A UNIFIED FRAMEWORK FOR MEASURING STEWARDSHIP PRACTICES APPLIED TO DIGITAL ENVIRONMENTAL DATASETS

Ge Peng[1*], Jeffrey L Privette[2], Edward J Kearns[2], Nancy A Ritchey[2], and Steve Ansari[2]

[1] NOAA's Cooperative Institute for Climate and Satellites, North Carolina State University (CICS-NC) and NOAA's National Climatic Data Center (NCDC), 151 Patton Avenue, Asheville, North Carolina, USA
**Email:* Ge.Peng@noaa.gov
[2]NOAA's NCDC, 151 Patton Avenue, Asheville, North Carolina, USA

ABSTRACT

This paper presents a stewardship maturity assessment model in the form of a matrix for digital environmental datasets. Nine key components are identified based on requirements imposed on digital environmental data and information that are cared for and disseminated by U.S. Federal agencies by U.S. law, i.e., Information Quality Act of 2001, agencies' guidance, expert bodies' recommendations, and users. These components include: preservability, accessibility, usability, production sustainability, data quality assurance, data quality control/monitoring, data quality assessment, transparency/traceability, and data integrity. A five-level progressive maturity scale is then defined for each component associated with measurable practices applied to individual datasets, representing Ad Hoc, Minimal, Intermediate, Advanced, and Optimal stages. The rationale for each key component and its maturity levels is described. This maturity model, leveraging community best practices and standards, provides a unified framework for assessing scientific data stewardship. It can be used to create a stewardship maturity scoreboard of dataset(s) and a roadmap for scientific data stewardship improvement or to provide data quality and usability information to users, stakeholders, and decision makers.

Keywords: Stewardship maturity, Data quality, Data interoperability, Digital Earth data, Environmental data, Preservability, Accessibility, Usability, Data integrity, Transparency

1 BACKGROUND

From commercial users in the private sector to researchers and educators in the public sector, digital environmental data users are asking for data to be dependable in terms of quality and production sustainability, to be from credible, secure, and authoritative sources, to be easily and publicly accessible online, and to be easily usable in a standard-based common data format with relevant documentation. With data volume rapidly increasing and a larger pool of products becoming available, many users want the ability to view spatial and temporal distributions of a given variable before requesting a subset of the data. Users are also requesting that documentation about the data be readily available online, including information on retrieving or deriving algorithms, data quality, and data characteristics such as climatology and uncertainty estimates. These user requirements are why ensuring and improving data quality, accessibility, usability, and production sustainability are important aspects of digital environmental data stewardship.

Climate observations and products are important assets in qualifying and monitoring climate change. Decisions about climate change monitoring and adaptation are complex with long-lasting economic, social, and political implications. It is therefore critical for such decisions to be based on consistent and credible data. Very often, decisions may need to leverage climate model projections. Modelers need to know and understand the upstream data quality management practices applied to their input datasets to help improve their model projections and better quantify the uncertainty associated with those projections, especially the uncertainty stemming from the quality of the observations (US CLIVAR Scientific Steering Committee, 2013).

Stewards of digital environmental data records are responsible for shepherding and safeguarding those valuable records to make sure that the data are being ingested, preserved, and served to users accurately and securely.

Reliable, robust, and transparent practices associated with these aspects of data stewardship are essential for meeting data and information system integrity requirements as well as security requirements imposed by U.S. laws and federal agency guidelines.

In response to the U.S. Information Quality Act (U.S. Public Law 106-554, 2001), the U.S. Office of Management and Budget (OMB) provided policy and guidelines for governmental agencies on data quality, objectivity, integrity, and transparency of information including digital data (OMB, 2002). Consequently, the National Oceanic and Atmospheric Administration (NOAA) issued Administrative Order 212-15 (NOAA, 2008a) on the management of environmental and geospatial data, requiring NOAA agencies, including its data centers, to "take appropriate steps to ensure acceptable accuracy, precision, representativeness, documentation, and long-term continuity of NOAA's quality data sets for the user community."

More recently, the Obama Administration has instituted a number of initiatives to promote the openness and availability of government data. The February 22, 2013 memorandum from the Office of Science and Technology Policy (OSTP, 2013) requires that digital scientific data funded by governmental agencies be made available to and useful for the public, industry, and the scientific community. The Open Data Policy memorandum issued by OMB (2013) requires U.S. governmental agencies "to collect or create information in a way that supports downstream information processing and dissemination activities". OSTP (2013) recommends that all federal agencies make plans to improve the public's ability to locate and access digital scientific data, encourage innovation in accessibility and interoperability, and measure and enforce compliance with federal regulations.

As with any process of improvement planning, agencies need to find out where they are in terms of their compliance to the federal regulations and what they need to do if any areas of non-compliance are identified. To this end, a unified framework would be beneficial for assessing the current stage of stewardship practices applied to individual datasets and for providing a roadmap that will guide future investments towards enhanced stewardship of environmental datasets.

Currently, there is no systematic framework to assess the vigor of stewardship practices applied to individual environmental datasets or to provide consistent information on data quality, data integrity and usability to users and stakeholders (Peng & Privette, 2014). One exception is the use of maturity assessment models for preservation. However, those preservation models steer towards assessing and improving processes within organizations associated with preservation of digital records (Kenney & McGovern, 2003; OCLC & CRL, 2007; Dollar & Ashley, 2009). The focus has been on assessing the level of maturity of digital repositories that range from academic institutional repositories to large data archives in terms of reliable storage, ingest, migration, and access to their collections instead of individual holdings. (A good overview of six preservation maturity models can be found in Bailey (2014).) To add complexity to the issue, datasets stewarded by the same institution are often governed differently. Furthermore, data products produced by the same organization are often in various levels of maturity in terms of their data quality, accessibility, and usability as well as the states of completeness of data quality metadata and documentation.

The existing preservation maturity models of digital data mostly cover the key functional entities that are the core of the digital preservation systems or organizations. For example, four levels of digital preservation are defined by the National Digital Stewardship Alliance (NDSA), organized into five functional areas: storage and geographic location, file fixity and data integrity, information security, metadata, and file format (Phillips, Bailey, Goethals, & Owens, 2013). Qualifying the preservation maturity level associated with a repository or an archive is important for addressing the question of whether the institution is trustworthy and for providing guidelines that the institution can use to develop its digital preservation programs (Kenney & McGovern, 2003). However, a maturity level of a repository does not necessarily yield practical information to users for the repository's individual datasets unless the repository has reached a certain level of the preservation maturity (e.g., level 3 or higher). It then implies consistent maturity in all its data holdings in those preservation functional areas. Currently, even a nationally or internationally credited archive tends to have a collection of data holdings in various stages of maturity with respect to stewardship practices. Furthermore, data usability and quality are not addressed explicitly in those preservation maturity models.
A critical and integral part of environmental data stewardship lies in scientific vigilance, i.e., oversight of data stewardship by scientists (NRC, 2005; 2007). This is analogous to business products where "data management is a shared responsibility between the business data stewards serving as trustees of enterprise data assets and technical

data stewards serving as the expert custodians and curators for these assets" (DAMA International, 2008; 2010). Successful long-term stewarding of scientific data products requires a shared responsibility of data stewards, technical professionals, and scientific stewards. Data stewards provide data management and governance knowledge and guidance. Technical professionals provide software development and system engineering support. Scientific stewards provide expert knowledge about the subject that the dataset is associated with, such as temperature or precipitation. Scientific stewards may provide information or guidance on data quality and characterization of the data to users and may also provide scientific oversight to ensure the accurate scientific representation of values. The meaning of those values is just as important as the accuracy of the values themselves, if not more so. Therefore, recognizing the role of scientific stewards in caring for scientific data is an important step forward in ensuring data quality and improving usability.

The product maturity assessment model described by Bates and Privette (2012) for individual climate data products is one of the few maturity models that explicitly address data quality. It measures the readiness of long-term climate data records for the transition from research to operation over six categories: software, metadata, documentation, product validation, public access, and utility. This product maturity model examines the stability of source code development associated with creating the product, the compliance of the code with the defined coding standards, the maturity of the product algorithm, the validation and application of the product, and the completeness of metadata and documentations. The product maturity model primarily assesses the maturity and utility of the products during the product development or refinement stage, but it also provides guidance on the product readiness in the areas of accessibility, usability, and transparency. This product maturity assessment model has been adapted by the NOAA's satellite Climate Data Records Program (CDRP) (Privette, Bates, Karl, Barkstrom, & Kearns, 2009, the current climate data records maturity matrix template can be accessed at http://www.ncdc.noaa.gov/cdr/guidelines.html) and the European Union's COordinating Earth observation data validation for RE-analysis for CLIMAte Services (CORE-CLIMAX) project (EUMETSAT, 2013).

