Data Identification, Citation and Tracking Best Practices

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

Leslie M. Smith
Thomas D. Kearney
Christopher Rutherford
Kristen Yarincik

June 30, 2019

Table of Contents

Executive Summary **2**

Scope **5**

Background **5**

Methodology **7**

Best Practices Research and Synthesis 7

Results & Discussion **9**

Data Citation Principles 9

Data Identification, Citation and Tracking Best Practices 10

Accessing Data Online 11

Assigning of Data Identifiers 12

Digital Object Identifiers 13

Providing Data Citation Guidance 14

Maintaining an Identifier through the Life Cycle of the Data 15

Versioning and Provenance 16

The Issue of Continuous, Streaming In Situ Data 17

Potential Solution: Coarse Granularity of PID Application 19

Methods of Data Citation Tracking 20

Data Citation Tracking Reports 22

Incentivizing Data Identification, Citation, and Tracking 22

Measuring Success 24

Conclusion/Recommendations **24**

References **27**

APPENDIX **28**

Best Practice Self-Assessment Tool 28

Steps for Using the Self-Assessment Tool 28

1. Best Practices List 28

2. Example Of Completed Best Practice Self-Assessment 28

3. Self Assessment Capability Scoring 29

4. Determine Maturity Levels 31

Executive Summary

Data identification and citation are critical for maintaining the standard of reproducibility in science as well as furthering scientific discovery by building off of the works of others. With advances in observation technology and cyberinfrastructure there has been a significant increase in data availability, volume and complexity openly available for scientific research. Often scientists are analysing datasets far too large to be curated by traditional means, i.e. within publications as tables and figures. As such, in order to provide sufficient information for another researcher to be able to access data used in a publication, proper citation of the source of the data and a unique identifier are needed to facilitate traceability back to the data.

In order to assess the current state and technology capabilities of data identification, citation, and tracking, research was conducted including both a literature review and review of websites of nine major observing systems and nine data aggregators. Research also included interviews with 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 identification, citation and tracking 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.

BP 1: Persistent data identifiers are associated with all data products.
Data identifiers can be “internal” meaning they are part of a system created and maintained by the observatory itself, or a Persistent Identifier (PID), which is created and maintained by a third party.

We recommend the use of PIDs as they are independent of the data generator, allowing the data identifier to persist through the lifetime of the dataset. Of the PID methods currently available, we recommend the use of the DOI system as it is a widely used system within the scientific community and provides the needed functionality (e.g. URL-based), flexibility (e.g. linking DOI’s across data versions), interoperability (e.g. use of existing internal identifiers in the

DOI suffix), and standardization (e.g. oversight by Registration Agencies to ensure proper metadata).

BP 2: Guidance is provided by an observatory for data citations.
Digital data are often served on a landing page which can provide critical metadata as well as information on how to properly cite the data. Guidance for proper data citations should be provided by an observatory to ensure that the user has sufficient information to make that citation when publishing the data.

It is our recommendation that the specific citation for each dataset be provided on its landing page as well as its metadata. This ensures that a) the user will see the citation without having to search through the website and b) removes any ambiguity of what was supposed to be included in the citation.

A data citation must contain sufficient information such that someone can use the citation to access the exact dataset that was used in an analysis. Thus, in addition to standard citation information (e.g. author and title), details must be provided such as date of download and a PID that directs back to the dataset. If the identifier itself is not able to take a user back to the exact dataset, additional details must be provided in the citation, including site/sensor/product/stream information and date range.

BP 3: Data identifiers are maintained throughout the life cycle of the data, including when observatory data are transferred to data aggregators.
It is important to maintain the original identifier for a dataset when it is incorporated into a data aggregator to ensures the original dataset can be accessed.

BP 4: Data versioning and provenance information is available and accessible.
One of the goals of data citation is to provide access to a dataset across its lifetime. Within that lifetime, however, corrections may be provided to those data (versioning) or the data may be downloaded, manipulated, and posted in new forms (provenance). It is important that no matter when a citation is accessed, a user can get to the data used at the time of the original analysis or go back to the original data. We recommend that versioning and provenance of the data are provided on both the data's landing page and its metadata.

BP 5: Processes are in place to track and report data usage.
Tracking of data usage and citations refers to the documentation of the number of times the data have appeared in a scholarly publication or been referenced by subsequent publications. Current interfaces that provide "automated" citation tracking, however, are limited in their search to those citations pre-loaded into their index. We recommend using tracking based on the DOI through a Registration Agency (e.g. DataCite), or using Google Scholar or Mendeley to search for keywords.

BP 6: Data usage and citation tracking metrics are provided to the funding agency/stakeholder community.

In order for these best practices to become an established practice within the scientific community, incentives are needed. For an observatory, an incentive could include using citation tracking information to provide a metric of scholarly impact and justification for future funding. Several funding agencies already require observatories to make their data publicly available and to provide proper citation of the data, we recommend that this become a common requirement from funding agencies. Additionally, several journals provide specific requirements for how data within a publication should be cited. These requirements help make citing data a common practice similar to the citation of other scholarly works.

While the implementation of citation and identification practices are fairly common for discrete datasets, they continue to be a challenge for continuous, streaming datasets. Streaming data are difficult to cite because they are large, continuously growing, and often served in made-to-order packages. As such, it is difficult to define the citable unit of the data. Assigning an identifier and providing a citation per data request would create an unmanageable number of datasets that an observatory would then need to provide landing pages for and archive.

Instead of focusing on identifying and citing continuous data at the dataset level, we recommend the application of a PID at the observatory or site level with additional information provided in the citation and metadata to guide a user to the original data.

Scope

This white paper on Data Identification, Citation, and Tracking examines the history and drivers for data citation, identifies current industry best practices, and provides recommendations.

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 identification,citation and tracking capabilities and maturity level.This tool can also be used to identify steps to achieve the next aspirational level. See Appendix.

Background

Scientific data and information has become increasingly digital, from the online archiving of individual “benchtop science” datasets to the digital generation of open access streaming *in situ* observations. This new age of digital science opens doors for greater collaboration, peer-review, and transparency, but also adds complexity to the issue of how to properly cite, identify, and track the use of these data (CODATA-ICSTI Task Group on Data Citation Standards and Practices 2013). No longer are scientists primarily publishing from their own datasets found in “binders of data” on shelves, 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.

A data citation is “a reference to data for the purpose of credit attribution and facilitation of access to the data” (CODATA-ICSTI Task Group on Data Citation Standards and Practices 2013). In our example of “binders of data”, a publication traditionally served as the citation for those data as both existed as core elements of the other. That is to say that the data presented in a paper were sufficiently finite that they could be displayed within the manuscript’s tables and figures in such a way that others could pull the data from the manuscript for re-use. The modern extension of this is to accompany publications with digital supplemental materials to more fully capture the data used in a publication. Supplemental materials can take the form of additional graphs, tables, or explanations of processing methods. Supplemental materials, however, do not meet the criteria for proper data citation as the data are archived on the publishers servers and are often only accessible by subscription (CODATA-ICSTI Task Group on Data Citation Standards and Practices 2013).

Regardless of the efficacy of data citation through publication or supplemental materials, the core issue with online data archives and streaming data is that datasets are now too large and complex to be captured by these traditional means alone (CODATA-ICSTI Task Group on Data Citation Standards and Practices 2013). New citation standards and methods are

> *“The use of published digital data, like the use of digitally published literature, depends upon the ability to identify, authenticate, locate, access, and interpret them.” (CODATA-ICSTI 2013)*

needed for large datasets and analyses that pool together datasets from different sources to fulfill the goals of credit attribution and facilitation of access. With this in mind, two key drivers have been identified for establishing and adhering to best practices of data identification, citation, and tracking

Key Driver #1: The proper citation and identification of data is critical to preserving scientific integrity and facilitating scientific progress (CODATA-ICSTI Task Group on Data Citation Standards and Practices 2013). Without proper access to data used in an analysis there is no way for another scientist to reproduce the research and verify conclusions. Access to data also facilitates future access and reuse of the data for additional analyses to extend the conclusions of one analysis.

