D-Lib Magazine

May/June 2016
Volume 22, Number 5/6

Scientific Stewardship in the Open Data and Big Data Era — Roles and Responsibilities of Stewards and Other Major Product Stakeholders

Ge Peng[1], Nancy A. Ritchey[2], Kenneth S. Casey[2], Edward J. Kearns[2], Jeffrey L. Privette[2], Drew Saunders[2], Philip Jones[3], Tom Maycock[1], and Steve Ansari[2]

[1]Cooperative Institute for Climate and Satellites-North Carolina, North Carolina State University and NOAA's National Centers for Environmental Information

[2]NOAA's National Centers for Environmental Information

[3]STG, Inc.

Corresponding Author: Ge Peng (ge.peng@noaa.gov)

DOI: 10.1045/may2016-peng

Abstract

Ensuring and improving quality and usability is an important part of scientific stewardship of digital environmental data products, but the roles of the responsible parties — those who manage quality and usability — have been evolving over time and have not always been clearly defined. Recognizing that in the Open Data and Big Data era, effective long-term scientific stewardship of data products requires an integrated and coordinated team effort of experts in multiple knowledge domains — data management, science, and technology — we introduce the following stewardship roles for each of these domains: data steward, scientific steward, and technology steward. This article defines their roles and high-level responsibilities as well as the responsibilities of other major product stakeholders, including data originators and distributors. Defining roles and formalizing responsibilities will facilitate the process of curating and communicating quality information to users. Clearly defined roles will allow effective cross-disciplinary communication and better resource allocation for data stewardship, supporting organizations in meeting the challenges of stewarding digital environmental data products in the Open Data and Big Data era.

Keywords: Scientific Data Stewardship, Information Quality, Data Steward, Scientific Steward, Technology Steward, Open Data, Big Data

1 Introduction

Ensuring data quality and improving availability and accurate representation of data and information are critical for informed, sound decision-making. Data and information quality management has always been a critical part of data management. However, for environmental data products that are produced or stewarded using federal funding, emerging non-functional requirements (such as correctness, usability, integrity, scalability, traceability, interoperability, etc.) and the advent of Big Data (characterized by massive data volumes, vast variety and complexity of data types, and low data latency) have dramatically changed the way digital datasets are being managed (Laney, 2001; Miller, 2013; Gurin, 2014; Shueh, 2014; Saey, 2015).

On the policy side, to be compliant with the U.S. Information Quality Act (U.S. Public Law 106-554, 2001), many U.S. federal agencies require their data providers to have comprehensive plans for managing and/or sharing non-restricted data and results in a timely manner, working with designated data centers or repositories (e.g., National Science Foundation (NSF), 2011; National Aeronautics and Space Administration (NASA), 2011; 2014; National Oceanographic and Atmospheric Administration (NOAA), 2011; U.S. Geological Survey (USGS), 2015). A summary of many federal funding agencies and their data access and sharing policies can be found here. (See Section 2 for definitions and scopes of terms used in this article.)

To promote openness and availability of government data, the U.S. Office of Management and Budget (OMB) and the White House Office of Science and Technology Policy (OSTP) have issued Open Data Policy and Increasing Access memoranda directing all federal government agencies to maximize public easy access to non-classified, federally-funded scientific data (OMB, 2013; OSTP, 2013), with an emphasis on ensuring and maximizing quality and utility of information (U.S. Public Law 106-554, 2001; OMB, 2002). (See Valen and Blanchat (2015) for an overview of each federal agency's compliance with the OSTP policies.)

The open data policy and data sharing requirements have brought closer than ever two groups of people — data producers and data managers — who are often at separate stages of the lifecycle of scientific data products. Previously, many data producers viewed data managers, especially at institutional archives, as a downstream resource. Many data managers at archives tended to simply accept, without any correspondence, whatever data producers would send to them. The ongoing, two-way communication and interactive relationship necessary for meeting the open data and data sharing requirements has greatly expanded the traditional scope of knowledge and expertise of people in each group to effectively communicate with each other. Data producers now must not only know their products but also gain basic knowledge of data management and preservation processes and standards. Similarly, data managers need to develop a basic understanding of the data products they are caring for, in addition to acquiring expert knowledge in data management. Meeting these increased knowledge and communication requirements may have been beneficial in the past but it is crucial in the Open Data and Big Data era. Those requirements are currently fulfilled by self-initiation of individuals, largely driven by need. However, in most cases, a gap remains in the presently defined or expected roles and responsibilities of these two groups. In the cases of well-defined roles and responsibilities, people in those two groups still face challenges in communicating effectively. Defining a role to bridge those two groups will help lead to more effective ways of interacting, resulting in more effective approaches for preserving and stewarding data products. For ensuring and improving the quality and usability of data products in the Open Data era, active two-way communication and timely information exchange is especially critical.

Although non-functional requirements (for example, constraints imposed by federal regulations and agency policies aimed at ensuring and maximizing quality and usability) are often well defined, functional requirements (what needs to be done to be compliant with those constraints) are not always clearly defined. Clearly and thoroughly defining functional requirements will help managers estimate and allocate sufficient resources for carrying out necessary tasks and, in return, help individuals, groups, or organizations be compliant with existing non-functional requirements. In this article, we identify roles to better facilitate the process of developing or updating functional requirements for ensuring and improving data product quality and usability.

On the Big Data side, the volume of world digital data has been growing at an astounding rate in recent decades. At turn of the century, Lyman *et al.* (2000) estimated that the digital world produced 1 to 2 exabytes of information annually. (See Table 1 below for data volume metric units definitions.) Eight years later, Swanson and Gilder (2008) predicted that the world's digital data volume could reach a zettabyte by 2015, a seemingly unreasonable estimate. However, the world volume surpassed the one zettabyte mark in 2010 and has increased by about 1 zettabyte per year ever since, reaching 2.8 zettabytes in 2012 (Gantz and Reinsel, 2012) and 4.4 zettabytes in 2013 (Turner *et al.*, 2014). This much faster than projected increase is in part due to surging digital data associated with digital video streaming, smart phone photo and video taking, airport surveillance, and internet surfing and indexing (Gantz and Reinsel, 2012). Turner *et al.* (2014) projected a 10-fold growth in the world's digital data from 2013 to the end of 2020.

Symbol	Metric Unit	Multiples of Bytes
KB	kilobyte	1000

MB	megabyte	1000^2
GB	gigabyte	1000^3
TB	terabyte	1000^4
PB	petabyte	1000^5
EB	exabyte	1000^6
ZB	zettabyte	1000^7
YB	yottabyte	1000^8

*Table 1: Multiples of bytes in the metric system**

*(*In computing, it is customary to use binary prefixes specified as powers of 2. See Prefixes for Binary Multiples for definitions of International System of Units (SI) of prefixes for binary multiples.)*

Concurrently, digital environmental data volume is also rapidly growing. For example, the data archive volume at the NOAA's National Centers for Environmental Information (NCEI) has increased more than 24-fold, to nearly 25 petabytes, since 2000 (Figure 1a), in large part due to the increase in satellite observations and numerical model data. The annual volume of data being served to users has also increased more than 11-fold, to about 6 petabytes per year, since 2008 (Figure 1b) and is expected to increase even faster with the emergence of cloud-based web services. For example, Amazon Web Services announced on 27 October 2015 that they would provide full access, for the first time, to the entire Level II data from the NOAA's Next Generation Weather Radar (NEXRAD) network. NEXRAD is a network of 160 high-resolution Doppler radar sites. The National Weather Service maintains the network and NCEI archives the Level II NEXRAD data from June 1991 to present. The Level II NEXRAD data can be used to retrieve precipitation. The entire Level II NEXRAD data collection is currently over 300 terabytes when compressed and is increasing at about 50 terabytes per year.

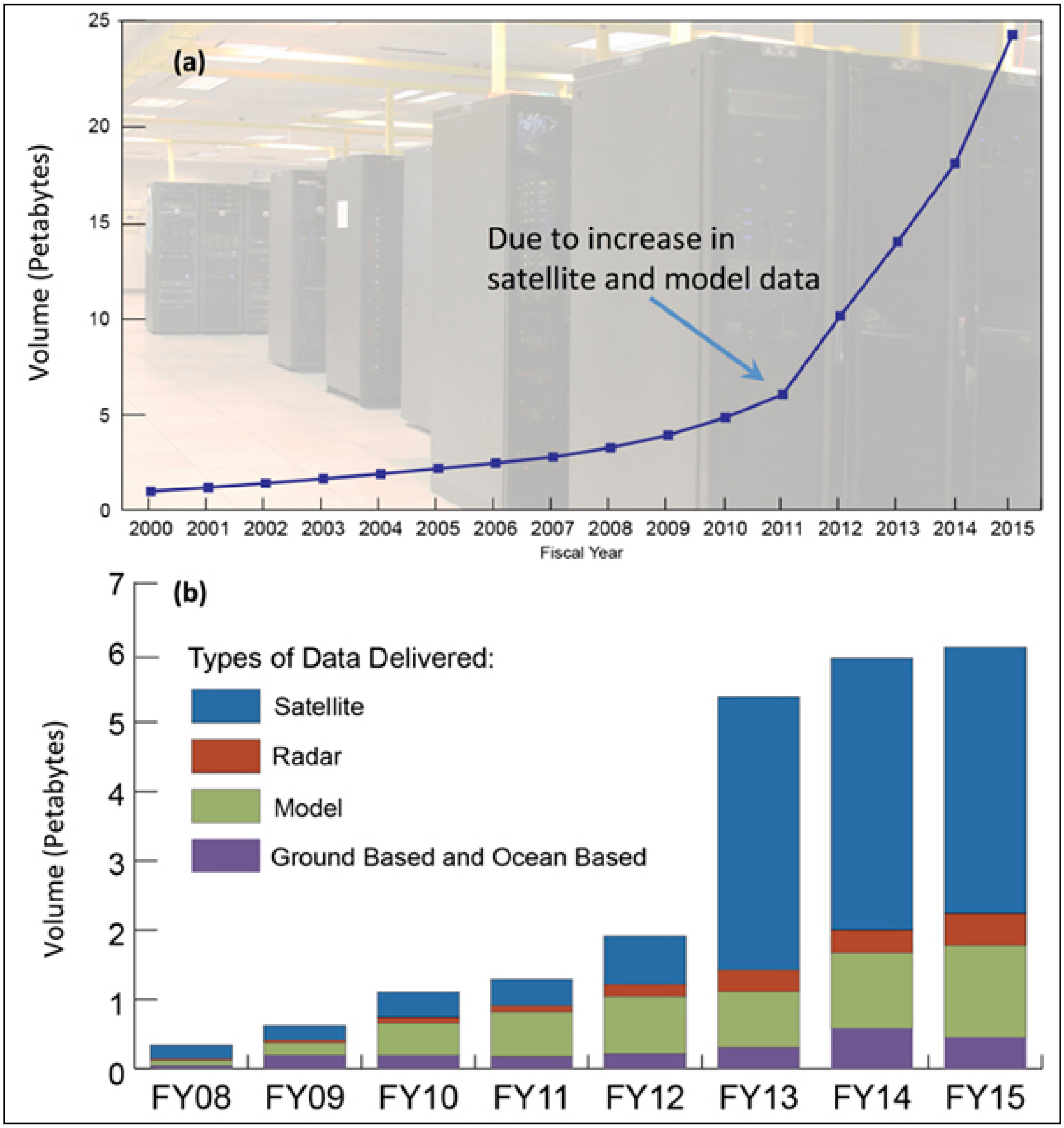


Figure 1: (a) Environmental data archive volume since year 2000 and (b) the user-requested data volume from the fiscal year (FY) 2008 to 2015 at the NOAA's National Centers for Environmental Information (NCEI). (Source: Karl (2015). Courtesy of Timothy Owen, NCEI.)