The value and quality of a dataset depends in part on the stewardship practices applied after its development and production. Therefore, a unified framework providing a holistic view of the quality of stewardship practices applied to individual datasets is beneficial to data stewards and users. In this paper, we present a stewardship maturity assessment model for digital environmental datasets. It follows a similar approach used in Bates and Privette (2012) but with a modified scale structure. Bates and Privette (2012) used a 6-level scale structure for their product maturity matrix. It is extremely difficult to be progressive with six levels in all components of our stewardship maturity matrix (SMM). After an examination of scale structures of various maturity models and mapping of stewardship maturity levels, we chose to adapt the naming conventions of the scale structures of the Capability Maturity Model Integration (CMMI) (SEI, 2010) and the Dollar and Ashley (2009) digital preservation capability maturity model. This provides us with a more progressive and representative 5-level scale structure.

2 THE SCOPE OF DATA TYPES

In this version of the scientific data stewardship maturity assessment model (version 1.0), data types will be limited to NOAA digital environmental and geospatial data products in an attempt to put bounds on its suitability and utility. The data product variables can be, but are not limited to, the essential climate variables (ECV) as defined by the World Meteorological Organization (WMO) for the Global Climate Observing System (WMO GCOS, 2010). However, care is taken for the stewardship maturity model to be readily applicable to similar datasets in other organizations and other digital geospatial variables.

Digital data, distinguished from physical records, such as paper weather reports, are represented in discrete numerical form that can be used by a computer or electronic device. Thus, preserving digital data will without doubt involve digital devices, the structure and nature of the data file formats, and applications applied to those files during data production, transferring, ingest, storage, access, and dissemination.

Environmental data are defined as the recorded and/or derived observations and measurements of the physical, chemical, biological, geological, or geophysical properties or conditions of the oceans, atmosphere, space environment, sun, and solid earth as well as correlative data and related documentation or metadata (NOAA, 2008a). For example, geoscience data products are those pertaining to the planet Earth. Media, including voice recordings and photographs, will not be included. In this paper, datasets are assumed to be publicly available without any restriction, although restricted datasets can be managed through methods such as user authorization and authentication. To this end, online means to be available and accessible online publicly unless mentioned otherwise.

Climate data are useful to consider as they are a subset of environmental data that is particularly sensitive to stewardship due to the length of its practical life cycle, breadth of scope, and the required consistency across its period of record. Climate is the historical behavior of atmosphere and ocean systems, weather is the day-to-day conditions of atmosphere and oceans in a region and their short-term (from minutes to weeks) variation (NASA, 2005). The WMO GCOS (2010) identified 50 ECV variables as "technically and economically feasible for systematic observation" and necessary to investigate and monitor climate change. This set of ECVs covers variables in atmospheric, oceanic, and terrestrial domains including temperature, wind, precipitation, salinity, water, ice caps and sheets, snow cover, and radiation.

Geospatial data describe the state and impact of environmental systems and include information on the geographic location and characteristics of constructed features and boundaries of the earth (EPA, 2005; NOAA, 2008a). Therefore, information about spatial and temporal characteristics of data products and support for spatial and temporal sub-setting will make it easier for users of various kinds to get and use the data products efficiently. It will also help users in selecting the most appropriate products for their needs and applying the products in a suitable way. For example, it is not suitable to use a daily product to examine the diurnal cycle or to use a product with a 2.5° x 2.5° horizontal resolution to investigate cloud or convective systems.

3 THE SCOPE OF SCIENTIFIC DATA STEWARDSHIP

Before key components of the stewardship maturity matrix and stewardship practices associated with each key component can be identified, the scope of the scientific data stewardship needs to be defined. What is scientific data stewardship? The National Research Council (NRC, 2007) has defined scientific data stewardship to encompass "all activities that preserve and improve the information content, accessibility, and usability of data and metadata." It also stressed that expert stewardship is required for both data and metadata.

Figure 1 outlines the scope of long-term scientific data stewardship for environmental data. The top level defines entities under which "non-functional" requirements are asserted on scientific data stewardship. The terms non-functional and functional requirements are often used in systems engineering to define, in a broad sense, what a system is supposed to be and to do. The term "non-functional requirements" is used here to refer to constraints imposed by U.S federal regulations and agency policies on the stewardship of environmental data. Non-functional requirements asserted on environmental data products can be deduced based on the aforementioned U.S. laws, agencies' guidelines, and expert bodies' recommendations (See Peng & Privette, 2014 for an example of those on NOAA's climate data records). Understanding the non-functional requirements is the first step in systematically determining the scope of scientific data stewardship (see the process flow diagram outlined in Figure 2).

Functional requirements, on the other hand, specify the functions associated with scientific data stewardship. In a general sense, policy sets guidelines under which procedures are developed and standards are defined in relevant functional areas. Practices are applied to datasets when tasks are carried out following steps of a defined procedure for data creation and stewardship.

Entities that pertain to scientific data stewardship are data preservability, accessibility/usability, sustainability, data quality, transparency/traceability/reproducibility, and information integrity (Figure 1). Data utility is not included in stewardship here as the discussion is ongoing as to whether it should be treated as an entity for service rather than for data stewardship.

Figure 1 also displays the major functional areas that enable non-functional requirements in the aforementioned stewardship entities to be met, through the second step outlined in Figure 2. It is very common for many of these major functional areas to be intertwined or overlapping. In order to help guide the identification of key components associated with measurable stewardship practices applied to individual datasets (the third step in Figure 2), subjective decisions are made to artificially separate and group them under the most relevant stewardship entity or entities. (Figure 1 also captures processes and services that are essential to successful scientific data stewardship, e.g., data quality management, metadata management, user service, and technical support.)

Major functional areas for preservability include product evaluation, product acquisition, data archive, and data governance. Product quality can be assessed using a product maturity assessment model (e.g., Bates & Privette, 2012). NOAA has a defined process, i.e., the NOAA Procedure for Scientific Records Appraisal and Archive Approval (SRAAA) (NOAA, 2008b), to ensure that archive resources are utilized to provide stewardship and long-term availability of the scientific records most desired by the scientific and user community. Therefore, only measureable stewardship practices in data archive and governance are examined for the preservability key component in the stewardship maturity matrix.

Functional areas for data accessibility and usability deal with the availability of data and how easy it is for users to find, get, and use data. Measureable stewardship practices in data search and discovery will be included in data accessibility while those in data format and product documentation are included in data usability.

Data operations and maintenance, product update, improvement and reprocessing are integral parts of product sustainability. They are all closely related to the availability and stability of funding resources. Therefore, for this paper, sustainability will focus on stability of resources by examining the level of commitment on product.

Ensuring data quality is critical during data creation (data quality assurance), after its creation (data quality assessment), and throughout the life cycle of data (data quality monitoring). Understanding what data quality practices have been applied to datasets is important to users, including modelers and decision-support tools developers. Information on data provenance, practices in data reference and citations are crucial for transparency and traceability, which will optimally lead to reproducibility.

Information integrity includes information security and data integrity. Information security includes security practices applied to both datasets and information systems that process, store, or host the datasets. Due to the potential difficulty in measuring security practices applied to information systems and due to the potential risk to the information systems in publicizing the practices applied to, we will only examine measureable practices associated with data integrity. More detailed discussion on information integrity is provided in Section 4.9.

The focus and rational of selected stewardship practices in the stewardship maturity matrix as the final step in Figure 2 will be described in more detail in Section 4.