Key Driver #2: Data citation is important to observatories generating large streaming datasets as it allows for traceability of an observatories scholarly impact which can be used as justification for continued funding (CODATA-ICSTI Task Group on Data Citation Standards and Practices 2013). Observatories can also use citation information if the need arises to determine which of their data are being used, providing justification to scale back or enhance portions of the observatory.

Data citation is a challenge for both those generating data and those using the data. Not only does there need to be a standardized way for scientists to attribute the use of data within a publication, but the data generator needs to provide a standard way to generate an identifier and citation information to enable this citation.

Though the scientific community has widely adopted established mechanisms for the citation of scholarly works, the citation of data has yet to be broadly implemented and has lagged behind the pace of digital data development (CODATA-ICSTI Task Group on Data Citation Standards and Practices 2013, Bourne et al. 2011). This lag in adoption may be an issue of cyberinfrastructure and technical capabilities on the side of the data generator, or an issue of social norms and a lack of implementation by the data user (CODATA-ICSTI Task Group on Data Citation Standards and Practices 2013).

This white paper presents a synthesis of industry best practices in data identification, citation, and tracking, 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 accessing data online, assigning persistent data identifiers, providing data citation guidance, maintaining identifiers when data are transferred to aggregators, providing versioning and provenance information, and the ongoing challenge of assigning persistent data identifiers to continuous streaming data.

Methodology

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

Best Practices Research and Synthesis

Data identification, citation, and tracking 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 two 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 research:

- Determine drivers for data identification, citation, and usage tracking
- Determine high level requirements for data identification and usage tracking
- Determine current state of industry capabilities to meet drivers

Figure 1 provided additional guidance for the research.

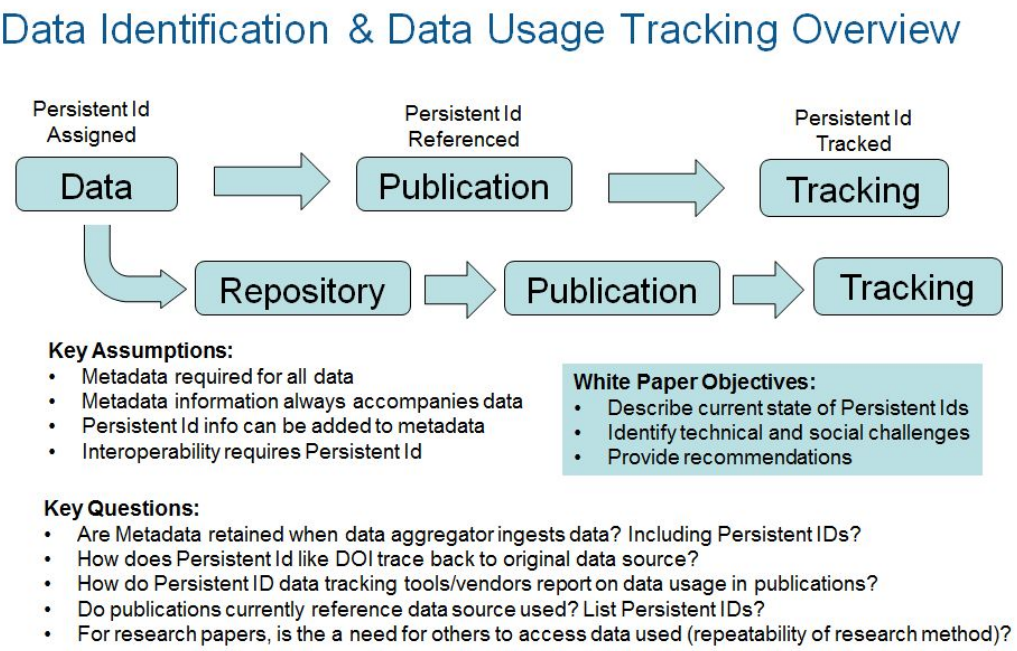

Figure 1. Data Identification & Data Usage Tracking Overview

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 level. It also provides context based on current industry wide best practice maturity levels.

The best practice tool enables an existing or new organization to assess their current data identification, citation, and tracking capabilities and maturity level. This tool can also be used to identify steps to achieve the next aspirational level. The best practice self-assessment tool 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 levels. 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 Appendix.

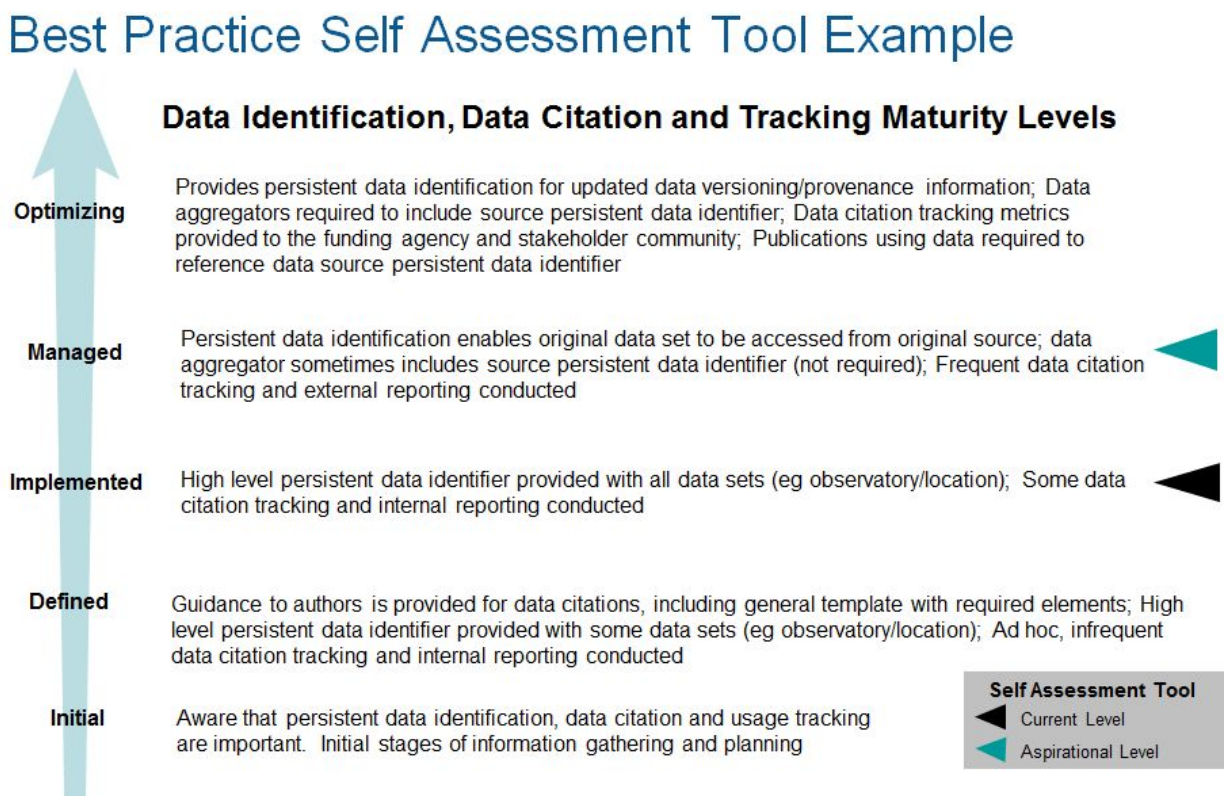

Figure 2. Best Practice Self Assessment Tool Example

Results & Discussion

> *"How much slower would scientific progress be if the near universal standards for scholarly citation of articles and books had never been developed....How many discoveries would never have been made if the titles of books and articles in libraries changed unpredictably, with no link back to the old title; if printed works existed in different libraries under different titles....How much less would we know...if the references at the back of most articles and books were replaced with casual mentions, in varying, unpredictable, and incomplete formats, of only a few of the works relied on?" (Altman & King 2006)*

Data Citation Principles

The hypothetical questions brought up by Altman and King (2006) can seem absurd when thinking about written scholarly works, but are all too accurate, even over a decade later, for the

treatment of the citation of data by the scientific community. There continues to be a clear need to establish best practices in data citation.