Data users, including digital environmental data users, have witnessed a shift from lacking access to enough data for well-informed decision-making, to having a surfeit of data available to them to analyze. For example, a study by International Data Corporation (IDC, 2014) estimated that while 25% of the World's digital data may contain valuable information if analyzed, less than 1% of the World's digital data are analyzed. It is not clear at this time to what extent the digital environmental data have been analyzed. However, we have observed that much of the available data do not currently have sufficient information on quality and usage to help users determine the most appropriate dataset or source for their needs. Thus, the explosion in available data poses a difficult challenge: for efficiently assessing data quality (e.g., Cai and Zhu, 2015); for effectively ensuring and improving data quality; and for fully capturing, describing, and providing sufficient and timely information about data quality to end-users. Without advances in scientific stewardship, the situation is likely to worsen. The importance of capturing and providing sufficient, traceable data quality information is undeniable in supporting application decisions or upholding the credibility of scientific findings.

Massive data volumes, vast data types and complex data structures (namely, Big Data), and strong demands for reduced data latency and improved accessibility and usability (namely, Open Data) will push the functional requirements, in terms of timeliness and scalability of environmental data management and services, beyond the capability of many data producers and data managers. Concurrently, those requirements have increased the need for expert knowledge about data products, data and system interoperability, and data management. This also leads to the increased demand on domain experts to communicate effectively with each other.

With an increasing number of people needed to meet growing demands in each domain at large archives, it is not feasible, efficient, or effective to require all people in the same domain to have the same level of domain knowledge and expertise, or the same ability to provide guidance to others and to communicate within or across domains. Therefore, defining a role of a centralized domain knowledge and communication hub would result in more effective use of organizational human resources.

These challenges have led to the need for an integrated stewardship team with at least three types of unique domain experts in: i) data management and preservation, i.e., data steward, ii) scientific data quality management and usability, i.e., scientific steward, and iii) system engineering and software development, i.e., technology steward. In this article, we will define their roles. We will also formalize their responsibilities and that of other product key players, including data originators and distributors, and other major product stakeholders (see product stakeholders definition in Section 2) within the context of shared responsibility for ensuring and improving data quality and product usability from the dataset-centric scientific data stewardship perspective.

The concept of shared responsibility for ensuring data product quality is not new. Data Management International (2010) defined two types of data stewards — business data stewards serving as trustees of business products and technical data stewards serving as the expert custodians and curators for these assets. However, as the lifecycle stages of scientific data products are quite different from those of business products, the roles and responsibilities of stewards for environmental data products will therefore need to be defined separately.

2 Terms and Definitions

A number of terms are used throughout this article. Their definitions and usage within the context of this article are described below in alphabetical order for clarity and reference:

- *Archive* is an organization that intends to preserve information for access and use by a designated community (CCSDS, 2012; ISO 14721, 2012).
- *Data distributors* are people or entities that provide access to data and/or information to consumers. This role may include data providers and data publishers. They may be a liaison between archives and data users. They may or may not be affiliated with archives or repositories. In some cases, users may obtain data directly from the data originators.
- *Data managers* are people who oversee the processing of data in an operational environment (Chisholm, 2014; NOAA, 2011). Data managers are oriented to working with data and ensuring the data are available, but not concerned with prompting good data governance within their domain (Chisholm, 2014).
- *Data originators* are people or entities that generate data products. They could be a data producer or a data provider. In a research environment, it is usually the principal investigator associated with a certain project or program. Data providers may be a liaison between data producers and archives. Sometimes, data originators may also serve as data distributors.
- *Data products* can be both original measurements and derived scientific products (adapted from the NCEI Glossary of Terms; see also Asrar and Ramapriyan, 1995; Committee on Earth Observation Satellites, 1999). However, data products are used in this article to denote post-processed and formatted science products created from either original measurements or derived data. Therefore, they refer to NOAA and NASA Level 2 to Level 4 data products defined in the Federal Geographic Data Committee (FGDC) Content Standard (FGDC, 2002). (For Level 0 and Level 1 data products, additional expertise in instruments and sensors used for measurements is crucial.)
- *Dataset* is defined as "identifiable collection of data" (ISO 19115, 2003) that may contain one or more data files in identical format, having the same geophysical variable(s) and product specification(s) such as the geospatial location or spatial grid. A dataset may contain original measurements or a derived product of a fixed version. Model output such as forecasts, projections, analyses, or re-analyses can be treated as a special case of derived products.

- *Designated data center* refers to an institution that intends to preserve information long-term for access and use by a designated community (CCSDS, 2012). *Designated national data center* is referred to as an archive that, in the United States, is required to be compliant with the standards and best practices of the National Archive and Records Administration (NARA) (NOAA, 2008). For example, NCEI is such an archive for Earth Science and geospatial data and information. On the other hand, *Non-designated data center* refers to a facility where extensive collections of environmental parameters are maintained because of individual research, institutional research, or operational requirements (e.g., the National Ice Center). A non-designated data center must still adhere to basic good stewardship practices, such as off-site backup and maintenance of adequate environmental control and security of their holdings, but may not be fully compliant with all of the NARA-accepted archival standards (NOAA, 2008).
- *Environmental data* are the recorded and/or derived geospatial 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 as defined by NOAA (2008).
- *Geospatial data* are observations or data products that 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, 2008) at a particular time or over a period of time. Therefore, information about spatial and temporal characteristics of data products and support for spatial and temporal subsetting will make it easier for end-users to get and efficiently use the data products.
- *Object* is defined as a digital data file, a paper record, an image, an article, or a collection of any or a mix of those items. An object and its accompanying metadata and documents may remain the same or, more likely, be modified somewhere between its submission and use. Additional metadata may be created and information about the object may be captured and made available to users during archival, stewardship, and service processes. This additional information may refer to, but is not limited to, descriptive and representative documents about the object, including retrieval algorithms, input data sources, and processing steps, for enhanced transparency and understandability; about the software and hardware systems used to generate the object for enhanced transparency and reproducibility; about the product quality procedures used to ensure product quality for enhanced data trustworthiness; and about how to get the data files and to use the product for enhanced data discoverability and usability. Capturing and conveying information about data, either through metadata or documentation, in a consistent manner will not only improve machine readability, namely, interoperability, but also make it easier for users to compare various products to determine the suitability for their applications.
- *Non-functional requirements*, in systems and software engineering, specify criteria that can be used to judge the operation of a system (ISO 25010, 2011; Chung, 1993). They are used in this paper to refer to constraints imposed on the preservation and stewardship of environmental data by federal laws, mandates, guidelines, and regulations (Peng *et al.*, 2015).
- *Repository* refers to a place where a large amount of something is stored (Merriam-Webster). The World Data System (WDS) of the International Council for Science (ICSU) has specified the repository types as: Domain or subject-based repository; Institutional repository; National repository system; Publication repository; Library/Museum/Archives; Research project repository (WDS-ICSU, 2015). In this article, repository refers to a facility that follows basic good stewardship practices including maintenance of adequate environmental control for its storage. A non-designated data center may sometime be referred to as a repository. For the sake of simplicity, unless mentioned otherwise to focus on the difference between designated/non-designated data centers and repositories, the term archive will be used in this article to denote an archive, a data center, or a repository.
- *Scientific quality* refers to the accuracy, precision, validity and suitability of product for intended applications (Ramapriyan *et al.*, 2015).
- *Stakeholder*, in terms of project management, is defined as "An individual, group, or organization which may affect, be affected by, or perceive itself to be affected by a decision, activity, or outcome of project" (PMI, 2013). Adapting this definition to product management, product stakeholder in this article refers to an individual, group, or organization that is involved in, has an interest in, or is potentially impacted by development, creation, preservation, stewardship, distribution, service, or application processes of the data product. Product key player refers to an individual, group, or organization that develops, produces, curates, stewards, publishes, or serves the data product to users. Other product stakeholders include but are not limited to product sponsors, project or program managers, and users.

- *Usability* is defined in a broad sense as "the ease of use and learnability of a human-made object" (Wikipedia). The International Standard (ISO 9241-11, 1998) defines usability as "The extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use." In this article, usability refers to product usability in terms of providing additional information about data products, including characteristics such as their statistical mean states and variability, uncertainty estimates, etc., to make it easier for users to understand and use the data product.

3 Defining the Roles of Stewards

To further understand the need for requiring and defining multiple roles for effective long-term data management, preservation, and stewardship of scientific data products, it is helpful to first describe the lifecycle stages of digital environmental datasets and entities in the archival information and stewardship systems.

3.1 Lifecycle stages of environmental datasets

Figure 2 presents a model of eight lifecycle stages of environmental datasets from the long-term preservation and stewardship perspective.

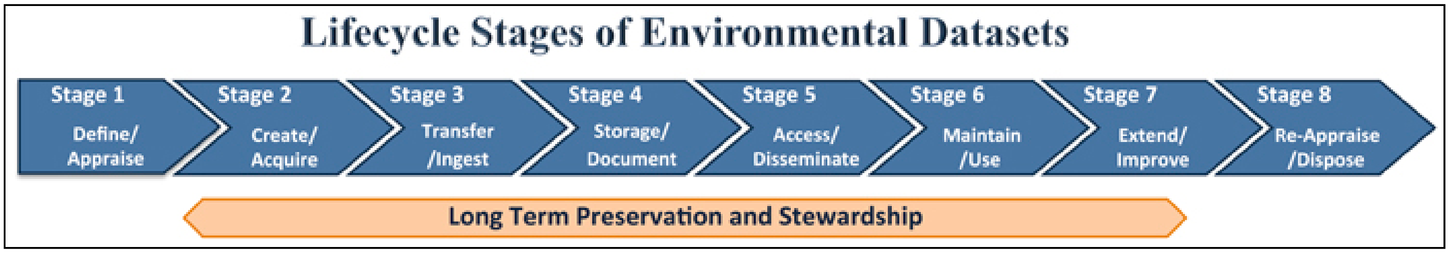

Lifecycle Stages of Environmental Datasets
Stage 1 Define/ Appraise
Stage 2 Create/ Acquire
Stage 3 Transfer /Ingest
Stage 4 Storage/ Document
Stage 5 Access/ Disseminate
Stage 6 Maintain /Use
Stage 7 Extend/ Improve
Stage 8 Re-Appraise /Dispose
Long Term Preservation and Stewardship

Figure 2: Diagram of lifecycle stages of digital environmental datasets for long-term data preservation and stewardship. Reference models for this dataset lifecycle stages model include Data Management International (2010), Federal Geographic Data Committee (2014), Digital Curation Centre (2012), and Data Documentation Initiative Alliance (2014).

