DATA SCIENCE JOURNAL

Peng, G, et al. 2019. Practical Application of a Data Stewardship Maturity Matrix for the NOAA *OneStop* Project. *Data Science Journal,* 18: 41, pp. 1–18. DOI: https://doi.org/10.5334/dsj-2019-041

PRACTICE PAPER

Practical Application of a Data Stewardship Maturity Matrix for the NOAA *OneStop* Project

Ge Peng[1], Anna Milan[2], Nancy A. Ritchey[2], Robert P. Partee II[3], Sonny Zinn[4], Evan McQuinn[5], Kenneth S. Casey[2], Paul Lemieux III[3], Raisa Ionin[3], Philip Jones[3], Arianna Jakositz[5] and Donald Collins[2]

[1] North Carolina State University, Cooperative Institute for Climate and Satellites – North Carolina (CICS-NC) at NOAA's National Centers for Environmental Information, Asheville, North Carolina, US

[2] NOAA's National Centers for Environmental Information, US

[3] Riverside Technology, Inc., US

[4] Earth Resources Technology, Inc., US

[5] Cooperative Institute for Research in Environmental Sciences (CIRES), US

Corresponding author: Ge Peng, Ph.D. (gpeng@ncsu.edu)

Assessing the stewardship maturity of individual datasets is an essential part of ensuring and improving the way datasets are documented, preserved, and disseminated to users. It is a critical step towards meeting U.S. federal regulations, organizational requirements, and user needs. However, it is challenging to do so consistently and quantifiably. The Data Stewardship Maturity Matrix (DSMM), developed jointly by NOAA's National Centers for Environmental Information (NCEI) and the Cooperative Institute for Climate and Satellites–North Carolina (CICS-NC), provides a uniform framework for consistently rating stewardship maturity of individual datasets in nine key components: preservability, accessibility, usability, production sustainability, data quality assurance, data quality control/monitoring, data quality assessment, transparency/traceability, and data integrity. So far, the DSMM has been applied to over 800 individual datasets that are archived and/or managed by NCEI, in support of the NOAA's *OneStop* Data Discovery and Access Framework Project. As a part of the *OneStop*-ready process, tools, implementation guidance, workflows, and best practices are developed to assist the application of the DSMM and described in this paper. The DSMM ratings are also consistently captured in the ISO standard-based dataset-level quality metadata and citable quality descriptive information documents, which serve as interoperable quality information to both machine and human end-users. These DSMM implementation and integration workflows and best practices could be adopted by other data management and stewardship projects or adapted for applications of other maturity assessment models.

Keywords: Data Stewardship Maturity Matrix; ISO Metadata Standards; Data Management and Stewardship; Data Preservation; Open Data; Transparency; Data Quality; Information Quality

1. Introduction

The National Oceanic and Atmospheric Administration (NOAA) is responsible for providing environmental intelligence to American citizens, businesses, and governments to enable informed decisions. Its mission is to understand and predict changes in climate, weather, oceans, and coasts; to share that knowledge and information with others; and to conserve and manage coastal and marine ecosystems and resources. NOAA collects and provides stewardship to geophysical measurements of more than two thousand diverse parameters. The data comes from a broad range of platforms, including but not limited to satellites, fixed and mobile radars, research aircraft, buoys, ships, land-based in situ surface and upper air networks, and weather and climate models (National Research Council 2007). NOAA's data capability is constantly improving with higher temporal and spatial resolutions, and improved observation systems. Data from new, unique, unconventional observing platforms such as unmanned water and air vehicles and instrumented animals are

becoming available and adding to the existing NOAA data portfolio. Derived products are created to help improve our understanding of the environment, or to meet new federal requirements or user demands. Data curation and stewardship need to keep up with this ever-changing landscape with 'an increased focus on information management standards and strategies to improve access, interoperability, and usability' (NOAA 2010).

With rapidly increasing data volume and elevated requirements for timely access of high-quality and readily usable and interoperable environmental data and information, effectively managing and providing the access to the NOAA data is a significant challenge. The *OneStop* project was initiated in 2015 as a pathfinder project to support NOAA's efforts to improve discovery and access services for NOAA's legacy data, by leveraging existing access technologies and infusing specific innovations. *OneStop* adopted the Common Framework for Earth Observation Data (CFEOD) which provides federal agencies with guidance on standards and practices to maximize data interoperability and allow for enhanced data discovery and access services (Casey et al. 2015; USGEO 2015). *OneStop* leverages and supports the NOAA Big Data Project (BDP) and the Big Earth Data Initiative (BEDI) (Casey 2016). Through implementation of these services in an open data framework at a massive scale, *OneStop* aims to enable broader use and reuse of NOAA's data in commercial and scientific applications.

The quality of a data product depends in part on the stewardship practices applied to it after its development and production. Assessing the current state of how individual datasets are preserved, documented, and disseminated to users is an essential part of managing NOAA data. It is a critical step towards meeting U.S. federal regulations, organizational requirements, and user needs, especially in the area of documenting and providing data quality information to establish or improve data trustworthiness. However, it is challenging to do so consistently and quantifiably. Therefore, a reference framework with measurable criteria that are applied to individual datasets is beneficial to data stewards and users (Peng et al. 2015; Casey 2016).

The data stewardship maturity matrix (DSMM), developed jointly by domain (data management, technology, and science) subject matter experts (SMEs) from NOAA's National Centers for Environmental Information (NCEI) and the Cooperative Institute for Climate and Satellites–North Carolina (CICS-NC), provides such a consistent framework. Leveraging institutional knowledge and community best practices and standards, the

Maturity Scale / Key Component	Level 1 - Ad Hoc Not Managed	Level 2 - Minimal Managed Limited	Level 3 - Intermediate Managed Defined, Partially Implemented	Level 4 - Advanced Managed Well-Defined, Fully Implemented	Level 5 - Optimal Level 4 + Measured , Controlled , Audit
Preservability	*The state of dataset being preservable*				
Accessibility	*The state of dataset being publicly searchable and accessible*				
Usability	*The state of data product being easy to understand and use*				
Production Sustainability	*The state of data production being sustainable and extendable*				
Data Quality Assurance	*The state of data product quality being assured/screened*				
Data Quality Control /Monitoring	*The state of data product quality being controlled and monitored*				
Data Quality Assessment	*The state of data product quality being assessed*				
Transparency /Traceability	*The state of data product being transparent, trackable, and traceable*				
Data Integrity	*The state of data integrity being verifiable*				

Figure 1: A conceptual model of the scientific data stewardship maturity matrix (DSMM) of National Centers for Environmental Information (NCEI) and the Cooperative Institute for Climate and Satellites–North Carolina (CICS-NC).

DSMM defines a graduated maturity scale for each of nine key components of scientific data stewardship to enable a consistent assessment of quantifiable stewardship practices applied to a given data product (Peng et al. 2015; see **Figure 1** for the conceptual model of the DSMM).

As a part of the vetting process, the DSMM has been applied to various different types of datasets managed by various different projects or different organizations (e.g., Ritchey and Peng 2015; Hou et al. 2015; Peng et al. 2016). These use case studies have improved the maturity of the DSMM. They have also helped identify gaps in organizational procedures or systems (e.g., Peng et al. 2016; Grace Peng, *personal communication*).

The criteria defined at each DSMM maturity level may be used to help individual projects or programs define stewardship requirements, monitor the project progress, and/or demonstrate compliance with specific stewardship practices. For example, if a project needs its data product to be at Level 3 for the Usability key component, making the product documents and source code publicly available should be one of the project requirements and on the product checklist. Due to the progressive and actionable nature of the criteria, the DSMM can also help create a roadmap for improving the stewardship maturity on a data set level (e.g., Peng et al. 2016).

Utilizing a reference framework to quantitatively assess dataset maturity is just a good starting point. Consistently and systematically capturing and integrating this data maturity information for machine and human end-users is an important part of improving data and information accessibility, usability, and interoperability. Both utilizing a maturity framework and consistently capturing dataset maturity information are fairly new for Earth Science data management and stewardship.

The *OneStop* project has aimed at improving the completeness and content of dataset-level metadata records via the *OneStop*-ready process as the solid first step towards achieving its project goals. **Table 1** outlines the *OneStop*-readiness requirements. To provide evidence-based data quality metadata, the stewardship maturity of individual datasets, namely, how datasets were preserved, documented, and disseminated, has been assessed utilizing the DSMM. So far, the DSMM has been applied to over 800 NOAA digital environmental datasets (**Figure 2**). The datasets represent many different data groups, including satellite-based oceanic, atmospheric, and cryospheric climate data records (CDR), digital elevation models (DEM), Global High-Resolution Sea Surface Temperature (GHRSST) products, Suomi National Polar-orbiting Partnership (S-NPP) data products, World Ocean Atlas (WOA), Water Column Sonar Data (WCSD), Level-2 Next Generation Weather Radar (NEXRAD) data, Geostationary Operational Environmental Satellite (GOES-R) series data and in situ global meteorological and hydrological data products (**Figure 2**). About 368 of those 800+ datasets have the DSMM ratings available to the general public via the *OneStop* portal.

