NOAA Environmental Data Management Framework

"NOAA is, at its foundation, an environmental information generating organization. Fundamental to ensuring that the wealth of environmental information generated by NOAA is effectively utilized now and for the long-term is an increased focus on information management standards and strategies to improve access, interoperability, and usability."

- From NOAA's Next Generation Strategic Plan (2010)

Revision History

Version	Date	Editor	Description
0.1	2012 Sep 7	Dr. Jeff de La Beaujardière, NOAA Data Mgmt Architect	First Draft sent for Committee review (EDMC, OSC, GISC, EAC) and others.
0.2	2012 Oct 31	"	Second draft for Committee review
0.3	2012 Nov 30	"	Third draft for NOSC and CIO Council review
0.4	2013 Jan 24	"	Fourth draft for NEP/NEC review
1.0	2013 Mar 14	"	Version 1.0 for presentation to SAB

Acknowledgements

This document was improved thanks to comments received from the following reviewers:

Anne Ball, NOS/CSC
Charles Baker, NESDIS
Julie Bosch, NESDIS/NODC
Deirdre Byrne, NESDIS/NODC
Leo Carling, OMAO
Kenneth Casey, NESDIS/NODC
Jennifer Clapp, NESDIS/IIAD
Don Collins, NESDIS/NODC
Jihong Dai, NMFS/ST
Darien Davis et al., OAR
Nancy DeFrancesco, NESDIS/CID
Larry Goldberg, NMFS
Peter Grimm, NESDIS/CID
Ted Habermann, NESDIS/NGDC
Karl Hampton, NESDIS/OSPO
Steve Hankin, OAR/PMEL
Scott Hausman, NESDIS/NCDC
Michelle Hertzfeld, NESDIS/IIAD
Paul Hirschberg, PPI
Christina Horvat, NWS
Thomas Karl, NESDIS/NCDC
Eric Kihn, NESDIS/NGDC
Tony Lavoi, OCIO/GIO
Clark Lind, NESDIS/NCDC
Roy Mendelssohn, NMFS
Lewis McCulloch, NESDIS/TPIO
Ana Pinheiro Privette, NESDIS/NCDC
Nancy Ritchey, NESDIS/NCDC
Steve Rutz, NESDIS/NODC
Jim Sargent et al., NMFS
Rebecca Shuford, NMFS
Matt Seybold, NESDIS/OSPO
Micah Wengren, NOS/OCS
Zdenka Willis, NOS/IOOS

Table of Contents

Executive Summary

This Environmental Data Management Framework defines and categorizes the policies, requirements, activities, and technical considerations relevant to the management of observational data and derived products by the US National Oceanic and Atmospheric Administration (NOAA). These data are an irreplaceable national resource that must be well-documented, discoverable, accessible, and preserved for future use. This Framework recommends that environmental data management (EDM) activities be coordinated across the agency, properly defined, and adequately resourced in order to ensure the usability, quality, and preservation of NOAA data.

The NOAA EDM Framework includes Principles, Governance, Resources, Standards, Architecture, and Assessment that apply broadly to many classes of data. The concept of the Data Lifecycle is introduced and separated into planning and production, data management, and data usage activities. Relevant NOAA policies, procedures, and groups are highlighted. Specific recommendations are enumerated in an Appendix.

The EDM Framework was developed in response to a recommendation from NOAA's Science Advisory Board (SAB) at their Spring 2012 meeting.[*] The transmittal letter from SAB Chair Raymond J. Ban to NOAA Administrator Dr Jane Lubchenco refers to "the urgent need to establish a NOAA-wide Environmental Data Management Framework ... that incorporates both access and archive elements of data management" in order to "integrate disparate environmental data management initiatives into an enterprise-wide environmental data management system meeting NOAA’s critical mission requirements as well as those of its constituents and users, over the long term."

[*] http://www.sab.noaa.gov/Reports/Reports.html

1. Introduction

1.1. Motivation

Accurate, timely, and comprehensive observations of the Earth and its surrounding space are critical to support government decisions and policies, scientific research, and the economic, environmental, and public health of the United States. Earth observations are typically produced for one specific purpose -- sometimes at great cost -- but are often useful for other purposes as well. It is important that these observations be managed and preserved such that all potential users can find, evaluate, understand, and utilize these data. The range of scientific and observation efforts at NOAA, and the resulting magnitude of data collections and diversity of data types, requires a systematic approach to data management that is broadly applicable yet can be tailored to particular needs.

This document establishes a conceptual Environmental Data Management (EDM) Framework of policies, organizational practices, and technical considerations to support effective and continuing access to Earth observations and derived products. The EDM Framework clarifies the expectations and requirements for NOAA projects and personnel involved in the funding, collection, processing, stewardship, and dissemination of environmental data. The goals of the Framework are (1) to promote a common understanding of data management policies and activities across NOAA, (2) to maximize the likelihood that environmental data are discoverable, accessible, well-documented, and preserved for future use, and (3) to encourage the development and use of uniform tools and practices across NOAA for handling environmental data. This Framework should guide and inform the development of program-specific data management plans and other NOAA activities to improve data management. Specific recommendations for activities in support of these goals are enumerated in Appendix A.

The NOAA Environmental Data Management Framework builds on ideas and recommendations from NOAA's *Next Generation Strategic Plan* (1), NOAA Administrative Order (NAO) 212-15 (2), the National Research Council (NRC) study *Environmental Data Management at NOAA* (3), the White House Office and Science and Technology Policy (OSTP) Interagency Working Group on Digital Data (IWGDD) report *Harnessing the Power of Digital Data: Taking the Next Step* (4), the US Group on Earth Observations (USGEO) *Exchanging Data for Societal Benefit* (5), the U.S. Chief Information Officer's *25 Point Implementation Plan to Reform Federal Information Technology Management* (6), and Open Government initiatives such as Data.gov. This Framework is also very well aligned with the draft US Office of Management and Budget (OMB) memorandum on "Managing Government Information as an Asset throughout its Life Cycle to Promote Interoperability and Openness."[*]

The NOAA EDM Framework was developed in response to a recommendation from NOAA's Science Advisory Board (SAB) at their March 2012 meeting. This Framework will be used and updated by NOAA's Environmental Data Management Committee (EDMC). EDMC activities, recommendations and directives

[*] Draft circulated for NOAA review the week of 2012-11-26; issuance date to be determined.

will be characterized in terms of the Framework. EDMC will periodically revise the Framework as needed. Concerns should be addressed to the EDMC Chair or Principal members.[*]

1.2. Key Concepts

Note: A list of acronyms may be found in Appendix B: Abbreviations.

Environmental Data: NAO 212-15 defines environmental data as "recorded and derived observations and measurements of the physical, chemical, biological, geological, and geophysical properties and conditions of the oceans, atmosphere, space environment, sun, and solid earth, as well as correlative data, such as socioeconomic data, related documentation, and metadata." For the purposes of this document, we use the terms "data" and "environmental data" interchangeably. This Framework focuses primarily on observations and derived products rather than numerical model outputs, but the latter are mentioned in several contexts. Non-digital media such as audio recordings or photographs are discussed only in the context of data rescue (see Section 3.2.7). Published papers, preserved geological or biological samples, and non-environmental data (personnel, budget, etc.) are outside the scope of this EDM Framework.

NOAA Data: Data collected directly by a NOAA entity or directly funded by a NOAA entity are the primary focus of this Framework. However, the NOAA National Data Centers archive data from a wide range of non-NOAA sources (e.g., international partners, commercial businesses, educational institutions and other federal agencies). Furthermore, many NOAA entities use data from non-NOAA sources to develop products. Some categories of externally-produced data may therefore need to be managed in the same manner as purely NOAA data.[†]

Observing System: Strictly speaking, an observing system is a set of one or more platforms (such as a satellite, buoy, radar, fixed instrument platform, ship, airplane, or autonomous vehicle), each containing one or more sensors. More generally, some observations may be completely or fully manual and involve human observations or sample gathering. This document uses the term "observing system" in a general sense and applies to both automatic and human observations.

1.3. Data Management Target State

Figure 1 illustrates conceptually the desired target state of NOAA data management activities. Not all activities are illustrated in this diagram, but it is useful as a high-level concept. The NOAA EDM Framework is intended to help guide NOAA activities toward such a target state. The modest expectations of this target state are appropriate for the medium term, and do not reflect the possible inclusion of advanced technologies in the longer term. Some NOAA datasets are nearly at this target state, but others are not; an assessment (see Section 2.6) will assist in determining the gaps. The

[*] https://www.nosc.noaa.gov/EDMC/

[†] An *External Data Usage Best Practice* document is in preparation in response to a another SAB recommendation.

Directive documents mentioned here, some of which are in preparation, are discussed more fully in Section 2.2.2.

Experimental observing systems might not achieve this target state, but should be aware of it to avoid decisions that would hinder its realization as the program matures or becomes operational. New program starts and technology refresh points should be taken as opportunities to maximize compatibility with the goals described in this Framework.

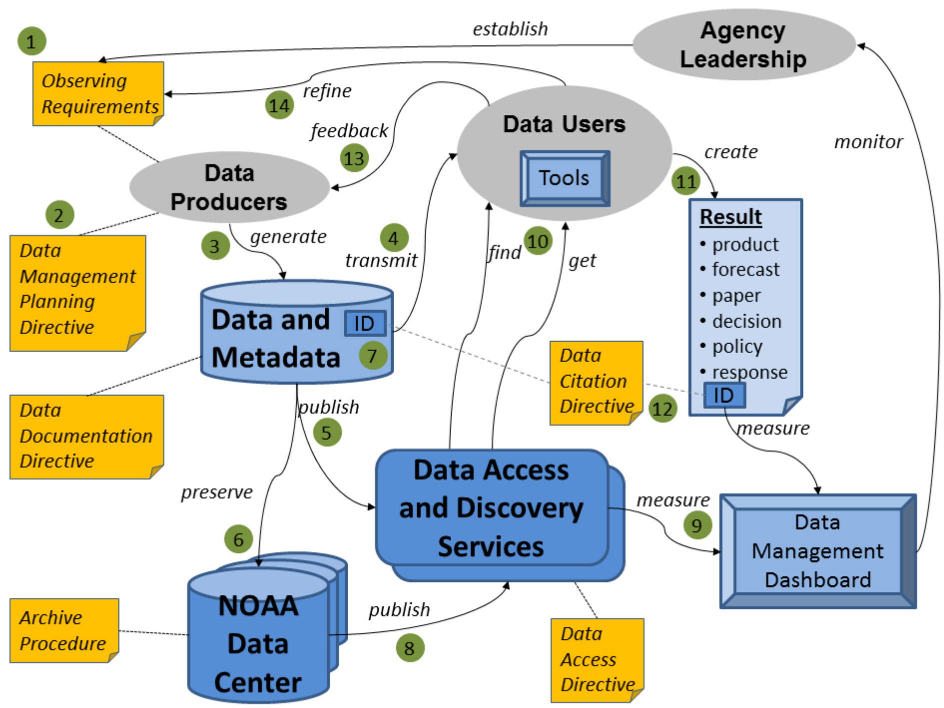

Figure 1: Conceptual overview of the desired target state of NOAA data management activities. Not all activities are illustrated. The numbers correspond to steps in the walk-through below.