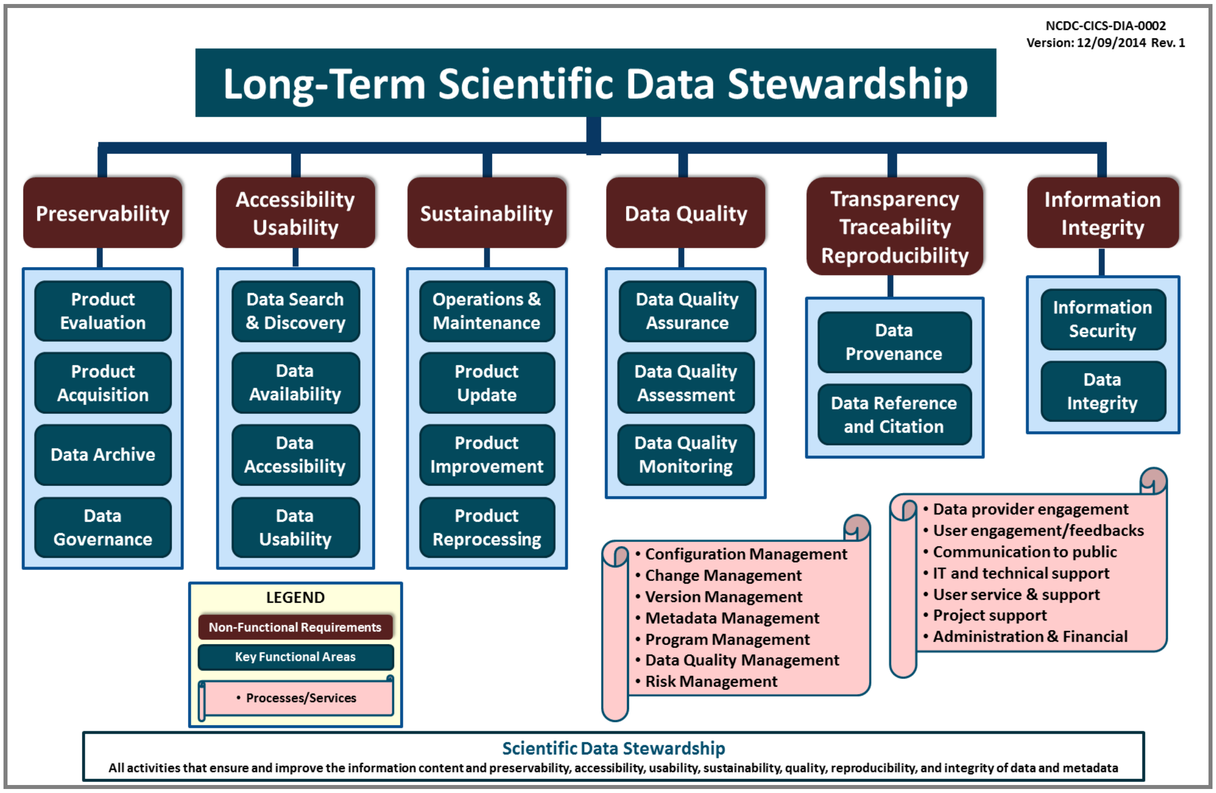

Figure 1. Diagram of the scope of long-term data stewardship for environmental data

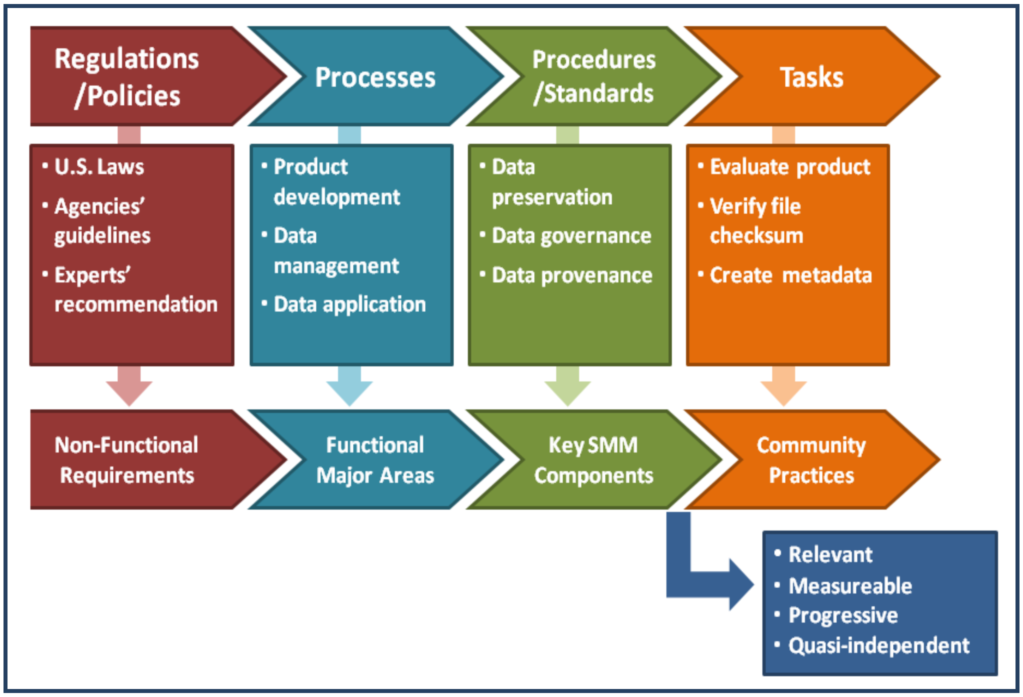

Figure 2. Flow diagram of the process of systematically identifying key components and defining levels of the scientific data stewardship maturity matrix (SMM) from regulations/policies, processes, procedures/standards to community practices

4 KEY COMPONENTS AND THE SCOPE OF EACH COMPONENT

Nine key components that are relevant to scientific data stewardship and measurable in a progressive fashion are identified in this baseline version of the stewardship maturity matrix (NCDC-CICS-SMM-0001-Rev.1) (Table 1). These components are preservability, accessibility, usability, production sustainability, data quality assurance, data quality control/monitoring, data quality assessment, transparency/traceability, and data integrity. Three key components on data quality are selected to reflect the importance of quality assurance during data creation, quality assessment after its creation, and quality monitoring throughout the life cycle of data.

For each component, a five-level progressive maturity scale is defined to assess stewardship practices applied to individual datasets, representing Ad Hoc, Minimal, Intermediate, Advanced, and Optimal stages. Level 1 is not managed with no procedures or standards defined. Level 2 is managed in a limited fashion. At level 3, procedures or standards are defined and managed but only partially implemented. Level 4 represents an advanced stage where procedures or standards are well-defined, managed, and fully implemented. The optimal stage, i.e., Level 5, represents the stage where the performances of the defined procedures are measured, controlled, and audited. For some key components, this level requires external audit. Level 5 also includes planning for future standards or technology changes to ensure that data are always credible, meaningful, and usable.

The best practices captured in this stewardship maturity matrix are not new but emerged from practices that are widely-used in the environmental data community. Assessing the stewardship maturity of a dataset touches on all aspects of stewarding for scientific data. It is usually beyond one person’s ability to carry out the task of defining a stewardship maturity assessment model as it requires a body of knowledge in multiple disciplines. The goals of our undertaking, which leverages institutional knowledge and community best practices, are to help alleviate the burden of data stewards from defining their own assessment models, to reduce incompatibility of stewardship maturity assessment results from individually defined models, and to provide a unified and holistic view of stewardship practice maturity of individual datasets to users and stakeholders. It is our hope to enable improvements in scientific data stewardship by providing the matrix to stewards and users. The optimal goal is to aid stewards with a tool that guides future improvements in their datasets and provide users with a consistent and easy to understand metric for individual datasets.

In this section, the scope and rationale for each key component and its five-level progressive maturity scale are outlined. When necessary, descriptions or definitions of terms used in the matrix may be given for clarity. In some instances, examples are also given to provide one of the various possible community-accepted best practices. However, we do not suggest that they are the only recommended standards. Community-accepted standards and guidelines may vary with different data user communities.

The good practices associated with implemented community processes and standards within an organization will provide a consistent framework for the organization to potentially achieve maturity level 3 for all its environmental data holdings. Maturity level 3, therefore, is the recommended stewardship maturity level for digital environmental and geospatial data products at nationally designated archives, such as NOAA’s data centers, especially for data sets of ECVs. Maturity levels 4 and 5 represent enhanced and optimal stages for each key component. The best practices at those levels may be too costly or time consuming to implement for most of the datasets. They are, however, recommended for high-impact and high-utility digital environmental datasets.

Having or adopting an open-standard will increase our ability to integrate and utilize multiple datasets. Enhanced interoperability will minimize data integration effort and reduce up-front and maintenance costs. The essential characteristics of an open-standard are free, publicly available, and data and vendor neutral, with a community consensus decision-making process. Recognizing the need for and benefits of open and consistent implementation of standards, an international collaborative effort was initiated by the United Nations on Global Geospatial Information Management (UN-GGIM). This is an on-going effort to develop open standards and to provide the global geospatial information community with much-needed guidance on adopting and implementing standards (UN-GGIM, 2014).