This paper presents the practicality of *OneStop* application of the DSMM as a part of the process of optimizing discoverability and usability of datasets within the *OneStop* framework. The paper covers the aspects

Table 1: *OneStop*-Readiness Requirements (Based on Delk and Milan 2018).

Category	Requirements
Collection* Level Metadata	• ISO 19115-2 metadata standards compliant; • Complete high quality and up-to-date content; • Adoption of the Global Change Master Directory (GCMD) keywords (Science, Place and Organization, Platform, Instrument, and Project); • URL to a browse graphic thumbnail
Granule Metadata**	*OneStop* ISO-Lite compliant metadata records (Li et al. 2017)
Data Formats	Data are in a standardized, non-proprietary machine-readable format
Data Stewardship Maturity Matrix (DSMM) Assessment	• Each collection shall have a DSMM Assessment; • The results shall be encoded into the collection-level metadata; • The DSMM ratings shall be displayed at the *OneStop* search and discovery portal; • The final assessment report should be published to the NOAA's central library repository
Data Access	Data are online with direct access options

* A collection is a grouping of environmental data or products that share common characteristics, is represented by a single metadata record, and consists of one or more granules. In this paper, a data collection refers to a minimum citable unit of data (Li et al. 2017), which often time refers to a dataset. We may use a data product and a dataset interchangeably.

** A granule is the smallest aggregation of data that can be independently managed (described, inventoried, and retrieved) in the *OneStop* system.

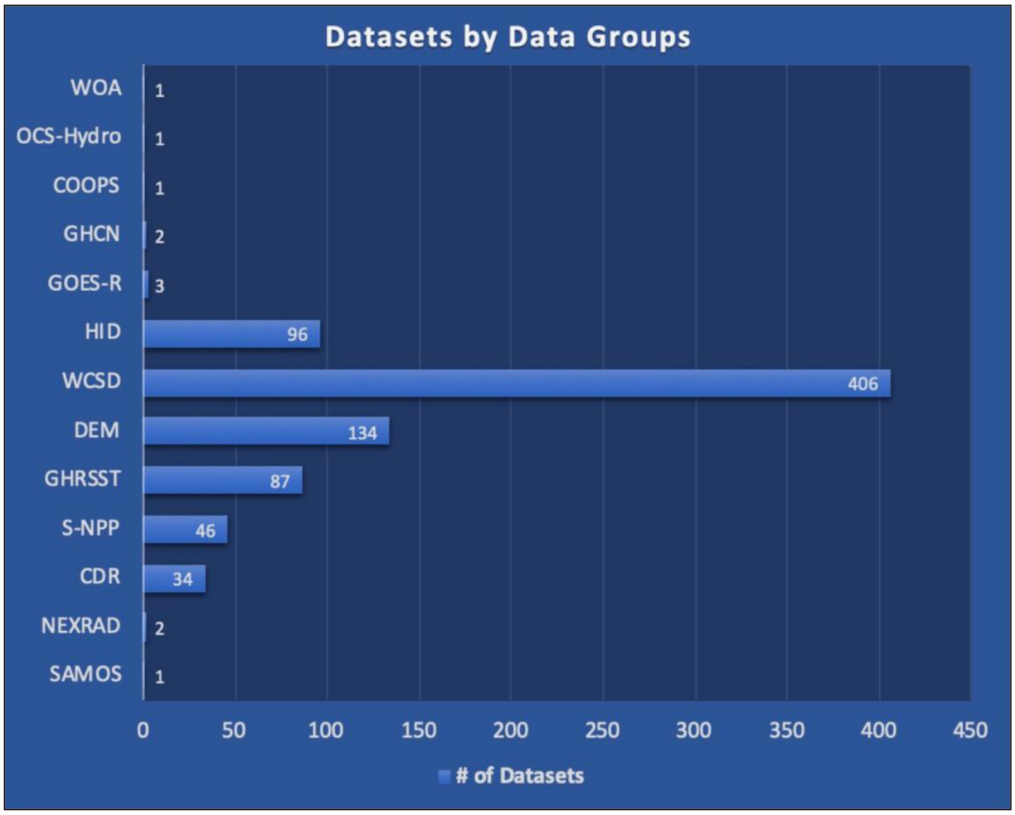

Figure 2: Datasets by data groups whose stewardship maturity has been assessed as of 6/30/2018. OCS-Hydro: NOAA's Office of Coast Survey-Hydrographic Survey Level-2 product; COOPS: CO-OPS National Water Level Observation Network (NWLON) and Physical Oceanographic Real-Time System (PORTS) data; GHCN: daily and monthly Global Historical Climatology Network (GHCN) products; HID: Hazard Images Database; SAMOS: Shipboard Automated Meteorological and Oceanographic System Quality-Controlled Underway Oceanographic and Meteorological Data. See other acronyms in Introduction.

of the entire lifecycle of the DSMM application, namely, evaluating, capturing, representing, baselining, integrating, and visualizing DSMM ratings.

2. Evaluation and Representation of DSMM Ratings

Applying the DSMM to individual datasets starts with training the *OneStop* metadata content editors. Carrying out the stewardship maturity evaluation of a dataset involves collecting relevant information, documenting evidence, and capturing the DSMM ratings for each key component. For example, criteria captured in the DSMM leverage the community best practices and standards including those for metadata. The best practices and standards may vary for individual disciplines. Individual projects may define their own standards, e.g., ISO metadata standards. Therefore, the information about those practices and standards applied to the individual datasets is useful to be captured as supporting evidence. The existing DSMM template developed by Peng (2015) was utilized to facilitate this process. A quick start-up user guide was first developed to provide high-level background information on the DSMM followed by a step-by-step guide on how to get the DSMM template, collect needed information, carry out stewardship maturity assessment of the dataset, and display consistently the assessment results.

Once an assessment has been carried out, the DSMM ratings need to be captured and integrated into other systems or tools so information about the stewardship maturity of the dataset can be conveyed, preferably in a consistent fashion, to both machine and human end-users. For the *OneStop* project, the DSMM ratings and supporting evidence, i.e., justifications, are captured into a centralized spreadsheet via a Google Form and/or a database via a web-based interface. The DSMM ratings and evidence are subsequently integrated into other systems or tools, which will be described in the next section, to generate ISO standard-based dataset-level quality metadata and data stewardship maturity reports (DSMRs), and to be utilized in the *OneStop* search and discovery algorithm. Some of the relevant aspects of representing DSMM assessment results are described in the following subsections.

Table 2: Description of fields in the data stewardship maturity report (DSMR) naming convention.

Field ID	Description
DatasetShortName	A short name that is descriptive of the data product which, preferably, is 30 or less characters containing letters, numbers, hyphen(s) and/or underscore(s) without any space or special characters. This short dataset name could include organization(s), data product abbreviation, and product type and/or version.
vnnrmm	The version and revision number of current maturity assessment. Two-digit integers are used for the version and revision numbers. For example, v01r00 will be used for the first baselined version. The version number only changes when the maturity ratings are modified. Changes to the revision number reflect other modifications to the assessment document, including update to justifications and/or error corrections.
yyyymmdd	Year, month, and day of the current maturity assessment version. For example, 20160408 for April 8, 2016.

2.1. File naming convention

An important aspect of the *OneStop* project's implementation of the DSMM is the concept of machine generating a human-readable and publishable data stewardship maturity report (DSMR) which is described in Section 2.5. When published, such DSMRs can be used to provide content-rich and knowledge-based dataset quality information. Anyone, however, could use the information in this and the following subsections to create such a DSMR manually or automatically.

We recommend publishing these DSMRs in a formal institutional repository with appropriate structured publication metadata to optimize their availability and utility. However, generation of DSMRs at scale has shown that following a file naming convention is useful for both internal management and external users of the reports. Further, a consistent file naming convention will help systematically publish and link this data quality descriptive information document. We, therefore, recommend the file naming convention of the resultant DSMR to be defined as:

<DatasetShortName>_MM-Stew_<vnnrmm>_<yyyymmdd>

MM-Stew is a maturity metadata tag described in section 2.4. The <…> denotes a field to be defined and completed by evaluators based on information pertaining to the dataset. Description for each field is provided in **Table 2**.

It is recommended to use standard or abbreviated variable names relevant to the user community. For example, environmental data vocabularies such as those from Climate Convention (CF; Eaton et al. 2014) or Observations for Model Intercomparisons (Obs4MIPs 2017) would allow the use of PMSIC for Passive Microwave Sea Ice Concentration. Another example would be to include abbreviations for institutions or programs such as EPA, USGS, NOAA, NASA, or S-NPP, preferably using community standard-based keywords such as Global Change Master Directory (GCMD) keywords (GCMD 2018). A hyphen could be used for more than one primary institution, for example, NOAA-NSIDC. It is recommended to consult, if possible, with scientific and/or data stewards for your choice of the product short name. Therefore, the file name:

NOAA-NSIDC_PMSIC_CDR-v2_MM-Stew_v02r01_20150623.pdf

will denote that it is the document containing the version v02r01 (i.e., version 2, revision 1) stewardship maturity assessment results as of June 23, 2015 for the version 2 NOAA/NSIDC Passive Microwave Sea Ice Concentration Climate Data Record.