Walk-through starting at the upper left of Figure 1:

1. Requirements for observational data are established by agency leadership and guide data producers in determining what NOAA observing systems to develop and deploy, and from what non-NOAA systems to acquire data.
2. Advanced planning based on the NOAA Data Management Planning directive addresses how the observed or acquired data will be handled and preserved.
3. Data producers generate data, and in accordance with the Data Documentation directive also ensure the creation of associated metadata that explains the nature, origin and quality of the data. This step implicitly includes quality control and product generation, which are not shown for simplicity.
4. Data are transmitted in near-real-time to operational data users.
5. Data are also made discoverable and accessible for other users via standardized online services per the Data Access directive.
6. Data and metadata are sent to a NOAA National Data Center (or other approved Archive facility) for long-term preservation.

7. Datasets are assigned a persistent identifier (ID) by the Data Center in accordance with the Data Citation directive.
8. The Data Center offers access and discovery of archived data using services compatible with those offered by the original data producers.
9. A Data Management Dashboard automatically measures statistics from metadata records and catalog holdings to enable leadership to assess the status of, and observe improvements in, data access, documentation, and preservation.
10. Data Users both in and out of NOAA can employ the software Tools of their choice to find, retrieve and decode data because NOAA metadata and services are well-defined and functional.
11. Users employ NOAA data to create a result such as a derived information product, forecast, scientific paper, decision, policy, or incident response.
12. The User can cite the data used by referencing its ID, so the agency can track usage and provide credit to data producers and managers.
13. Users have the opportunity to provide feedback regarding data quality and other attributes.
14. Finally, Users help refine the requirements for new or improved observations.

2. The Environmental Data Management Framework

The basic elements of the Environmental Data Management Framework are illustrated in Figure 2. The EDM Framework includes Principles, Governance, Resources, Standards, Architecture, and Assessment that apply broadly to many classes of data, and individual Data Lifecycles for particular data collections. This Section discusses the over-arching themes. Section 3 introduces the concept of the Data Lifecycle and discusses the interrelated activities that occur during the life of a particular dataset.

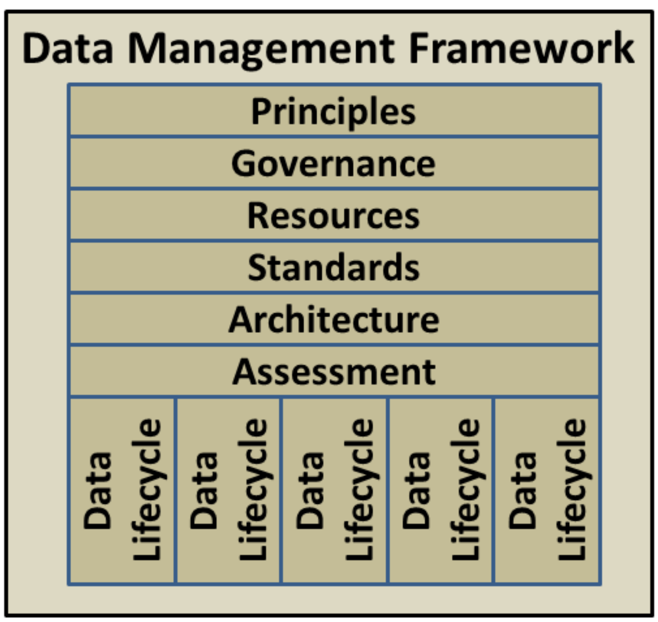

Figure 2: The Environmental Data Management Framework includes Principles, Governance, Resources, Standards, Architecture, and Assessment that apply broadly to many classes of data, and individual Data Lifecycles for particular data collections.

2.1. Principles

The following basic principles generally apply to all NOAA environmental data, though there may be exceptions for particular datasets on a case-by-case basis (such as proprietary or confidential data).

Full and Open Access: NOAA data should be made fully and openly available to all users promptly, in a non-discriminatory manner, and free of charge (or at minimum cost).
Long-Term Preservation: NOAA data should be managed as an asset and preserved for future use.
Information Quality: NOAA data should be well documented and of known quality.
Ease of Use: NOAA observations should be transformed into relevant products for end users that are made discoverable and accessible online using interoperable services and standardized formats to encourage the broadest possible use.

These principles are further explained in the following subsections.

2.1.1. Full and Open Access

In general, data managed or paid for using federal funds should be available to the public as soon as possible after collection, in a non-discriminatory manner, and at minimum cost. It is not necessary to distribute data to the public directly from the operational data processing systems as long as data are made available at an appropriate point downstream. Exceptions to this principle should be rare and explicitly justified on a case-by-case basis. (For example, data may contain confidential or personally-identifiable information; data purchased from commercial vendors may not be redistributable; data distribution may be restricted by Memorandum or other agreement; open access may not apply to every part of a satellite data stream handled by NOAA because we may be operating satellites owned by other organizations or there may be NOAA instruments on non-NOAA satellites.)

- **Timeliness**: NOAA data should be made publicly available with minimum time delay after capture. The timeliness may not be the same in all cases -- for example, routine, ongoing observations by automated sensors will be more promptly available than the results of sporadic, labor-intensive data collection. Data calibration, processing, and quality control processes should be automated whenever possible to minimize any delays. In limited circumstances, some scientific investigations may permit a temporary data hold (typically not more than 1-2 years) before distribution.
- **Non-discrimination**: NOAA data should be made publicly available to the widest community possible. NOAA data should be approved for general release and distributed in a manner that does not unfairly hinder access unless a specific exemption has been granted. Possible exceptions to open access include data whose public dissemination is prohibited by law (e.g., personally identifiable or proprietary information), by commercial agreement, or for reasons of national security (e.g., classified information).
- **Minimum cost**: NOAA data should be made available free of charge to the greatest extent possible, and certainly free of profit. Data should be made available and accessible online via web services or other internet-based mechanisms whenever possible. In limited circumstances, the cost of

reproduction may be charged to the user when it is necessary to ship data on physical media or when specialized or certified products must be created to satisfy a particular request.

2.1.2. Long-Term Preservation

Earth observations are not reproducible after the moment of measurement has passed, and are often acquired using costly technologies such as satellites, ships, aircraft, advanced sensors, open-ocean buoys, autonomous vehicles, and human observers. These observations should be managed as agency and national assets, preserved for future use, and protected from unintended or malicious modification. Data should not only be preserved in their original form but should be actively stewarded to ensure continuing usability.

2.1.3. Information Quality

Environmental data and metadata should be of known quality, and ideally of good quality. Explanations of quality control (QC) processes, and the resulting quality assessment itself, should be included or referenced in data documentation. See Sections 3.2.3 and 3.2.4 for further information regarding QC and Data Documentation.

Raw data may be distributed in (near) real time before QC and documentation have been completed, but it must be clearly communicated to prospective users that the quality may not be known when data are provided on an “as-is” basis.

2.1.4. Ease of Use

To encourage the broadest possible use of NOAA data, users should be able to find observations and derived products easily through search engines, catalogs, web portals, or other means. Data should typically be made available and accessible via web services or other internet-based mechanisms rather than by shipping physical media or by establishing dedicated or proprietary linkages. These services should comply with non-proprietary interoperability specifications for geospatial data. Data should be offered in formats that are known to work with a broad range of scientific or decision-support tools. Common vocabularies, semantics, and data models should be employed. Feedback from users should be gathered and should guide usability improvements. Users should be able to unambiguously cite datasets, both for later reuse and to provide credit and traceability to the originator. These topics are discussed in more detail in Sections 3.2 and 3.3.

2.2. Governance

2.2.1. NOAA bodies with policy or technical authority over data management

Figure 3 illustrates the agency bodies that play a direct role in governance of environmental data management at NOAA. We discuss their activities in this section.

The **Environmental Data Management Committee** (EDMC) [*] is a nexus of EDM governance activities at NOAA. EDMC was established in 2010 by NOAA Administrative Order (NAO) 212-15 (2), and reports to both the **Chief Information Officers (**CIO**) Council**[†] and the **NOAA Observing Systems Council** (NOSC)[‡]. EDMC is a voting body with representatives from NESDIS, NMFS, NOS, NWS, OAR, OMAO, PPI, the NOAA Data Management Architect (DMA), and the NOAA Enterprise Architect (EA).

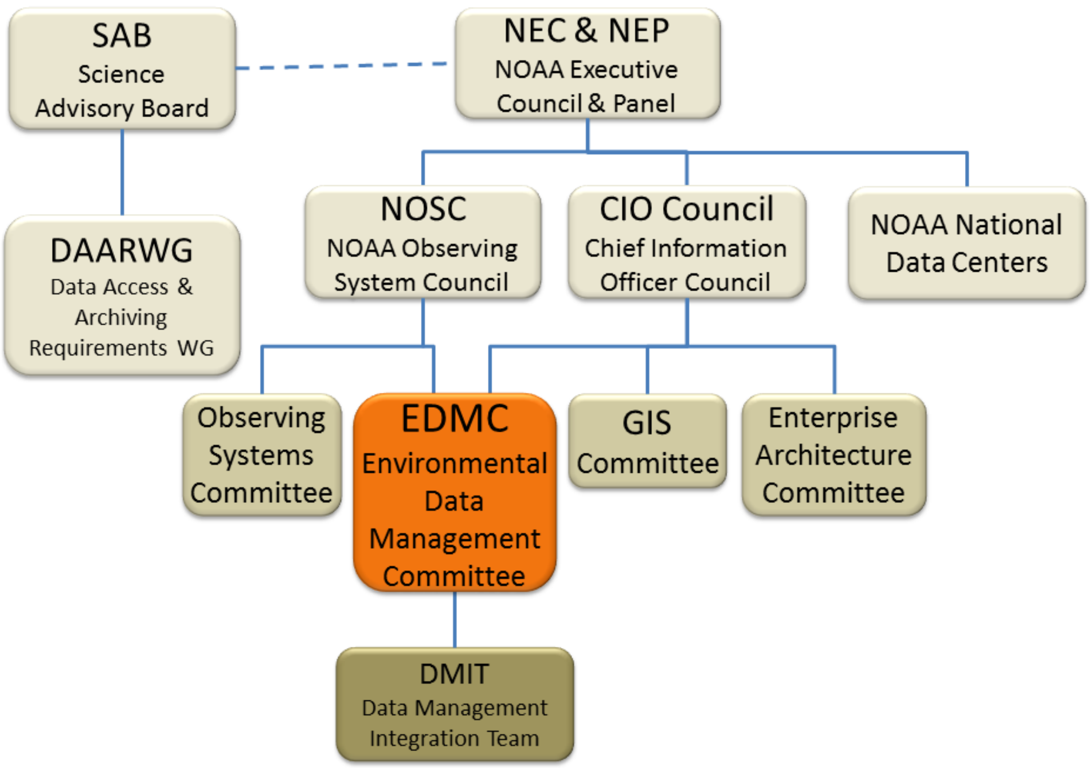

Figure 3: Governance structure for environmental data management at NOAA. Solid lines indicate reporting authority; dashed lines indicate liaison or advisory relations. The NOAA National Data Centers are technically within NESDIS but operate on behalf of the entire agency, and are therefore shown as reporting to NEC & NEP for simplicity.