In 2010, an international task group was created to address the issue of creating practical and consistent data citation standards. This group was created as a collaboration between the Committee on Data for Science and Technology (CODATA) and the International Council for Scientific and Technical Information (ICSTI). As part of this effort, the CODATA-ICSTI Task Group on Data Citation Standards and Practices collaborated with the U.S. National Academies of Sciences to host an international workshop (National Research Council, 2012). Since that workshop, the Task Group set out to assess the "current state of practice for data citation and attribution, noting emerging trends, successes, and challenges." As part of that effort, the Task Group created a set of ten "first principles" for data citation (CODATA-ICSTI Task Group on Data Citation Standards and Practices 2013). These principles have been embraced by the international community as they can be seen clearly echoed in the Force 11 Joint Declaration of Data Citation Principles (Data Citation Synthesis Group 2014), which has itself been endorsed by scientific societies (e.g. 2015 AGU Position Statement[1]).

These data citation principles are enumerated in their original form below:

1. **Status of Data:** Data citations should be accorded the same importance in the scholarly record as the citation of other objects.
2. **Attribution:** Citations should facilitate giving scholarly credit and legal attribution to all parties responsible for those data.
3. **Persistence:** Citations should be as durable as the cited objects.
4. **Access:** Citations should facilitate access both to the data themselves and to such associated metadata and documentation as are necessary for both humans and machines to make informed use of the referenced data.
5. **Discovery:** Citations should support the discovery of data and their documentation.
6. **Provenance:** Citations should facilitate the establishment of provenance of data.
7. **Granularity:** Citations should support the finest-grained description necessary to identify the data.
8. **Verifiability:** Citations should contain information sufficient to identify the data unambiguously.
9. **Metadata Standards:** Citations should employ widely accepted metadata standards.
10. **Flexibility:** Citation methods should be sufficiently flexible to accommodate the variant practices among communities but should not differ so much that they compromise interoperability of data across communities.

Data Identification, Citation and Tracking Best Practices

From the secondary and primary research conducted, two key concepts emerged:

[1] https://sciencepolicy.agu.org/files/2013/07/AGU-Data-Position-Statement-Final-2015.pdf

1. Data citations must be ubiquitous, standardized, and persist through the lifetime of the dataset;
2. Data citations must point to a dataset in such a way that the data can be directly accessed and any derivations or manipulations of that dataset clearly understood.

Building upon these two concepts, six best practices for Data Identification, Citation and Tracking were synthesized from the secondary and primary research.

- **BP 1:** Persistent data identifiers are associated with all data products
- **BP 2:** Guidance is provided by an observatory for data citations
- **BP 3:** Data identifiers are maintained throughout the life cycle of the data, including when observatory data are transferred to data aggregators
- **BP 4:** Data versioning and provenance information is available and accessible
- **BP 5:** Processes are in place to track and report data usage
- **BP 6:** Data usage and citation tracking metrics are provided to the funding agency/stakeholder community

Each of these best practices are discussed in detail, accompanied by context and literature references in the remainder of the white paper. A best practice Self-Assessment Tool is presented in the Appendix.

Accessing Data Online

Data can be accessed online through a variety of modes, e.g. machine-to-machine interface, Erdapp server, thredds server, or a graphical user interface (GUI) created by the observatory or data aggregator. In many cases, the GUI provides a “landing page” for datasets that contains critical metadata and information about the dataset as well as access points to download data from the other modes (e.g. Erdapp, etc).

For the purposes of this paper, metadata is defined as the “documentation of data [that] serves the purpose of making data discoverable, usable, and understandable[2]”. The International Organization for Standardization (ISO) has developed a standard for metadata (ISO 19115). Metadata is a critical element for every level of data identification, citation, and tracking and will be discussed throughout this paper. As it is such a key element, it is critical for metadata to be included on the data landing page so when accessed there is no ambiguity of what those data represent.

The landing page can be thought of as the Universal Resource Locator (URL) designation for the dataset. Simply providing someone with this URL should give them all of the information needed to access and understand the data. Typical information on a landing page includes:

[2] https://www.ncddc.noaa.gov/metadata-standards/

dataset title, description, author, geographic location (either via map or coordinates), data collection period, and information or links for data download.

Based on the secondary research of 18 observatories and aggregators, 16 serve data and were examined for this paper. Of these 16, 11 had landing pages for their data. Though landing pages are frequently used within the community, the information provided on each is not standardized. We suggest that landing pages, in addition to general description information, also include a data identifier, suggested citation, and versioning/provenance information.

Assigning of Data Identifiers

As noted earlier in the paper, prior to open access, digitally archived datasets, the identification of a dataset was the paper within which the data were first presented. As digital archiving and open access online datasets have become a prominent aspect of modern science, novel methods of dataset identification have become necessary.

BP 1: Persistent Identifiers are associated with all data products.

In some cases, data are identified with codes provided internally (Ariani et al., 2014), for example, alphanumeric sequences to denote specific data products or data streams on certain assets within a program. Or internal accession numbers that can be entered into the search field within a database to pull up a dataset. Though internally effective, these identifiers do not meet the best practice needs for proper data citation as they are not ubiquitous, standardized, and persistent. Internal identifiers are only functional in the short term as long as they are maintained by the observatory.

Persistent Identifiers (PIDs) are a key element of a good citation (Ariani et al., 2014, CODATA-ICSTI Task Group on Data Citation Standards and Practices 2013, Gries et al, 2018). PIDs refer to a unique, web-compatible alphanumeric code assigned to a data set that is able to be preserved long-term (CODATA-ICSTI Task Group on Data Citation Standards and Practices 2013, Gries et al, 2018). PIDs should directly link the user to the dataset as well as any associated metadata (Ariani et al., 2014, CODATA-ICSTI Task Group on Data Citation Standards and Practices 2013, Gries et al, 2018).

Perhaps the most well known PID in the observational community is the Digital Object Identifiers (DOI). The DOI system was first launched in 2000 (International DOI Foundation 2012) and was accepted as an ISO Standard in 2010 (ISO 26324-2012)[3]. The DOI system is governed by a broader International DOI Foundation which is made up of a federation of Registration Agencies that provide the “central hub” for DOI system operation. These Registration Agencies ensure that DOI names are created using a consistent set of standards and that adequate metadata are

[3] https://www.iso.org/standard/43506.html

provided with each submission (International DOI Foundation 2012). There is a license fee associated with assigning a DOI name, but once assigned, the DOI name can be freely and openly utilized.

Prior to the development of the DOI system in the mid-1990s, several other data identifiers were created. These included the 1) Uniform Resource Name (URN), perhaps best known for the ISBN identifier for library books, 2) Archival Resource Key (ARK)[4] managed by the California Digital Library, and 3) the Handle[5] operated by the Corporation for National Research Initiatives (CNRI), which provides the technical basis for what has now become the DOI system (International DOI Foundation 2012).

An alternative system to the DOI system is the Universally Unique Identifier (UUID), sometimes also referred to as the Globally Unique Identifier (GUID). Similar to DOI names, UUIDs are persistent and unique identifiers (IETF RFC 4122). Unlike the DOI system, however, the generation of UUIDs do not depend on a central registration authority. Anyone can generate a UUID as long as they follow the standard methods (ISO/IEC 9834-8:2005) generated by the Internet Engineering Task Force[6].

Of the 16 observatories/aggregators used in this analysis, three provided no identifiers, eight exclusively provided an internal identifier, and five provided a PID either by itself or in addition to an internal identifier. An additional group provided PIDs for a small subset of their data that could be downloaded as discrete datasets, but not to the bulk of their data which was streaming. Of the groups that utilized PIDs, four used DOIs and one used UUIDs.