2.2. Reporting formats

It is beneficial to capture and represent DSMM ratings using a consistent format for both human and machine end-users. For human end-users, having a consistent layout of diagrams with a standardized color scheme in a consistent layout DSMR helps people readily understand and interpret the assessment results. Consistency is also helpful for visually comparing ratings from different datasets. A progressive, green-scale color scheme defined in the DSMM template is recommended (see **Table 3**).

To convey and represent DSMM ratings consistently, two different types of standardized DSMM graphics were developed (**Figures 3** and **4**). **Figure 3** is referred to as a scoreboard and **Figure 4** as a rating diagram. They both are essentially presenting the same stewardship maturity rating information but from two different perspectives. While **Figure 4** presents a high-level, abstractive view of DSMM ratings, **Figure 3** also captures the definitions, allowing for a more detailed description of criteria used for assessments. If two cells in a scoreboard are filled, it is an indication that only a partial rating at the higher level is satisfied (**Figure 3**). The same information is conveyed in a rating diagram by a lighter shaded star (**Figure 4**). Both the scoreboard and rating diagram, if included in a maturity document like DSMR, will display maturity levels of the latest assessment version and will not change with revisions.

For machine end-users, combination of the consistent color scheme, diagrams, document naming convention and layout will make the DSMM ratings and stewardship maturity information more integrable and interoperable.

2.3. Review and baseline process

A previous pilot NCEI DSMM use case study revealed that it is very beneficial to carry out assessment and review by a team – consisting of members from the Integrated Product Team (IPT), the Data Steward, and the DSMM SME (Peng et al. 2016). Within NCEI, an IPT usually consists of a data product SME or POC (Point-Of-Contact), an archive specialist, and an operation or access specialist.

Table 3: DSMM scale definition and RGB color scheme for representing DSMM ratings.

Maturity Scale	Definition	Color Code	R	G	B	Color
Level 1	Ad Hoc/Unknown; Not managed	Lighter green	229	244	224	
Level 2	Minimal; Managed; Not or limited defined	Light green	203	224	192	
Level 3	Intermediate; Managed; Defined, partially implemented	Green	176	223	161	
Level 4	Advanced; Managed; Well-defined, fully implemented	Dark green	85	168	57	
Level 5	Optimal; Level 4 + measured, controlled, audit	Darker green	56	112	38	

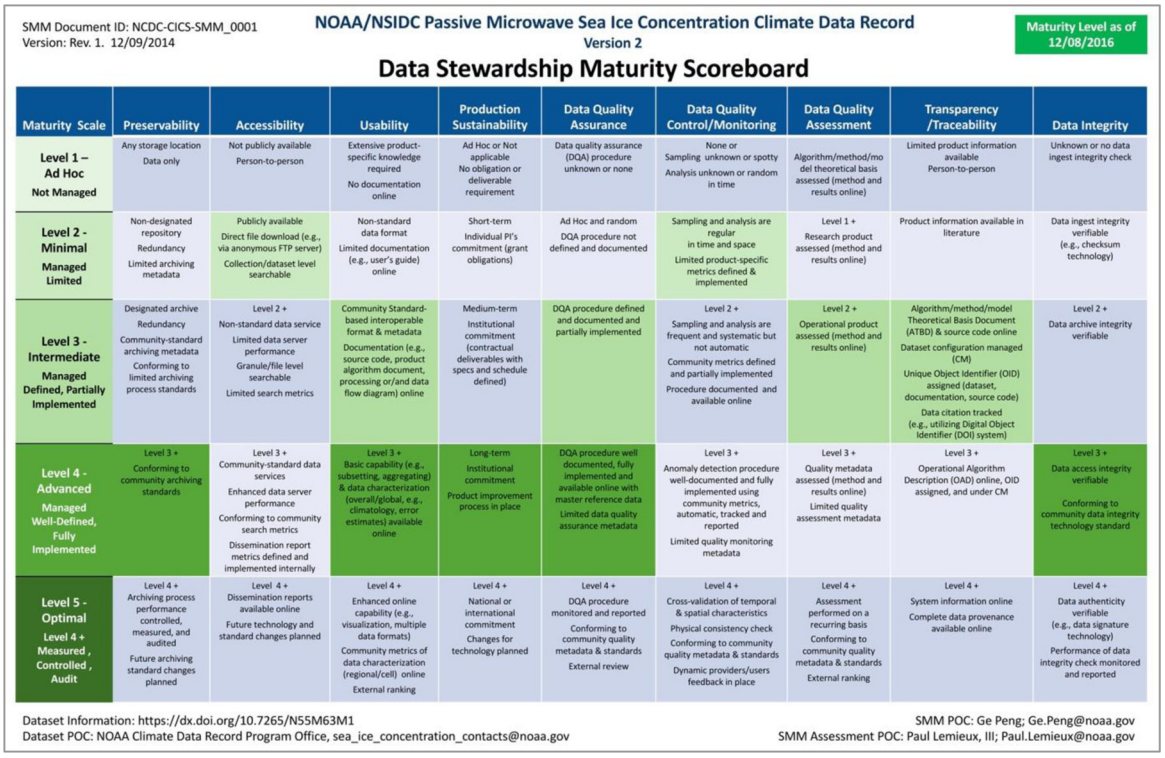

Figure 3: Data stewardship maturity scoreboard for NOAA-NSIDC_PMSIC_CDR-v2 as of December 8, 2016. If two vertical cells are shaded by greens, it is an indication that only a partial rating at the higher level is satisfied. From Lemieux et al. (2017).

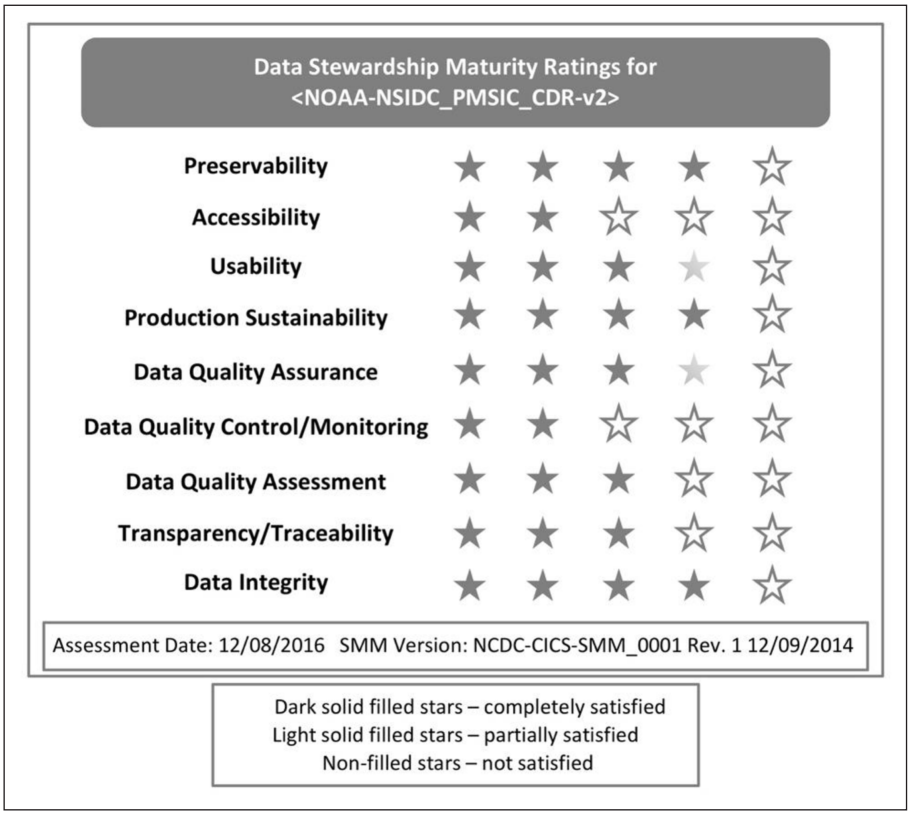

Figure 4: Data stewardship maturity rating diagram for NOAA-NSIDC_PMSIC_CDR-v2 as of December 8, 2016. Dark/light/none solid filled stars denote the criteria are completely/partially/not satisfied. From Lemieux et al. (2017).

Self-evaluation results can be used internally within an organization or for personal purpose. For an organization entity, such as a repository, data center, or program, we recommend establishing a review and baseline process for DSMM assessments to ensure their quality.

To indicate the current status of the assessment to other team members or major stakeholders, the levels are recommended to progress in the following order:

- Not yet Assessed
- Preliminary Assessment
- Initial Assessment Draft (complete assessment but before the first (team) review)
- Revised Assessment Draft
- Final Assessment Draft
- Baselined

The number of interim drafts and their reviews before the first baseline varies for each individual dataset. It usually depends on the current availability and accessibility of the information about stewardship practices to the evaluators. It is recommended to version a draft of the assessment document as v00rxx, where xx denotes a two-digit revision number, for example, for the first draft:

<DatasetShortName>_MM-Stew_<v00r01>_<yyyymmdd>

The baselined version of the stewardship maturity assessment document should contain the complete metadata of the DSMM assessment, ratings, and justifications (**Figure 5**). It may also contain the scoreboard and rating diagrams.