The **Data Management Integration Team** (DMIT)[§] is a cross-NOAA group composed of technical experts in web services, metadata, archiving, and other relevant fields. DMIT members provide guidance and support via a mailing list and telecons. All Data Centers and significant data-producing or data-management projects should have a DMIT representative.

The **NOAA National Data Centers** -- the National Climatic Data Center (NCDC), National Geophysical Data Center (NGDC), and National Oceanographic Data Center (NODC) -- have policies and procedures

[*] https://www.nosc.noaa.gov/EDMC/

[†] http://www.cio.noaa.gov/IT_Groups/noaa_cio_CIOCouncil.html

[‡] https://www.nosc.noaa.gov/

[§] https://geo-ide.noaa.gov/wiki/index.php?title=Category:Data_Management_Integration_Team for long-term preservation. This Framework is written in accordance with those policies. Data Center policies are influenced by the US National Archives and Records Administration (NARA) policies and have broad application to NOAA EDM practices.

Individual programs and projects are also responsible for sound data management practices. Leaders of these programs have some discretion regarding technical implementation, but are encouraged to maximize compatibility and reduce development and maintenance costs by coordinating with each other, with the Data Centers, and with EDMC and DMIT.

The **Science Advisory Board** (SAB)[*], particularly through its standing **Data Access and Archiving Requirements Working Group** (DAARWG)[†], performs an external oversight role regarding data management activities. The development of this EDM Framework was recommended by DAARWG and SAB.[‡]

2.2.2. NOAA policies and documents relating to data management

NOAA's ***Next Generation Strategic Plan*** (NGSP) (1) makes numerous references to the need for good data management practices. The NGSP declares that NOAA's Mission is Science, Service and Stewardship, where "Service is the communication of NOAA's research, data, information, and knowledge for use by the Nation's businesses, communities, and people's daily lives." One of NOAA's Objectives is "Accurate and reliable data from sustained and integrated Earth observing systems." The NGSP states:

> NOAA will research, develop, deploy, and operate systems to collect remote and *in situ* observations, and manage and share data through partnerships and standards… Fundamental … is an increased focus on information management standards and strategies to improve access, interoperability, and usability of NOAA's environmental information resources... Evidence of progress includes … Improved data interoperability and usability through application and use of common data management standards.

The **Annual Guidance Memorandum** (AGM), **AGM Implementation Plans**, and the **Corporate Portfolio Analysis** (CPA) Decision Memorandum, all part of NOAA's Strategic Execution and Evaluation (SEE) process, provide general direction regarding priorities and budget for all NOAA activities including those involving data management. Corporate issues and activities relating to environmental data management are codified in the NGSP Implementation Plan of the Enterprise Objective on Reliable Data from Integrated Earth Observing System. The EDMC will implement activities resulting from SEE decisions via NOSC direction.

---

[*] http://www.sab.noaa.gov/

[†] https://www.nosc.noaa.gov/EDMC/DAARWG/index.php

[‡] http://www.sab.noaa.gov/Reports/Reports.html

NOAA Administrative Order (NAO) 212-15 (2) establishes environmental data management policy for NOAA and provides high-level guidance for procedures, decisions and actions regarding EDM. NAO 212-15 provides the EDMC with the authority to develop and approve Procedural Directives (PDs). Four PDs have been issued, and two others are currently in development:

- ***Data Management Planning Procedural Directive*** (7)**:** Directs managers of all data production projects and systems to plan in advance for data management, and contains a planning template with questions to be addressed by data production projects.
- ***Procedure for Scientific Records Appraisal and Archive Approval*** (8): Defines the process used to identify and appraise scientific records for NOAA archiving.
- ***Data Documentation Procedural Directive*** (9)**:** States that all NOAA data collections, and products derived from these data, and services that provide NOAA data and products, shall be documented. Establishes a metadata content standard (International Organization for Standardization [ISO] 19115 Parts 1 and 2) and a recommended representation standard (Extensible Markup Language [XML] formatted per the ISO 19139 schema) for documenting NOAA's environmental data and information.
- ***Data Sharing for NOAA Grants Procedural Directive*** (10)**:** States that all NOAA Grantees must share data produced under NOAA grants and cooperative agreements in a timely fashion, except where limited by law, regulation, policy or security requirements. Grantees must address this requirement formally by preparing a Data Sharing Plan as part of their grant project narrative, and by sharing data from funded projects within not more than two years. Specific language has been approved by NOAA Office of General Counsel for inclusion in announcements of opportunity and notices of award.
- ***Data Access Procedural Directive*** (in preparation): States that all NOAA environmental data shall be made accessible via the Internet, except in limited circumstances, and discusses appropriate services and formats. (Expected to be issued in 2013.)
- ***Data Citation Procedural Directive*** (in preparation): States that NOAA datasets shall be assigned a persistent identifier, with a corresponding documentation page maintained by a NOAA Data Center. Urges data users to cite datasets used in papers, decisions and other products, and recommends a citation format including the identifier. (Expected to be issued in 2013.)
- ***External Data Usage Best Practice*** (in preparation in response to another SAB recommendation[*]): provides a worksheet of potential issues to consider when using non-NOAA data. (To be delivered at March 2013 SAB.)

NAO 212-15 and EDMC Procedural Directives are high level. More detailed implementation guidance and best practices are recorded in the ***NOAA Environmental Data Management Wiki*** (11). Project-specific technical documentation is also more detailed.

NAO 212-13 (12) establishes requirements for the protection of all NOAA IT resources, including data and information.

[*] http://www.sab.noaa.gov/Reports/Reports.html

NOAA's **Guiding Enterprise Architecture Principles** (13) states that NOAA data are a corporate resource to be managed appropriately throughout their life cycle, and calls for technical solutions that are applicable NOAA-wide, standardized, interoperable, and secure.

2.2.3. National or inter-agency policies and documents

There are a number of US national or inter-agency policies and documents relevant to the governance of NOAA data management practices.

OMB Circular A-16 (14) "provides direction for federal agencies that produce, maintain or use spatial data either directly or indirectly in the fulfillment of their mission," and defines the National Spatial Data Infrastructure (NSDI) as "the technology, policies, standards, human resources, and related activities necessary to acquire, process, distribute, use, maintain, and preserve spatial data."

The ***Digital Government Strategy*** (15) is intended to "unlock the power of government data to spur innovation" by enabling "an increasingly mobile workforce to access high-quality digital government information and services anywhere, anytime, on any device." The Strategy directs agencies to architect systems for interoperability and openness, to modernize content-publication models, and to deliver better, device-agnostic digital services at a lower cost.

The **25 Point Implementation Plan to Reform Federal IT** (6) calls for consolidation of surplus or underutilized data centers and establishes a "Cloud-first" policy for acquisition of new computing capability. The **Federal Cloud Computing Strategy** (16) lays out the Cloud approach in greater detail.[*]

The draft OMB Memorandum on "Managing Government Information as an Asset throughout its Life Cycle to Promote Interoperability and Openness"[†] states that " management of information resources must begin at the earliest stages of the planning process, well before information is collected or created" and directs federal agencies to use open standards, to design systems for interoperability and information accessibility, and to create and maintain a data inventory. The alignment between the draft OMB Memo and this NOAA EDM Framework is nearly complete, except that the Memo covers personally-identifiable information in greater detail.

2.2.4. External Coordination

NOAA is not the only organization that produces and uses environmental data. In order to maximize compatibility of NOAA observations with other data it is important that there be awareness of and coordination with external bodies regarding standards and technical approaches. Furthermore, many NOAA-sponsored observations are tied to significant national and international components and activities. NOAA programs that participate in international observing activities should, where possible, influence those international structures to align with and benefit from NOAA data management

[*] See also Appendix C: Cloud Computing of this Framework.

[†] Draft circulated for NOAA review the week of 2012-11-26; issuance date to be determined.

practices but might not be held to same level of compliance as purely in-house systems. Relevant external bodies include, among others:

- World Meteorological Organization (WMO)
- Committee on Earth Observing Satellites (CEOS)
- Coordination Group for Meteorological Satellites (CGMS)
- Intergovernmental Oceanographic Commission (IOC)
- Group on Earth Observations (GEO) and USGEO
- Federal Geographic Data Committee (FGDC)
- Open Geospatial Consortium (OGC)
- International Organization for Standardization Technical Committee 211 for Geographic Information and Geomatics (ISO/TC211)
- International Committee for Information Technology Standards - Geographic Information Services (INCITS/L1)
- White House Office of Science and Technology Policy (OSTP) National Science and Technology Council (NSTC) committees and working groups
- Air Force Weather Agency (AFWA)
- US Integrated Ocean Observing System (IOOS) Program
- Selected programs in National Science Foundation (NSF), National Aeronautics and Space Administration (NASA), US Geological Survey (USGS), Environmental Protection Agency (EPA) and others.
- Unidata
- Federation of Earth Science Information Partners (ESIP)

2.2.5. Monitoring and Enforcement

Formal authority and responsibility to enforce NOAA data management policy and directives resides with Line Office (LO) and Staff Office (SO) leadership and their designees. LO/SO representatives to the NOSC, CIO Council, and EDMC should ensure their leadership is aware of and understands NOAA policies and procedures for data management. The EDMC reports progress on implementation of Procedural Directives to NOSC and CIO Council on a periodic basis. EDMC members are expected to report implementation status for their Offices to the EDMC and to their Assistant CIOs and NOSC representatives.