Digital Object Identifiers

Of the PID methods currently available, we recommend the use of the Digital Object Identifiers (DOI) system because it is widely used within the scientific community and provides the needed 1) functionality, 2) flexibility, 3) interoperability, and 4) standardization to satisfy the citation and identification best practices.

1. Functionality. When assigning a DOI and registering it through a Registration Agency, three pieces of information must be provided - name, URL, and metadata. In this way, DOI names are permanently assigned to a dataset, or another object, with a link that will take a user to the dataset as well as information about the dataset (ISO 26324). The URL should direct a user to a landing page where the object referred to by the DOI can be accessed. This object could be a specific cruise, an observatory, a dataset, a glider mission, a publication, or any number of objects. This allows information about the dataset to be updated throughout time, but the link to the dataset and the DOI name will remain unchanged (ISO 26324).

[4] http://n2t.net/e/ark_ids.html
[5] http://www.handle.net/
[6] https://www.ietf.org/

2. Flexibility. DOI names have the capability of being linked together. As such, different versions of the data could have different DOIs which would then be linked together, allowing a user to trace backward or forward through the versions.

3. Interoperability. A key element of DOI names is that they do not need to replace an existing internal identifier; rather, they can be used in conjunction with the identifier. For example, an internal identifier can be incorporated into the suffix of a DOI (ISO 26324). For interoperability, it is recommended that DOI suffixes contain the same alphanumeric sequence used internally to identify the object (observatory, site, dataset, etc). For example, if a DOI was associated with salinity data that internally was referred to as "SAL01234," that same code should be incorporated into the suffix of the DOI.

> *"An identifier such as a DOI is of no value without some related metadata describing what it is that is being identified." DOI Handbook 4.3*

4. Standardization. Metadata is a critical component of the registration process as it is this information that identifies the object as a separate entity (International DOI Foundation 2012). Typical information in the metadata include "names, identifiers, descriptions, types, classifications, locations, times, measurements, relationships" (International DOI Foundation 2012). Registration Agencies ensure that metadata provided meets DOI requirements prior to assigning a DOI name.

Additionally, as PIDs are applied to data by a third party they move with the dataset regardless of whether the dataset changes servers or owners during its lifetime. Though being a third party assigned identifier is a benefit for longevity if the data originator changes, it does raise the concern of what happens if that third party registry is not maintained (Ariani et al., 2014). Because the DOI system is governed by an international federation of independent agencies, their viability does not depend on one entity alone, helping to guarantee its persistence.

Providing Data Citation Guidance

Publications are the primary mechanism to disseminate research to the scientific community and archive it within the scientific record (Bourne et al. 2011). As such, it is important to provide mechanisms to incorporate citations of datasets within publications so that they are similarly disseminated and archived.

> *"The primary purpose of citation has been to support an argument with evidence, though over the years it has also become a mechanism for attribution, discovery, quality assurance, and provenance." (CODATA-ICSTI 2013)*

> *BP2: Guidance is provided by an observatory for data citation*

In order for researchers to properly cite a data source, it is important that they are provided with the information needed to create that

citation. Data citations should include enough information for a user to trace from the citation to the data source such that the data can be reused (CODATA-ICSTI Task Group on Data Citation Standards and Practices 2013). To ensure that datasets are properly cited, it is advised that citations for each dataset be provided with the data rather than generally on the observatory's website.

Data citation information also needs to be included in the metadata (CODATA-ICSTI Task Group on Data Citation Standards and Practices 2013). For example, within the ISO 19115 standard, citation information goes within the CI_Identification field[7]. The identifier associated with the data, either the internal identifier or PID, is added to the MD_Identifier field. It is important to include both so this information stays with the data, independent of the researcher re-visiting an observatory's website to find the reference information.

Based on the secondary research of 18 observatories and aggregators, 16 serve data and were examined for this paper. Four of these did not provide any citation guidance (only data are provided), three provided overarching citation guidance, and nine provided specific citation information with individual datasets, either on the landing page of the dataset or in the metadata. Typical information provided in a citation included: Name of data originator (e.g., the observatory), dataset or data product name/title, URL of the data landing page, date of download, data distributor name (if different from originator), and persistent identifier. In cases where the name of a dataset does not provide sufficient granularity to determine what data were downloaded, additional information is required such as a site ID and date range of data downloaded. In some cases, year information is provided, either as the year the metadata were published or the year a dataset was added to an aggregator.

One interesting addition to this guidance from an observatory was guidance for instances when numerous (greater than six) datasets from one observatory are used in a publication. In this case, they suggest using a more general citation accompanied by a table with information on each dataset. With observatories serving hundreds of data products and promoting large-scale interdisciplinary research, this is a necessary caveat to provide so that the work of citing the data does not become onerous on the researcher.

Maintaining an Identifier through the Life Cycle of the Data

The lifespan of a dataset is longer than the research project that created it, and often longer than the lifespan of the researchers themselves. As such, it is important to preserve access to data throughout its whole life cycle so it can continue to be used (CODATA-ICSTI Task Group on Data Citation Standards and Practices 2013). This can be accomplished by properly citing and assigning PIDs to data in such a way that facilitates their portability.

[7] https://geo-ide.noaa.gov/wiki/index.php?title=ISO_Citations

A key transition in the life of a dataset is the handoff of the dataset from data generator to data aggregator to user. It is important that, as the data changes hands, information of metadata and provence as well as access to the original data are retained (Gries et al. 2018). This can be accomplished by maintaining the original data identifier or linking identifiers of different versions across each exchange. It should be noted that the onus for maintaining the identifier when data are incorporated into an aggregator is both on the part of the data generator and the data aggregator. The data generator needs to first provide the identifier, but the aggregator also needs to have the capability to receive and display the identifier.

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

Eleven of the sixteen organizations whose websites were reviewed for this paper serve data generated from external sources, making them data aggregators. Of these, only four retain the original identifiers from the source data; three of these retain internal identifiers and one retains DOI names (when present). Two additional organizations retain the URL for the original data, though in some cases the URL simply directs back to the main project landing page. It should be noted that the lack of ID retainment in many cases may not be due to the data aggregator, but rather the original source did not provide identification information.

Data aggregators can play a key role in ensuring that data receive identifiers and proper citation instructions if they require identifiers for all data they serve, or provide a mechanism for identifier creation. In the case of one of the aggregators that did not display identifiers with the data, identifiers were not a required field on the data submission form to add data to the site. Conversely, with other data aggregators, in order to ensure that all data received an identifier, they either applied their own internal identifier (that could exist in conjunction with a DOI if present) or assigned a DOI.

Versioning and Provenance

Going back to the “binders of data” metaphor, there was little need to note versioning of these data as final data were printed in hard copy, published, put on a shelf, and sat unchanged collecting dust. In the new model of large, online datasets with data, in some cases, streaming real-time, versioning of data is a necessity (CODATA-ICSTI Task Group on Data Citation Standards and Practices 2013). In this new reality, if someone were to follow a specific citation to a dataset they still might end up with different data than was used in the original analysis as the data now being served online was updated in some way (e.g., a new, more sophisticated processing algorithm applied, an error in a calibration coefficient). It is for this reason that a date of download must be provided in a citation and information on data versioning and provenance must be included

BP4: Data versioning and provenance information is available and accessible.

with each dataset, preferably both on the dataset landing page and within the metadata.

The extent of information to include about data changes within each version is a challenge for the observatory to determine. The overall goal of providing information on versioning is so that someone tracing a citation back to the data is able to replicate the science and for a researcher to understand how the data was changed after its original download. As such, the level of detail that needs to be included will likely depend on the data product, the types of changes made, and the needs of that specific user community.

In this paper, versioning of data 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. Provenance refers to changes made by external users and is defined as “the chain of ownership of an object and the history of transformations applied to it” (CODATA-ICSTI Task Group on Data Citation Standards and Practices 2013). For example, a data aggregator may apply a conversion to the dataset, or apply their own corrections to it upon integration into their system.