Metadata help establish the context of data. Metadata capture and describe information about data, ranging from the time frame and spatial extent of the data to data processing algorithms and steps. They have been traditionally considered as a separate element in maturity assessment models. However, as one will see in this section, there are certain elements of metadata that are more relevant to one key component than to others. The completeness of metadata in one area, such as preservability, does not necessarily provide sufficient information in other areas, such as data usability or traceability. Therefore, in this version, metadata elements have been categorized into nine key components of the stewardship maturity matrix. At this time only a crude and high-level definition will be provided along with a recommendation of elements for some matrix key components to provide some guidance for implementation.

The standards or technologies may change over the lifetime of many datasets. For example, the types of media for storing data or the technology to retrieve files from file systems and to serve data to users could change over time. The computer systems and their operating systems used to process or reprocess datasets may also change. Although the exact change may not be known at the current time, for long-term viable stewardship, it is important to be vigilant about the prospect of future change and to have a procedure in place. In this version, this issue is explicitly addressed in the maturity level 5 for the preservability, accessibility, and production sustainability components.

Table 1. The stewardship maturity matrix for digital environmental data products (Documentation ID: NCDC-CICS-SMM-0001; Version: 12/09/2014 Rev. 1)

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	Any storage location Data only	Non-designated repository Redundancy Limited archiving metadata	Designated archive Redundancy Community-standard archiving metadata Conforming to limited archiving standards	Level 3 + Conforming to community archiving standards	Level 4 + Archiving process performance controlled, measured, and audited Future archiving standard changes planned
Accessibility	Not publicly available Person-to-person	Publicly available Direct file download (e.g., via anonymous FTP server) Collection/dataset level searchable online	Level 2 + Non-standard data service Limited data server performance Granule/file level searchable Limited search metrics	Level 3 + Community-standard data service Enhanced data server performance Conforming to community search metrics Dissemination report metrics defined and implemented internally	Level 4 + Dissemination reports available online Future technology and standard changes planned
Usability	Extensive product-specific knowledge required No documentation online	Non-standard data format Limited documentation (e.g., user's guide) online	Community standard-based interoperable format & metadata Documentation (e.g., source code, product algorithm document, processing or/and data flow diagram) online	Level 3 + Basic capability (e.g., subsetting, aggregating) & data characterization (overall/global, e.g., climatology, error estimates) available online	Level 4 + Enhanced online capability (e.g., visualization, multiple data formats) Community metrics of data characterization (regional/cell) online External ranking
Production Sustainability	Ad Hoc or Not applicable No obligation	Short-term Individual PI's commitment	Medium-term Institutional commitment (contractual deliverables with	Long-term Institutional commitment Product improvement process	Level 4 + National or international commitment

	or deliverable requirement	(grant obligations)	specs and schedule defined)	in place	Changes for technology planned
Data Quality Assurance	Data quality assurance (DQA) procedure unknown or none	Ad Hoc and random DQA procedure not defined and documented	DQA procedure defined and documented and partially implemented	DQA procedure well documented, fully implemented and available online with master reference data Limited data quality assurance metadata	Level 4 + DQA procedure monitored and reported Conforming to community quality metadata & standards External review
Data Quality Control/ Monitoring	None or Sampling unknown or spotty Analysis unknown or random in time	Sampling and analysis are regular in time and space Limited product-specific metrics defined & implemented	Level 2 + Sampling and analysis are frequent and systematic but not automatic Community metrics defined and partially implemented Procedure documented and available online	Level 3 + Anomaly detection procedure well-documented and fully implemented using community metrics, automatic, tracked and reported Limited quality monitoring metadata	Level 4 + Cross-validation of temporal & spatial characteristics Physical consistency check Conforming to community quality metadata & standards Dynamic providers/users feedback in place
Data Quality Assessment	Algorithm/method/model theoretical basis assessed (methods and results online)	Level 1 + Research product assessed (methods and results online)	Level 2 + Operational product assessed (methods and results online)	Level 3 + Quality metadata assessed Limited quality assessment metadata	Level 4 + Assessment performed on a recurring basis Conforming to community quality metadata & standards External ranking
Transparency /Traceability	Limited product information available Person-to-person	Product information available in literature	Algorithm Theoretical Basis Document (ATBD) & source code online Dataset configuration managed (CM) Unique Object Identifier (OID) assigned (dataset, documentation, source code) Data citation tracked (e.g., utilizing Digital Object Identifier (DOI) system)	Level 3 + Operational Algorithm Description (OAD) online, OID assigned, and under CM	Level 4 + System information online Complete data provenance online

Data Integrity	Unknown or no data ingest integrity check	Data ingest integrity verifiable (e.g., checksum technology)	Level 2 + Data archive integrity verifiable	Level 3 + Data access integrity verifiable Conforming to community data integrity technology standard	Level 4 + Data authenticity verifiable (e.g., data signature technology) Performance of data integrity check monitored and reported

4.1 Preservability

The preservation of digital objects is a fairly mature area with well-defined processes, standards, and best practices. One of these is the Open Archival Information System (OAIS) Reference Model (RM), developed with broad international participation and adopted by the International Standards Organization (ISO) in 2003 (Lavoie, 2000; ISO 14721, 2012; CCSDS, 2012). The OAIS RM provides a conceptual framework applicable to any long-term digital archiving organization and offers a common ground with shared concepts and terminology relevant to preservation and access of digital objects over the long term (ISO 14721, 2012).

The focus under the preservability key component is fairly narrow compared to other preservation maturity assessment models – it only focuses on assessing practices associated with data storage for resilience requirements, i.e., backup or a duplicate copy (redundancy) in a physically separate facility for disaster recovery, and on compliance to community-accepted archive practices and metadata standards.

Endorsed by the Federal Geographic Data Committee (FGDC) and adapted by NOAA (NOAA EDMC 2011; Habermann, 2014a), ISO 19115 and 19115-2 define a metadata standards framework for describing geographic data (ISO 19115, 2003; ISO 19115-2, 2009) with an implementation schema provided in ISO 19139 (2007).

Future changes for preservability include changes in archive practices and standards, storage media format, and applications (both hardware and software) to store or retrieve those records.

Care is taken to make the maturity scale of the preservability key component consistent with the OAIS reference model and the NDSA's digital preservation maturity model when it is appropriate. It is not entirely identical because both OAIS and NDSA models are organization-oriented while the stewardship maturity model is product-oriented. In addition, data quality practices and data characteristics are not explicitly captured or described in the OAIS and NDSA models but will be addressed in the data quality and usability components of the stewardship maturity matrix (e.g., section 4.3).

A number of classifications for storage or repositories have been used in the preservability maturity levels and are described in more detail below.

The phrase "any storage location" at preservability maturity level 1 denotes a storage media owned or operated by individuals or institutions that are not held to the National Archives and Records Administration (NARA) archiving standards and either do not conform to good stewardship practices or information about the storage condition is not available. "Data only" is defined as a sequence of bits without any additional structural and semantic information (OCLC & RLG, 2002).

Non-designated repositories like the NOAA "Centers of Data" are facilities where extensive collections of environmental parameters are maintained because of individual research, institutional research, or operational requirements (e.g., the National Ice Center). The Centers of Data, which are not held to all of the NARA-accepted archive standards, must still adhere to basic good stewardship practices, such as off-site data backup and maintenance of adequate environmental control and security for their holdings (NOAA, 2008a). Some information about the data, i.e., metadata, is included in the dataset. Therefore, datasets preserved by "non-designated repositories" are treated as being more mature than those stored at "any storage location" as "data only".

Designated archives like the NOAA National Data Centers are major archives that maintain, process, and distribute retrospective environmental and geospatial data. The NOAA National Data Centers provide long-term stewardship for most of NOAA's environmental and geospatial data and a broad range of user services. The Centers serve as Agency Record Centers and are subject to all of the NARA-accepted archive standards (NOAA, 2008a).

Physical data storage protection is extremely critical. However, it is primarily associated with the responsibility of the system owners (either computer system or storage facility) rather than practices applied to individual datasets by data stewards. Therefore, physical data protection is not included in the stewardship matrix.

Preservability metadata are necessary for file storage and retrieval purposes only. They include a unique identifier for the dataset, file naming convention, file size, data volume, and, if available, a unique identifier for the collection-level metadata record associated with the dataset. The file naming convention is treated in this paper as being more relevant to preservability; however, the implication of defined file naming conventions on data usability and interoperability should also be considered when determining file naming conventions.

4.2 Accessibility

As pointed out by Pennock (2006), the continual development of computing hardware and software poses the biggest risk to the accessibility of digital objects as many digital file formats are dependent on computing environments. The community-accepted and machine-independent file formats will be helpful in enhancing accessibility. However, we feel that file formats are even more critical for usability and are therefore assessed in the usability key component.