2.3. Resources

NOAA data cannot be adequately managed without proper resources, including personnel, budget and other supporting elements. Lack of resources is often a factor leading to data that are poorly documented, inaccessible, or improperly preserved.

2.3.1. Personnel

Competent and motivated personnel are the key to proper management of environmental data. NOAA has many such individuals across the agency. Their work is more effective when they can exchange knowledge and work together. Such collaboration is supported in part by participation in the groups mentioned in Section 2.2.1. One intent of this document is to provide a conceptual framework and common understanding of their work.

Significant improvements in NOAA data management cannot be made on the basis of volunteer efforts. Employees responsible for any aspect of data management should have that role clearly stated in their performance plan, and should have the authority and means to carry out that role. Too often, activities such as creating and maintaining metadata, making data available to other users, or ensuring data are properly transmitted to an archival facility are treated as tasks that are ancillary to an employee's regular duties. These tasks typically are not included in employee performance plans, are not acknowledged as important by supervisors, and are not rewarded by the agency. Making good data management a part of NOAA's core business practices would help provide acceptance and recognition of these efforts. Data usage tracking and citation may also help (see Sections 3.2.9 and 3.3).

NOAA personnel should be informed of need for good data management principles. Relevant staff should be offered training in data management practices. Data-related knowledge of departing staff should be captured as part of exit procedures.

2.3.2. Budget

The cost of producing observations is typically much greater than the cost of properly managing the resulting data. Satellites, radars, ship and aircraft time, and field campaigns are expensive and labor-intensive, and without proper planning may consume the entire project budget while leaving little for proper data management. The *Data Management Planning Procedural Directive* (7) is intended in part to address this problem. Data-producing projects are required to consider how they will store, transmit, document and archive their data. Program managers, project leaders, and technical personnel should work together to adequately plan and budget for data management. (See also Section 0.)

With constrained budgets, NOAA cannot improve everything at once. Therefore, the following approach is suggested:

- Build new systems right the first time.
- Take advantage of tech refresh points to improve existing systems.
- Bring existing high-value datasets and systems into compliance over time, prioritizing key datasets such as those from NOAA Observing Systems of Record or those used in the National Climate Assessment.

2.3.3. Other Resources

Other resources include Data Centers, pilot projects, teams, conferences, documentation, and software. Some examples are listed below.

- **Data Centers:** The NOAA National Data Centers (NCDC, NGDC, and NODC) are among the world's premier facilities for long-term preservation and stewardship of environmental data. NOAA projects, guided by the *Procedure for Scientific Records Appraisal and Archive Approval* (8), can work with these facilities to ensure their data are properly archived. Each Data Center is also establishing a catalog service to enable discovery of its holdings.
- **Pilot Projects:** Pilot projects are designed to test the implementation of new technologies prior to operational adoption. Examples include the NOAA National Data Centers Cloud Pilot[*] and the NOS Shared Hosting capability[†] to allow projects to host datasets for public access without needing to operate their own servers.
- **Teams:** Various cross-NOAA groups of personnel involved in EDM activities provide mutual support and guidance. The Data Management Integration Team (DMIT) is one such group with a mailing list and monthly telecons. More broadly, the Federation of Earth Science Information Partners (ESIP)[‡] is an open networked community, originally founded by NASA and NOAA, that brings together practitioners in science, data management, and IT.
- **Conferences:** The EDMC organizes an annual NOAA Environmental Data Management Conference for agency-wide exchange of relevant knowledge, successes, and problems. This event is typically held in June in the Silver Spring, Maryland area. Other workshops and meetings occur throughout the year.
- **Documentation:** The NOAA EDM Wiki (11) includes Best Practices and other guidance. This resource is publicly readable, and can be edited by NOAA personnel who request an account. NOAA projects are encouraged to consult and contribute to this Wiki. The Wiki also includes a repository of EDM Plans[§] submitted in compliance with the *Data Management Planning Procedural Directive.*
- **Software**: Good data management does not necessarily require writing new code. Open-source and commercial software packages exist for editing metadata or providing user-facing (public) services for data discovery, access, or visualization. Some recommended software is listed on the EDM Wiki.

2.4. Standards

Different types of standards are applicable in various phases of the Data Lifecycle. These include common vocabularies, standards for data quality, metadata standards that specify the content and structure of documentation about a dataset, data models and format standards that specify the content and structure of the digital data itself, and interface standards that specify how services are invoked. Some standards are general-purpose and may require specialization for particular data types. Adoption

[*] https://www.nosc.noaa.gov/EDMC/documents/edmcon/2012_breakout_sessions/Casey-CLASS_Cloud_Access_pilot_16May2012.pdf

[†] https://sites.google.com/a/noaa.gov/noaa-open-source-gis/noaa-hpcc-shared-hosting

[‡] http://esipfed.org/

[§] https://geo-ide.noaa.gov/wiki/index.php?title=Category:Data_Management_Plans of common standards supports interoperability, which enables diverse data, tools, systems, and archives to be combined without writing custom software to handle every data link. The broad use of a small set of common data, metadata, and protocol standards across NOAA, especially using international standards where possible, will decrease the cost of making and using NOAA observations, enhance the utility of the data, and help avoid redundant technical development. Existing data exchange agreements with NOAA, domestic and international partners must be upheld, but NOAA practices should be introduced appropriately in international coordination groups to foster compatibility of data management approaches.

2.5. Architecture

2.5.1. Infrastructure

NOAA infrastructure involved in environmental data management includes the observing platforms and systems themselves, data collection and processing systems, the archival data centers (NCDC, NGDC, NODC) and their associated systems for data ingest, storage and stewardship, other NOAA centers of data, dedicated data links such as the WMO Global Telecommunication System (GTS) and Satellite Broadcast Network (SBN), general-purpose network infrastructure, high-performance computing systems, and other computing resources. NOAA partners also operate infrastructure for data that NOAA may ingest.

These infrastructure components are expensive to acquire and maintain. Costs can be reduced over the long term by avoiding project-specific systems built from scratch. Instead, gradual adoption of commodity hardware and software, and the establishment of enterprise systems that provide functionality for multiple projects or the entire agency, are preferable. Adoption of interoperability standards (see Section 2.4) will support and simplify information exchange among NOAA systems and between NOAA and external data providers .Costs may be reduced by using commercial or NOAA-operated Cloud services (shared, pay-as-you-go information technology (IT) resources such as storage, processing, or software that can be scaled up or down based on demand).[*]

2.5.2. Service-Based Approach

NOAA environmental data must be available to users both inside and outside of NOAA. It is more efficient to make a given dataset accessible from a single authoritative source than to have users download, maintain, and possibly redistribute multiple copies, because the timeliness and accuracy of duplicative collections becomes increasingly uncertain. NOAA data and metadata should therefore be delivered through services -- that is, through web-based interfaces that can be invoked by software applications. These services can offer functions such as searching for data, retrieving a copy or a subset of data, visualizing data (e.g., producing a colored map or a time-series graph), or otherwise transforming data (e.g., converting to other formats or other coordinate systems). Rather than

[*] See Appendix C: Cloud Computing for further discussion.

establishing vertically-integrated "stovepipes" that only provide services for specific users and customers, a shared-services architecture, as illustrated in Figure 4, is recommended.

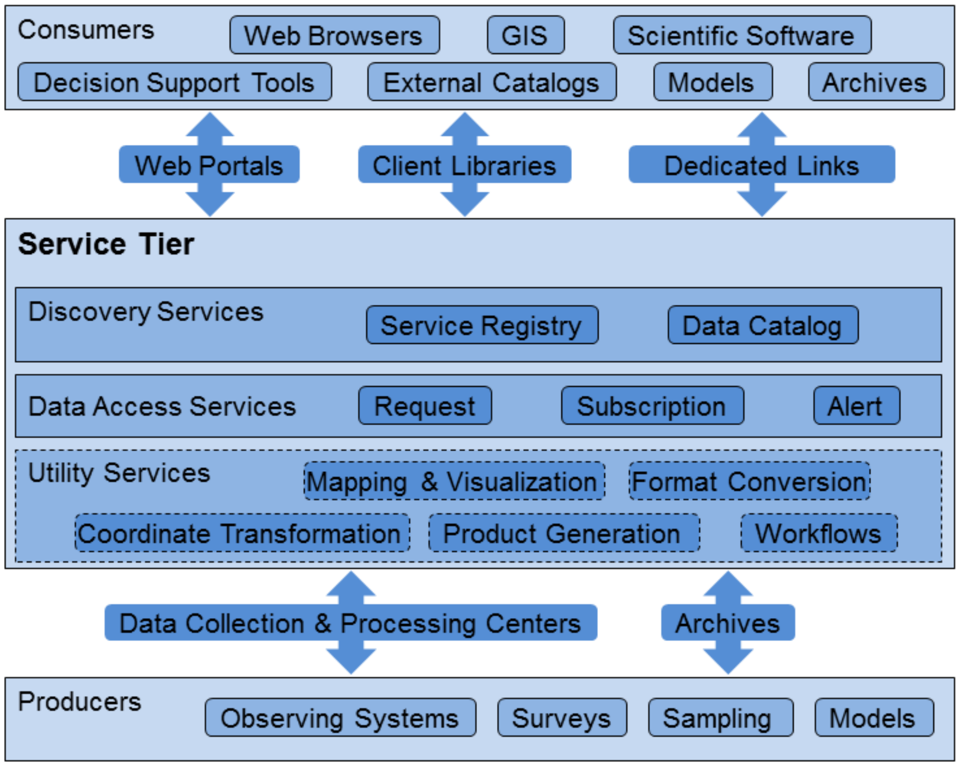

Figure 4: Schematic of shared-services architecture. Rather than explicitly linking individual data producers to specific customer applications, data management services and tools are generalized and decoupled as much as possible. Shared services can be established at an agency level (e.g., for data catalogs), and compatible services (e.g., based on the same pre-approved software) can be established at the program level where needed.

Services should be as consistent and standardized as possible to simplify the programming of applications that can integrate information from multiple sources. Such applications currently exist for a variety of well-known service types. New or enhanced applications can be written by NOAA, our partners, and the private sector as needed. The Digital Government Strategy (15) states that "We must enable the public, entrepreneurs, and our own government programs to better leverage the rich wealth of federal data to pour into applications and services by ensuring that data [are] open and machine-readable by default."