2.4. Quality metadata implementation

The following maturity metadata (MM) tags are defined as possible identifiers of data quality metadata in the dataset-level metadata records. These MM tags systematically indicate different perspectives for potential linkage to different defined maturity assessment models when such a system becomes available in the future:

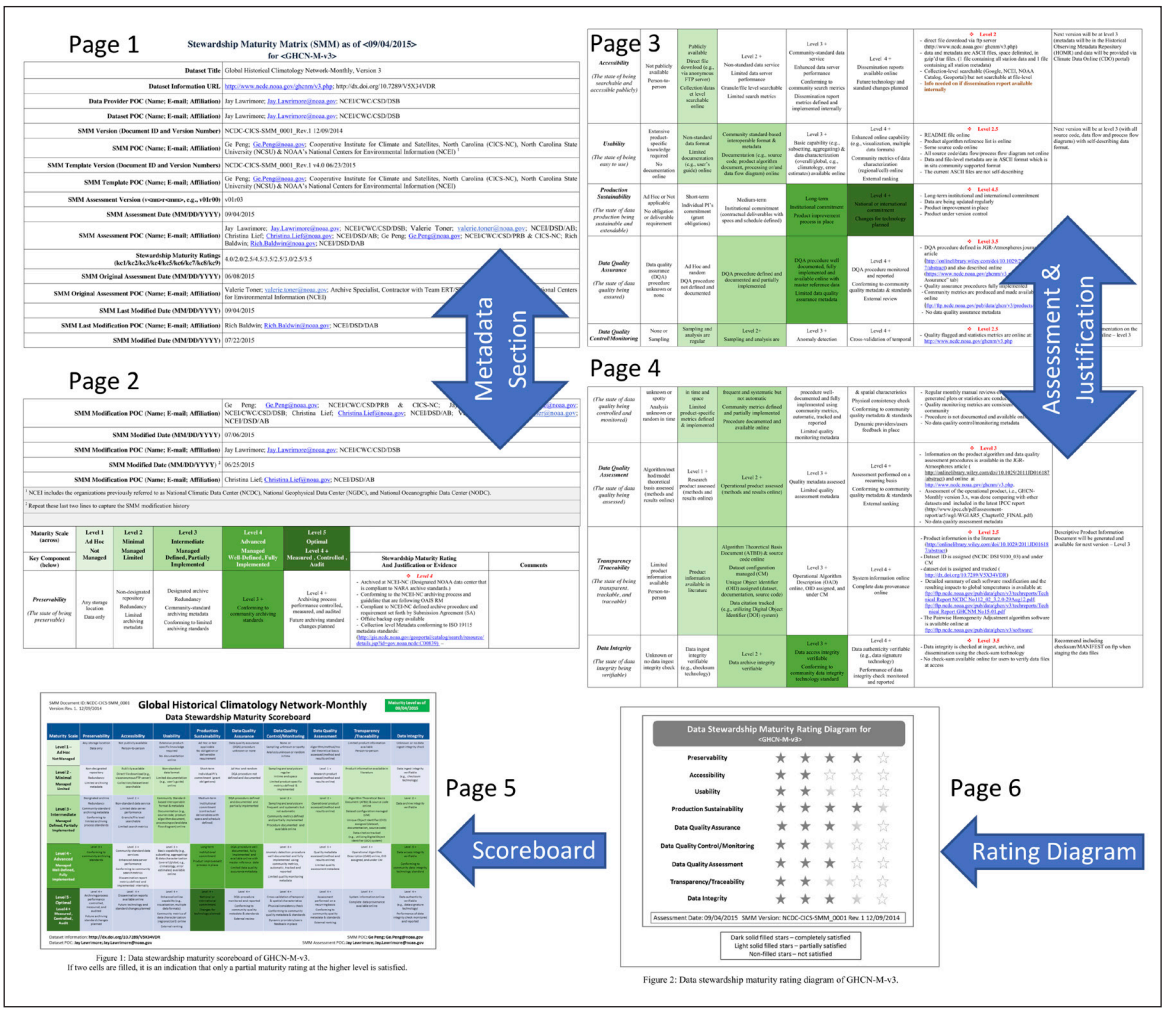

Figure 1: Data stewardship maturity scoreboard of GHCN-M-v3.
If two cells are filled, it is an indication that only a partial maturity rating at the higher level is satisfied.

Figure 2: Data stewardship maturity rating diagram of GHCN-M-v3.

Figure 5: Page view of the stewardship maturity assessment document (Lawrimore et al. 2015; 6 pages in total) for NCEI GHCN-Monthly v3 data product to show the general layout of the document. GHCN is the acronyms for Global Historical Climatology Network. This image is not intended to display the content of the document (see Lawrimore et al. 2015 for the content). The maturity ratings and lessons learned from the DSMM GHCN-Monthly use case study can be found in Peng et al. (2016).

- MM-Scie for science maturity information
- MM-Prod for product maturity information
- MM-Stew for stewardship maturity information
- MM-Serv for service maturity information

The best practices for implementing the DSMM ratings into the ISO standard-based dataset-level quality metadata records are developed by the *OneStop* Metadata Team in collaboration with the NOAA Metadata Working Group (Ritchey et al. 2016). The conceptual framework is outlined in **Figure 6** and **Table 4**. The current xml representation for encoding the DSMM assessment ratings into an ISO metadata record is provided in Appendix A.

2.5. Data Stewardship Maturity Report

The DSMR is formatted consistently with other NOAA technical memoranda. It captures the revision history, provides a recommended citation for the report, consists of an introduction, assessment results including the rating scoreboard and diagram, and a reference list of works cited within the assessment (see an example in Lemieux et al. 2017).

In collaboration with the NOAA Central Library, the *OneStop* Metadata Team developed a process to create and publish a NOAA Technical Information Series as defined in NOAA Administrative Order 201-32G: Scientific and Technical Publications for the NESDIS line office (NOAA 1993). Having DSMRs publicly available helps the *OneStop* project be compliant with the Information Quality Act (US Public Law 106–554,

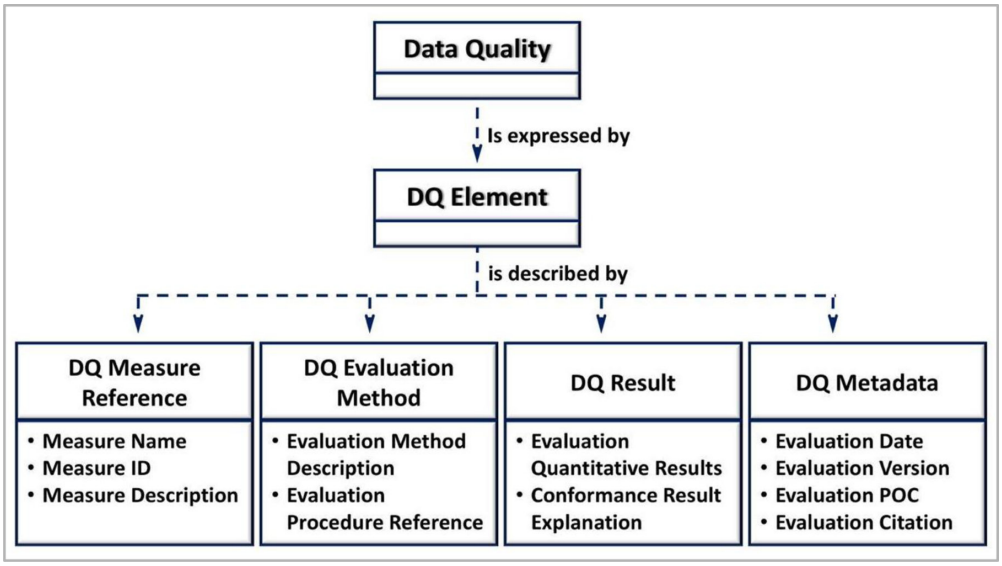

Figure 6: Implementation best practices for adopting ISO data quality metadata standards.

Table 4: The Conceptual Framework for Implementing DSMM ratings into ISO Data Quality Metadata. Examples of DSMM assessment ratings and a maturity report are highlighted in bold in the second column.