> *“Citations should include sufficient fixity and other administrative provenance metadata to verify that a data object later accessed is equivalent to the data object originally cited.” (CODATA-ICSTI 2013)*

Versioning can be handled in a couple ways. Each new version of data could receive their own identifier and therefore their own landing page. Alternatively, the same identifier could be used across versions as long as additional information is provided to unambiguously identify which version of data were used and the history of data versions is described.

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 prior to that date. This is clearly an area that could see significant value from enhancement.

The Issue of Continuous, Streaming In Situ Data

Within the environmental observatory community, challenges continue in the process of implementing data citation practices, namely PIDs, for continuous, streaming data.

One of the inherent requirements of data citation is that datasets must be available to the user in a citable unit (Gries, et al 2018). That is to say the citation must represent a discrete, coherent collection of data that can be identified and then accessed. Another way to think of this

is of the dataset as a file or a zipped folder full of files that can be downloaded as unit. Within the DOI system, each DOI name can only refer to one object and it is recommended that each object only be given one DOI name (International DOI Foundation 2012). This establishes a one-to-one relationship between an object and a DOI name (CODATA-ICSTI Task Group on Data Citation Standards and Practices 2013). These two elements make adding identifiers to continuous, streaming *in situ* data extremely challenging as the data "package" is continually changing with new values added every second and chunks of the data are generated upon request for download.

Long-term, continuous observations are often 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. While this is a great feature for the data user, particularly with complicated data products, it creates a situation where it is impossible to guarantee a one-to-one relationship between data and a DOI (or other PID). For example, User #1 could download a stream of data from 5am Tuesday to 6am Monday. User #2 could download from the same data product, but only from 7am Tuesday to 3am Monday. The second download would be fully encompassed in the first but receive a separate PID. Or say, User #1 in 2011 downloaded a stream of data that encompassed January 1, 2010 to December 31 2010, then in 2017 User #2 downloaded that same data set. They would receive separate PIDs for each download even though the data were identical.

Another issue with streaming data is one of the creation of an unmanageable number of identifiers. Take the example of a machine-to-machine interface wherein users can set scripts to access observatory data and automatically pull data from the system at set intervals. Say a script were to pull data every hour on the hour for a year, that would add up to almost 8,800 separate data requests. If each request received a PID that would also mean that one user would now how 8,800 PIDs associated with that one data stream they were downloading. How would an observatory manage that number of PIDs? Would the user then need to apply all 8,800 PIDs in a citation of those data? Currently 175 million DOIs have been assigned (DOI FAQ page). This seems like a very large number, but some observatories can receive up to 30 million data requests in a quarter across a machine-to-machine interface.

Aside from the issues of duplicity, redundancy, and the sheer volume of PIDs that could be created, there is the added complication of how to provide access to the datasets represented by the PIDs. One of the key requirements for proper citation is that the citation must provide access to that data (CODATA-ICSTI Task Group on Data Citation Standards and Practices 2013), but if datasets are generated on a per request basis it is not technically feasible to store a copy of that data and maintain access to each of those requests within the observatory's cyberinfrastructure. In fact, that defeats much of the purpose of the made-to-order data generation. The observatory would need to store a very large amount of duplicate data in order

to make this possible, and have a sophisticated interface for users to then access each of those datasets via individual landing pages.

As discussed previously, five of the groups reviewed used persistent identifiers with all of their data. Of those groups, three exclusively served discrete datasets and were able to create a DOI for each of the dataset packages that a user could download. The two groups with continuous datasets chose different methods to create citable units of their data. In one case, an observatory took monthly “snapshots” of the entire observatory and served those data as individual NETCDF files each with its own DOI. This organization also provided the option for utilizing one DOI to cite the entire observatory. In the other case, the aggregator assigned a DOI to each observatory network it served. Each network DOI directed users to a landing page where data from that network could then be accessed. In a couple cases, organizations with continuous data also served some discrete datasets. These were either packets of preliminary data or a derived data packet. In one case DOIs were provided to the discrete dataset pages; in the other UUIDs were used.

Several groups with continuous data also made a note on their webpages that they were currently working on applying DOIs to their datasets. One of these observatories describes a plan to eventually apply UUIDs by creating packages of discrete datasets out of their continuous datastream. Thus, the desire seems to be there, but the implementation is challenging. As a short term solution, these observatories are instead using detailed citations with internal data identifiers and observatory URLs.

Potential Solution: Coarse Granularity of PID Application

PIDs must support the finest-level of granularity necessary to effectively identify the data (CODATA-ICSTI Task Group on Data Citation Standards and Practices 2013). For example, the identifier may be as broad as one identifier for an observatory as a whole, or as narrow as a unique identifier for each unique set of data downloaded. In the case of a broadly defined persistent identifier, additional data citation notation is required in order to unambiguously identify the portion of data used in a publication or analysis (CODATA-ICSTI Task Group on Data Citation Standards and Practices 2013).

As discussed in the previous section, if a PID were assigned to every dataset downloaded from an observatory with streaming data -- i.e. each made-to-order dataset -- there would be the potential need for an observatory to have PIDs on the order millions. Though it is possible to create an infinite number of DOI names, it is not technically practical to maintain them -- e.g. landing pages and access to data -- nor is it practical for a user to have to cite potentially thousands of datasets in their publication.

With the added complexity that comes from associating persistent identifiers with streaming datasets, it is important to seek alternative methods for citation than the traditional one identifier for one dataset download package. We suggest that this solution comes from determining the

appropriate level of granularity needed for a user to access a dataset, and an observatory to maintain that dataset. If the DOI name is assigned too broadly and is not accompanied by ancillary citation information, a user would not be able to access the subset of data used in an analysis. If the DOI name is assigned at too fine a scale then there is the risk of producing more DOI names than can be effectively managed and tracked. In order to tackle issues of granularity, NOAA has created a list of questions for an entity (e.g., an observatory or data aggregator) to ask in order to determine the appropriate level of granularity to assign[8]. These questions are listed below with slight re-wording for the observatory/data aggregator application.

- Can the [observatory] maintain a landing page for each DOI assigned?
- How does the [data aggregator] accept and manage submitted datasets?
- How does the [observatory or data aggregator] expose datasets to users?
- What does the user community consider a complete and sufficient citation in their field?
- Will the granularity support a reasonably accurate starting point for subsequent research?
- What do you want users to cite?
- What do you want to get citation metrics on?

Several large organizations that serve continuous data have explored these options and have concluded that to apply DOI names to their continuous data, they must be applied at coarse levels of granularity. For example, NOAA recommends that DOI names be assigned at the level of an observing network[9] for continuous data streams from buoys. Similarly, the International Federation of Digital Seismograph Networks (FDSN) has implemented a system where DOI names are assigned at the level of a seismic network[10] (Evans et al., 2015).

Not included in the NOAA list are gliders or other autonomous vehicles. As these often conduct missions in differing locations between servicing intervals depending on the needs of the observatory - e.g. glider #547 is in the North Atlantic from January to March, and then moved to a site in the NE Pacific for September to November - they must be treated differently than fixed assets. We recommend that instead of associating an identifier with a specific glider, identifiers should be associated with specific glider missions, similar to a vessel survey in the NOAA guidance.

Using the DOI per observatory approach is a very coarse approach to assigning identifiers. The primary limiting factor for DOI application is whether the observatory can maintain the necessary landing pages and data access needed to associate the DOI. As such, it is conceivable that an observatory could assign DOI's at the array, site, or mooring level depending on their maintenance capacity.

[8] https://geo-ide.noaa.gov/wiki/index.php?title=Data_Citation_Granularity
[9] https://geo-ide.noaa.gov/wiki/index.php?title=Data_Citation_Granularity
[10] http://www.fdsn.org/services/doi/

Methods of Data Citation Tracking

Tracking of data citations refers to the documentation of the number of times the data has appeared in a scholarly publication.