In Stage 1, scientific quality requirements for a data product are defined based on user needs and product requirements as a part of the product development process, often by principal investigators or data producers. The uniqueness and intrinsic value and scientific quality of a data product may also be assessed in this stage, usually carried out by an archive in collaboration with science subject matter experts as a part of the data product acquisition process.

During Stage 2, the data product is produced. Quality information such as error sources or characteristics should be documented. In some instances, an operational version of mature research data products may be produced through a research-to-operation process. One example is the climate data records (CDRs) managed under the NOAA's CDR Program to ensure product sustainability and improve the product maturity for climate study and monitoring (NRC, 2007; Bates *et al.*, 2015). Efforts with targeted use and standardized formats for data, metadata, and documentation such as the Observations for Climate Model Intercomparison (Obs4MIPs) also aim to create well-established and documented datasets for climate model evaluations (Teixeira *et al.*, 2014). Procedures and practices defined or adapted during Stages 1 and 2 dictate the scientific maturity of the data product.

In Stage 3, data files of the product are transferred and ingested, often from data producer or provider to an archive. Data producers or providers will work with the archive and follow the proper procedure to help ensure data integrity during the data transfer process. Product quality information should also be conveyed to archives. Stage 4 involves storing data files and creating metadata and relevant documentation for the product as part of the archival process. This is often done by an archive with defined procedures for ensuring data integrity during data ingest and archive. Stage 5 handles staging and disseminating the data along with metadata and documents to consumers. Procedures are often defined to ensure data integrity during data retrieval for staging and data dissemination. Procedures and practices defined and adapted during

Stages 3 to 5 dictate the stewardship maturity of the data product, while those defined or adapted during Stages 1 and 2, such as data assurance procedures, may directly influence its stewardship maturity.

Stage 6 is about maintaining and using the data product, while Stage 7 pertains to extending and improving the data product. Stage 8 — the last stage — reappraises the uniqueness and intrinsic value of the dataset, possibly with a decision to reuse or dispose of the dataset if deemed necessary.

Improvements to a data product and its documentation can occur at any stage in this model, not just in Stage 7. This can involve circling back from any subsequent stage when necessary (not shown). Continuous data product re-evaluation for quality characteristics and improvement is an important part of the long-term preservation and stewardship process but is beyond the scope of this article.

It is easy to observe that the responsibilities of long-term preservation and stewardship could span multiple disciplines and roles, even when they take place within the same institution.

3.2 A conceptual diagram for data product development-stewardship-application process

We introduce here a conceptual diagram illustrating how the roles and responsibilities of product key players fit into the data product development-stewardship-application process. This diagram is organized into processes, functional entities, information packages, organization entities, and roles (Figure 3). It is constructed and modified based on the concepts and framework of an Open Archival Information System (OAIS) reference model (CCSDS, 2012) that is adapted to an international standard (ISO 14721, 2012).

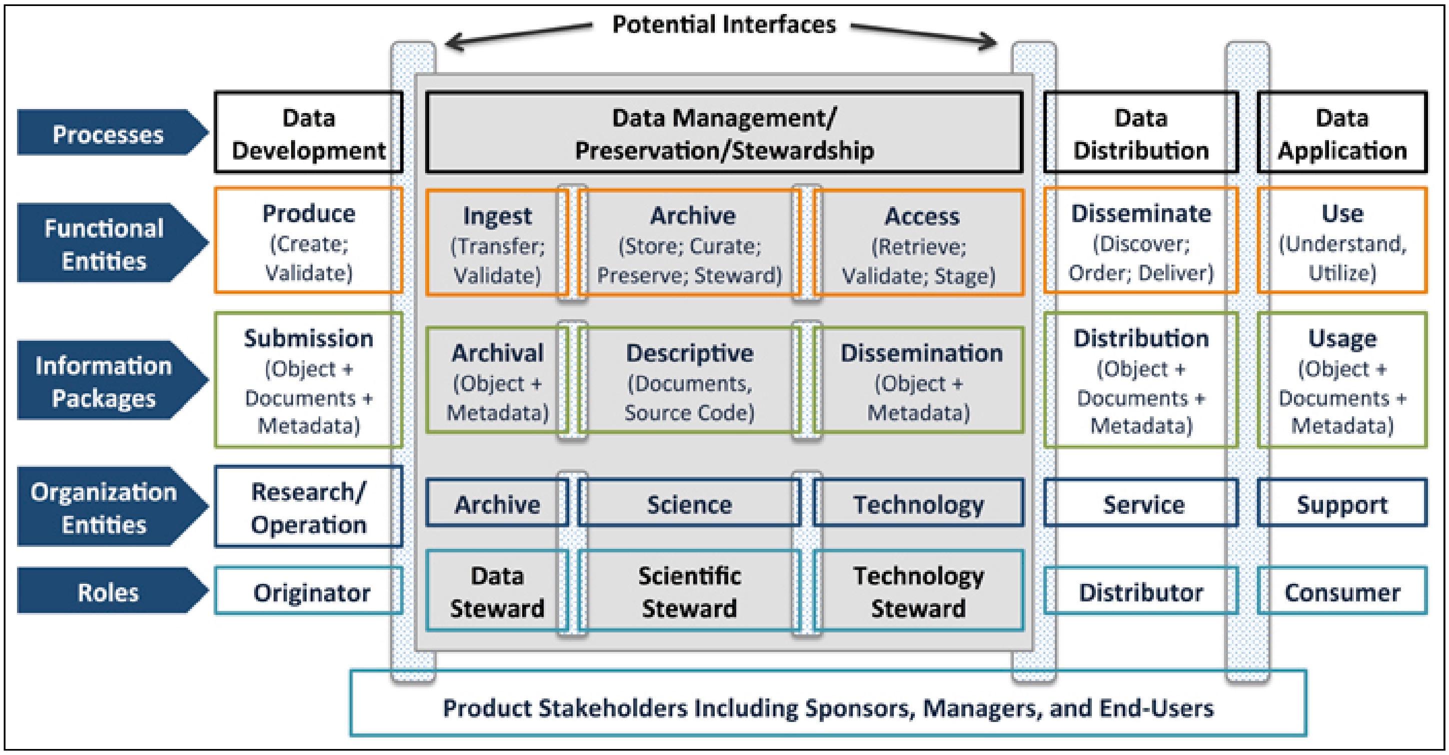

Potential Interfaces
Processes
Data Development
Data Management/ Preservation/Stewardship
Data Distribution
Data Application
Functional Entities
Produce (Create; Validate)
Ingest (Transfer; Validate)
Archive (Store; Curate; Preserve; Steward)
Access (Retrieve; Validate; Stage)
Disseminate (Discover; Order; Deliver)
Use (Understand, Utilize)
Information Packages
Submission (Object + Documents + Metadata)
Archival (Object + Metadata)
Descriptive (Documents, Source Code)
Dissemination (Object + Metadata)
Distribution (Object + Documents + Metadata)
Usage (Object + Documents + Metadata)
Organization Entities
Research/ Operation
Archive
Science
Technology
Service
Support
Roles
Originator
Data Steward
Scientific Steward
Technology Steward
Distributor
Consumer
Product Stakeholders Including Sponsors, Managers, and End-Users

Figure 3: Conceptual diagram of an end-to-end data development-stewardship-application process. It is organized as processes, functional entities, information packages, roles, and organization entities of data development; data management, preservation, and stewardship; data distribution; and data application systems. Constructed and modified from the concepts and frameworks of ISO 14721 standards (CCSDS, 2012). Dot-shaded columns denote potential interfaces between different functional entities, roles, and organization entities.

Traditionally, data management processes are concerned with defining, executing, and supervising of activities or functions that focus on controlling, protecting, and delivering the data as well as enhancing the value of data by curating or improving associated metadata and content information (Data Management International, 2010).

Data preservation processes focus on the activities or functions necessary to ensure long-term access and use of the data, beyond the limits of media or technology change (e.g., IFDO Data Federation). The OAIS reference model provides a common framework of the terms, archival concepts, and system architecture for long-term digital information preservation and access (CCSDS, 2012; ISO 14721, 2012; Lavoie, 2000). Originally developed for space agencies by the Consultative Committee for Space Data System (CCSDS, 2012), this model was adopted by the International Organization for Standardization in 2003 via efforts of international collaborations and then revised in 2012 (ISO 14721, 2012). This reference model aims to provide recommendations for long-term preservation and access activities (CCSDS, 2012) and has been adopted by many large archival institutions such as NCEI.

Scientific data stewardship is defined as all activities to preserve or improve the information content, accessibility, and usability of environmental data (NRC, 2007). Stewardship is explicitly added in this conceptual diagram to emphasize the activities or functions associated with providing expert oversight to ensure the quality and consistent use of the product and/or to provide value-added information (Peng *et al.*, 2015; Tech Target, 2015).

Descriptive Information Package (DIP) is primarily defined as "Package Descriptions" to support finding, ordering, and retrieving of data holdings in the OAIS reference model (CCSDS, 2012). It is extended here to include product descriptive and representation documentation such as Algorithm Theoretical Basis Documents (ATBD) and product processing software packages that will help meet requirements for traceability, transparency, and reproducibility with enhanced product usability.

3.3 Roles of stewards

Roles within the long-term data management, preservation, and stewardship processes are separated into data, scientific, and technology stewards. Stewards in this article are roles assigned to domain subject matter experts (SME). SMEs are people with extensive knowledge and experiences in their fields. The role of SME is gained and not assigned (Chisholm, 2014). Stewards need to have a mindset of caring for other people's data and need to be concerned with how users are doing with the data in a broader domain (Chisholm, 2014; Information Management, 2014; Peng, 2015). Therefore, not all SMEs are capable of becoming a steward.

Stewards are considered to be at the highest rank in their own domain knowledge and expertise hierarchy, while the other roles in the same domain hierarchy may be simply defined as a point of contact (POC), a specialist, or a subject matter expert. Overall, stewards need to be aware of federal policies and mandates and governmental guidelines, help define functional requirements to meet those non-functional requirements, define procedures and provide domain best practices guidance to others. Therefore, stewards serve as a centralized domain knowledge and communication hub.