Measure Name	Data Stewardship Maturity Assessment
Measure ID	MM-Stew
Measure Description	The Data Stewardship Maturity Matrix (DSMM) is a unified framework that defines criteria for each of nine components based on measurable practices, which can be used to apply a progressive, 6-level rating to an individual dataset, representing stewardship maturity stages rated as Not Assessed or Not Available (Level 0), Ad Hoc (Level 1), Minimum (Level 2), Intermediate (Level 3), Advanced (Level 4), and Optimal (Level 5).
Evaluation Description	Data Stewardship Maturity Assessment was evaluated by the metadata content editor for the NOAA *OneStop* project using the Scientific Data Stewardship Maturity Assessment Model Template v4.0.
Procedure Reference	Peng, Ge. The Scientific Data Stewardship Maturity Assessment Model Template. 2015-06-23. doi:10.6084/m9.figshare.1211954
Date of Measurement	2016-12-08
Quantitative Results	
Preservability	**advanced**
Accessibility	**minimum**
Usability	**advanced**
Production Sustainability	**advanced**
Data Quality Assurance	**advanced**
Data Quality Control/Monitoring	**minimal**
Data Quality Assessment	**intermediate**
Transparency/Traceability	**intermediate**
Data Integrity	**advanced**
Conformance Results Explanation	Data stewardship maturity assessment was carried out by NOAA *OneStop* metadata content editor, in collaboration with subject matter experts of the product and the maturity matrix.
Reference	**Lemieux, P., G. Peng, and D.J. Scott, 2017: Data Stewardship Maturity Report for NOAA Climate Data Record (CDR) of Passive Microwave Sea Ice Concentration, Version 2. figshare, doi:10.6084/m9.figshare.5279932**

2001, Section 515), Office of Management and Budget and NOAA guidance, and projects requirements on information sharing and transparency.

3. Workflows, Templates, and Tools

To facilitate and automate the process of applying the DSMM to a dataset, capturing the detailed supporting evidence, providing the provenance of the assessment, and integrating the DSMM ratings, a variety of workflows, templates, and tools were developed and are described in this section (**Figure 7**, also Ritchey et al. 2016).

3.1. Data Stewardship Maturity Questionnaire

DSMM evaluates data management and stewardship practices from a dataset being acquired, archived, and disseminated. Obtaining all necessary information on the practices is a laborious task. The most challenging part of the DSMM assessment is obtaining data quality information as it is often not readily available publicly or available in a consistent, machine-readable format. These components require information on practices associated with data quality assurance, control, and assessment, which is best provided by data producers or product stewards. At the present time, the data quality information is collected and derived by dedicated *OneStop* metadata content editors based on the available literature and from direct communication with product stewards. Ideally, the relevant information would be collected and documented early in the data lifecycle and used to automatically map to the DSMM ratings. Different domain experts could contribute to relevant key components of the dataset stewardship maturity assessment to provide a thorough and accurate assessment.

To alleviate some of the current burden for both product stewards and metadata content editors, a Data Stewardship Maturity Questionnaire (DSMQ) was developed to streamline the DSMM assessment process. The DSMQ consists of a set of standardized questions and pre-selected answers, aiming to provide an easy mechanism for collecting information that may not be available publicly (Partee et al. 2018; see **Figure 8** for an example). Some questions allow for linking to publicly available information which, if it is machine readable, would allow for automated field population of the DSMM results. The DSMQ will also help improve the scalability of the assessment process. A web-based user-interface has been developed to guide NOAA data managers through this assessment process using the DSMQ (see *OneStop* Metadata Content Editor Team 2018 for more details on how to sign in and use the tool). The assessment results are captured in a database and encoded into the ISO dataset-level metadata record and the *OneStop* search and discovery algorithm.

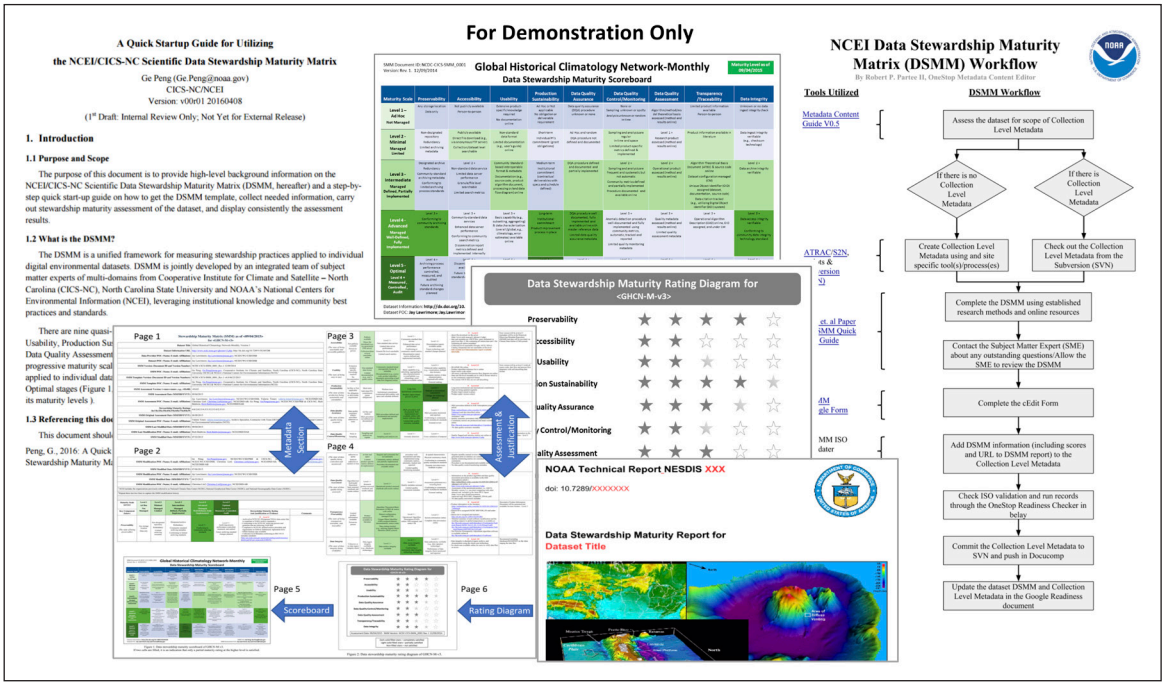

Figure 7: Example DSMM guidance, workflow, templates, and tools developed to facilitate data stewardship maturity assessments. From Ritchey et al. (2016).

3.2. Automation tools and workflows

Tools are developed to automatically generate the DSMM graphics and draft data stewardship maturity reports (DSMRs). **Figure 9** outlines the flow of DSMM data and the automation process of DSMR generation. A template of a web-based form was developed to facilitate the DSMM assessment results collection and integration process. This template builds on an existing NCEI web application, CEdit. CEdit provides a graphical user interface for creating, updating, and exploring DSMM assessment results in the XML format. However, for a bulk operation, it is convenient to use its RESTful API.

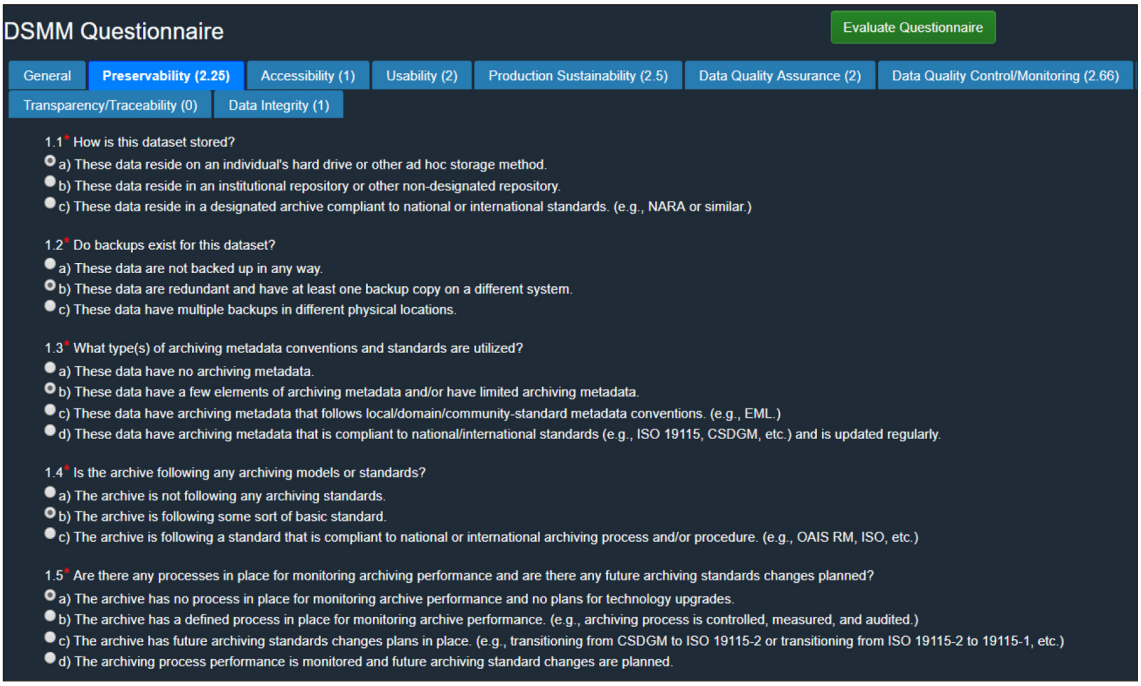

Figure 8: A screenshot of Data Stewardship Maturity Questionnaire (DSMQ) highlighting the questions and possible answers for the Preservability component of the DSMM. Note, the numbers following the component names in the tabs are scores updated in real time as a user proceeds through the survey.