The maturity of the accessibility key component will focus on whether users can easily find and access data online. It measures whether a dataset is searchable and discoverable for collection only or to the granule level; the latter is considered to be more mature. The relative performance of data services or data servers with respect to support for customization and for processing of user orders and data downloading is assessed from accessibility maturity level 3 and higher. Dissemination statistics and reports are expected to be defined and available internally at maturity level 4 and dynamically generated and available online at level 5.

“Direct file download” refers to basic file transfer from one host to another such as via an anonymous FTP (File Transfer Protocol) or an HTTP (Hypertext Transfer Protocol) server.

Data service refers to a method of machine-to-machine communication designed to be used by scripts, programs, or applications other than the web browser such as a web service. Web Services are designed for automation and enable the integration of data access into custom applications or workflow using standards-based data structures and attributes like OPeNDAP (Open-source Project for a Network Data Access Protocol) or the OpenGIS Web Map Service (WMS). A description of the OPeNDAP software and information on using a client or setting up a server can be found at: http://docs.opendap.org/index.php/UserGuide (see also Cornillon, Gallagher, & Sgouros, 2003). A WMS is a standard protocol for serving geo-registered map images over the internet that has been approved as an ISO standard (ISO 19128, 2005).

Data server refers to a middleware application that sits between the data and users, providing access services such as visualization, subsetting, and/or data translation capabilities. Examples include THREDDS (Thematic Real-time Environmental Distributed Data Services) Data Server (TDS) (http://www.unidata.ucar.edu/publications/factsheets/2010sheets/thredds_factsheet.pdf) and ERDDAP (Environmental Research Division's Data Access Program) which builds on an OPeNDAP server for gridded and tabular data and offers a wide range of output data formats (http://coastwatch.pfeg.noaa.gov/erddap/index.html).

Future changes of technology or metadata standards are primarily associated with search and discovery.

Access metadata pertain to information for search and discovery such as title, abstract, and keywords.

4.3 Usability

Usability deals with how easily users are able to use the data and learn whether the data are suitable for their own data requirements. It is closely tied to online documentation availability (e.g., Quick starter guide, Users’ guide, or Readme file), file format for interoperability (machine-independent and scalable), online data customization (e.g., subsetting or aggregating) and visualization capability, and data characterization (e.g., spatial and temporal resolution, mean and standard deviation, spectral distributions, uncertainty characterization and estimates). It strives to alleviate the users’ burden of learning about and understanding the data.

The usability key component, therefore, focuses on the availability of knowledge about data and the ease of viewing, customizing, and using the data for the whole or a sub-spatial domain or for a whole or a part of the temporal period.

Data update frequency and latency may be important to users in terms of data usefulness and suitability for their application requirements. Data frequency and latency are more closely related to data utility, i.e., usefulness, than data usability, i.e., easy to use. They are to be included in a service maturity assessment model that is under development. However, they could be treated as a part of the metrics for data characterization if it is deemed necessary.

Interoperability is the ability for two or more systems to communicate and exchange data (syntactic interoperability) and information (semantic interoperability). Consistent information system architecture is essential to achieving the optimal data and information interoperability, but it is the responsibility of program and system owners and therefore is not included here. The Program Manager-Information Sharing Environment (PM-ISE) Information Interoperability Framework (I^2F) (PM-ISE, 2014) and architecture implementation pilot done by the Group Earth Observations (GEO, 2010) provide architecture guidelines and accepted practices for achieving interoperability through geospatial system integration and solution development.

Enhanced maturity in data accessibility and usability, e.g., conforming to community-defined standards on file formats and file and variable naming conventions, will increase the interoperability of datasets. This may reduce operational costs associated with maintaining multiple systems and tools for data operators or initial investments associated with making use of the data in their derived products for commercial users.

A community-accepted, machine-independent, self-describing data format, such as Network Common Data Form (NetCDF), will help improve not only data interoperability but also data integrity and reliability by accurately rendering the presentation of their content independent of computing environment.

Usability metadata include information for making the dataset easier to utilize properly, such as spatial and temporal extent information.

4.4 Production sustainability

Sustained and consistent data products are crucial to observing, monitoring, and understanding climate variability and change (Houghton, Townshend, Dawson, Mason, Zillman, & Simmons, 2012). Production sustainability is addressed in terms of various degrees of commitment for and associated requirements on the product. Here, it is assumed that the commitment is backed by the necessary financial support. At maturity level 5, changes for technology in data production should be routinely incorporated into planning for continued, sustained stewardship.

Ensuring the sustainability of observing systems is critical for product sustainability (NOAA NESDIS Archive Task Team, 2002). However, it ties more closely with the responsibility of the programs or organizations rather than individual stewards or producers.

A dataset with a limited record period is a special case where production sustainability may be set to maturity level 1 in order to indicate low production sustainability or be defined as not applicable.

Metadata for production sustainability capture information about project/program and financial supporting resources.

4.5 Data quality assurance

Data quality assurance (DQA) is a set of activities or procedures focused on defect prevention to be followed in order to ensure product quality during development. Data quality screening (DQS) is a set of activities intended to ensure the source data are clean. DQS is a commonly used procedure for identifying missing or redundant records, outliers (ranges and variations), and checking for normality (shape and skewness) and linearity (consistency).

DQA combined with DQS when appropriate ensures that the products meet the requirements (i.e., building it the right way). It is proactive, process-oriented, and commonly used in industrial, software development, and service sectors (Tennant, 2001; SEI, 2010). Statistical methods are usually employed to identify defects. Information about the associated procedures and methods, however, is important to users and therefore is the focus in this key component. The maturity levels are assessed by whether there is a DQA procedure and if it is documented, defined, implemented, monitored, or externally reviewed.

Quality assurance metadata capture the methods and results associated with data quality assurance/screening procedures and practices. ISO 9000 (2009) provides a consistent terminology for quality management systems. ISO 19157 (2013) provides a good framework for capturing methods and results in data quality. A good overview of ISO 19157 data quality metadata elements can be found in Habermann (2014b). Technical and scientific oversight is important for ensuring the accuracy of data quality assurance metadata.

4.6 Data quality control/monitoring

Data quality control (DQC) is a set of activities taken to evaluate the product to ensure that it conforms to the required specifications. It is product-oriented and focuses on data anomaly detection. It is usually carried out after the product is created or at each major milestone of the product development and processing cycle. It often employs statistical tools with well-established metrics for the user community.

Data quality monitoring (DQM) is DQC performed in a continuous way throughout the life cycle of data.

The maturity of this key component will be measured on sampling coverage and frequency, whether the procedure is carried out systematically and automatically and whether procedures for cross-validation of temporal and spatial characteristics and physical consistency checks (e.g., mass or momentum conservation) are defined and performed. A dynamic interaction between data producers and users is recommended for the optimal stage of data quality monitoring and anomaly detection.

Quality control/monitoring metadata capture the methods and results associated with data quality control/monitoring procedures or practices applied to the dataset. Technical and scientific oversight is important for ensuring the accuracy of data quality control/monitoring metadata.

4.7 Data quality assessment

Data quality assessment is a set of activities designed to ensure that the products are scientifically sound (i.e., building the right thing), by carefully evaluating the product, usually by comparison with similar well-established and validated observations or data product(s). It is important to enlist the expertise and oversight of data providers, scientific stewards, and users. Utilize community standards is recommended when defining metrics for assessment.

The matrix of data quality assessment is scaled based on whether product algorithm, research product, operational product, or quality metadata has been assessed and whether assessment methods and results are available online. Level 5 requires recurring assessment, external ranking of like-products, and conformance to data quality metadata standards.

The difference between a research and operational product lies in the maturity of the product, e.g., maturity level 3 or higher based on the product maturity model developed by Bates and Privette (2012) and whether processing procedures for the product are defined, managed, and monitored.

Data quality assessment metadata capture and provide information on methods or procedures used for evaluating and validating data products and the results from those analyses. Scientific oversight is important for ensuring the accuracy of data quality assessment metadata.

4.8 Transparency/traceability

This key component measures the degrees of transparency and traceability via availability of information on data provenance and data processing systems. Optimal levels of transparency and traceability provide a high degree of reproducibility.

The focus here is on the level of availability of information about the product and how it was created, the level of practices associated with management of documents, source code, and system information, and whether data and publication citations were tracked, such as by utilizing a Digital Object Identifier (DOI) system.