NOAA data exist in many heterogeneous systems managed by multiple independent operators. National and NOAA activities in support of data center consolidation are designed to reduce the total number of computing facilities with dedicated power and cooling, and often with underutilized capacity, as a cost-saving measure. However, consolidation is unlikely to result in completely merging all diverse NOAA systems for distributing and archiving data into a single master system. Even if such a target state were achievable within NOAA, other agencies, other nations, and the private sector will retain their own systems. A federated systems approach, as illustrated in Figure 5, is therefore necessary to leverage and harmonize multiple legacy, modern, and future systems that have evolved separately and are managed independently. A federated system is a collection of project-specific or agency-wide information systems that are independently managed and loosely coupled in a way that provides the behavior of a single system while enabling each organization to remain the steward of its own information. When ingesting data from non-NOAA systems, factors such as IT security and the availability and integrity of the data must be considered. These factors and others are discussed in *External Data Usage Best Practice* (in preparation).

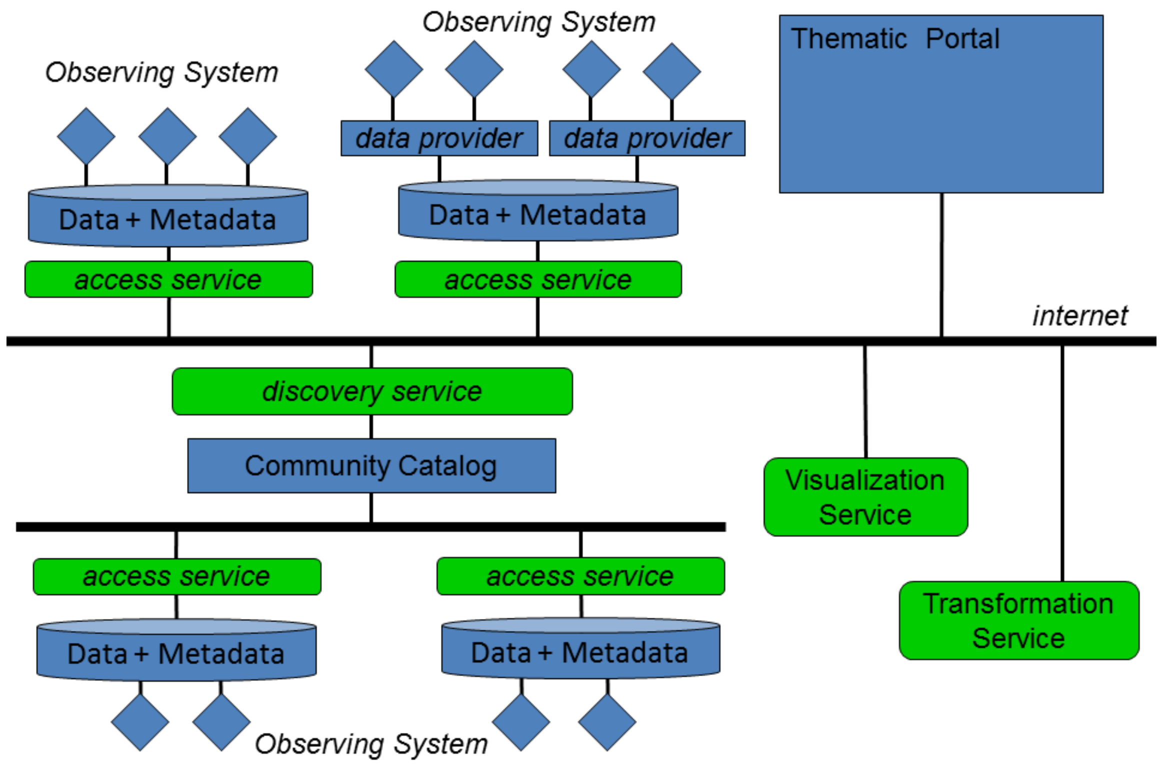

Figure 5: Schematic of service-based approach to providing access to data and metadata from observing systems. Data are stored in databases or file systems. Data access is mediated by services that provide security (limiting direct interaction with the back-end system), convenience (providing a table of contents and allowing customized subsets to be requested), and standardization (making access methods and formats compatible even if the internal storage differs). Catalogs can be built from these data access services, and can provide a discovery service to enable users to search for data. Value-added services such as visualization or other transformations can be provided, either by the original data holders or by third parties. Thematic portals can be constructed to present a unified access point to related datasets from multiple sources.

2.5.3. Designing for Flexibility

Innovations in IT and engineering are frequent and may offer significant benefits in cost or efficiency. NOAA should strive for modular and flexible architectures for observing systems, data management systems, and IT infrastructure in order to allow emerging technologies to be readily implemented. Custom-built, vertically integrated systems guided by inflexible design methodologies should be avoided because they are difficult to modify and may lock NOAA into old technologies or specific vendors..

2.6. Assessment

Assessment of NOAA data management activities includes estimating the current state, measuring progress, and getting feedback from users and implementers. The attributes we can assess include completeness of EDM planning, quality of metadata, level of data accessibility, and successful preservation for the long term.

- **Estimating the current state** of NOAA EDM: The Technology, Planning and Integration for Observation (TPIO)[*] program is assessing how data from NOAA Observing Systems of Record are managed. This will provide a baseline status.
- **Measuring progress:** Line-office representatives report on the implementation of Procedural Directives at meetings of the EDMC. The EDMC chair reports progress to NOSC and CIO Council several times per year. TPIO and the NGDC Enterprise Data Systems Group have begun prototyping a Data Management Dashboard intended to show current values and trends in metrics such as metadata quality and data accessibility.
- **Feedback:** NOAA personnel and contractors involved in EDM are invited to contact the EDMC and the DMIT regarding successes, failures, lessons learned and suggestions concerning this EDM Framework, EDMC Procedural Directives, and related activities. NOAA data providers can seek and respond to feedback from users. The US Paperwork Reduction Act imposes some limitations on methods for gathering feedback.[†]

3. The Data Lifecycle

The *Data Lifecycle* includes all the activities that affect a dataset before and during its lifetime. Different datasets may have somewhat different lifecycles, but this model is intended to be general. The use of the term "lifecycle" includes long-term preservation and is not meant to imply a finite lifetime or limited period of usefulness. We divide lifecycle activities into three groups, as shown in Figure 6:

- **Planning and Production**, which includes all activities up to and including the moment that an observation is captured by an observing system or data collection project;
- **Data Management**, which includes all activities related to processing, verifying, documenting, advertising, distributing and preserving data;
- **Usage**, which includes all activities performed by the consumer of the data (these activities are often outside the direct control of data managers).

[*] TPIO resides within NESDIS but performs NOAA-wide functions including supporting the NOAA Data Management Architect; serving as Executive Secretariat for the EDMC, NOSC, and DAARWG; and maintaining and analyzing a database of observing systems and requirements. See https://www.nosc.noaa.gov/tpio/.

[†] http://www.cio.noaa.gov/Policy_Programs/pracust.html

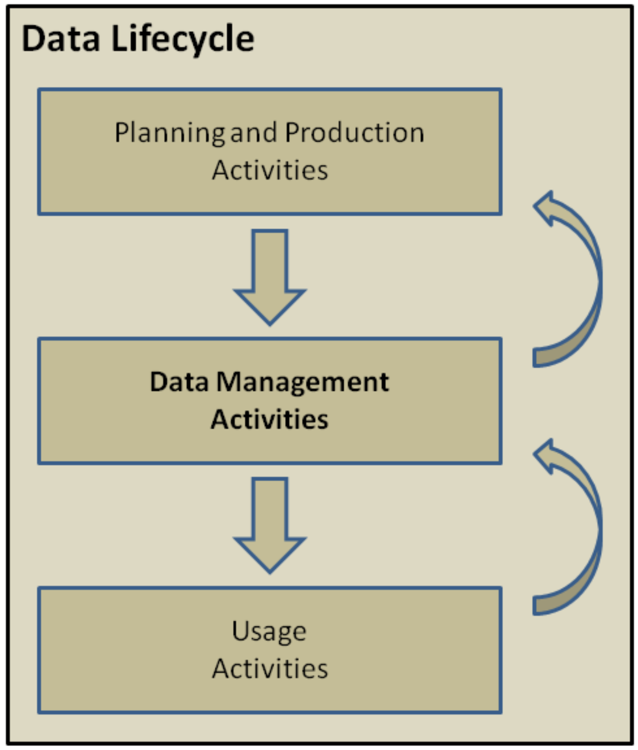

Figure 6: Overview of the Data Lifecycle, showing a decomposition into Planning and Production, Data Management, and Usage activities. The block arrows suggest the normal flow of information from planning towards usage, and the curved arrows indicate that the process may be cyclical, with conceptually "later" activities feeding back to or triggering "earlier" activities.

Figure 7 is a more detailed view of the Data Lifecycle, including all of the activities mentioned in this Section. The Data Lifecycle is a dynamic process rather than a linear sequence. That is, the steps in the lifecycle are not independent, but rather depend on and influence actions taken at other steps. For example, inadequate documentation at an early stage can prevent later use; generation of products from original data may yield new derived data that must also be collected and managed; user feedback regarding data may change or augment the documentation about data. Likewise, because data may go through multiple cycles of use and reuse by different entities for different purposes, effective management of each step, and coordination across steps in the lifecycle, are required to ensure that data are reliably preserved and can be accessed and used efficiently.