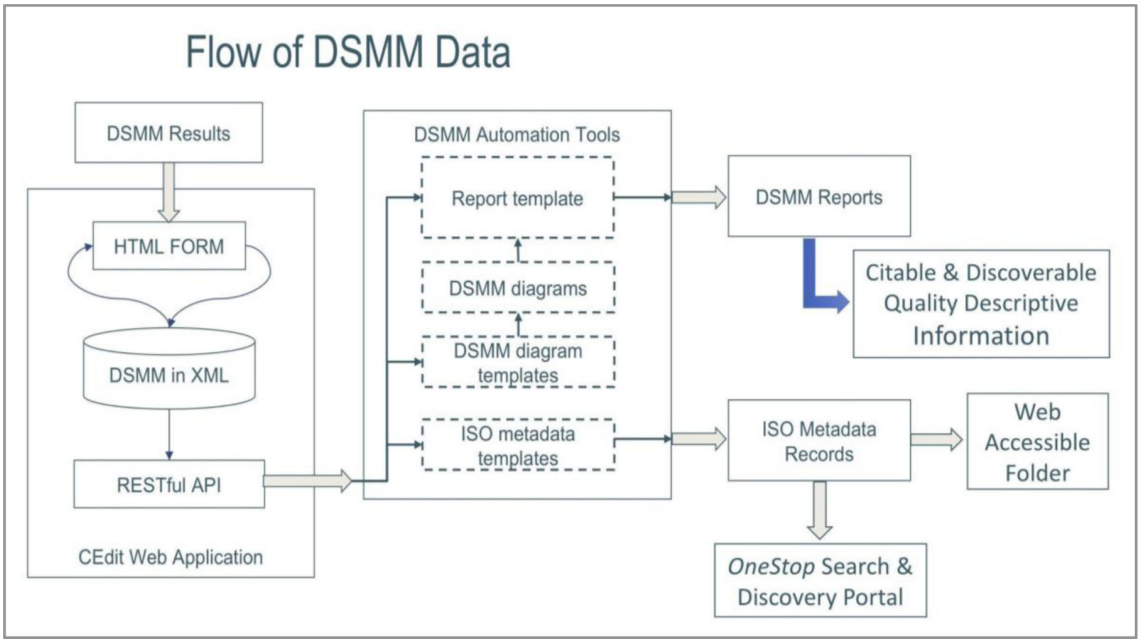

Figure 9: DSMM data flow chart with CEdit implementation diagram (left box area). Adapted from Zinn et al. (2017).

4. Integration to Other Systems and Tools

Figure 10 shows the DSMM assessment integration workflow at NCEI. It portrays the process and outlines the tools used by metadata content editors when assessing the stewardship of metadata records. The process begins by assessing the scope of the dataset. After identifying the structure of the dataset (what are collections, granules, etc.), the process splits among two varying paths based on that analysis. Datasets that have collection-level metadata move along to the next step, while datasets that do not have collection-level

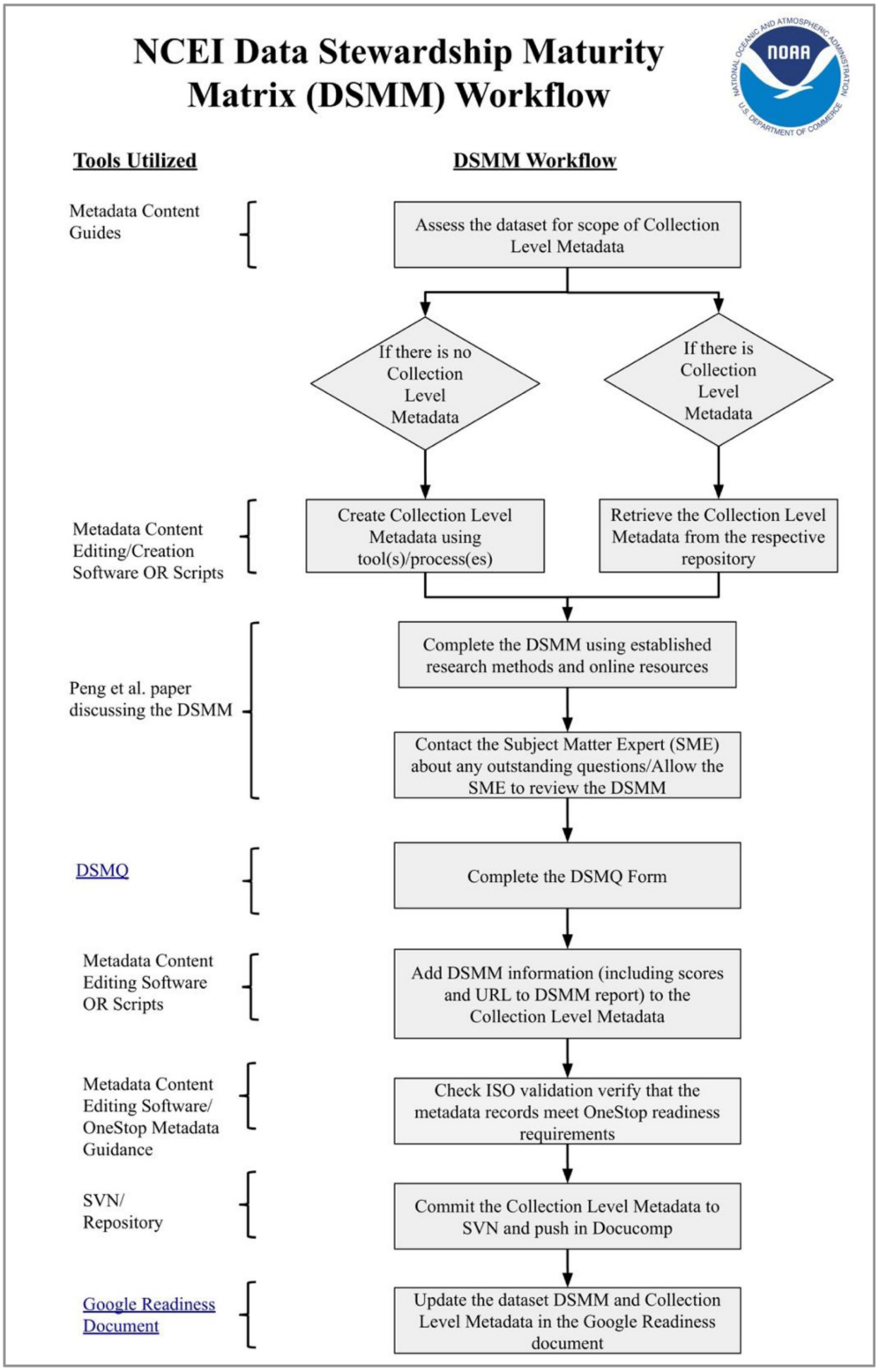

Figure 10: DSMM assessment integration workflow.

metadata must have these records created for them using a preferred metadata editor. Next, the metadata content editor completes the DSMM assessment using established research methods, online resources, and subject matter expert consultations. All of this information is then placed in a web-based form based on Restful API (i.e., CEdit) as well as within the collection-level metadata. Last, the metadata record must be validated to ensure that it meets ISO compliance before being committed back to the subversion (SVN) and pushed to a metadata component management system.

4.1. Integration to ISO dataset-level metadata

The numerical DSMM ratings are systematically encoded into the dataset-level metadata record as a 6-level rating system to each of the nine components of the DSMM, using the ISO data quality metadata standards and XML implementation described in Section 2.4 and Appendix A. Tools are being developed to perform the integration from either the centralized Google spreadsheet or the database.

4.2. Integration to OneStop search and discovery algorithm

Using the 6-level rating system defined in the ISO dataset-level metadata, *OneStop* assigns a numerical value correlated to the given score (integer between 0 and 5, inclusive) to each of the nine components upon ingesting and parsing a data set's metadata record. A mean average value for the components is then calculated and stored in the underlying search index with the dataset's parsed information.

When a user initiates a search against *OneStop*, either via the API directly or through the website, results are given individual scores based on their level of relevancy to the given search text and any provided filters. In addition to this, a separate score is calculated based on the value of the DSMM average, specifically the logarithm of 1 plus its DSMM score (i.e., log(1 + 0) or zero for a dataset without any DSMM score and log(1 + 5) or approx. 0.78 for one with a score of 5). This secondary score is then added to the query score yielding a higher overall result ranking for datasets with higher DSMM averages. The end result is that DSMM ratings serve as a tie-breaker in the *OneStop* search and discovery algorithm for items that would otherwise be equally relevant.

5. Displaying DSMM Ratings

On the *OneStop* portal, a dataset's collection view (see **Figure 11**) visually displays the DSMM average score as fully or partially filled stars. The actual numerical average is displayed by clicking the information icon next to the stars.

6. Summary and Discussion

The combination of rapidly increasing data volume and variety, along with elevated requirements for timely access to high-quality and readily usable and interoperable environmental data and information has posed a significant challenge for effectively managing and servicing the NOAA data. The *OneStop* project was initiated to support NOAA's efforts to improve discovery and access services for NOAA's legacy data.