Within the scientific community, there are only a few interfaces that automatically provide this tracking. Most well known may be the Web of Science Data Citation Index (DCI), originally created by Thomson Reuters and currently maintained by Clarivate Analytics.[11] There are limitations with this interface, however, as it only includes research data from recognized data repositories; currently only 350 repositories are included.[12] An additional shortcoming is that it is a subscription based service. As such, though it is likely for most large universities, institutions, and management agencies to have subscriptions, individuals and small non-profits likely do not.

The OpenCitations Index provides an open access alternative to the Web of Science DCI.[13] Similar to the DCI, citations must be manually added to the index by its creators in order to be accessed. As of February 24, 2019, Open Citations Index had ingested over 300,000 bibliographic resources which contained almost 14 million citation links and 7.5 million cited resources. Though these are impressive numbers, this is still not a comprehensive search tool.

If an observatory applies DOI names through DataCite, the most common Registration Agency used for scientific data, it is possible to track the occurrence of that DOI using tracking provided by DataCite.[14] While this is a free service, it is limited to only citations that employ DOI names.

Though critical, the development of tracking software has lagged behind the increased use of persistent identifiers, which as noted has itself lagged behind the surge of big data in scientific research. As such, for many organizations, the tracking of data uses in scholarly works is accomplished via manual means. Independent of a citation algorithm, citation search engines - e.g. Google Scholar, Mendeley, Web of Science - can be used by observatories to search via key terms for the use of their data in publications (CODATA-ICSTI Task Group on Data Citation Standards and Practices 2013).

BP 5: Processes are in place to track and report data usage

These tracking methods all rely on the data user following through on their obligation to include a citation of the data, either through an identifier, or following the proper citation format described by an observatory, as well as on aggregators maintaining identifier and citation information when data are migrated to their system. Research into citation success rates does not paint a promising picture, however, with studies finding that authors often fail to cite the data they use, or when focusing on one dataset that articles citing that dataset provide

[11] https://clarivate.com/products/web-of-science/web-science-form/data-citation-index/
[12] https://clarivate.com/products/web-of-science/web-science-form/data-citation-index/
[13] http://opencitations.net/
[14] https://stats.datacite.org/

incomplete citation details (CODATA-ICSTI Task Group on Data Citation Standards and Practices 2013). This further emphasizes the need to incentivize citation and utilize methods developed by the community to ensure widespread adaptation (CODATA-ICSTI Task Group on Data Citation Standards and Practices 2013).

Seven of the observatories/aggregators surveyed tracked publications. Of those, two provided reference lists outdated by a couple of years, and the other five provided lists that appeared current and consistently maintained. In the case of observatories/aggregators with discrete datasets, the tracking is on a dataset by dataset basis. For smaller projects and observatories with continuous data, tracking is at the observatory level with one master list maintained. In many cases the method of tracking was not provided, but when it was it was either tracking via Google Scholar or Mendeley.

Though tracking publications is the most common method used amongst observatories, it should be noted that depending on the audience of the observatory data products (managers, policy makers, educators, etc) solely tracking scientific publication may leave out significant amounts of grey literature and other resources generated using the data. Additionally, moving beyond publications, Altmetric[15] is a tool that searches for any reference to a project, grant, or dataset that has a URL associated with it. Whether that reference is in a publication, a blog post, or a Twitter post. This may prove to be the future of data citation tracking, but it may not be worth the effort at the moment as it is unlikely to be a fast social switch in the relative value of a scholarly publication compared to a tweet or a blog post.

Data Citation Tracking Reports

Data citation tracking reports are the synthesis and analysis of the results of data tracking. It is more than simply a list of publications utilizing a particular dataset, but rather provides insight into the trends of citation. Reports could take many forms, either as part of an observatory's annual report, a summary description or graph on an observatory's website, or a report generated by a tracking mechanism such as Google Scholar that is linked to from the observatory's website.

While several of the observatories track data citations by presenting online lists of publications, few synthesized that information into reports. Only in one case were these data included in a report that was available online. In two other cases, publication information was clearly synthesized and put on the website in the form of an informal "report". It is not clear if more reporting of these metrics occurred behind the scenes.

Incentivizing Data Identification, Citation, and Tracking

Data citation, though slow in implementation, is core to the facilitation of reproducible scholarship by making data accessible (Data Citation Synthesis Group 2014). Reproducible

[15] https://www.altmetric.com/

> *"In particular, to obtain the benefits that networked knowledge promises, we have to put in place reward systems that encourage scholars and researchers to participate and contribute." (Bourne et al. 2011)*

scholarship not only furthers additional scientific discovery, it also provides a mechanism for accountability and transparency (NRC 2012) facilitating an enhanced peer-review process in which someone could replicate the work to check the findings (CODATA-ICSTI Task Group on Data Citation Standards and Practices 2013).

Tracking scientific data citations is critical for sustaining these best practices as it provides an incentive for that citation. Tracking data citations could show the impact of a dataset, effectively assigning datasets with a similar currency to publications. This information could then be viewed similarly to publications in terms of tenure and promotion decisions (very loosely NRC 2012 p207) or for future funding sources. There is a need to create "new methods and metrics for evaluating quality and impact that extend beyond traditional print outputs to embrace the new technologies." (Bourne et al. 2011)

> *"Citations support a research infrastructure to provide the necessary recognition and reward of data work, in addition to providing attribution detail, facilitating future access, and fostering cross-collaboration and investigation" (CODATA-ICSTI 2013)*

With increased focus on the need for transparency of federally funded research, many agencies have new requirements that data are made accessible and properly cited (Socha 2013, NRC 2012). Providing citations and identifiers can provide this proof of open data sharing (NRC 2012, CODATA-ICSTI Task Group on Data Citation Standards and Practices 2013). Additionally, funding agencies may require usage statistics to justify future funding. For example, NASA archives have a senior review every 2-3 years to assess the extent that the data are being used in publications (NRC 2012).

> *BP 6: Data usage and citation tracking metrics are provided to the funding agency and stakeholder community*

While it is clear that several funding agencies require that data be made public from research they are funding, the requirements are less specific in terms of citations and PID methods. From the organizations surveyed, there was no clear connection between citation guidance and funding agency, with the exception of the organizations examined that were funding entities themselves that generated their own IDs and citation information.

In addition to clear guidance, there must be some mode of incentive or requirement to ensure data are actually cited. This can come from requirements by

funding organizations or by requirements from publications for proper data citation, much the same way a journal requires proper citation of scholarly works. For example, the Coalition on Publishing Data in the Earth and Space Sciences (COPDESS) made up of major publishers and repositories have committed to requiring that all data to be included in publications be made publicly available (Stall 2017). However, in practice the application of this commitment has been uneven within the community as there is no standard method to define how to include the data (Stall 2017).

A survey of well known scientific journals (6) yielded mixed results. Across the board, journals pressed for all data in the manuscript to be made sufficiently accessible such that others may be able to repeat the work. However, in most cases the requirements were either vague on how authors needed to do this, or relied on the use of supplemental materials. Two of the journals surveyed requested the addition of DOIs for publicly accessible data, but only one journal provided information about what else should be included in that citation. For that journal, the requirements for citation included author(s), title, publisher (repository name), identifier (DOI), and date of deposition.

In order for data citation to become more prevalent, citation requirements must be more clearly articulated and enforced by journals. In turn, data providers must provide adequate information to users such that they are able to meet these citation requirements by journals. So far, observatories that are providing citation guidance are in-line with what journals that provide specific guidance require.

Measuring Success

Quantifying effectiveness of data identification and citation comes through traditional tracking of the number of scientific publications acknowledging the data and their associated citation factors, such as the *h* index (Hirsh 2005). These metrics signify when an observatory or findings based on observatory data were used to help further scientific discovery. If an observatory has successfully laid the foundation for their data to be identified and cited by users then they will see higher numbers, and more easily trackable citations of those data. Note that these metrics also enable Community Engagement and Science Impact Performance Metrics as the currency of science is based on publications and citations.