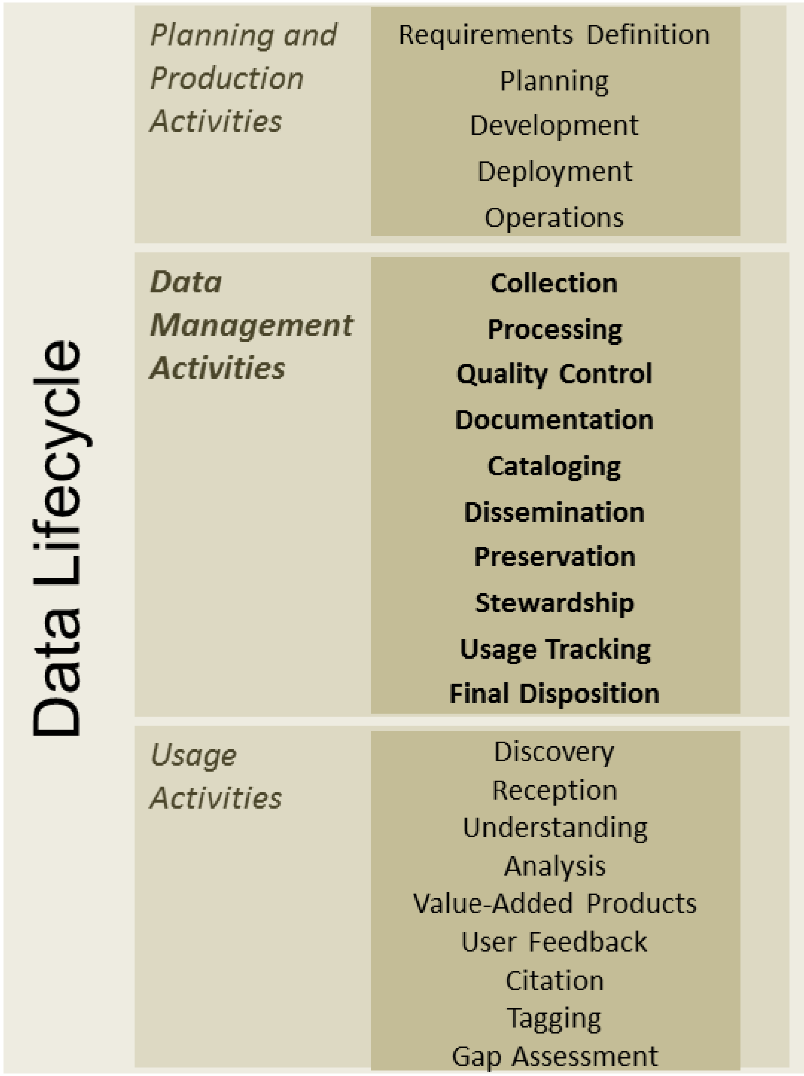

Figure 7: Activities in the Data Lifecycle. The Data Management Activities block is the focus of this Framework.

A lifecycle data management process ensures that observing systems are based on requirements, that the resulting data are properly stewarded, and that data can be used both for their original purpose and in novel ways.

Each phase of the Data Lifecycle is described in the following sub-sections.

3.1. Planning and Production Activities

The first phase of the Data Lifecycle is Planning and Production activities, which comprise:

- Requirements Definition
- Planning
- Development
- Deployment

- Operations

These include such tasks as assessing the need and requirements for a new observing system, planning how to meet those requirements and how to manage the resulting data, developing any necessary sensors, deploying the observing system, and operating and maintaining the observing system.

The Planning activity includes preparing for management of the resulting data. The *Data Management Planning Procedural Directive* (7) requires such planning and provides a template of questions to be considered. This planning should be done before data are collected, but existing projects without adequate plans should also address the issues. The Data Management Plan should be flexible and updated as needed because matters not considered in the original plan, or changes in technology, may emerge as data are acquired, processed, distributed, and archived. Program managers, project leaders, and technical personnel should work together and with NOAA EDM groups to plan data management in ways that maximize data compatibility and reduce overall costs.

The other activities in this phase are largely outside the scope of this EDM Framework, which focuses instead on the management of actual data once observations are collected. Nevertheless, activities that occur later in the Data Lifecycle may influence this phase. For example, a calibration error discovered during quality control may lead to changes in the operating procedure, and gap analysis may reveal new requirements.

3.2. Data Management Activities

The second phase of the Data Lifecycle is Data Management Activities, which include:

- Data Collection
- Processing
- Quality Control
- Documentation
- Cataloging
- Dissemination
- Preservation
- Stewardship
- Usage Tracking
- Final Disposition

3.2.1. Data Collection

Data Collection typically refers to the initial steps of receiving raw data from an environmental sensor or an observing campaign. Collection may also include purchasing commercial datasets, negotiating arrangements for access to data from foreign systems, issuing contracts for data collection, and issuing research grants that may result in the creation of environmental data. NOAA grantees are required by the *Data Sharing for NOAA Grants Procedural Directive* (10) to include a data sharing plan with their proposal and to share data in a timely fashion if funded. NOAA projects that use non-NOAA data should follow the *External Data Usage Best Practice* (in preparation) to ensure that relevant risks are considered.

3.2.2. Data Processing

Data Processing includes all the steps necessary to transform raw data into usable data records and to generate the suite of routine data products. Such processing is typically performed by specialized systems that have their own internal data management controls. Users do not normally have direct access to the processing system. However, the design of these systems can have a great impact on the cost to the agency and on the timeliness, preservation, and quality of the resulting data records and products. Processing systems should not be built from scratch for each observing system, because this does not enable the agency to leverage past investments or existing resources.

3.2.3. Quality Control

NOAA data should be of known quality, which means that data documentation includes the result of quality control (QC) processes, and that descriptions of the QC processes and standards are available. QC tests should be applied to data, including as appropriate automated QC in near-real-time, automated QC in delayed-mode, and human-assisted checks. Quality-assurance (QA) processes should be applied to provide validation that observations meet their intended requirements throughout the Data Lifecycle. QA may also include intercalibration of data from sensors on multiple systems. All QC and QA checks should be publicly described. The results of these checks should be included in metadata as error estimates or flagging of bad or suspect values. Raw data that have not undergone QC should be clearly documented as being of unknown quality.

3.2.4. Documentation

Data documentation provides information about the spatial and temporal extents, source, lineage, responsible parties, descriptive attributes, quality, accuracy, maturity, known limitations, and logical organization of the data. Formal, structured documentation is known as metadata. Metadata are critical for documenting and preserving NOAA's data assets. Standardized metadata support interoperability with catalogs, archives, and data analysis tools to facilitate data discovery and use. Correct and complete metadata are essential to ensuring that data are used appropriately and that any resulting analyses are credible.

The core metadata standards for NOAA environmental data are ISO 19115 (content) and ISO 19139 (Extensible Markup Language [XML] schema), as established by the *Data Documentation Procedural Directive* (9). Some older metadata records use the more limited Federal Geographic Data Committee (FGDC) Content Standard for Digital Geospatial Metadata (CSDGM); these should be converted to ISO and then improved using approaches and tools described on the EDM Wiki[*]. Conversion of well-structured metadata (e.g., in FGDC XML) to ISO is relatively straightforward, but non-standard or free-form documentation is more problematic.

[*] https://geo-ide.noaa.gov/wiki/index.php?title=Category:Metadata_Tools

Forecast model run collections should likewise be documented to enable discovery and understanding.

3.2.5. Cataloging

"Cataloging" is used here in a general sense to refer to all mechanisms established by data providers to enable users to find data. The word "Discovery" is employed below (Section3.3) to refer to the user's act of finding data-- Cataloging enables Discovery.

NOAA environmental data should be readily discoverable because modern research and decision-making depend critically on the ability to find relevant data from multiple agencies and disciplines.

Cataloging methods include enabling commercial search engines to index data holdings, establishing formal standards-based catalog services, and building web portals that are thematic, agency-specific, or government-wide. General web searching is often the first step for potential users, so this activity should be supported. However, advanced searching based on location, time, semantics or other data attributes requires formal catalog services.

The proliferation of portals such as data.gov, geo.data.gov, ocean.data.gov, NASA Global Change Master Directory (GCMD), Group on Earth Observations (GEO) Portal and others means that data providers are asked to register multiple times in different sites. This is not scalable and leads to redundant effort and duplicated cataloging of datasets. Data providers should be able to register their service in a single catalog and have other catalogs and portals automatically become aware of the new data. Some of the recommendations in Appendix A address this.

3.2.6. Dissemination

"Dissemination" is used here to mean both actively transmitting data and, more typically, enabling users to access data on request. NOAA environmental data should be readily accessible to intended customers as well as other potential users. Many users prefer direct access to online data via internet services that allow customized requests rather than bulk download of static files or delayed access via ordering services for near-line data. For high-volume data collections requiring near-line storage, NOAA data managers should carefully consider cloud hosting strategies and caching algorithms based on usage tracking to maximize the likelihood of popular data being online. Online services should comply with open interoperability specifications for geospatial data, notably those of OGC, ISO/TC211, and Unidata.

Actively transmitting data to operational customers is necessary in some cases. However, establishing new data conduits that are proprietary or duplicative should be avoided. Existing distribution channels should be shared where possible. Commodity hardware and software should be used in preference to custom-built systems. Government-funded open-source technologies[*] should be considered.

Data should be offered in formats that are known to work with a broad range of scientific or decision-support tools. Common vocabularies, semantics, and data models should be employed.

[*] E.g., Unidata Internet Data Distribution/Local Data Manager (IDD/LDM) system.

Numerical model outputs are often disseminated to users. Wherever possible, services and formats compatible with the observational data should be used to facilitate integration or comparison of data and model outputs from several sources.

3.2.7. Preservation and Stewardship

Data preservation ensures data are stored and protected from loss. Stewardship ensures data continue to be accessible (for example, by migrating to new storage technologies) and are updated, annotated or replaced when there are changes or corrections. Stewardship also includes reprocessing when errors or biases have been discovered in the original processing.

The NOAA National Data Centers -- NCDC, NGDC, and NODC -- are operated by NESDIS but perform data preservation and stewardship on behalf of the entire agency. NOAA data producers must establish a submission agreement with one of these data centers as described in the *Procedure for Scientific Records Appraisal and Archive Approval* (8), and must include archiving costs in their budget. To ensure data produced by grantees are archived, new Federal Funding Opportunities (FFOs) should arrange and budget in advance with a NOAA Data Center for archiving of data to be produced by the funded investigators.

Because an observation cannot be repeated once the moment has passed, all observations should be archived. Not only the raw data but also the accompanying information needed for understanding current conditions of the observation (e.g., satellite maneuver, instrument reports, change history *in situ* instruments, etc.) should be preserved. In some cases, especially the case of high resolution satellite imagery, strict compliance with this principle would result in substantial additional costs to telecommunications networks and data storage systems. In those cases where that cost is not budgeted, following a cost/benefit analysis, the issue will be brought to the NOSC for guidance as to whether additional funds should be requested through the budget process.

The representation of data that needs to be preserved and stewarded for the long-term should be negotiated with the Data Center and identified in the relevant data management plan. Key derived products, or the relevant versions of software necessary to regenerate products that are not archived, should also be preserved. The *Procedure for Scientific Records Appraisal and Archive Approval* (8) defines a process and includes a questionnaire to determining what to archive.

Some numerical model outputs should be preserved. These outputs are often voluminous or ephemeral, and what subset to archive should be carefully considered. The criteria for such decisions are outside the scope of this Framework.

Data rescue refers to the preservation of data that are at risk of loss. Such data include information recorded on paper, film, or obsolete media, or lacking essential metadata, or stored only in the scientist's computer. Data rescue is expensive--much more expensive than assuring the preservation of current datasets. NOAA datasets at risk should be registered with the International Council for Science (ICSU) Committee on Data for Science and Technology (CODATA) Data at Risk Task Group (DARTG).[*†]

Data that has been sent to a NOAA Data Center should also be discoverable and accessible as described in the preceding sections. Ideally, the mechanisms for cataloging and disseminating archival data should be interoperable with those for near-real-time data.