An Object Identifier or OID is an identifier used to name an object, formally defined in ISO/IEC 8825 (2002) to enable distributed computing systems to uniquely identify an object with a reasonable confidence whereas the main purpose of the DOI system is "to make a collection of identifiers actionable and interoperable" (http://en.wikipedia.org/wiki/Digital_object_identifier). The Universally Unique Identifier (UUID), standardized by the Open Software Foundation (OSF), is a commonly used identifier standard.

An Algorithm Theoretical Basis Document (ATBD) contains the physical and mathematical description of the algorithm used in product generation while an Operational Algorithm Description (OAD) describes the operational algorithm used to create the data product from a systems and software engineering perspective. It usually describes in detail the source data, the "as-built" software architecture, and the operational environment needed to reproduce the product.

Traceability metadata record the data's origin and where they have moved over time to reach their current state. Data lineage is one of the entities in the ISO 19115 metadata standard that can be used for this purpose. ISO 26324 (2012) provides a DOI standard for DOI systems.

4.9 Data integrity

Data integrity refers to the validity of data, i.e., the accuracy and consistency of data over its entire lifecycle. The Data integrity component in this version of the stewardship maturity matrix primarily assesses the practices applied to datasets to ensure the data files are free of intentional or unintentional corruption during data transfer, ingest, storage, and dissemination and to ensure data authenticity at access. Commonly utilized technology includes check-sum and digital signature technology. A data integrity check at the data ingest is viewed as the first essential step and the foundation for ensuring data integrity. It ensures that data are not corrupted during the transfer from data providers/producers to archives/repositories and usually requires a close coordination between the two entities.

In addition to the data integrity check performed during ingest and dissemination, regular data integrity checks are recommended for the data being stored or archived, especially for national archives and repositories. However, such checks are not explicitly included in the maturity matrix as they are more closely related to data management decisions made by organizations.

Information security is "protecting information and information systems from unauthorized access, use, disclosure, disruption, modification, or destruction" as stated in the Federal Information Security Management Act, U.S. Public Law 107-347 (2002). Security standards, controls, and guidelines are developed by the National Institute of Standards and Technology (NIST) for federal information systems and organizations in order to provide a unified framework to protect operations and assets from hostile cyber attacks, natural disasters, structural failures, and human errors (NIST, 2013).

From the data stewardship perspective, information integrity deals with both data integrity and information system integrity. It measures the stewardship practices used to ensure that the data are being transferred, ingested, stored, and disseminated accurately with a proper level of user authorization and security practices for information systems in the network that host or store those datasets. Commonly-used practices attempt to protect datasets from corruption and modification (intentional or non-intentional) and to protect systems from unauthorized access and structural failure.

Baseline security requirements to ensure integrity are applied to the U.S. federal information systems based on the their impact level (i.e., low, moderate, and high) as determined by the Federal Information Processing Standards (FIPS) Publication 199, *Standards for Security Categorization of Federal Information and Information Systems* (NIST, 2004). The information system integrity and security touches on three key areas – confidentiality, availability, and security (U.S. Public Law 107-347, 2002).

As this matrix is intended to be utilized for publicly available datasets, requirements for confidentiality will not be included. However, caution should be taken by U.S. data providers and stewards to ensure that restricted information is not publicly available and compliant with regulations like the U.S. Department of State's International Traffic in Arms Regulations (ITAR) and the Department of Commerce's Export Administration Regulations (EAR).

The availability of the information systems is also not included here as it tends to be closely associated with the strength Information Technology (IT) support in an organization as a whole rather than with practices applied to individual datasets although it is crucial for data accessibility.

Data security practices, such as anti-virus scans, can be utilized as recommended procedures during data ingest, especially for datasets of small data volumes. For data of extremely high volume, decisions about utilizing anti-virus scans can be difficult and involve balancing resources, risk, and data latency requirements. The current practice is to focus on secured transfer between secured data providers and ingest systems. As a result, data security practices are not included explicitly in the data integrity maturity level but are recommended to be utilized when possible as a part of data integrity check during ingest.

Due to the nature of information system security requirements, they tend to be more closely associated with processes and procedures that each organization defines and implements as a whole. We recognize the potential difficulty in measuring the quality of procedures that are in place and potential risks to those information systems posed by publicizing the procedures and practices. As a result, assessing the maturity of practices for security and integrity of information systems will not be included in the matrix. However, this omission is not intended to undermine their importance. Stewards, data managers, and technical professionals are encouraged to become familiar with aforementioned documents and to closely work with security officers in their organizations to make sure the security and integrity of their information systems conform to the level of impact that their datasets require.

5 CONCLUSION AND DISCUSSION

A stewardship maturity assessment model for NOAA digital environmental and geospatial data products is presented and described. This model, which is in the form of a stewardship maturity matrix, is consistent with the published guidance from expert bodies but distinguishes itself from most of the existing preservation maturity models in the following aspects:

- It is dataset-oriented as opposed to process-oriented, providing a unified framework to assess the robustness of quantifiable stewardship practices that are applied to individual environmental geospatial datasets.
- It stresses data quality and the scientific oversight in data and metadata quality and usability that are critical to climate environmental data products and their users and stakeholders.

Related to scientific oversight, the concept of scientific stewards is introduced. Scientific stewards can be data providers or subject matter experts at data centers or repositories. They have a shared responsibility with data stewards and technical professionals for ensuring data and metadata quality and improving data usability.

The goal of defining the stewardship maturity matrix for data stewards is to provide a holistic, consistent, quantifiable, and scalable measure of stewardship maturity for data users and stakeholders including data providers and decision-support system users. It is our hope that this undertaking will help alleviate the burden that data stewards face in defining their own assessment models and reduce incompatibility of stewardship maturity assessment results from individually defined models. Effort was taken to generalize the maturity levels of this stewardship assessment model to be applicable to diverse digital environmental data products in various scientific and user communities. The underlying best practices and standards in the stewardship maturity matrix are intended to be community-accepted to allow the flexibility of its implementation.

Utilizing this data stewardship maturity assessment model will help data stewards get a consistent and quantifiable measure of an organization's data holdings. The current stage of stewardship maturity ratings will help management validate its compliance to federal regulations on stewarding digital environmental geospatial data. The results can be used to identify potential areas for improvement, especially for high-impact and high-utility datasets. It can be further used to create a roadmap forward for enhancing stewardship maturity of selected datasets in the identified areas by following community-accepted best practices. Furthermore, an evaluation of the data stewardship maturity of a product can be used to build a stewardship cost model for planning purpose – based on the difference between the current maturity levels of key components and relevant stewardship requirements – prior to beginning the archive and data governance process.

This stewardship maturity model can also be used as the basis for a ranking system for a collection of multiple datasets with the same variable but generated by different groups. The ranking system can be used by the general public or businesses seeking to make an educated choice on utilizing a dataset from the product collection. In addition, due to the numerical nature of the stewardship maturity matrix, the results from all assessed datasets can easily be integrated into search algorithms.

The maturity matrix can be utilized by data providers or scientific stewards seeking to evaluate and improve the quality and usability of their products. The results can be used by climate modelers, decision-support system users, and scientists to better understand the upstream data and data quality management practices applied to their input datasets.

It is anticipated that stewards who wish to utilize this maturity matrix will define the standards and metrics in the various aspects of scientific data stewardship tailored to their own holdings. To assist consistent implementation across national and international agencies, collaborations have been initiated with the Federation of Earth Science Information Partners (ESIP) data stewardship committee, the National Geospatial Data Asset (NGDA) lifecycle maturity assessment working group, and the European Organisation for the Exploitation of Meteorological Satellites (EUMETSAT) Climate Product group. These collaborations should help ensure consistency of the maturity level definitions and establish and provide guidance on best practices and standards for key components to the Earth data community.

It is also anticipated that this scientific data stewardship maturity assessment model be a living document with levels of the maturity matrix for each component and, potentially, the choice of key components being refined or modified over time. The incorporation of constructive comments and suggestions from the data stewardship community is vital to the adoption and ultimate utility of the model.

An assessment template using the latest version of the stewardship maturity matrix will be maintained at: http://dx.doi.org/10.6084/m9.figshare.1211954. We encourage stewards who utilize this template and carry out self-evaluations of stewardship maturity of their datasets to document their justifications in detail and make these available to data users. This will allow for transparency and feedback from the users and should help improve the objectiveness of stewardship maturity assessment results.