Data usage and citation tracking metrics are increasingly being requested by the funding agency/stakeholder community. Funding agencies more commonly require usage statistics to justify future funding. For example, NASA archives have a senior review every 2-3 years to assess the extent that the data are being used in publications (NRC 2012). Providing data usage reports and metrics provides documentation of data tracking and usage.

See the Observatory Performance Metrics Best Practice white paper for additional science impact metric examples.

Conclusion/Recommendations

While the practice of applying PIDs has begun to be integrated into the curation of discrete datasets, observatories with continuous, streaming, *in situ* data have struggled across the board to incorporate them. In the few cases where a PID was applied to a continuous dataset, it was only able to be done by subsetting the continuous data into discrete datasets, i.e. monthly snapshots of the data. Where observatories have lacked in PIDs, they have succeeded in providing clear guidance for alternative means of citing the data as well as ensuring that all data downloaded are accompanied by metadata.

The solution to adding PIDs to continuous data is not a simple one. Even if the technical challenge of how to automatically generate a DOI string for each download were to be achieved, it would be extremely difficult to properly maintain that DOI. Each DOI must point to a landing page where the data that that DOI represents can be found. Not only is it not feasible for an observatory to create that many landing pages, it is also not feasible for an observatory to store that many discrete datasets to be displayed on those landing pages.

In the absence of providing one-to-one DOIs per datasets, as is possible with traditional investigator-driven data collection into discrete datasets, a clear set of universally accepted best practices is needed for observatories with continuous, streaming in situ data. For example, NOAA has a set of rules regarding the level of granularity required for their DOIs.[16] As a result of our research and analysis, we propose the following recommendations for what that guidance should entail.

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: DOIs should be the PID applied to all observatory data as they are the globally accepted standard metric (ISO 26324-2012). For interoperability, it is recommended that DOI suffixes contain the same alphanumeric sequence used internally to the object being identified (observatory, site, dataset, etc). For example, if a DOI was associated with salinity data that internally was referred to as “SAL01234” then that same code should be incorporated into the suffix of the DOI. (Best Practice 1)

Recommendation #2: DOIs should be applied at varying levels of granularity depending on the type of data generated. For continuous, streaming in situ data on moorings or seafloor packages, DOI’s should be provided at the array or site level. Ship-based data, for example

[16] https://geo-ide.noaa.gov/wiki/index.php?title=Data_Citation_Granularity

from maintenance cruises, should be assigned a DOI per cruise. Glider data should be treated the same as ship-based data, with a DOI assigned per glider mission. Note that this is only necessary if the same glider could be deployed at multiple locations during its lifespan. If a glider stays in the same location, and runs the same mission track, it can be treated similar to a mooring. Note that each DOI assigned will require its own landing page with access to that data. As such, ultimately the level of granularity of DOI used will be dependent on an observatory's ability to generate and maintain these landing pages. If multiple pages are not able to be maintained, DOIs should be applied at the observatory level. (Best Practice 1)

Recommendation #3: Guidance for data citation should be clearly provided on the observatory's website. The citation should include basic information about the dataset, such as the dataset generator (i.e. the observatory), title of the dataset, distributor (if different from the observatory), and the DOI. The citation also should include all information necessary to take a user from the DOI to the exact data stream used in an analysis. For example, this information could include: Site ID, Sensor/Product/Stream ID, Date Range, and Date Downloaded. It is recommended that if more than six data products/streams are used in a publication that this information be provided in the form of a table. (Best Practice 1,2)

Recommendation #4: Data identifiers should be maintained when observatory data are integrated into data aggregators. The aggregator can apply their own identifier as long as the DOI is also used and when the aggregator identifier is called up, the DOI is clearly visible on the landing page. Until the maintenance of DOIs is a standard practice among data aggregators, it must be a requirement of the observatory that aggregators cannot incorporate their data unless the DOI is correctly included and maintained. It is also recommended that observatories that have implemented DOIs for their data reach out to aggregators currently serving their data without DOIs and request that those DOIs be added. (Best Practice 3)

Recommendation #5: Versioning and provenance data should be included on both the dataset's landing page as well as its metadata. This will ensure that a user can access the data used at the time of the original analysis regardless of when they access the citation or go back to the original data. (Best Practice 4)

Recommendation #6: All data should be provided with associated metadata. These metadata can either be displayed on a dataset's landing page, or in the instance of datasets being created made-to-order from a continuous data stream, the metadata should be included with the data download. Full dataset citation information as well as versioning and provenance information must also be provided in the metadata associated with a dataset. (Best Practice 2,4)

Recommendation #7: Once DOIs are applied, data citations should be tracked via native DOI tracking tools, for example through DataCite. Until proper citation practices become universally used across the community, observatories should also monitor citations through key words using tools such as Google Scholar or Mendeley. (Best Practice 5)

Recommendation #8: In order to ensure compliance with data citation and identification practices by the scientific community, these best practices should be incentivised as requirements from both funding agencies and scientific publications. Since there is general agreement that all data in a manuscript must be made sufficiently accessible such that others may be able to repeat their work, journals need to provide clear guidance for how data must be cited in their publications and the editors and peer reviewers must ensure that these practices are upheld. (Best Practice 6)

References

Altman, M. and G. King. 2006. A proposed standard for the scholarly citation of quantitative data. DLib Magazine 13(3/4).

Ariani, A., A.J. Barton, J. Brase, F. Brown, T. Demeranville, P. Herterich, L. McAvoy, L. Paglione, S. Ruiz, and G. Thorisson. 2014. Workflow for interoperability. ORCID and DataCite Interoperability Network Deliverable D4.2.

Bourne, P. E., T. Clark, R. Dale, A. de Waard, I. Herman, E. Hovy, and D. Shotton editors. 2011. Improving future research communication and-scholarship, The FORCE 11 Manifesto. Available from http://www.force11.org.

CODATA-ICSTI Task Group on Data Citation Standards and Practices. 2013. Out of cite, out of mind: the current state of practice, policy, and technology for the citation of data. Data Science Journal 12:CIDCR1–CIDCR75.

Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. Martone M. (ed.) San Diego CA: FORCE11; 2014 https://doi.org/10.25490/a97f-egyk

Evans, P. L., A. Strollo, A. Clark, T. Ahern, R. Newman, J. F. Clinton, H. Pedersen, and C. Pequegnat. 2015. Why seismic networks need digital object identifiers, Eos, 96, doi:10.1029/2015EO036971.

Gries, C, 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

IETF RFC 4122. Network Working Group. 2005. A Universally Unique IDentifier (UUID) URN Namespace. The Internet Society. Memo. https://tools.ietf.org/html/rfc4122#ref-3

International DOI Foundation. 2012. DOI Handbook. doi.org/10.1000/182. Date Accessed February 2019.

ISO/IEC 9834-8:2005. Information technology -- Open Systems Interconnection -- Procedures for the operation of OSI Registration Authorities: Generation and registration of Universally Unique Identifiers (UUIDs) and their use as ASN.1 Object Identifier components. Accessed https://www.iso.org/standard/36775.html

National Research Council 2012. For Attribution: Developing Data Attribution and Citation Practices and Standards: Summary of an International Workshop. Washington, DC: The National Academies Press. https://doi.org/10.17226/13564.

Paulk, M., W. Curtis, M.B. Chrissis, and C. Weber. 1993. *Capability Maturity Model for Software (Version 1.1)* (CMU/SEI-93-TR-024). Retrieved March 14, 2019, from the Software Engineering Institute, Carnegie Mellon University website: http://resources.sei.cmu.edu/library/asset-view.cfm?AssetID=11955

Stall, S. (2017), Enabling findable, accessible, interoperable, and reusable data, *Eos, 98,* https://doi.org/10.1029/2018EO081907.