The DSMM was utilized to provide evidence-based dataset stewardship maturity information for search and discovery as a part of the *OneStop*-ready process. The dataset maturity information is useful to end users to help them make informed decisions for their unique data use requirements.

During the end-to-end application of the DSMM for NOAA datasets, challenges continued to emerge. Initially, DSMM assessments including the justifications were captured manually, using the Peng (2015) template. The results were then entered manually via a Google Form and collected in a Google spreadsheet. Next, the DSMM ratings were integrated into the metadata records through a script. It became quite clear early on that the scalability of assessing, representing, and integrating DSMM ratings needs to be improved and automation is a must in order to apply the DSMM to hundreds to thousands of NOAA datasets. A web-based tool was developed to capture assessment results and automatically generate the draft of the data stewardship maturity reports following a consistent template. The workflows were created to input the ratings into a database for them to be readily integrated into the collection-level metadata records. As this is the first time for NOAA to explicitly curate ISO quality metadata. The implementation best practices are developed by the NCEI Metadata Working Group with help from the NOAA Enterprise Metadata Working Group. The xml representation of ISO DSMM quality metadata is included in Appendix A. The DSMQ approach was developed and implemented to further lessen the burden of collecting relevant information and help improve consistency of the assessments. With the DSMQ approach, the DSMM ratings are automated calculated based on the pre-populated answers to the questionnaire, which are stored in the database and then integrated into the ISO metadata – all are done under an automated workflow.

Figure 11: An example of displaying DSMM rating on the *OneStop* portal.

The best practices developed for the application of DSMM to NOAA datasets, such as those for evaluation of data stewardship maturity, the integration workflow of DSMM assessment results, ISO data quality metadata implementation template, and data maturity report template, help NOAA meet U.S. federal regulations, organizational requirements, and user needs on information quality, including transparency, interoperability, accessibility and usability. These best practices are helpful for improving the scalability of the DSMM application. They can be beneficial to any organization that wishes to utilize the DSMM.

Different maturity assessment models may be developed to assess different quality attributes (see an overview by Peng 2018). DSMM may be adapted by an organization to include additional quality attributes such as the Stewardship Maturity Matrix for Climate Data (SMM-CD; WMO SMM-CD Working Group 2019) developed under the High-Quality Global Data Management Framework (HQ-GDMFC) initiative of the World Meteorological Organization (WMO). The application workflows and best practices described in this paper can also be adapted for practical application of other types of maturity matrices. In that case, modifications may be necessary to account for different key components and scale structure of these maturity matrices.

The DSMM approach is different from an audit certification approached at an archive level, e.g., the OAIS standard ISO 16363 (2012). ISO 16363 (2012) establishes comprehensive audit metrics for what a repository must do to be certified as a trustworthy digital repository. The OAIS certification focuses on the capability of the archive. Three important qualities of trustworthiness are integrity, sustainability, and support for the entire range of digital repositories in three different aspects: organizational infrastructure, digital object management, and infrastructure and security risk management (ISO 16363 2012). The DSMM, on the other hand, focuses on the stewardship practices applied to individual datasets. The two approaches overlap primarily in the area of digital object management. A synergy of the DSMM criteria and the World Data System core trustworthy data repositories requirements (Edmunds et al. 2016) has been observed and will be examined in a future study (Wendy Gross, *personal communication*).

It is anticipated that the practical application of DSMM will evolve. We encourage constructive comments and suggestions from the Earth Science data stewardship community. Workflows or practices may be improved over time. A latest version of this article will be maintained via the Open Science Framework at DOI:10.31219/osf.io/fp3js.

To provide a gradual way for people to get relevant information on or to get started with the DSMM, a consolidated resource in the form of a flow diagram with clickable links has been developed (Peng 2017). The latest version can be downloaded at: https://figshare.com/articles/Getting_To_Know_And_To_Use_DSMM/5346343.

Additional File

The additional file for this article can be found as follows:

- **Appendix A.** The current xml representation for encoding the DSMM assessment ratings as quality metadata into an ISO metadata record. DOI: https://doi.org/10.5334/dsj-2019-041.s1

Acknowledgements

We are grateful for support from the management of NCEI's Data Stewardship Division and Center for Weather and Climate. David Pezdirtz provided input on documentation versioning best practices and Valerie Toner provided feedback on the first draft of the DSMM User Guide. The NOAA Metadata Working Group participated in the development of and reviewed the ISO Data Quality template for DSMM (XML) and the MM-Stew Codelist (XML). Feedback from Kathy Martinolich, Jacqueline Mize, and Yuanjie Li were beneficial. We thank Richard Fozzard and Martin Aubrey for help with CEdit. Mara Sprain coordinated with the NOAA Central Library for developing a workflow to publish DSMRs in the institutional repository with assigned Digital Object Identifier (DOI) for each DSMR and also reviewed the paper. We thank many scientific stewards and product SMEs for providing beneficial information about their datasets and/or reviewing the draft DSMM assessments of their datasets, especially Barry Eakins, Korak Saha, Lihang Zhou, Tom Ryan, Steve Ansari, Axel Graumann, Kenneth Knapp, Melissa Zweng, Tim Boyer, and Matthew Menne. Comments from two DSJ anonymous reviewers are beneficial in improving the quality of this paper.

Funding Information

This work was supported by the NOAA *OneStop* Data Discovery and Access Framework Project.

Competing Interests

The authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

Author Contributions

Anna Milan and Philip Jones led the development of the ISO Data Quality metadata template (XML) and the MM-Stew Codelist (XML) for encoding DSMM ratings into ISO dataset-level quality metadata, which is displayed in Appendix A. Milan also contributed to the content of Table 4 and the manuscript draft.

Nancy Ritchey contributed to and oversaw the application of NCEI/CICS-NC scientific data stewardship maturity matrix (DSMM) to the NOAA *OneStop* Project. She generated Figure 7 and contributed to Table 4 and the manuscript draft.

Kenneth Casey, along with Milan and Ritchey, defined the *OneStop*-ready process and contributed to the manuscript draft.

Robert Partee, Paul Lemieux, and Raisa Ionin carried out and documented the DSMM assessments of NCEI datasets for the *OneStop* project, with beneficial input or feedback from Jones and Casey. They designed the layout of data stewardship maturity reports (DSMRs) and developed the data stewardship maturity questionnaire (DSMQ). Partee generated Table 2 and Figure 10, Lemieux provided Figure 8, and both contributed to the manuscript draft.

Sonny Zinn designed and implemented software tools to automatically capture and integrate DSMM ratings to other *OneStop* systems and to automatically generate the draft of DSMRs. He also implemented a web-based user-interface for DSMQs. All were done in collaboration with Milan, Partee, Lemieux, and Ionin. Zinn also generated Figure 9.

Don Collins led, together with Ionin, the effort of defining the procedure of publishing by and archiving the DSMRs at the NOAA Central Library, with the participation of Partee and Lemieux.

Peng trained the *OneStop* metadata content editors, contributed to or participated in all above efforts. She designed the DSMM graphics, i.e., the DSMM rating scoreboard and diagram. She also drafted the manuscript and generated Figures 1–5 and Table 3, and contributed to Table 4.

Evan McQuinn and Arianna Jakositz contributed to the design and implementation of the *OneStop* search and discovery algorithm. Arianna also provided Figure 11 and contributed to the manuscript draft.

All authors reviewed the manuscript and approved the submission.

Author Information

Dr. Ge Peng is a Research Scholar at the Cooperative Institute for Climate and Satellite-North Carolina (CICS-NC) between North Carolina State University and NOAA's National Centers for Environmental Information (NCEI). Dr. Peng holds a Ph.D. in meteorology and is experienced in assessing and monitoring quality of Earth Science data products. She has extensive knowledge of digital data management and has been leading the cutting-edge research on scientific data stewardship. Dr. Peng is the leading author of the NCEI/CICS-NC scientific data stewardship maturity matrix (DSMM) and participated in the NOAA *OneStop* project as the subject matter expert of the DSMM.

Nancy A. Ritchey is the Archive Branch Chief at NOAA's National Centers for Environmental Information (NCEI) and the lead for the *OneStop* Metadata and Data Improvement Team. She is responsible for preserving NCEI's extensive holdings of environmental data for future generations. Nancy holds a M.S. degree in Atmospheric Science and has extensive knowledge and experience in digital and physical data management, related standards and leading practices. She is a co-author of the DSMM and involved in national and international activities related to data preservation and standards.

Evan McQuinn is a software engineer who works for the Cooperative Institute for Research in Environmental Sciences (CIRES) as an affiliate of NCEI. He has been involved in the architecture and implementation of several enterprise software systems in NCEI, including the *OneStop* search engine and user interface which leverage this work.

Dr. Kenneth S. Casey is the Deputy Director of the Data Stewardship Division in the NOAA National Centers for Environmental Information (NCEI) and the project manager for the NOAA *OneStop*. In these roles, Dr. Casey provides leadership and guidance to NCEI staff and sets the technical direction of division activities, projects, and programs. He coordinates across NCEI and with the broader community to promote NCEI as a responsible citizen of the global environmental data management community, leveraging from and contributing to relevant activities of that community.