6 ACKNOWLEDGMENTS

G. Peng is supported by NOAA’s Satellite Climate Data Record Program (CDRP) through the Cooperative Institute for Climate and Satellites – North Carolina (CICS-NC) under Cooperative Agreement NA09NES4400006. Many subject matter experts on metadata, archive, access, user service, system engineering and architecture, software engineering and architecture, information system security, data management, configuration management, project management, data quality management, ground-based and remote-sensing data product development and validation, and research-to-operation transition, who are at or affiliated with NOAA’s National Climatic Data Center (NCDC), have provided information on what they do or have reviewed the scope of the long-term scientific data stewardship with beneficial suggestions for its content and layout. Some of them have also reviewed and helped refine the definitions of maturity levels of one or more key components. We would like to thank them all.

Peng would like to thank NCDC, CICS-NC, and CDRP management for their continuous support for this research project and the following people for beneficial discussions and suggestions: Ken Knapp, Drew Saunders, John Keck, M. Scott Koger, Otis Brown, John Bates, Daniel Wunder, Bryant Cramer, W. Jesse Glance, Richard Kauffold, Derek Arndt, Jason Cooper, Ana Privette, Jason Symonds, David Bowman, Terry McPherson, Tom Zhao, John Stachniewicz, and Linda Copley.

The ESIP Data Stewardship Committee members, Ken Casey, NOAA’s National Oceanographic Data Center

(NODC), Dan Kowal, NOAA's National Geophysical Data Center (NGDC), Grace Peng, National Center for Atmospheric Research (NCAR), Ruth Duerr, National Snow and Ice Data Center (NSIDC), Donna Scott, NSIDC, Sky Bristol, U.S. Geological Survey (USGS), Curt Tilmes, National Aeronautics and Space Administration (NASA), Linda Musser, Penn State University (PSU), Lorri Peltz-Lewis, U.S. Forest Service (USFS), Wendy Blake-Coleman, Environmental Protection Agency (EPA), Philip Jones, NCDC, Derek Arndt, NCDC, Drew Saunders, NCDC, Louis Vasquez, NCDC, Jörg Schulz, EUMETSAT, Viju John, EUMETSAT, Rob Roebeling, EUMETSAT, Kaiser-Weiss Andrea, Deutscher Wetterdienst (DWD), Anne Wilson, Colorado University, Xiaogang Ma, Rensselaer Polytechnic Institute (RPI), have all been invited to review the beta version of the stewardship maturity matrix. Peng would like to thank them all, especially Dan Kowal, Ken Casey, Donna Scott, Derek Arndt, Grace Peng, and Philip Jones for beneficial comments and suggestions.

We thank NCDC internal reviewers, Data Science Journal external reviewers, Linda Copley, Donna Scott, Derek Yu, and Christina Lief for reviewing the manuscript. Their constructive comments and suggestions have improved the paper's clarity.

John Bates, NCDC, has recommended and provided the references for the CORE-CLIMAX CDR maturity matrix and assessment instruction manual. Jeffrey Booth, U.S. Department of Homeland Security, has recommended and provided the UN-GGIM reference on essential standards guide and PM-ISE references on interoperability framework and reference architecture. M. Scott Koger, NCDC, has recommended the references on information systems security.

Megan O'Connor, Tom Maycock, Jason Yu, and Katy Vincent have read and edited the manuscript, which has increased its clarity and readability.

Disclaimer: Any opinions or recommendations expressed in this manuscript are those of the author(s) and do not necessarily reflect the views of NOAA or CICS-NC.

7 REFERENCES

Bailey, J. (2014) I review 6 digital preservation models so you do not have to. Blog paper. Retrieved Jun 19, 2014 from the World Wide Web: http://www.jeffersonbailey.com/i-review-6-digital-preservation-models-so-you-dont-have-to/.

Bates, J. J. & Privette, J. L. (2012) A maturity model for assessing the completeness of climate data records. *EOS, Transactions of the AGU*, pp 44, 441.

Becker, C., Antunes, G., & Barateiro, J. (2011) A capability model for digital preservation. *Proc. of 7th International Conference on Preservation of Digital Objects*, Nov. 1–4, 2011, Singapore.

CCSDS (The Consultative Committee for Space Data Systems) (2012) Reference Model for an Open Archival Information System (OAIS), Recommended Practices, Issue 2. Version: CCSDS 650.0-M-2. 135 pp. Retrieved January 27, 2014 from the World Wide Web: http://public.ccsds.org/publications/archive/650x0m2.pdf.

Cornillon, P., Gallagher, J., & Sgouros, T. (2003) OPeNDAP: Accessing data in a distributed, heterogeneous environment. *Data Science Journal 2*, pp 164 – 174.

DAMA International (2008) DAMA-DMBOK Functional framework. Version 3.2. Mosley, M (Ed.). 19 pp. Retrieved Jan 28, 2014 from the World Wide Web: http://www.dama.org/files/public/DAMA-DMBOK_Functional_Framework_v3_02_20080910.pdf.

DAMA International (2010) Guide to the Data Management Body of Knowledge (DAMA-DMBOK). Mosley, M., Brackett, M., & Earley, S., (Eds.) Technics Publications, LLC, New Jersey, USA. 2nd Print Edition. 406 pp.

Dollar, C. & Ashley, L. (2009) Digital Preservation: A New Framework Based on a Capability Maturity Model. Presentation at the *Managing Electronic Records Conference*, Cohasset.

EPA (U.S. Environmental Protection Agency) (2005) EPA's National Geospatial Data Policy. 11 pp. Retrieved January 28, 2014 from the World Wide Web: http://www.epa.gov/geospatial/docs/National_Geospatial_Data_Policy.pdf.

EUMETSAT (2013) CORE-CLIMAX Climate Data Record Assessment Instruction Manual. Version 2, 25 November 2013. The latest version can be accessed at: http://www.eumetsat.int/website/home/Data/ClimateService/index.html
.

GEO (Group on Earth Observations) (2010) End to end discovery and access engineering report - GEO architecture implementation pilot, phase 2. Version 1.0. OGC Document # 09-182r1. 30 pp. Retrieved Jul 23, 2014 from the World Wide Web: http://www.opengeospatial.org/standards/per?utm_source=emailcampaign307&utm_medium=phpList&utm_content=textemail&utm_campaign=OGC+publishes+Testbed+10+Cross-Community+Interoperability+Engineering+Reports.

Habermann, T. (2014a) Metadata evaluation and improvement.. Retrieved August 9, 2014 from the World Wide Web: http://figshare.com/articles/MetadataMeasurementAndImprovement_pdf/1133879.

Habermann, T. (2014b) ISO 19157 – A framework for progress on data quality. Presented at the Federation of Earth Science Information Partners (ESIP) Documentation Cluster, May, 2014.

Houghton, J., Townshend, J., Dawson, K., Mason, P., Zillman, J., & Simmons, A. (2012) GCOS at 20 years: the origin, achievement and future development of the Global Climate Observing System. *Weather 67*, pp 227-235, doi: 10.1002/wea.1964.

ISO 14721 (2012) Space data and information transfer systems – Open archival information system – Reference model. Version: ISO 14721:2012. Geneva, Switzerland.

ISO 19115 (2003) Geographic Information – Metadata. Version: ISO 19115:2003(E), 150 pp. Geneva, Switzerland.

ISO 19115-2 (2009) Geographic Information – Metadata. Part 2: extension for imagery and gridded data. Version: ISO 19115-2:2009(E), 50 pp. Geneva, Switzerland.

ISO 19128 (2005) Geographic information – Web map server interface. Version: ISO 19128:2005, 76 pp. Geneva, Switzerland.

ISO 19139 (2007) Geographic Information – Metadata – XML Schema Implementation. Version: ISO 19139:2007, 111 pp. Geneva, Switzerland.

ISO 19157 (2013) Geographic Information – Data Quality. Version: ISO 19157:2013. Geneva, Switzerland.

ISO 26324 (2012) Information and documentation – Digital object identifier system. Version ISO 26324:2012. Geneva, Switzerland.

ISO/IEC 8825 (2002) Information technology – ASN.1 encoding rules: Specification of Basic Encoding Rules (BER), Canonical Encoding Rules (ER) and Distinguished Encoding Rules (DER). Version: ISO/IEC 8825-1:2002 or ITU-T X.690. Retrieved October 16, 2014 from the World Wide Web: http://www.itu.int/ITU-T/studygroups/com17/languages/X.690-0207.pdf.

ISO 9000 (2005) Quality management systems – Fundamentals and vocabulary. Version: ISO 9000-2005. Geneva, Switzerland.