3.2.8. Final Disposition

Each NOAA National Data Center already has a records retention schedule that documents the length of time it will retain particular classes of data and product. Each data producer should also have a records retention schedule indicating when their data should be transferred to a Data Center for long term preservation. As IT resource consolidation and reduction occurs, it will become increasingly necessary to transfer custody of data records from local servers and services to NOAA Data Centers.

Retirement and eventual removal of archived material requires resources to update metadata, to request and respond to public comments, and to provide public notification of removal. The metadata record might be preserved indefinitely.

3.2.9. Usage Tracking

Usage tracking refers to NOAA's ability to measure how often datasets are being used. Crude estimation can be made by counting data requests or data transmission volumes from Internet servers. However, such statistics do not reveal whether data that was obtained was actually used, or if used whether it was helpful, or whether the initial recipient redistributed the data to other users.

More sophisticated means of assessing usage while preserving the anonymity of users are desirable. NOAA data producers, in collaboration with a NOAA Data Center, should assign persistent identifiers to each dataset, and include the identifier in every metadata record and data file. The *Data Citation Procedural Directive* (in preparation) will address this topic. Researchers and other users will be encouraged to cite the datasets they use (see Section 3.3).

3.3. Usage Activities

The third phase of the data lifecycle is Usage. These activities are typically outside the scope of data manager influence -- once a user has obtained a copy of the desired data, what he or she does with it may be unknown or uncontrolled. However, the ability to obtain and use data is certainly a by-product of a good lifecycle data management process, and information from or about users may influence or improve the data management process. NOAA is the biggest user of its own data, so improvements in data management could reduce cost and complexity within the agency.

---

[*] http://ils.unc.edu/~janeg/dartg/

[†] One example of NOAA data at risk is analog tide gauge data recorded on paper (marigrams) stored in over 1000 boxes at the US National Archives.

Activities related to Usage in the Data Lifecycle include:

- Discovery
- Reception
- Analysis
- Value-Added Product Generation
- Feedback
- Citation
- Tagging
- Gap Analysis

Users must be able to **Discover** and **Receive** data they want. These activities are enabled by NOAA Cataloging and Dissemination activities (Sections 3.2.5 and 3.2.6).

Analysis is defined broadly to include such activities as a quick evaluation to assess the usefulness of a dataset, or the inclusion of a dataset among the factors leading to a decision, or an actual scientific analysis of data in a research context, or data mining. Such activities are only possible if the data have been well-documented (Section 3.2.4) and are of known quality (Section 3.2.3).

Users of NOAA environmental data may create derived or **Value-Added Products**. These new products may themselves constitute a new dataset that merits its own lifecycle data management process. NOAA or NOAA-funded projects that routinely create new products should establish and follow a data-management plan and ensure the products they generate are discoverable, accessible, and archived. New products should be linked back to the original source data via appropriate documentation and citation of dataset identifiers (see Section 3.2.9).

Data users should have a mechanism to provide **Feedback** to NOAA regarding usability, suspected quality issues, and other aspects of its data. Agency point-of-contact information should be included in the metadata. Any feedback received should be acted upon if possible and included in the metadata if appropriate in order to help future users. Limited mechanisms for user feedback, notably Help Desks at each data center, have been established. These require that the user have obtained the data from the Data Center and be willing to engage in dialog. Possible additional approaches include mailing lists or social media.

Citation refers to the ability to unambiguously reference a dataset that was used as input to a model, decision, scientific paper, or other result. This is an emerging topic of broad interest[*] that will be addressed by NOAA's *Data Citation Procedural Directive* (in preparation). The Earth Science Information Partnership (ESIP) Federation also provides citation guidelines.[†] The core concepts are (1) persistent

[*] Workshops in 2011 include *Geo-Data Informatics: Exploring the Life Cycle, Citation and Integration of Geo-Data*[*] (Broomfield, Colorado, March 2011) sponsored by the National Science Foundation (NSF) and *Developing Data Attribution and Citation Practices and Standards*[*] (Berkeley, California, August 2011) sponsored by the National Academy of Sciences (NAS) Board on Research Data and Information (BRDI).

[†] http://wiki.esipfed.org/index.php/Interagency_Data_Stewardship/Citations

identifiers are assigned to each dataset and (2) the identifier and ancillary information are included in the reference list of the paper or other work. This is analogous to citing a book by its International Standard Book Number (ISBN) and indicating the title and page numbers used. (See also Section 3.2.9.)

Tagging refers to the ability to identify a dataset as relevant to some event, phenomenon, purpose, program, or agency *without* needing to modify the original metadata. Existing examples of tagging include the ability of NOAA users of Google Drive to assign documents to multiple collections without modifying the folder hierarchy, or of Facebook users to tag individuals in a photo without editing the file-level metadata. The ability to tag is essential because the current practices of (a) creating new Catalogs and asking people to re-register a relevant subset of their data there, or (b) asking people to add new metadata tags such that an external project can detect them (e.g., the GEOSS DataCORE activity in 2011), are not scalable because they require additional work and lead to the proliferation of duplicate datasets and metadata records. No specific solution is proposed here, but an appropriate use of collection-level catalogs may support tagging.

Gap Analysis refers to the determination by users of data or decision-makers that additional data are needed to satisfy operational requirements or to understand a phenomenon -- for example, more frequent coverage, improved spatial or spectral resolution, or observations of other quantities. Gap analysis may also address continuity of observations to meet operational requirements or enable long-term trend analysis. Such a determination influences the Requirements Definition activity, which is the start of a new Data Lifecycle.

4. Summary

NOAA data constitute an irreplaceable national resource that must be well-documented, discoverable, accessible, and preserved for future use. Good data management should be part of NOAA's core business practices, and employees and leadership should be aware of their roles and responsibilities in this arena. The NOAA Environmental Data Management Framework recommends that EDM activities be coordinated across the agency, properly defined and scoped, and adequately resourced. The Framework defines and categorizes the policies, requirements, and technical considerations relevant to NOAA EDM in terms of Principles, Governance, Resources, Standards, Architecture, Assessment, and the Data Lifecycle. The Framework enumerates specific recommendations in Appendix A.

NOAA thanks the Science Advisory Board for its recommendation in March 2012 that an Environmental Data Management Framework be developed.

Appendix A: Recommendations

The following is a partial list of recommendations that would advance the goals of improved environmental data management at NOAA. They are grouped according to who would be primarily responsible for implementing them.

Data Producers and Observing System owners:

1. Write Data Management Plans (DMPs) and submit them to the EDMC DMP repository.
2. Allocate an appropriate percentage of project funds to managing the resulting data.
3. Ensure data producers Initiate the negotiation of submission agreements, including relevant budget requirements, with a NOAA Data Center in advance of data collection.
4. Solicit feedback from users regarding the accessibility, usability and quality of NOAA data, make improvements if appropriate, and report improvements or issues to EDMC or DMIT.
5. Support the Observing System of Record[*] Data Management Assessment.
6. Ensure that observing requirements and capabilities are included and validated in the NOAA Observing System Architecture (NOSA) and Consolidated Observing Requirements List (CORL) databases maintained by TPIO.[†]
7. Produce ISO metadata natively for new environmental data.
8. Transition metadata from legacy standards (FGDC CSDGM), non-standard formats, and unstructured documentation to correct and complete ISO metadata records, focusing especially on high-value datasets and observing systems of record.
9. Leverage tools already developed for metadata transformation and quality assessment.

Data Management Integration Team and other technical staff

10. Use existing domestic and international data, metadata, and protocol standards wherever suitable in preference to *ad hoc* or proprietary methods. If existing standards seem not suitable, provide feedback to EDMC or relevant standards body.
11. Coordinate adoption of interoperability standards by working with cross-NOAA groups such as EDMC and DMIT, and with external coordination groups.
12. Document best practices, experiences, examples, useful software, teams, events, etc on the EDM Wiki.
13. Coordinate enhancements to open-source software via DMIT or other cross-NOAA teams to avoid duplication of effort.
14. Publish on the NOAA EDM Wiki (11) the conventions, profiles and examples adopted to specialize standards for particular data types.
15. Develop a NOAA Cloud Strategy to address deployment scenarios, IT security issues, and procurement mechanisms.

[*] https://www.nosc.noaa.gov/OSC/sor.php

[†] https://www.nosc.noaa.gov/tpio/

16. Establish a federated search capability across multiple distributed catalogs and metadata sources that can be queried both by data users and by external or thematic catalogs.
17. Determine whether tagging (see Section3.3) can help triage available datasets as suitable for inclusion in various portals and external catalogs.

EDMC:

18. Support development of the Metadata Rubric and Data Management Dashboard.
19. Review data management plans of projects that seek funding approval from the IT Review Board (ITRB).*
20. Identify projects that do not properly document, share, or archive their data. Assist them in adopting good data management practices. Bring them to the attention of NOAA Leadership if necessary.

CIO community:

21. Pre-approve IT security Certification and Accreditation (C&A) for standard software packages to maximize compatibility and minimize the administrative hurdles involved in setting up new servers.
22. Promote reusable software and modular systems for reduced development and maintenance cost.
23. Assess investments in new or upgraded infrastructure components prior to approval regarding use of commodity technologies, ability to support multiple projects, and interoperability.
24. Continue and expand efforts for shared hosting of small datasets.
25. Maintain legacy data exchange mechanisms as needed, but consider adoption of common standards as part of technology refresh cycle.
26. Promote implementation of modern data access services for all NOAA data collections.
27. Revise IT security policies to make Cloud deployments routine and easier to approve than in-house systems.

NOAA Leadership:

28. Decline or postpone projects seeking approval from the ITRB for IT funding if data management planning and budgeting are inadequate.
29. Empower Line Offices to designate an EDM Officer (similar to IT Security Officer) with the authority and responsibility to oversee and enforce EDM compliance within their Office. Include such duties in the individuals' performance plans.
30. Update individual performance plans of all employees who produce, document, or manage data to permit, acknowledge and empower their work.
31. Ensure that personnel responsible for environmental data understand the need for data management and are trained in good EDM practices.
32. Identify or establish process for transferring program or project funds as needed to the designated long term archival repository or other appropriate data management entities.
33. Ensure that Federal Funding Opportunities (FFOs) plan for archiving of grant-produced data at a NOAA Data Center.

---

* The NOAA ITRB name and description are currently under revision. Existing description and terms of reference are at http://www.cio.noaa.gov/IT_Groups/noaa_cio_nitrb.html.