3.3.1 Role of data stewards

The role of data stewards has been previously defined as leading governance practices and providing guidelines on governance (Khatibloo *et al.*, 2014; Information Management, 2014; Chatfield and Selbach, 2011). From the scientific data stewardship perspective, data stewards are responsible for ensuring compliance with data management standards, including community standards on data quality metadata and policies such as the U.S. Information Quality Act (U.S. Public Law 106-554, 2001) and Open Data Policy (OMB, 2013). They also need to provide data management guidance and help define data management requirements to other stewards, documentation and metadata team members, and other key stakeholders.

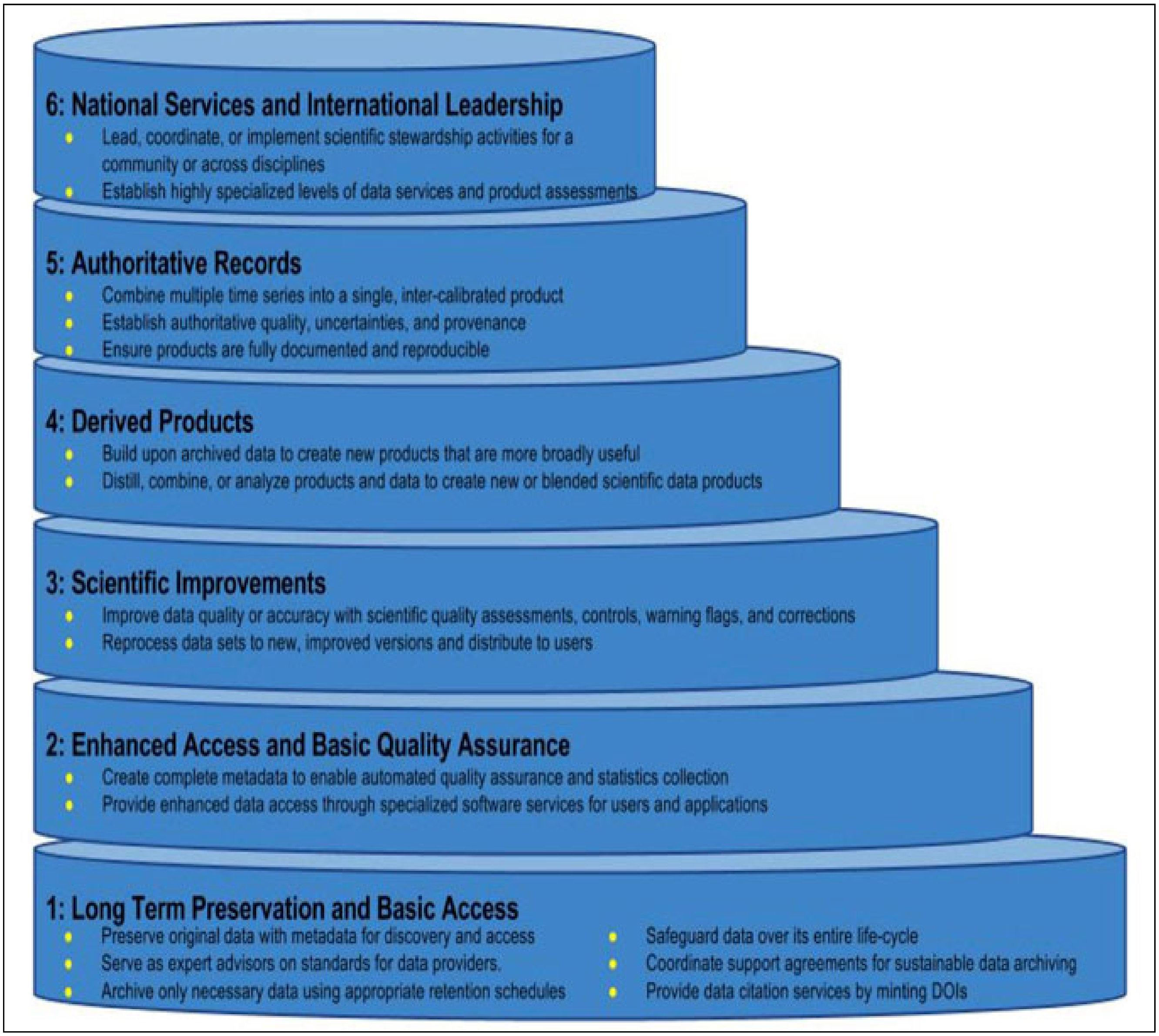
Someone currently fulfilling the role of data manager with extensive knowledge in data management and preservation could be assigned the role of a data steward. It is, however, important for the person to expand his or her general knowledge in technology and scientific domains and to have the mindset of promoting good data management practices beyond the normal community for which the person is generally responsible.

3.3.2 Role of scientific stewards

For environmental and geospatial data, precision and accuracy of the data itself is vital, but having complete, correct metadata and other relevant information about the data (e.g., spatial, temporal, and spectral characterizations, uncertainty sources and estimates) is equally important for effective long-term preservation and use of the data. Expert bodies (NRC, 2005; 2007) have established the need for and emphasized the importance of scientific oversight for environmental data products. The responsibility of ensuring data quality and improving data usability traditionally fell on the shoulders of data producers but is migrating to that of data managers, in part, as a result of the requirements for making data accessible in an open and timely fashion, driven by user needs. However, effectively and accurately capturing, describing, and conveying data quality information in a timely manner can be beyond the scope or capability of many data producers and data managers or even data stewards, when the tasks are formed alone. To address the need to fill this capacity gap, Peng *et al*. (2015) introduced the concept of scientific steward.

The role of scientific stewards is to provide expert knowledge about the subject that the dataset is associated with, such as temperature or precipitation; to provide scientific oversight to ensure the accurate scientific representation of data and metadata values, namely, scientific integrity; to provide information or guidance on data quality and characterization (Peng *et al*., 2015); and to help define data quality and usability requirements to other stewards, data producers, and other key stakeholders.

While it is important to have scientific stewards participate and oversee the basic stewardship services, such as the first two levels of tiered data stewardship service defined by NCEI (2014), shown in Figure 4, the role of scientific steward becomes essential for achieving or ensuring higher levels of stewardship maturity and service (see Peng, 2015 for definitions of stewardship maturity levels for individual datasets). This is particularly true for functional areas associated with evaluating and monitoring product quality and with improving product usability by providing or promoting the availability of data characteristics, such as spatial and temporal means and their variability, data error sources, and uncertainty estimates (Figures 4 and 5).



6: National Services and International Leadership
- Lead, coordinate, or implement scientific stewardship activities for a community or across disciplines
- Establish highly specialized levels of data services and product assessments

5: Authoritative Records
- Combine multiple time series into a single, inter-calibrated product
- Establish authoritative quality, uncertainties, and provenance
- Ensure products are fully documented and reproducible

4: Derived Products
- Build upon archived data to create new products that are more broadly useful
- Distill, combine, or analyze products and data to create new or blended scientific data products

3: Scientific Improvements
- Improve data quality or accuracy with scientific quality assessments, controls, warning flags, and corrections
- Reprocess data sets to new, improved versions and distribute to users

2: Enhanced Access and Basic Quality Assurance
- Create complete metadata to enable automated quality assurance and statistics collection
- Provide enhanced data access through specialized software services for users and applications

1: Long Term Preservation and Basic Access
- Preserve original data with metadata for discovery and access
- Serve as expert advisors on standards for data providers.
- Archive only necessary data using appropriate retention schedules
- Safeguard data over its entire life-cycle
- Coordinate support agreements for sustainable data archiving
- Provide data citation services by minting DOIs

Figure 4: Tiers showing levels of stewardship services for NOAA's environmental data products. (Source: NCEI (2014). Courtesy of Kenneth Casey, NCEI.)

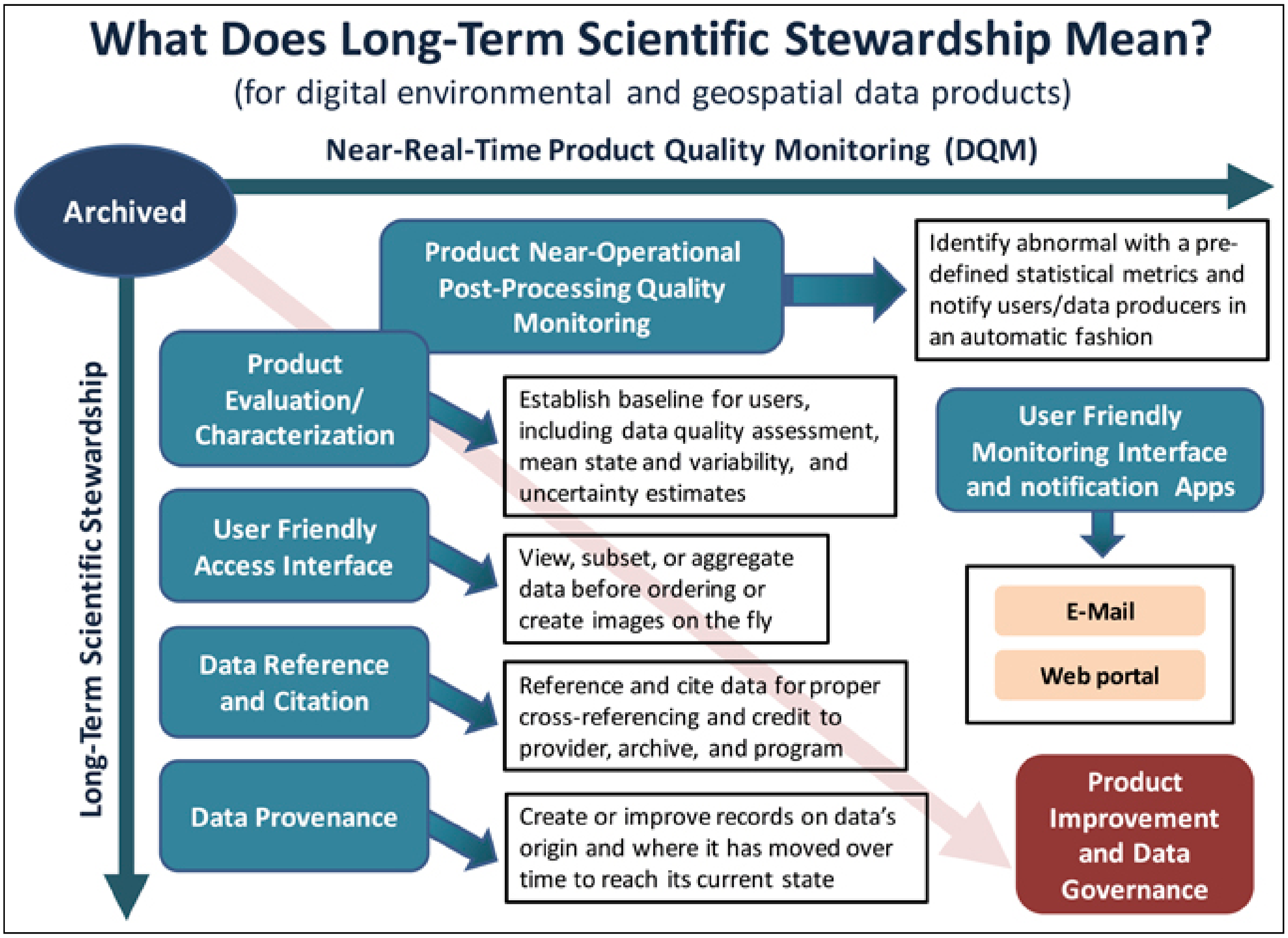

What Does Long-Term Scientific Stewardship Mean?
(for digital environmental and geospatial data products)
Near-Real-Time Product Quality Monitoring (DQM)
Archived
Long-Term Scientific Stewardship
Product Near-Operational Post-Processing Quality Monitoring
Identify abnormal with a pre-defined statistical metrics and notify users/data producers in an automatic fashion
Product Evaluation/ Characterization
Establish baseline for users, including data quality assessment, mean state and variability, and uncertainty estimates
User Friendly Monitoring Interface and notification Apps
User Friendly Access Interface
View, subset, or aggregate data before ordering or create images on the fly
E-Mail
Web portal
Data Reference and Citation
Reference and cite data for proper cross-referencing and credit to provider, archive, and program
Data Provenance
Create or improve records on data's origin and where it has moved over time to reach its current state
Product Improvement and Data Governance

Figure 5: Diagram of functional areas (cyan-filed boxes) for scientific stewardship of digital environmental data products.

