reported

MD BP 1: Self-Assessment Capability Scoring:

- Metadata procedures (SOPs) developed and publicly accessible – 1 point
- Metadata SOPs align with a community recognized standard – 2 points
- Metadata performance standards defined, quality tracked, metrics reported – 3 points

MD BP 2: Sufficient metadata always accompanies data products. Examples include: Simple Dublin Core, set of 15 standard metadata fields designed to cover the most useful items of information. A more robust set of metadata could add relevant secondary and tertiary information such as instrument model, software and service history, instrument pre and post calibration coefficients, sample data verification information, data identification (e.g. DOI), data QA/QC information and current points of contact.

MD BP 2: Self-Assessment Capability Scoring:

- All data products come with a minimum set of metadata – 1 point
- All data products come with a robust set of metadata including data DOI * – 2 points
  * See Data Identification, Citation and Tracking best practice white paper

MD BP 3: Validation of the metadata has been performed. Examples include: Using an open source validation test to ensure the metadata in XML file format are compliant and were correctly imported into the appropriate fields.

MD BP 3: Self-Assessment Capability Scoring:

- Metadata provided is periodically manually checked for completeness – 1 point
- Metadata provided is checked using automated file validations – 2 points

MD BP 4: Metadata file creation processes are automated. Examples include: Data pulls from metadata sources such as vendor calibration coefficients files, sample data verification files, instrument service history logs, cruise deployment/recovery files. Metadata file creation process are defined and publicly available.

MD BP 4: Self-Assessment Capability Scoring:

- Significant (>50%) portion of metadata files creation is automated – 1 point
- All metadata files creation is automated, process defined and available – 2 points

4. Determine Maturity Levels

1. Add up your Metadata capability score points from above to determine your current capability maturity level.
2. Identify your aspirational Metadata maturity level by selecting a desired best practice capability score. Add up your desired capability score points to determine your aspirational capability maturity level

Initial Level	0 points
Defined Level	1-2 points
Implemented Level	3-4 points
Managing Level	5-8 points
Optimizing Level	8+ points

Interoperability Self-Assessment Tool

Steps for Using the Self-Assessment Tool

1. Review Best Practices List
2. Review Example of a completed best practice self-assessment
3. Determine Self-Assessment Capability Scoring
4. Determine Maturity Levels

1. Best Practices List

Interoperability Best Practices:

IO BP 1: Controlled vocabularies & defined ontologies are used that adhere to community standards.
IO BP 2: Community recognized and supported data format standards are used.
IO BP 3: Dataset provenance is clearly documented and available to users.
IO BP 4: Third-party systems are synchronized with evolving resources.
IO BP 5: Clear data and community software reuse statement (license) is provided.
IO BP 6: Community aligned usability design concepts are employed for user interfaces and tools.
IO BP 7: Actively participate in interoperability best practices communities

Interoperability Supporting Best Practices Previously Defined:

DI BP 1: Persistent data identifiers are associated with all data

DI BP 3: Data identifiers are maintained throughout the life cycle of the data, including when observatory data are transferred to data aggregators

DI BP 4: Information about data versioning and provenance is available and accessible

DQ BP 8: Versioning of modified datasets are available and accessible

MD BP 1: Metadata aligns with community recognized standards.

MD BP 2: Sufficient metadata always accompanies data products.

MD BP 3: Validation of the metadata has been performed.