Appendix B: Abbreviations

Abbreviation	Meaning
AFWA	Air Force Weather Agency
CEOS	Committee on Earth Observing Satellites
CIO	Chief Information Officer
CSDGM	FGDC Content Standard for Digital Geospatial Metadata
DMA	Data Management Architect
DMIT	NOAA Data Management Integration Team
EA	Enterprise Architect
EDM	Environmental Data Management
EDMC	NOAA Environmental Data Management Committee
FGDC	US Federal Geographic Data Committee
FFO	Federal Funding Opportunity
GCMD	NASA Global Change Master Directory
GEO	Group on Earth Observations
GIS	Geographic Information System
GTS	WMO Global Telecommunication System
IOC	Intergovernmental Oceanographic Commission
IOOS®	US Integrated Ocean Observing System
ISO	International Organization for Standardization
IWGDD	Interagency Working Group on Digital Data
NAO	NOAA Administrative Order
NCDC	National Climatic Data Center
NESDIS	National Environmental Satellite Data and Information Service
NGDC	National Geophysical Data Center
NGSP	NOAA Next Generation Strategic Plan
NMFS	National Marine Fisheries Service
NOAA	National Oceanic and Atmospheric Administration
NODC	National Oceanographic Data Center
NOS	National Ocean Service
NOSC	NOAA Observing Systems Committee
NRC	National Research Council
NWS	National Weather Service
OAR	Office of Atmospheric Research
OCIO	Office of the CIO
OGC	Open Geospatial Consortium
OMAO	Office of Marine and Aviation Operations
OMB	Office of Management and Budget
OSPO	Office of Satellite and Product Operations
OSTP	Office of Science and Technology Policy
PDs	EDMC Procedural Directives
PPI	NOAA Office of Program Planning and Integration
SAB	NOAA Science Advisory Board
SBN	Satellite Broadcast Network
TC211	ISO Technical Committee 211 for Geographic Information
TPIO	NOAA Technology Planning and Integration for Observations
USGEO	US Group on Earth Observations
WMO	World Meteorological Organization

Appendix C: Cloud Computing

Cloud computing refers to the use of shared information technology (IT) resources such as storage, processing or software. Cloud resources can be scaled up or down based on demand. Multiple projects can share resources without each needing to have surplus capacity for the maximum expected load. Projects can acquire and pay for IT resources on an as-needed basis without maintaining in-house computing facilities. These shared IT resources can be operated either externally by commercial Cloud service providers or internally by one division on behalf of the entire agency. Cloud computing is a fundamental shift from the traditional approach of having each project procure and operate dedicated, in-house IT resources.

The US Chief Information Officer has issued a "Cloud-first" policy (6) and a *Federal Cloud Computing Strategy* (16). NOAA is required to consider Cloud-based approaches in favor of building or maintaining dedicated IT systems. Within NOAA, the Google Unified Messaging System (UMS) contract for email, calendars, and document sharing is an example of migration to the Cloud. Possible Cloud deployment scenarios for environmental data include:

- The master copy of a NOAA dataset is retained internally at a NOAA Data Center, while a public copy is sent via one-way push to a publicly accessible commercial Cloud where external customers (the private sector, the general public, foreign governments) can obtain data and perhaps invoke additional services (subsetting, visualization, transformation, etc). A digital signature (checksums or hashes) is produced and compared where appropriate to confirm the authoritativeness of the public copy.
- Non-real-Time Processing: climate product generation, satellite data reprocessing, and other non-real-time computation are performed on commercial cloud resources. The resulting products are also disseminated via the Cloud. The input data may already reside in the same Cloud.

Such scenarios would reduce the load on NOAA servers and allow capacity to be quickly ramped up during periods of high demand.

Costs and procurement mechanisms must be assessed carefully in Cloud deployments. There are monthly charges based on data storage, data retrieval, and computing cycles that must be budgeted for and payable across the fiscal year boundaries.

Appropriate IT security must be considered when NOAA data are hosted on commercial Cloud services. NOAA servers must comply with *NAO 212-13: NOAA Information Technology Security Policy* (12). Cloud deployments may reduce information technology (IT) security risks to NOAA systems by placing public-facing servers outside the NOAA security boundary. The General Services Administration (GSA) Federal Risk and Authorization Management Program (FedRAMP)* was established to ensure secure cloud computing for the federal government. Only vendors authorized by FedRAMP may be used. [Note: as of

* http://www.gsa.gov/portal/category/102371

the start of FY 2013, no Cloud service providers have formally met FedRAMP requirements or been granted a provisional authorization.

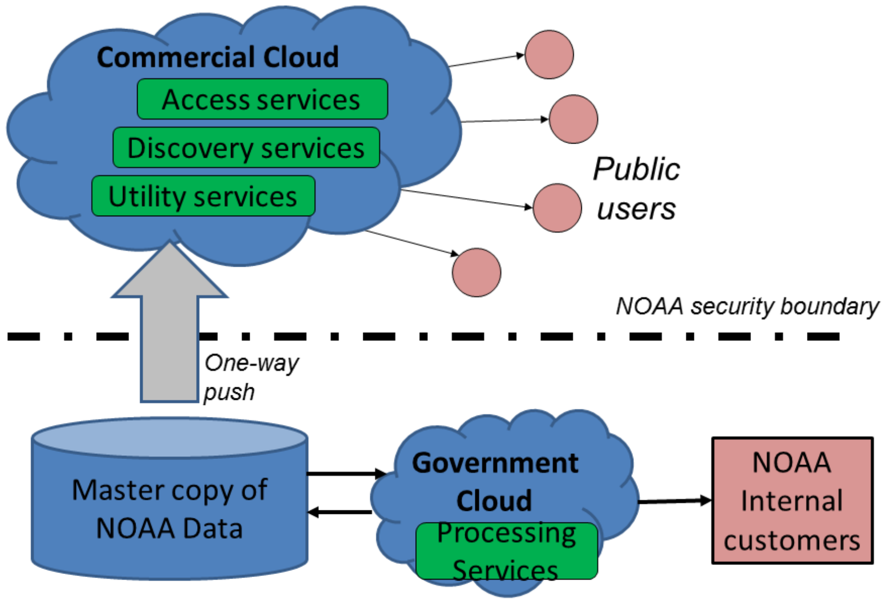

Figure 8: Potential Cloud deployment scenario for NOAA data.

Appendix D: References

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

2. —. *NAO 212-15: Management of Environmental Data and Information.* Silver Spring, MD : US National Oceanic and Atmospheric Administration, 2010. http://www.corporateservices.noaa.gov/ames/administrative_orders/chapter_212/212-15.html.

3. **NRC.** *Environmental Data Management at NOAA: Archiving, Stewardship, and Access.* Washington : National Research Council, 2007. http://www.nap.edu/catalog.php?record_id=12017.

4. **IWGDD.** *Harnessing the Power of Digital Data: Taking the Next Step.* Washington : Interagency Working Group on Digital Data, 2011. http://www.cendi.gov/publications/pub_CENDI-2011-1.pdf.

5. **USGEO.** *Exchanging Data for Societal benefit: An IEOS Web Services Architecture.* Washington : United States Group on Earth Observations, 2008. http://usgeo.gov/images/USGEOMain/exchangingdataforsocietalbenefit.pdf.

6. **US CIO.** *25 Point Implementation Plan to Reform Federal Information Technology Management.* Washington : White House Office of Management and Budget (OMB), 2010. http://www.cio.gov/documents/25-Point-Implementation-Plan-to-Reform-Federal%20IT.pdf.

7. **NOAA EDMC.** *Data Management Planning Procedural Directive.* Silver Spring, MD : US National Oceanic and Atmospheric Administration, 2011. https://www.nosc.noaa.gov/EDMC/PD.all.php.

8. —. *Procedure for Scientific Records Appraisal and Archive Approval.* Silver Spring, MD : US National Oceanic and Atmospheric Administration, 2008. https://www.nosc.noaa.gov/EDMC/PD.all.php.

9. —. *Data Documentation Procedural Directive.* Silver Spring, MD : US National Oceanic and Atmospheric Administration, 2011. https://www.nosc.noaa.gov/EDMC/PD.all.php.

10. —. *Data Sharing for NOAA Grants Procedural Directive.* Silver Spring, MD : US National Oceanic and Atmospheric Administration, 2012. https://www.nosc.noaa.gov/EDMC/PD.all.php.

11. —. *NOAA Environmental Data Management Wiki.* [Online] https://geo-ide.noaa.gov/wiki/.

12. **NOAA.** *NAO 212-13: NOAA Information Technology Security Policy.* Silver Spring, MD : US National Oceanic and Atmospheric Administration, 2003. http://www.corporateservices.noaa.gov/ames/administrative_orders/chapter_212/212-13.html.

13. **NOAA OCIO.** *Guiding Enterprise Architecture Principles.* Silver Spring, MD : US National Oceanic and Atmospheric Administration, 2008. https://secure.cio.noaa.gov/secure_docs/EA_Documentation/Guiding_EA_Principles.pdf.

14. **US OMB.** *Circular No. A-16: Coordination of Geographic Information and Related Spatial Data Activities.* Washington DC : US Office of Management and Budget, 2002. http://www.whitehouse.gov/omb/circulars_a016_rev.

15. **US EOP.** *Digital Government: Building a 21st Century Platform to Better Serve the American People.* Washington DC : US Executive Office of the President, 2012. http://www.whitehouse.gov/sites/default/files/omb/egov/digital-government/digital-government.html.

16. **US CIO.** *Federal Cloud Computing Strategy.* Washington : White House Office of Management and Budget, 2011. https://cio.gov/wp-content/uploads/downloads/2012/09/Federal-Cloud-Computing-Strategy.pdf.

17. **NOAA OCIO.** *NOAA IT Review Guidance.* Silver Spring, MD : US National Oceanic and Atmospheric Administration, 2012. http://www.cio.noaa.gov/IT_Groups/NISN-3.007_NITRB_Guidance.pdf. NISN 3.007.

18. **NOAA DMIT.** *Global Earth Observation Integrated Data Environment (GEO-IDE) Concept of Operations.* Silver Spring, MD : US National Oceanic and Atmospheric Administration, 2006.

19. **NOAA NWS.** *Obtaining Environmental Data from External Parties.* Silver Spring, MD : National Oceanic and Atmospheric Administration, 2009. http://www.nws.noaa.gov/directives/001/001.htm. NWSPD 1-1201.

20. **NOAA OCIO.** *NOAA Information Quality Act Guidelines.* Silver Spring, MD : US National Oceanic and Atmospheric Administration, 2012. http://www.cio.noaa.gov/Policy_Programs/IQ_Guidelines_011812.html.