It is possible that a data originator, such as a principal investigator, can act as a scientific steward. However, it is essential for that person to gain general knowledge of data management, to be familiar with tools used for archive and access, to have a basic understanding of user requirements, and to be willing to work closely with data and technology stewards.

3.3.3 Role of technology stewards

In the Big Data era, increased data volumes and variety, complex data structures, and low data latency requirements have made it difficult to manually assess and monitor data quality of all data holdings. For ensuring data quality, the development and maintenance of tools for monitoring product quality becomes an important part of scientific stewardship (Figure 5). Currently, a gap often exists between managing data quality and defining requirements for software and system development. To fill the gap, either data managers or technical professionals must gain the application or scientific knowledge required to define the appropriate requirements for the tools.

As data are increasingly treated as valuable assets for decision-makers, decision support based on fast data analysis has made ensuring data quality a critical but challenging task. Therefore, having tools available is not just helpful but a necessity for effectively stewarding and serving digital scientific data. Those tools allow data and scientific stewards to effectively capture, describe, and convey data quality information. Tools help monitor data quality, in addition to supporting data preservation and access processes. To develop tools that are useful and usable to data and scientific stewards, software developers must be able to understand and capture the data use and stewardship requirements and define their implementation requirements. Tools are also beneficial to end-users, such as those allowing users to view data products before requesting aggregated or subsetted data for their unique applications.

The role of technology stewards is defined in this article to fulfill such a need. A technology steward has domain knowledge including, but not limited to, software development, database management, web service application development, and system integration. Technology stewards need to have general knowledge of data and metadata management and of the general requirements of users of digital environmental data and information.

The role of a technology steward in ensuring and improving data quality and usability rests with providing software and system guidance, ensuring compliance of community interoperability standards, ensuring data integrity during system and technology upgrades, and defining system requirements for other stewards, development team members, and other key stakeholders.

The role of technology steward is likely to be assigned to a software or system developer or engineer. Again, it is crucial for the technology steward to gain general knowledge of data management and science domains and to have a mindset for promoting good data interoperability and usability practices to a broader domain.

In short, now and into the future, successful and effective long-term management, preservation, and scientific stewardship of digital environmental data products requires an integrated and coordinated effort of a team of stewards — subject matter experts in three different domains. They are data stewards, scientific stewards, and technology stewards. It is recommended that all three types of stewards learn the basic knowledge of the others to be most effective in communicating with each other and with other product stakeholders.

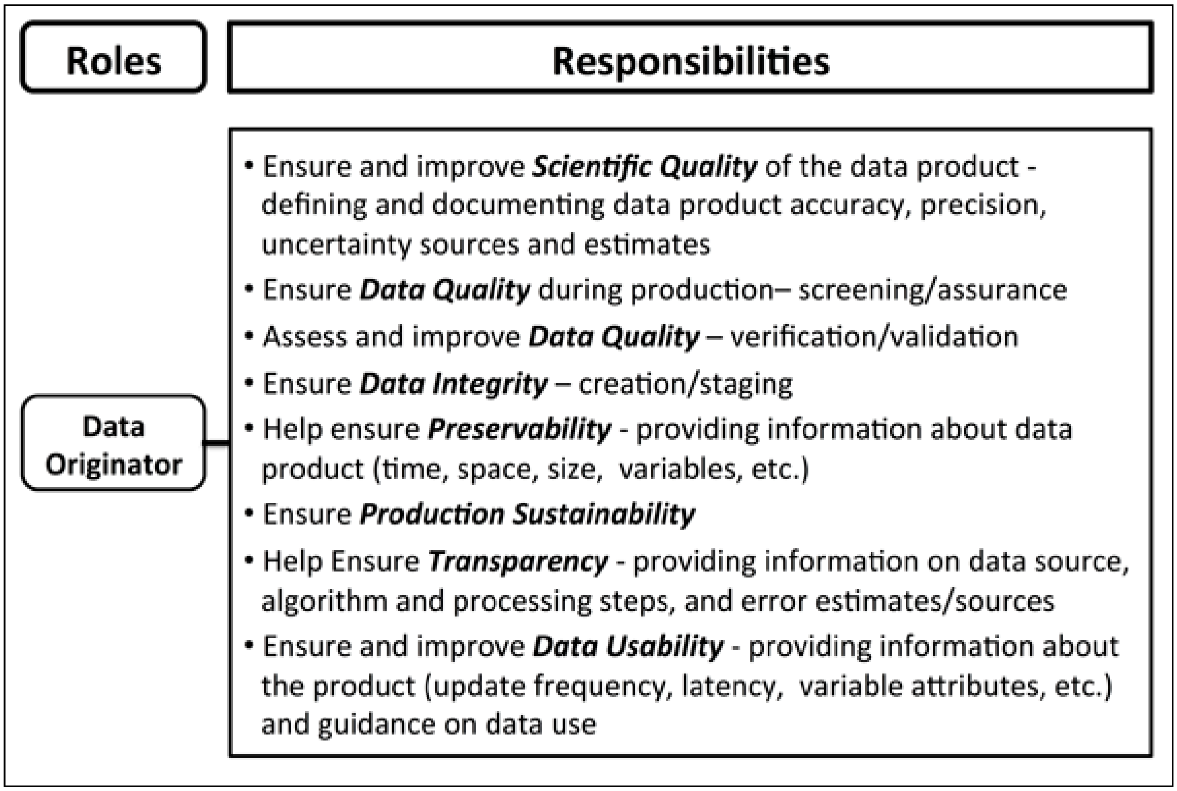
4. Responsibilities of Key Players and Other Major Stakeholders

Ensuring and improving data quality is an end-to-end process, from defining the requirements for the data product, developing and producing the product, ingesting and storing data files, creating metadata and relevant documentation, staging and disseminating the data and associated metadata and documentation, to using the data product. Errors can be made or overlooked in any of these stages or between stages, potentially by anyone. Although one may not completely eliminate the possibility of errors, a well-defined data quality management process and vigorous monitoring and oversight will help identify and address potential data quality issues. Defining the responsibilities of all product key players will facilitate this process and effectively minimize the chance of an error being made or overlooked and ensure that errors that do occur are addressed promptly.

4.1 Data originators

Data originators, including data producers, are at the forefront of ensuring and improving quality of their data products. They are responsible for defining and documenting product accuracy and precision, namely science quality (Ramapriyan *et al.*, 2015), and ensuring product sustainability. They have the responsibility for adapting the community best practices for data screening, assurance and quality control. Information about product quality, including sources of errors and uncertainty estimates, also needs to be established and documented via verification and validation, possibly combining efforts of scientific stewards and scientific users.

To help with data preservation and use, data originators are responsible for providing information about the data product, such as temporal and spatial extent, file size and data volume, variable attributes, and data latency and frequency, as well as data sources, retrieval or processing algorithm and steps, and error source and uncertainty estimates (Figure 6). It is recommended that data originators adopt a community-standard-based, self-describing, and machine-independent data format for enhanced data accessibility and usability and improved data interoperability.



Roles

Responsibilities

Data Originator

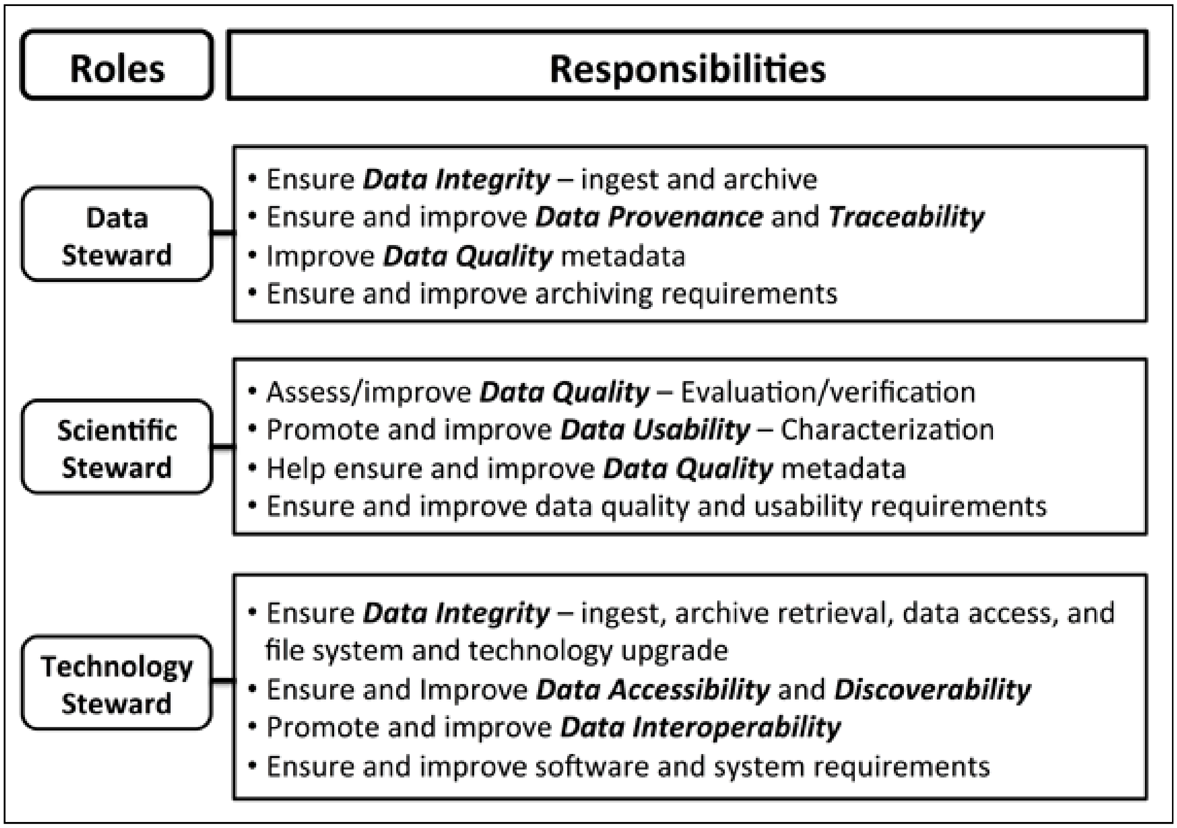
- Ensure and improve ***Scientific Quality*** of the data product - defining and documenting data product accuracy, precision, uncertainty sources and estimates
- Ensure ***Data Quality*** during production– screening/assurance
- Assess and improve ***Data Quality*** – verification/validation
- Ensure ***Data Integrity*** – creation/staging
- Help ensure ***Preservability*** - providing information about data product (time, space, size, variables, etc.)
- Ensure ***Production Sustainability***
- Help Ensure ***Transparency*** - providing information on data source, algorithm and processing steps, and error estimates/sources
- Ensure and improve ***Data Usability*** - providing information about the product (update frequency, latency, variable attributes, etc.) and guidance on data use