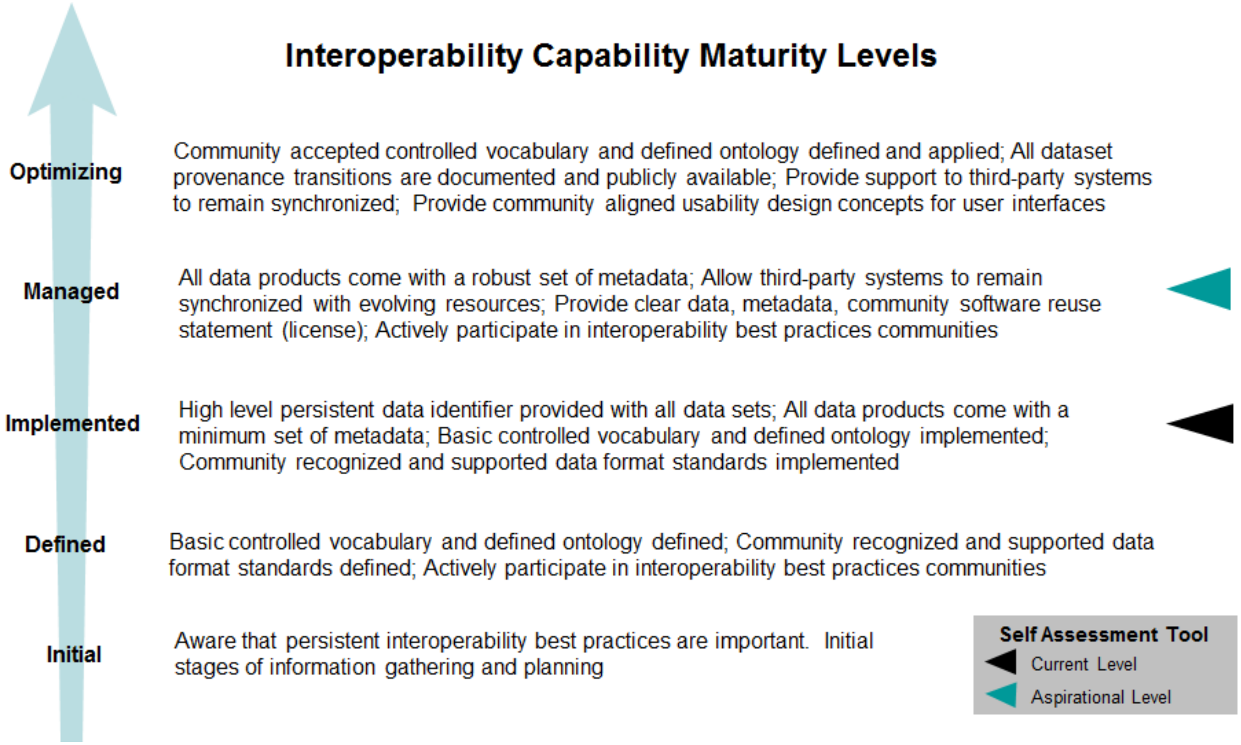
MD BP 4: Metadata file creation processes are automated.

DS BP 3: Enhanced data user services are enabled by unique user registration.

DS BP 5: Users are provided with the ability to access and execute saved community code and API functions.

2. Example Of Interoperability Best Practice Self-Assessment

The example below displays one potential combination of capabilities, which results in maturity levels for a hypothetical observatory. Each observatory will have different combinations of capabilities, which aggregate to a certain maturity level. For example, one observatory may excel at tracking and reporting data citations, whereas another may excel at providing data citation guidance. A simplified capability scoring method is described in the next step.



Interoperability Capability Maturity Levels

Optimizing Community accepted controlled vocabulary and defined ontology defined and applied; All dataset provenance transitions are documented and publicly available; Provide support to third-party systems to remain synchronized; Provide community aligned usability design concepts for user interfaces

Managed All data products come with a robust set of metadata; Allow third-party systems to remain synchronized with evolving resources; Provide clear data, metadata, community software reuse statement (license); Actively participate in interoperability best practices communities

Implemented High level persistent data identifier provided with all data sets; All data products come with a minimum set of metadata; Basic controlled vocabulary and defined ontology implemented; Community recognized and supported data format standards implemented

Defined Basic controlled vocabulary and defined ontology defined; Community recognized and supported data format standards defined; Actively participate in interoperability best practices communities

Initial Aware that persistent interoperability best practices are important. Initial stages of information gathering and planning

3. Data Support Services - Self Assessment Capability Scoring

For each best practice, determine the capability maturity score for your observatory. Only select one capability score per best practice. It is assumed each capability score is inclusive of prior score. Note: Score assumes if capability maturity not present, score is 0.

Interoperability best practices incorporate best practices previously defined in this white paper, as well as the Data Identification, Citation and Tracking white paper. These previously mentioned supporting best practices are integral to Interoperability best practices. The Self-Assessment Tool in the Appendix incorporates these. They are listed together here for ease of reference.