Paul A. Lemieux III is a Metadata and Data Archiving Specialist with Riverside Technology, Inc., in support of NOAA's National Centers for Environmental Information (NCEI). He holds a BA in Geography and a MS in

Information Sciences and specializes in geospatial technologies including metadata and analysis. He also has experience in academic and federal libraries. Paul is active in the NOAA Enterprise Metadata Working Group, ESIP (Earth Science Information Partners) Documentation Cluster and the USGS (United States Geological Survey) Community for Data Integration.

Raisa Ionin is a *OneStop* Metadata Content Editor with Riverside Technology, Inc., in support of NOAA's National Centers for Environmental Information (NCEI). She holds a Master of Science in Information and Library Science (MLIS) from the University of Maryland and has worked as an academic research librarian, web developer, data analyst, and knowledge management professional in past positions for various libraries and government agencies.

References

Casey, K. 2016. The NOAA *OneStop* data discovery and access framework project. Version: June 3, 2016. [Available online at: https://ioos.noaa.gov/wp-content/uploads/2016/06/OneStop-IOOS-DMAC-03-June-2016.pdf].

Casey, K, Fischman, D, Hausman, S and **Relph, J.** 2015. NOAA *OneStop* Project Charter. *NOAA's National Centers for Environmental Information*. USA, 8 Version: 1.2 10/02/2015.

Delk, Z and **Milan, A.** 2018. Enhancing discoverability and access for your data in *OneStop*. *NCEI Branch Seminar*, September 11, 2018.

Eaton, B, Gregory, J, Drach, B, Taylor, K, Hankin, S and **others.** 2014. NetCDF Climate and Forecast Metadata Conventions. Version: 1.7.2. 28 March 2014. [Available online at: http://cfconventions.org/Data/cf-conventions/cf-conventions-1.7/build/cf-conventions.html].

Edmunds, R, L'Hours, H, Rickards, L, Trilsbeek, P and **Vardigan, M.** 2016. Core Trustworthy Data Repository Requirements. [Available online at: https://zenodo.org/record/168411-.WV_NQBPyuSN]. DOI: https://doi.org/10.5281/zenodo.168411

GCMD. 2018. Global Change Master Directory (GCMD) Keywords, Version 8.6. Greenbelt, MD: Global Change Data Center, Science and Exploration Directorate, Goddard Space Flight Center (GSFC) National Aeronautics and Space Administration (NASA). [Available online at: https://earthdata.nasa.gov/about/gcmd/global-change-master-directory-gcmd-keywords].

Hou, C-Y, Mayermik, M, Peng, G, Duerr, R and **Rosati, A.** 2015. Assessing formation quality: Use case studies for the data stewardship maturity matrix. Poster. *2015 AGU Fall meeting*, 14–18 December 2015, San Francisco, CA, USA.

ISO 16363. 2012. Space data and information transfer systems – Audit and certification of trustworthy digital repositories. Version: ISO 16363:2012. Geneva, Switzerland.

Lawrimore, J, Toner, V, Lief, C, Peng, G and **Baldwin, R.** 2015. The Stewardship Maturity Assessment Document for the Global Historical Climatology Network (GHCN) Monthly Product, Version 3. Document ID: GHCN-M-v3_MM-Stew. Version: v01r03 20150904. *Figshare*. DOI: https://doi.org/10.6084/m9.figshare.7195715

Lemieux, P, III, Peng, G and **Scott, DJ.** 2017. Data Stewardship Maturity Report for NOAA Climate Data Record (CDR) of Passive Microwave Sea Ice Concentration, Version 2. *Figshare*. DOI: https://doi.org/10.6084/m9.figshare.5279932

Li, Y, Milan, A and **Jones, P.** 2017. Light under ISOLite. *2017 NOAA Environmental Data Management Workshop*, 9–10 January 2017, Bethesda, MD. [Available online at: https://nosc.noaa.gov/EDMW_2017/2017-EDMW-presentations/3B/3B.3%20EDMW-2017-NCEI-1.pptx].

National Research Council. 2007. Environmental data management at NOAA: Archiving, stewardship, and access. 116 Washington, DC: The National Academies Press. DOI: https://doi.org/10.17226/12017

NOAA. 1993. NOAA Administrative Order 201–32G: Scientific and Technical Publications for the NESDIS line office. Version: 02/04/1993. [Available online at: http://www.corporateservices.noaa.gov/ames/administrative_orders/chapter_201/201-32G.html].

NOAA. 2010. *NOAA's Next Generation Strategic Plan*. Silver Spring, MD: US National Oceanic and Atmospheric Administration, 2010. [Available online at: http://www.ppi.noaa.gov/ngsp/].

Obs4MIPs. 2017. Observations for Climate Model Intercomparisons. [Available online at: https://www.earthsystemcog.org/projects/obs4mips].

***OneStop* Metadata Content Editor Team.** 2018. Data Stewardship Maturity Questionnaire (DSMQ) User's Guide. *NOAA OneStop Project*. [Available online at: https://geo-ide.noaa.gov/wiki/index.php?title=Data_Stewardship_Maturity_Questionnaire_(DSMQ)_User's_Guide].

Partee, RP, II, Lemieux, P, III, Ionin, R and **Peng, G.** 2018. A Streamlined Approach to Assessing the Stewardship Practices of NOAA Datasets. Poster. *NOAA Environmental Data Management Workshop*, April 23–24, 2018. Silver Spring, MD, USA.

Peng, G. 2015. NCDC/CICS-NC Scientific Data Stewardship Maturity Matrix Self-Evaluation Template. *Figshare*. Version: NCDC-CICS-SMM-0001-Rev.1 v4.0 06/23/2015. [The latest version is available online at: https://figshare.com/articles/NCDC_CICSNC_SDSMM_Template/1211954]. DOI: https://doi.org/10.6084/m9.figshare.1211954

Peng, G. 2017. Getting to know and to use NCEI/CICS-NC Data Stewardship Maturity Matrix (DSMM). *Figshare*. DOI: https://doi.org/10.6084/m9.figshare.5346343

Peng, G. 2018. The state of assessing data stewardship maturity – an overview. *Data Science Journal*. DOI: https://doi.org/10.5334/dsj-2018-007

Peng, G, Lawrimore, J, Toner, V, Lief, C, Baldwin, R, Ritchey, N, Brinegar, D and **Delgreco, SA.** 2016. Assessing Stewardship Maturity of the Global Historical Climatology Network-Monthly (GHCN-M) Dataset: Use Case Study and Lessons Learned. *D.-Lib Magazine*, 22. DOI: https://doi.org/10.1045/november2016-peng

Peng, G, Privette, JL, Kearns, EJ, Ritchey, NA and **Ansari, A.** 2015. A unified framework for measuring stewardship practices applied to digital environmental datasets. *Data Science Journal*, 13. DOI: https://doi.org/10.2481/dsj.14-049

Ritchey, N and **Peng, G.** 2015. Assessing stewardship maturity: use case study results and lessons learned. *2015 AGU Fall meeting*, 14–18 December 2015, San Francisco, CA, USA.

Ritchey, NA, Peng, G, Jones, P, Milan, A, Lemieux, P, Partee, R, Ionin, R and **Casey, KA.** 2016. Practical Application of the Data Stewardship Maturity Model for NOAA's *OneStop* Project. Abstract #IN43D-08. *2016 AGU Fall Meeting*, 12–16 December 2016, San Francisco, CA, USA.

USGEO (The U.S. Group on Earth Observations). 2015. Common Framework for Earth-observation data. Version: 02 December 2015. 39 [Available online at: https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/common_framework_for_earth_observation_data_draft_120215.pdf].

US Public Law 106–554. 2001. Information Quality Act. *Publ. L. 106–554*. 101. [Available online at: http://www.gpo.gov/fdsys/pkg/PLAW-106publ554/html/PLAW-106publ554.htm].

WMO SMM-CD Working Group. 2019. The guidance booklet on the WMO-Wide Stewardship Maturity Matrix for Climate Data. Document ID: WMO-SMM-CD-0002. Version: v03r00 20190131. *Figshare*. DOI: https://doi.org/10.6084/m9.figshare.7002482

Zinn, S, Relph, J, Peng, G, Milan, A and **Rosenberg, A.** 2017. Design and implementation of automation tools for DSMM diagrams and reports. *2017 ESIP winter meeting*, January 11–13, 2017, in Bethesda, MD. [Available online at: http://commons.esipfed.org/sites/default/files/Zinn_etal_OneStop_DSMM_ESIP%2020170113_0.pdf].

How to cite this article: Peng, G, Milan, A, Ritchey, NA, Partee, RP, II, Zinn, S, McQuinn, E, Casey, KS, Lemieux, P, III, Ionin, R, Jones, P, Jakositz, A and Collins, D. 2019. Practical Application of a Data Stewardship Maturity Matrix for the NOAA *OneStop* Project. *Data Science Journal,* 18: 41, pp. 1–18. DOI: https://doi.org/10.5334/dsj-2019-041

Submitted: 18 October 2018 **Accepted:** 09 August 2019 **Published:** 23 August 2019