Figure 6: Summary diagram of responsibilities of data originators in ensuring and improving data quality and usability. A data originator could be a principal investigator, data producer, or data provider. Bold, italic, and capitalized terms refer to non-functional requirements.

As the first step toward ensuring data integrity in the long-term data preservation process, data originators or producers are responsible for creating and providing data delivery and integrity information for individual data files. The information, for example, may include checksums that are created using standards-based technology and description of the technology. They are also responsible for providing guidance, including limitations on data use, and are encouraged to collaborate with stewards and user service and engagement teams to ensure effective long-term preservation, stewardship, and use of the data product.

4.2 Data stewards

Data stewards have the responsibility of ensuring and improving data provenance and traceability, and defining and providing data archiving requirements to data producers (Figure 7). They are responsible for collecting, capturing, and conveying data quality information. Data stewards, along with technology stewards, are responsible for ensuring data integrity during data transfer, ingest, and storage. They need to collaborate with data producers and scientific stewards to capture and convey data quality information. Data stewards must also work with technology and scientific stewards to create tools to facilitate the collection of data quality and usability information.



Roles	Responsibilities
Data Steward	• Ensure ***Data Integrity*** – ingest and archive • Ensure and improve ***Data Provenance*** and ***Traceability*** • Improve ***Data Quality*** metadata • Ensure and improve archiving requirements
Scientific Steward	• Assess/improve ***Data Quality*** – Evaluation/verification • Promote and improve ***Data Usability*** – Characterization • Help ensure and improve ***Data Quality*** metadata • Ensure and improve data quality and usability requirements
Technology Steward	• Ensure ***Data Integrity*** – ingest, archive retrieval, data access, and file system and technology upgrade • Ensure and Improve ***Data Accessibility*** and ***Discoverability*** • Promote and improve ***Data Interoperability*** • Ensure and improve software and system requirements

Figure 7: Same as Figure 6 except for data stewards (top), scientific stewards (middle), and technology stewards (bottom).

4.3 Scientific stewards

Scientific stewards are responsible for ensuring and improving data product quality and usability (Figure 7), working closely with data producers. They are responsible for ensuring the accuracy of what the data and information represent. Scientific stewards define data quality and usability requirements, collaborating with data stewards and data producers. They may also assess product quality and provide product characterization, such as spatial and temporal distributions and variability, uncertainty sources and estimates, in collaboration with data producers/providers. It is recommended that scientific stewards work closely with data producers or providers to address issues and feedback from users to improve product quality and usability.

4.4 Technology stewards

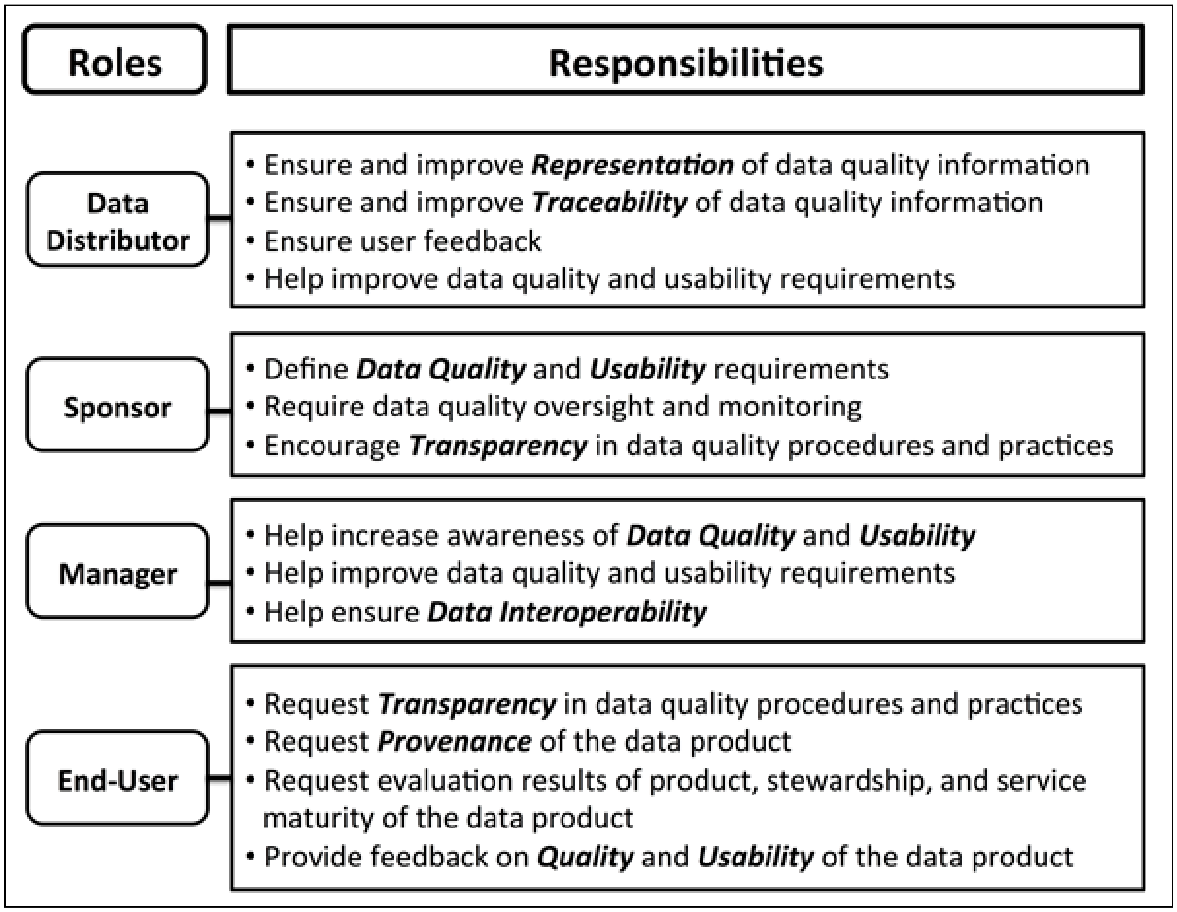
Technology stewards are responsible for ensuring data integrity during the data transfer, archive, retrieval, and dissemination processes, and especially for data staging and access, working closely with data stewards. They have a responsibility for providing community-standards-based guidance and requirements to data producers and other stewards on data accessibility and interoperability, including data format and variable-naming conventions. Technology stewards are also responsible for defining system requirements for tools developed for data management and stewardship including quality monitoring.

4.5 Data distributors

Data distributors make data products available to users. In this article, they include data providers and publishers. In many situations, data providers may work between data originators, such as data producers from an organization that produces the data products, and data managers/stewards from archives, or between archives and users. Data providers between archives and users may also act as data publishers and distributors (Figures 3 and 8). They are increasingly playing an important role in improving trackability and traceability of data products and in collecting user feedback. If the role of data provider/publisher/distributor is assigned, the entity is responsible for ensuring and improving representation of data quality information, working with data producers, stewards, and scientific users. Data publishers are encouraged to provide an efficient way of collecting and analyzing feedback from users on data quality and usability, in collaboration with data producers and scientific stewards.

4.6 Other major product stakeholders

Successful long-term stewardship of environmental data products requires involvement from all stakeholders, including experienced end-users, managers, and decision-makers. Sponsors, product managers and decision-makers such as project or program management and institutional management need to be data-centric and advocate for expert stewardship (Figure 8). They have a responsibility for supporting or encouraging community best practices in ensuring data quality and improving data usability. Sponsors from many major national and international programs (such as the Group on Earth Observations (GEO); the NOAA's CDR Program; the NASA's Making Earth Science Data Records for Use in Research Environments (MEaSUREs) Program; the European Union's COordinating Earth observation data validation for RE-analysis for CLIMAte ServiceS (CORE-CLIMAX) Project; Data Observation Network for Earth (DataONE); etc.) are already requiring data providers to follow community best practices in ensuring data quality during production and to describe the data quality assurance procedures and provide error sources and uncertainty estimates to users. We encourage all product sponsors to define such a requirement and to require their data producers to document in detail data quality management practices and provide them to archives along with their datasets. Archives are encouraged to develop policy guidelines and tools to standardize procedures and documentation to ensure consistency and efficiency of the information collection and curation processes (e.g., USGS, 2011; Jones *et al.*, 2011). Product sponsors should also encourage data producers and stewards to assess maturity of the data quality practices applied to the data products utilizing consistent framework(s), such as maturity models for assessing the readiness and completeness of climate data records and product systems (Bates and Privette, 2012; EUMETSAT, 2015), and a maturity model for assessing stewardship practices applied to individual datasets (Peng *et al.*, 2015). To allow for transparency and traceability, the quality information and maturity ratings should be provided to users with detailed justifications.



Roles	Responsibilities
Data Distributor	• Ensure and improve ***Representation*** of data quality information • Ensure and improve ***Traceability*** of data quality information • Ensure user feedback • Help improve data quality and usability requirements
Sponsor	• Define ***Data Quality*** and ***Usability*** requirements • Require data quality oversight and monitoring • Encourage ***Transparency*** in data quality procedures and practices
Manager	• Help increase awareness of ***Data Quality*** and ***Usability*** • Help improve data quality and usability requirements • Help ensure ***Data Interoperability***
End-User	• Request ***Transparency*** in data quality procedures and practices • Request ***Provenance*** of the data product • Request evaluation results of product, stewardship, and service maturity of the data product • Provide feedback on ***Quality*** and ***Usability*** of the data product

Figure 8: Same as Figure 6 except for data distributors, sponsors, project and/or program managers, and end-users. A data distributor could be a data provider or publisher.