Kenney, A.R. & McGovern, N.Y. (2003) The Five Organizational Stages of Digital Preservation, in *Digital Libraries: A Vision for the Twenty-first Century*, a festschrift to honor Wendy Lougee on the occasion of her departure from the University of Michigan. Hodges, P., Bonn, M., Sandler, M., & Wilkin, J. P. (Eds.). DOI: http://dx.doi.org/10.3998/spobooks.bbv9812.0001.001.

Lavoie, B. (2000) Meeting the challenges of digital preservation; OAIS reference model. Retrieved Apr 1, 2014 from the World Wide Web: http://oclc.org/research/publications/library/2000/lavoie-oais.html.

NASA (the National Aeronautics and Space Administration) (2005) What's the difference between weather and climate? (http://www.nasa.gov/mission_pages/noaa-n/climate/climate_weather.html.)

NIST (National Institute of Standards and Technology) (2004) Standards for Security Categorization of Federal Information and Information Systems. Federal Information Processing Standards Publication 199, 13 pp. Retrieved Feburary 27, 2014 from the World Wide Web: http://csrc.nist.gov/publications/fips/fips199/FIPS-PUB-199-final.pdf.

NIST (2013) Security and privacy controls for Federal Information Systems and organizations. Natl. Inst. Stand. Technol. Spec. Publ. 800-53 Revision 4, 460 pp, doi: http://dx.doi.org/10.6028/NIST.SP.800-53r4. Retrieved Feburary 26, 2014 from the World Wide Web: http://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-53r4.pdf.

NOAA (2008a) NOAA Administrative Order 212-15 - Management of environmental and geospatial data. 4 pp. Retrieved March 3, 2013 from the World Wide Web: http://www.corporateservices.noaa.gov/ames/administrative_orders/chapter_212/212-15.pdf.

NOAA (2008b) NOAA procedure for scientific records appraisal and archive approval. 28 pp. Retrieved November 24, 2009 from the World Wide Web: https://www.ngdc.noaa.gov/wiki/images/0/0b/NOAA_Procedure_document_final.pdf.

NOAA EDMC (Environmental Data Management Committee) (2011) NOAA data documentation procedural directive. Version 1.0. Retrieved January 14, 2015 from the World Wide Web: https://www.nosc.noaa.gov/EDMC/PD.DD.php.

NOAA EDMC (2013) NOAA recommended practice for the use of external data. Version 1.0. Mar 27, 2013. 8 pp. Retrieved Mar 26, 2013 from the World Wide Web: https://www.nosc.noaa.gov/EDMC/documents/NOAA_RP_for_the_Use_of_External_Data_v1.0.pdf.

NOAA NESDIS (National Environmental Satellite, Data, and Information Service) Archive Task Team (2002) Scientific data stewardship – Executive summary consensus. Version: December 9, 2002.

NRC (National Research Council) (2005) Review of NOAA's plan for the scientific stewardship program. 37 pp. Washington, DC: The National Academies Press. Retrieved Jan 29, 2013 from the World Wide Web: www.nap.edu/catalog/11421.html.

NRC (2007) Environmental data management at NOAA: Archiving, stewardship, and access. 116 pp. The National Academies Press, Washington, D.C. Retrieved Jan 29, 2013 from the World Wide Web: www.nap.edu/catalog/12017.html.

OCLC (Online Computer Library Center) & CRL (the Center for Research Library) (2007) Trustworthy repositories audit & certification: Criteria and checklist. Version 1.0. 94 pp. Retrieved Jan 27, 2014 from the World Wide Web: http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf.

OCLC & RLG (Research Libraries Group) (2002) A metadata framework to support the preservation of digital objects. Version: June 2002. 54 pp. Retrieved Jul 1, 2014 from the World Wide Web: http://www.oclc.org/content/dam/research/activities/pmwg/pm_framework.pdf?urlm=161391.

OMB (Office of Management and Budget) (2002) Guidelines for Ensuring and Maximizing the Quality, Objectivity, Utility, and Integrity of Information Disseminated by Federal Agencies. Federal Register, 67(36). OMB Notice February 22, 2002. Retrieved August 23, 2013 from the World Wide Web: http://www.whitehouse.gov/sites/default/files/omb/fedreg/reproducible2.pdf.

OMB (2013) Open Data Policy – Managing Information as an Asset. Version: OMB Memorandum May 9, 2013. Retrieved August 13, 2014 from the World Wide Web: http://www.whitehouse.gov/sites/default/files/omb/memoranda/2013/m-13-13.pdf.

OSTP (Office of Science and Technology Policy) (2013) Increasing access to the results of federally funded scientific research. Version: OSTP Memorandum February 22, 2013. Retrieved December 2, 2013 from the World Wide Web: http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf.

Peng, G. & Privette, J. L. (2014) Stewardship maturity matrix – a unified framework for assessing data quality and usability practices applied to individual digital environmental data products. *2014 Federation of Earth Science Information Partners (ESIP) summer meeting*, 7 – 11 July 2014, Copper Mountain, CO, USA.

Peng, G. & Privette, J. L. (2014) Non-functional requirements on climate data records. Figshare. http://dx.doi.org/10.6084/m9.figshare.1149985.

Pennock, M. (2006) Digital preservation - Continued access to authentic digital assets. The Joint Information Systems Committee (JISC), Nov 2006. 4 pp. Retrieved Jul 22, 2014 from the World Wide Web: http://www.jisc.ac.uk/media/documents/publications/digitalpreservationbp.pdf.

Phillips, M., Bailey, J., Goethals, A., & Owens, T. (2013) The NDSA levels of digital preservation: An explanation and uses. 7 pp. Retrieved January 21, 2014 from the World Wide Web: www.digitalpreservation.gov/ndsa/working-groups/documents/NDSA_Levels_Archiving_2013.pdf

Privette, J.L., Bates, J., Karl, T., Barkstrom, B., & Kearns, E. (2009) NOAA's approach to provide CDRs within a new interagency initiative. *5th Symposium on future national operational environmental satellite systems*, the 89th AMS annual meeting, 11 – 16 January, 2009. Phoenix, AZ, USA.

PM-ISE (Program Manager-Information Sharing Environment) (2014) Information Interoperability Framework (I^2F) - National Security Through Responsible Information Sharing, Version 0.5, March 2014. 109 pp. Retrieved August 10, 2014 from the World Wide Web: http://ise.gov/ise-information-interoperability-framework.

PM-ISE (2014) Geospatial Interoperability Reference Architecture (GIRA) - National Security Through Responsible Information Sharing, Version: July 11, 2014. 226 pp.

SEI (Software Engineering Institute) (2010) CMMI for Service, Version 1.3, November 2010. Carnegie Mellon University. 520 pp.

Tennant, G. (2001) Six Sigma: SPC and TQM in manufacturing and services. Gower Publishing, Ltd.

UN-GGIM (2014) A guide to the role of standards in geospatial information management. Version: 8 August, 2014. Prepared by Open Geospatial Consortium (OGC), The Internatioanl Organization for Standards (ISO), Technical committee 211 (TC 211) Geographic Information/Geomatics, and The International Hydrographic Organization (IHO). 27 pp. Retrieved 11 August 2014 from the World Wide Web: http://ggim.un.org/docs/meetings/GGIM4/E-C20-2014-8_Essential%20Standards%20Guide%20for%20UNGGIM.pdf.

U.S. CLIVAR Scientific Steering Committee (2013) US Climate Variability and Predictability Program Science Plan. Report 2013 – 7, US CLIVAR Project Office, Washington, DC 20005. 97 pp.

USGS (United States Geological Survey) (2014) USGS guidelines for the preservation of digital scientific data. 6 pp. Retrieved Jun 30, 2014 from the World Wide Web: http://www.digitalpreservation.gov/ndsa/working_groups/documents/USGS_Guidelines_for_the_Preservation_of_Digital_Scientific_Data_Final.pdf.

U.S. Public Law 10 (2001) Information Quality Act. Publ. L. pp 106-554. Retrieved August 29, 2013 from the World Wide Web: www.gpo.gov/fdsys/pkg/PLAW-106publ554/html/PLAW-106publ554.htm

U.S. Public Law 107-347 (2002) Federal Information Security Management Act. Pub.L. pp 107-347. Retrieved February 26, 2014 from the World Wide Web: http://www.gpo.gov/fdsys/pkg/PLAW-107publ347/html/PLAW-107publ347.htm.

WMO GCOS (Global Climate Observing System) (2010) WMO GCOS Essential Climate Variable. Retrieved May 7, 2014 from the World Wide Web: https://www.wmo.int/pages/prog/gcos/index.php?name=EssentialClimateVariables.

(Article history: Received 28 October 2014, Accepted 4 January 2015, Available online 27 January 2015)