APPENDIX

Best Practice Self-Assessment Tool

The best practice self-assessment tool enables an existing or new organization to assess their current data identification, citation, and tracking capabilities and maturity level. 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 for an organization to identify and plan for improvements in people, process, and technology that support data identification, citation, and tracking.

Steps for Using the Self-Assessment Tool

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

1. Best Practices List

DI BP 1: Persistent data identifiers are associated with all data products
DI BP 2: Guidance is provided by an observatory for data citations
DI BP 3: Data identifiers are maintained throughout the life cycle of the data, including DI DI when observatory data are transferred to data aggregators
DI BP 4: Data versioning and provenance information is available and accessible.
DI BP 5: Processes are in place to track and report data usage
DI BP 6: Data usage and citation tracking metrics are provided to the funding agency/stakeholder community

2. Example Of Completed 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 levels. 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.

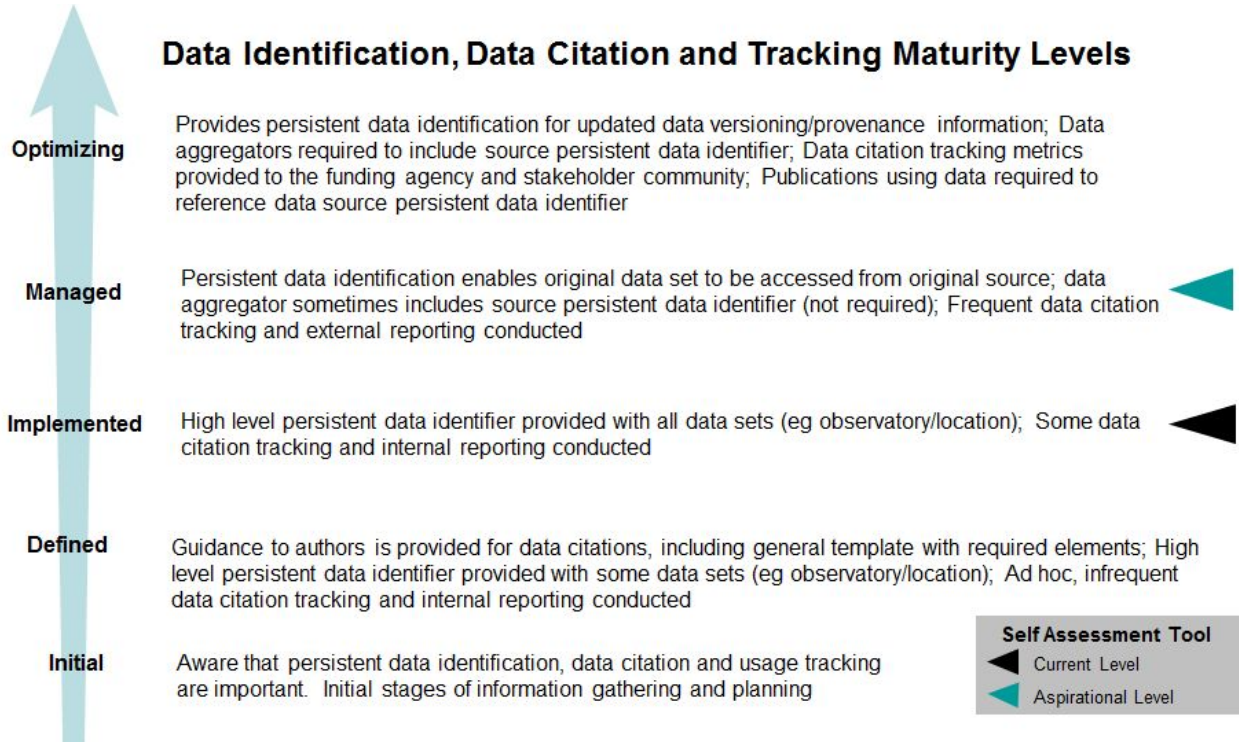

Figure 1: Example of a completed best practice self-assessment

3. 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.

DI BP 1: Persistent data identifiers are associated with all data products. Examples: Data identifiers refer to a unique, web-compatible alphanumeric code assigned to a data set that is able to be preserved long-term. The identifier can have varying levels of granularity, but in order to be effective they must support the finest-level necessary to effectively identify the data (Socha 2013). Data source identifier providers include: Digital Object Identifiers (DOI) managed by DataCite, NOAA NCEI Accession Numbers, Archival Resource Key (ARK) managed by the California Digital Library.

- Data products have high level (observatory/location) PID – 1 point
- 50% of the data products have an individual PID – 2 points
- All data products have an individual PID - 3 points

DI BP 2: Guidance is provided by an observatory for data citations. Examples: Guidance for proper data citations provided by an observatory. Details must be provided such as date of download and a PID that directs back to the dataset. Data citation information can be included

in the metadata (CODATA-ICSTI Task Group on Data Citation Standards and Practices 2013). For example, within the ISO 19115 standard, citation information goes within the CI_Identification field. The identifier associated with the data, either the internal identifier or PID, is added to the MD_Identifier field. It is important to include both so this information stays with the data, independent of the researcher re-visiting an observatory's website to find the reference information.

- General guidance provided for citing data – 1 point
- Suggested data citation template provided with required elements – 2 points
- Suggested data citation is provided on dataset's landing page or in dataset metadata file - 3 points

BP 3: Data identifiers are maintained throughout the life cycle of the data, including when observatory data are transferred to data aggregators. Examples: The onus for maintaining the original data identifier or linking identifiers of different versions across each exchange. It should be noted that the onus for maintaining the identifier when data are incorporated into an aggregator is both on the part of the data generator and the data aggregator. The data generator needs to first provide the identifier, but the aggregator also needs to have the built in capability to receive and display the identifier.

- Data provider and aggregator sometimes includes source persistent identifier (not required) - 1 point
- Data provider and aggregator required to include source persistent identifier - 2 points

BP 4: Data versioning and provenance information is available and accessible. Examples: During data lifetime, corrections may be provided to those data (versioning) or the data may be downloaded, manipulated, and posted in new forms (provenance). It is an important date of download be provided in a data citation and b) information on data versioning and provenance included with each dataset, preferably both on the dataset landing page and within the metadata. The extent of information to include about data changes within each version is a challenge for the observatory to determine. As such, the level of detail that needs to be included will likely depend on the data product, the types of changes made, and the needs of that specific user community.

- Minimal Data Versioning and provenance included - 1 point
- Comprehensive Data Versioning and provenance included – 2 points

BP 5: Processes are in place to track and report data usage. Examples:Tracking of data citations refers to quantifying the number of times the data have appeared in a scholarly publication, or been referenced by subsequent publications. Example tracking methods include through a Registration Agency (e.g. DataCite), or using Google Scholar or Mendeley to search for keywords. Data usage tracking also includes non-sharlary publications, for example Altmetric

is a tool that searches for any reference to a project, grant, or dataset that has a URL associated with it. Whether that reference is in a publication, a blog post, or a Twitter post.

- Ad hoc informal data citation tracking and internal reporting conducted – 1 point
- Formal process for data citation tracking and external reporting conducted – 2 points

BP 6: Data usage and citation tracking metrics are provided to the funding agency/stakeholder community. Examples: Funding agencies more commonly require usage statistics to justify future funding. For example, NASA archives have a senior review every 2-3 years to assess the extent that the data are being used in publications (NRC 2012). Providing data usage reports and metrics provides documentation of data tracking and usage.

- Ad hoc informal data usage and citation tracking metrics are sometimes provided to funding agency/stakeholders – 1 point
- Formal data usage and citation tracking metrics regularly provided to funding agency/stakeholders – 2 points

4. Determine Maturity Levels

Add up your capability score points to determine your current maturity level:

Initial Level 0	0-1 points
Defined Level 1	2-5 points
Implemented Level 2	6-7 points
Managing Level 3	8-9 points
Optimizing Level 4	9+ points

Identify your aspirational maturity level by selecting a desired best practice capability score. Add up your desired capability score points to determine your aspirational maturity level.