For the purposes of Self-Assessment scoring, only two supporting best practices are included in capability/maturity scoring (DI BP 1 and MD BP 2). They are listed first below. The remaining capability/maturity scoring list are Interoperability best practices.

DI BP 1: Persistent data identifiers are associated with all data (see Data Identification, Citation and Tracking white paper for details)

DI BP 1: Self-Assessment Capability Scoring:

- All data have high level (observatory/location) PID – 1 point

MD BP 2: Sufficient metadata always accompanies data products (see Metadata section).

MD BP 2: Self-Assessment Capability Scoring:

- All data products come with a minimum set of metadata – 1 point
- All data products come with a robust set of metadata including data DOI – 2 points

IO BP 1: Controlled vocabulary and defined ontology that adheres to community standards. Examples include: Uniform Resource Identifier (URI) that unambiguously identifies a particular resource; Defined ontology that explains relationships between a set of concepts and categories in a subject area or domain.

IO BP 1: Self-Assessment Capability Scoring:

- Basic controlled vocabulary and defined ontology defined and applied – 1 point
- Community accepted controlled vocabulary and defined ontology defined and applied – 2 points

IO BP 2: Community recognized and supported non-proprietary data format standard. Examples include: OPeNDAP: Open-source Project for a Network Data Access Protocol; SOS: Sensor Observation Service; SWE: Sensor Web Enablement; O&M: Observations and Measurements; WMS: Web Map Service; WCS: Web Coverage Service; ERDDAP: Environmental Research Division Data Access Program.

IO BP 2: Self-Assessment Capability Scoring:

- Community recognized and supported data format standards implemented – 1 point

IO BP 3: Dataset provenance is clearly documented and available to users. Examples include: Internal provenance transitions between raw data, processed data, derived data. External provenance transformations and transitions between data repositories. Descriptive material includes: System diagrams, process flows.

IO BP 3: Self-Assessment Capability Scoring:

- Most dataset provenance transitions are documented – 1 point
- All dataset provenance transitions are documented and publicly available – 2 points

IO BP 4: Allow third-party systems to remain synchronized with evolving resources. Examples include: Application programming interfaces (API) that support various data alignment needs (API) Resource Sync, OAI-PMH, Linked Data Platform (LDP), VOSpace, Open Archives Initiative Object Reuse and Exchange (OAI-ORE). See Research Data Repository Interoperability Primer DOI: 10.15497/RDA00020 and www.openarchives.org for additional examples and more details.

IO BP 4: Self-Assessment Capability Scoring:

- Allow third-party systems to remain synchronized with evolving resources – 1 point
- Provide support to third-party systems to remain synchronized – 2 points

IO BP 5: Provide clear data and community software reuse statement (license). Examples include: Providing reuse statement (license) on web landing page, in metadata, in community software tool documentation. May also include attribution and citation requirements.

IO BP 5: Self-Assessment Capability Scoring:

- Provide clear data, metadata, community software reuse statement (license) – 1 point

IO BP 6: Community aligned usability design concepts for user interfaces and tools. Examples include: When user moves easily from one user data source interface to another. Usability is the ease of use of software tool interface with effectiveness, efficiency, and satisfaction.

IO BP 6: Self-Assessment Capability Scoring:

- Community aligned usability design concepts for user interfaces – 1 point

IO BP 7: Actively participate in interoperability best practices communities. Examples include:

- Research Data Repository Interoperability WG https://www.rd-alliance.org/groups/research-data-repository-interoperability-wg.html
- Open Archives Initiative https://www.openarchives.org/
- DataONE Users Group https://www.dataone.org/dataone-users-group

IO BP 7: Self-Assessment Capability Scoring:

- Actively participate in interoperability best practices communities – 1 point

4. Determine Maturity Levels

1. Add up your Interoperability capability score points from above to determine your current capability maturity level.
2. Identify your aspirational Interoperability maturity level by selecting a desired best practice capability score. Add up your desired capability score points to determine your aspirational capability maturity level

Initial Level	0 points
Defined Level	1-2 points
Implemented Level	3-6 points
Managing Level	7-8 points
Optimizing Level	8+ points