On the other end of the stakeholder spectrum, end-users of data products are encouraged to voice their demand for product transparency in terms of readiness, completeness, and availability of data quality information. Feedback from users is an integral part of ensuring and improving data quality and usability, especially from scientific or experienced end-users who have in-depth domain or product knowledge. Gathering and addressing feedback often demands close collaboration among data producers or providers, scientific stewards, and scientific users (e.g., Peng *et al.*, 2014). An open and continuous communication among product key players and stakeholders and an established process to prioritize and address issues identified by data users are essential.

In summary, an integrated and coordinated effort from a team of subject matter experts with at least three different types of domain knowledge, active participation from end-users with an effective way of obtaining and addressing their feedback, and high-level awareness of and support from management are all important parts of ensuring or improving quality and usability of digital environmental data and information. Continuous oversight from all stewards and open and continuous communication among product key players and stakeholders are essential for effective long-term scientific stewardship of digital environmental data and information in the Open Data and Big Data era.

5 Summary and Discussion

Information is considered "a valuable national resource and a strategic asset to the Federal Government" (OMB, 2013). Digital environmental data products in particular are increasingly treated as important assets for both scientific and business communities. This has imposed more rigorous requirements on data quality management for sound, informed decision-making. While Big Data is defined by size and variety and Open Data by its timely public access and use (e.g., Gurin, 2014), the combination of Open Data and Big Data has increased the need for domain knowledge integration while leading to more domain expertise division. In this article, we suggest that an integrated team of domain experts, consisting of a data steward, scientific steward, and technology steward, is necessary for effective long-term scientific stewardship of digital environmental data products. We have defined the roles and high-level responsibilities of these stewards within the context of ensuring and improving data quality and usability.

Ensuring data quality is an end-to-end development-stewardship-application process, so procedures performed during each stage and between stages of the dataset lifecycle shown in Figure 2 will likely affect the quality of the product. The responsibilities of data producers and other major stakeholders, including product sponsors, scientific or experienced end-users, project and program managers, and decision-makers, are also described.

We recommend that all designated national data centers or archives define these three types of stewardship roles, at least at the institutional level. Given the overlapping nature of the responsibilities of data, scientific, and technology stewards and the resource constraints, it may not be necessary or feasible to have three separate individuals for these three roles — one person could be assigned more than one stewardship role. The key qualities of appropriate candidates for the role(s) are expertise in the required domain(s) and the mindset of promoting good practices for data quality management within and across domains. For individual products, it may be possible to define a role of product steward to oversee the overall scientific data stewardship of the products, leveraging an integrated team of specialists in each of three domains, and to facilitate domain knowledge exchange and communication with data producers, product management, data distributors, data service support, and users.

The concept of the long-term scientific stewardship of environmental and geospatial data is still evolving. High-level responsibilities of stewards and other major product stakeholders in this article are formulated based on our current understanding and will likely be revised or expanded over time. Defining roles and formalizing responsibilities of stewards and other major product stakeholders will help managers and stewards understand their responsibilities in ensuring and improving the quality and usability of environmental data products. Doing so will allow effective cross-disciplinary communication and efficient resource allocation for data stewardship, supporting organizations in better meeting the challenges of stewarding digital environmental data products in the Open Data and Big Data era.

Acknowledgements

Part of this work is supported by NOAA's Climate Data Record Program under Cooperative Agreement NA14NES432003. Ge Peng thanks John Keck and Ken Knapp for beneficial discussions in the early stage of this study on the need of science SMEs for the long-term stewardship of scientific data products and Jason Cooper on the need to re-evaluate and potentially dispose datasets as a part of dataset lifecycle stages.

Timothy Owen has provided the images and Danny Brinegar and Stephen Delgreco have provided the original data used in Figure 1. Sara Veasey and Deb Misch have helped improve the presentation of Figs. 2 and 3. Wendy Gross provided the WDS-ICSU (2015) reference. Discussions on project stakeholders with Derek Yu have been beneficial. Jason Yu has read and edited the manuscript, which has improved its clarity and readability.

Christina Lief and Jay Lawrimore have reviewed the manuscript as the part of the NCEI internal review process. Justin Goldstein, Hampapuram Ramapriyan, Chung-Yi Hou, Curt Tilmes, and Denise Hills have reviewed the manuscript as the part of the community-wide external review effort in association with the Data Stewardship Committee and Information Quality Cluster of the Federation of Earth Science Information Partners (ESIP). Their edits and suggestions have improved the clarity and conciseness of this article.

Disclaimer

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

References

[1] Asrar, G. R., and H. K. Ramapriyan, 1995: Data and information system for mission to planet earth. *Remote Sensing Reviews*, 13, 1-25. http://doi.org/10.1080/02757259509532294

[2] Bates, J. J. and Privette, J. L., 2012: A maturity model for assessing the completeness of climate data records. *EOS, Trans. AGU*, 93(44), 441. http://doi.org/10.1029/2012EO440006

[3] Bates, J. J., J. L. Privette, E. J. Kearns, W. J. Glance, and X. Zhao, 2015: Sustained production of multidecadal climate records — Lessons from the NOAA Climate Data Record Program. *Bull. American Meteoro. Soc.* http://doi.org/10.1175/BAMS-D-15-00015.1

[4] Cai, L. and Y. Zhu, 2015: The challenges of data quality and data quality assessment in the big data era. *Data Science Journal*, 14. http://doi.org/10.5334/dsj-2015-002

[5] CCSDS (The Consultative Committee for Space Data Systems), 2012: Reference model for an open archival information system (OAIS) — Recommendation for Space Data System Practices. Version CCSDS 650.0-M-2 June 2012. 135 pp.

[6] Chatfield, T., and R. Selbach, 2011: Data management for data stewards. *Data Management Training Workshop*. Bureau of Land Management.

[7] Chisholm, M., 2014: Data stewards versus Subject Matter Experts and Data Managers. *Information Management*. Version May 28, 2014. Access date: October 6, 2014.

[8] Chung, L., 1993: Representing and using non-functional requirements: A process-oriented approach. *Ph.D. Thesis. Univ. of Toronto.*

[9] Committee on Earth Observation Satellites, 1999, Interoperable Catalogue Systems (ICS) Collections Manual (CM), version 1.3.

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

[11] Digital Curation Centre (DCC), 2012: The DCC Curation Lifecycle Model.

[12] Data Documentation Initiative Alliance, 2014: Data Documentation Initiative lifecycle.

[13] EPA (the U.S. Environmental Protection Agency), 2005: EPA's National Geospatial Data Policy. Version: 24 August 2005.

[14] EUMETSAT, 2013: CORE-CLIMAX Climate Data Record Assessment Instruction Manual. Version 2, 25 November 2013.

[15] Federal Geographic Data Committee (FGDC), 2002: Content standard for digital geospatial metadata — extension for remote sensing data. FGDC-STD-012-2002. *Federal Geographic Data Committee*. Washington, D.C.

[16] Federal Geographic Data Committee (FGDC), 2014: Stages of the geospatial data lifecycle. *Federal Geographic Data Committee*, Version 31 March 2010.

[17] Gantz, J. and D. Reinsel, 2012: The digital universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East. *International Data Corporation*.

[18] Gurin, J., 2014: Big data and open data: What's what and why does it matter? *The Guardian*. Version: 15 April 2014.

[19] Information Management, 2014: 5 roles of the data stewards.

[20] IDC (International Data Corporation), 2014: Discover the digital universe of opportunities: Rich data and the increasing values of the internet of things. *The EMC 7th Digital Universe Study*.

[21] ISO 19115, 2003: Geographic Information — Metadata. I*nternational Organization of Standards*. Version: ISO 19115:2003.

[22] ISO 14721, 2012: Space data and information transfer system — Open archival information system — Reference model. *International Organization of Standards*. Version: ISO 14721:2012.

[23] ISO 25010, 2011: Systems and software engineering — Systems and software quality requirements and evaluation (SQuaRE) — System and software quality models. Version: ISO/IEC 25010:2011.

[24] ISO 9241-11, 1998: Ergonomic requirements for office work with visual display terminals (VDTs) — Part 11: Guidance on usability. *International Organization of Standards*. Version: ISO 9241-11:1998.

[25] Jones, P., K. Roberts, and N. Ritchey, 2011: An introduction to the Advanced Tracking and Resources tool for Archive Collection (ATRAC). *27th Conference on Interactive Information Processing Systems*. AMS 91st Annual Meeting, 22 - 27 January 2011, Seattle, WA, USA.

[26] Karl, T. R., 2015: Key challenges for environmental data and information as viewed from NCEI. *3rd Meeting of Department of Commerce Data Advisory Council*, 29 - 30 October 2015. Boulder, CO, USA.

[27] Khatibloo, F., D. Frankland, A. Smith, and W. Arellano, 2014: Building data stewardship is a new customer intelligence imperative. Forrester. Published on February 22, 2013 and updated February 28, 2014.

[28] Laney, D., 2001: 3D data management: Controlling data volume, velocity, and variety. *Meta Group*. Version: 6 February 2001.

[29] Lavoie, B, 2000: Meeting the challenges of digital preservation; OAIS reference model. *Online Computer Library Center*.

[30] Lyman, P., H. R. Varian, J. Dunn, A. Strygin, and K. Swearingen, 2000: How much information?

[31] Miller, R. J., 2013: Big Data Curation. *DIMACS/CCICADA Workshop on Big Data Integration*. Rutgers, New York, June 20 - 21, 2013.

[32] National Aeronautics and Space Administration (NASA), 2011: Guidelines for development of a data management plan. Earth Science Division. NASA Science Mission Directorate.

[33] NASA, 2014: NASA Plan for Increasing Access to the Results of Scientific Research.

[34] National Oceanic and Atmospheric Administration (NOAA), 2011: NOAA Environmental Data Management Committee Procedural Directive — NOAA data sharing policy for grants and cooperative agreements. Version 1.0.

[35] NOAA, 2008: NOAA Administrative Order 212-15 — Management of environmental and geospatial data.

[36] NOAA's National Centers for Environmental Information (NCEI), 2014: Ties of data stewardship service. *National Centers for Environmental Information*. Version: 20141124.

[37] NRC (National Research Council), 2005: Review of NOAA's plan for the scientific stewardship program. 37 pp. *The National Academies Press*. Washington, D.C. http://doi.org/10.17226/11421.

[38] NRC, 2007: Environmental data management at NOAA: Archiving, stewardship, and access. 116 pp. *The National Academies Press*, Washington, D.C. http://doi.org/10.17226/12017

[39] National Science Foundation (NSF), 2011: Directorate of Mathematical and Physical Sciences: Advice to PIs on Data Management Plans. *National Science Foundation*. (The NSF policy on dissemination and sharing of research results can be found here.)

[40] 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.

[41] OMB, 2013: Open Data Policy — Managing Information as an Asset. OMB Memorandum May 9, 2013.

[42] OSTP (The White House Office of Science and Technology Policy), 2013: Increasing access to the results of federally funded scientific research. Version: OSTP Memorandum February 22, 2013.

[43] Peng, G., J.-R. Bidlot, H.P. Freitag, C.J. Schreck, III, 2014: Directional bias of TAO daily buoy wind vectors in the central equatorial Pacific Ocean from November 2008 to January 2010. *Data Science Journal*, 13, 79-87. http://doi.org/10.2481/dsj.14-019

[44] Peng, G., J.L. Privette, E.J. Kearns, N.A. Ritchey, and S. Ansari, 2015: A unified framework for measuring stewardship practices applied to digital environmental datasets. *Data Science Journal*, 13. http://doi.org/10.2481/dsj.14-049.

[45] Peng, G., 2015: A new paradigm for ensuring and improving data quality and usability — Roles and responsibilities of key players and stakeholders. 2015 ESIP Summer Meeting, Jul 14 - 17, 2015, Pacific Cove, CA, USA.

[46] PMI (Project Management Institute), 2013: A guide to the project management body of knowledge. 5th Edition. ANSI/PMI 99-001-2013.

[47] Ramapriyan, H., D. Moroni, and G. Peng, 2015: Improving information quality for Earth Science data and products — An overview. #IN14A-01. AGU 2015 Fall Meeting, San Francisco, CA, USA.

[48] Saey, T. H., 2015: Big data studies come with replication challenges. *Science News*, 187. No. 3, Feb 7, 2015. p.22.

[49] Shueh, J., 2014: Open data: What is it and why should you care? *Government Technology*. Version: 17 March 2014.

[50] Swanson, B. and G. Gilder, 2008: Estimating the Exaflood — The impact of video and rich media on the internet - A 'zettabyte' by 2015? *Discovery Institute*, version: 29 January 2008.

[51] Tech Target, 2015: Essential Guide — Building an effective data governance framework.

[52] Teixeira, J., D. Waliser, R. Ferraro, P. Gleckler, T. Lee and G. Potter, 2014: Satellite Observations for CMIP5: The Genesis of Obs4MIPs. *Bull. Amer. Meteor. Soc.*. http://doi.org/10.1175/BAMS-D-12-00204.1

[53] Turner, V., J. F. Gantz, D. Reinsel, and S. Minton, 2014: Data growth, business opportunities, and the imperatives. Executive summary for "The digital universe of opportunities: Rich data and the increasing values of the internet of things". *International Data Corporation*.

[54] U.S. Geological Survey (USGS), 2015: U.S. Geological Survey Instructional Memorandum. IM OSQI 2015-03.

[55] USGS Manual, 2011: In Chapter 502.2 — Fundamental science practices: Planning and conducting science research. *U.S. Geological Survey*. Version: 16 December 2011.

[56] U.S. Public Law 106-554, 2001: Information Quality Act. Publ. L. 106-554.

[57] Valen, D., Blanchat, K. 2015: Overview of OSTP Responses. *figshare*. http://doi.org/10.6084/m9.figshare.1367165

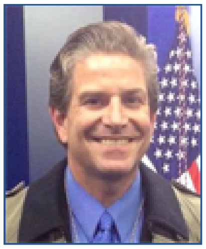
[58] WDS-ICSU, 2015: Catalogue of common requirements. DSA-WDS Partnership Working Group. V2.1 25/08/2015.

About the Authors

Ge Peng is a Research Scholar at the Cooperative Institute for Climate and Satellite-North Carolina (CICS-NC) of North Carolina State University and affiliated with the 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 experience in working with metadata specialists and software developers. She is currently leading the effort on evaluation of NOAA sea ice and surface flux climate data records and application of the NCEI/CICS-NC Scientific Data Stewardship Maturity Matrix.

Nancy Ritchey is the Archive Branch Chief at NOAA's National Centers for Environmental Information (NCEI). She is responsible for preserving NCEI's extensive collection of environmental data for future generations. Nancy has extensive knowledge and experience in digital and physical data management and related standards and leading practices. She's involved in national and international activities related to data preservation and standards. Nancy holds a M.S. degree in Atmospheric Science.

Ken Casey is the Deputy Director of the Data Stewardship Division in the NOAA National Centers for Environmental Information (NCEI). In this role, 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.



Edward J. Kearns is the Chief of the Weather Science Division in the Center for Weather and Climate at the NOAA's National Centers for Environmental Information (NCEI) in Asheville, NC. Dr. Kearns attended the University of Miami (B.S. Physics) and the University of Rhode Island (Ph.D. Physical Oceanography). He has worked for NOAA, the University of Miami, and the National Park Service on satellite products, integrated observing systems, data management, and coastal ecosystem restoration. Dr. Kearns is currently the technical lead for NOAA's Big Data Partnership and is leading NCEI's environmental assessment and monitoring projects.

Jeffrey Privette is Deputy Director of the Center for Weather and Climate, part of NOAA's National Centers for Environmental Information (NCEI). Dr. Privette started his career at NASA where he developed methods to measure the land surface using satellites. He also led the MODIS and JPSS VIIRS Land Validation Programs, the CEOS/WGCV Land Product Validation Subgroup, and was NASA's Deputy Project Scientist for Suomi-NPP satellite. He joined NOAA in 2006, where he has managed the Climate Data Record Program, and remote sensing and climate services divisions. He received his PhD in Aerospace Engineering Sciences from the University of Colorado.

Drew Saunders is the Software Engineering Support Branch Chief at NOAA's National Centers for Environmental Information (NCEI). He is responsible for requirements, design, development and testing of system infrastructure and tools for data ingest and metadata collection. Research topics include reengineering of legacy applications for reduction of technical debt and maintainability. He is also interested in open source solutions for ingest, archive, search, and dissemination capabilities for Big Data. He received his B.S. in Computer Science from North Carolina State University.

Philip Jones is a digital archive and metadata specialist with STG Inc. at the NOAA National Centers for Environmental Information (NCEI) in Asheville (previously NOAA NCDC). He collaborates on several data management and science data system projects at NCEI for all kinds of environmental data including satellite-based, land station-based and model. His research focuses on digital data preservation, metadata management/applications and process improvement. He earned his M.S. in Atmospheric Science from the University of Alabama in Huntsville.

Tom Maycock is the Science Public Information Officer for the Cooperative Institute for Climate and Satellites-North Carolina (CICS-NC), and an editor and project coordinator for the Assessments Technical Support Unit at NOAA's National Centers for Environmental Information (NCEI) in Asheville, NC. Tom has a Bachelor of Arts degree from Northwestern University, with a double major in Physics in English Literature.

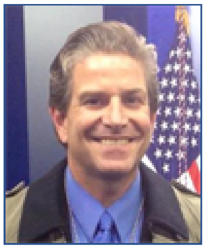
About the Authors

Ge Peng is a Research Scholar at the Cooperative Institute for Climate and Satellite-North Carolina (CICS-NC) of North Carolina State University and affiliated with the 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 experience in working with metadata specialists and software

developers. She is currently leading the effort on evaluation of NOAA sea ice and surface flux climate data records and application of the NCEI/CICS-NC Scientific Data Stewardship Maturity Matrix.

Nancy Ritchey is the Archive Branch Chief at NOAA's National Centers for Environmental Information (NCEI). She is responsible for preserving NCEI's extensive collection of environmental data for future generations. Nancy has extensive knowledge and experience in digital and physical data management and related standards and leading practices. She's involved in national and international activities related to data preservation and standards. Nancy holds a M.S. degree in Atmospheric Science.

Ken Casey is the Deputy Director of the Data Stewardship Division in the NOAA National Centers for Environmental Information (NCEI). In this role, 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.



Edward J. Kearns is the Chief of the Weather Science Division in the Center for Weather and Climate at the NOAA's National Centers for Environmental Information (NCEI) in Asheville, NC. Dr. Kearns attended the University of Miami (B.S. Physics) and the University of Rhode Island (Ph.D. Physical Oceanography). He has worked for NOAA, the University of Miami, and the National Park Service on satellite products, integrated observing systems, data management, and coastal ecosystem restoration. Dr. Kearns is currently the technical lead for NOAA's Big Data Partnership and is leading NCEI's environmental assessment and monitoring projects.

Jeffrey Privette is Deputy Director of the Center for Weather and Climate, part of NOAA's National Centers for Environmental Information (NCEI). Dr. Privette started his career at NASA where he developed methods to measure the land surface using satellites. He also led the MODIS and JPSS VIIRS Land Validation Programs, the CEOS/WGCV Land Product Validation Subgroup, and was NASA's Deputy Project Scientist for Suomi-NPP satellite. He joined NOAA in 2006, where he has managed the Climate Data Record Program, and remote sensing and climate services divisions. He received his PhD in Aerospace Engineering Sciences from the University of Colorado.

Drew Saunders is the Software Engineering Support Branch Chief at NOAA's National Centers for Environmental Information (NCEI). He is responsible for requirements, design, development and testing of system infrastructure and tools for data ingest and metadata collection. Research topics include reengineering of legacy applications for reduction of technical debt and maintainability. He is also interested in open source solutions for ingest, archive, search, and dissemination capabilities for Big Data. He received his B.S. in Computer Science from North Carolina State University.

Philip Jones is a digital archive and metadata specialist with STG Inc. at the NOAA National Centers for Environmental Information (NCEI) in Asheville (previously NOAA NCDC). He collaborates on several data management and science data system projects at NCEI for all kinds of environmental data including satellite-based, land station-based and model. His research focuses on digital data preservation, metadata management/applications and process improvement. He earned his M.S. in Atmospheric Science from the University of Alabama in Huntsville.

Tom Maycock is the Science Public Information Officer for the Cooperative Institute for Climate and Satellites-North Carolina (CICS-NC), and an editor and project coordinator for the Assessments Technical Support Unit at NOAA's National Centers for Environmental Information (NCEI) in Asheville, NC. Tom has a Bachelor of Arts degree from Northwestern University, with a double major in Physics in English Literature.

Steve Ansari is a Physical Scientist at NOAA's National Centers for Environmental Information (NCEI). He is responsible for software development, data management, data access and visualization supporting the diverse archive of NCEI environmental data. He is also the project manager for the operational hosting and development team of Climate.gov. He has B.S. in Physics from the University of North Carolina at Asheville.

Steve Ansari is a Physical Scientist at NOAA's National Centers for Environmental Information (NCEI). He is responsible for software development, data management, data access and visualization supporting the diverse archive of NCEI environmental data. He is also the project manager for the operational hosting and development team of Climate.gov. He has B.S. in Physics from the University of North Carolina at Asheville.

PRINTER-FRIENDLY FORMAT Return to Article