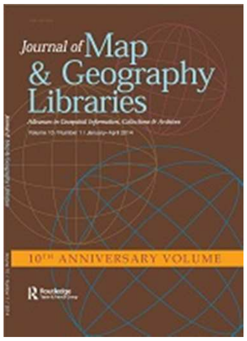


NOAA Environmental Data Management

Journal:	*Journal of Map & Geography Libraries*
Manuscript ID:	WMGL-2015-0015.R1
Manuscript Type:	Original Article
Keywords:	data management, earth observations, metadata, open data, cloud computing

NOAA Environmental Data Management

Abstract

The US National Oceanic and Atmospheric Administration (NOAA) operates over 100 observing systems and numerical models providing information and forecasts about the planetary environment from the bottom of the ocean to the surface of the Sun. NOAA data constitute an irreplaceable resource that must be well-documented, discoverable, accessible, and preserved for future use. Good data management should therefore be part of NOAA's core business practices, and employees and leadership should be aware of their roles and responsibilities in this arena. NOAA has developed an Environmental Data Management (EDM) Framework document that discusses Principles, Governance, Resources, Standards, Architecture, Assessment, and the Data Lifecycle, and which enumerates specific recommendations. The NOAA EDM Committee has issued Directives pertaining to data management planning, archiving, data access, metadata, and other topics. A Data Catalog has been established, and a project to assign persistent, citable identifiers to archival data is well underway. Numerous groups at NOAA are performing technical work related to data access, usability, and preservation.

The purpose of this paper is to describe these documents and activities in order to share our experiences and to provide guidance and encouragement for improved data management at other organizations.

1. Introduction

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 nation and the world. Earth observations are typically produced for specific purposes -- 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 the data. The US National Oceanic and Atmospheric Administration (NOAA) operates more than a hundred observing systems and numerical models providing information and forecasts about the planetary environment. Data from individual observing systems serve their primary customers quite effectively and are often very well managed at NOAA through established data processing and dissemination systems and archival data centers. However, 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 to ensure all data are equally-well managed and are made available in a comparable fashion to maximize secondary use and to address societal benefit areas requiring data from multiple sources.

This paper discusses NOAA activities in support of optimal environmental data management (EDM). We present a conceptual target state, and describe a Framework of Environmental Data Management (EDM) policies, organizational practices, and technical considerations to support effective and continuing access to Earth observations and derived products. We highlight recent policy and implementation activities. Finally, we conclude with some recommendations for optimal data management. Although this paper focuses on NOAA activities, much of the content is broadly applicable to other Earth-observing organizations.

We define environmental data using terminology from NOAA Administrative Order (NAO) 212-15 (NOAA 2010) 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." For the purposes of this paper, we use the terms "data" and "environmental data" interchangeably. This paper focuses primarily on observations and derived products, but also touches on numerical model outputs. Non-digital data, documents, preserved geological or biological samples, and non-environmental data (personnel, budget, etc.) are outside the scope of this discussion.

2. 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 summary. NOAA EDM activities are intended to help guide all NOAA data toward such a target. The modest expectations depicted here 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. The Directive documents mentioned here, some of which are still in preparation, are discussed more fully below.

Figure 1: GOES HERE

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 *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 the NOAA National Centers for Environmental Information (NCEI) for long-term preservation, in accordance with the *Scientific Records Appraisal and Archive Approval* procedure.
7. Datasets are assigned a persistent identifier (ID) in accordance with the *Data Citation* directive.
8. NCEI 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. Users can cite data they utilize by referencing persistent identifiers, 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.

3. The Environmental Data Management Framework

The *NOAA Environmental Data Management Framework* (NOAA EDMC 2013) document defines and categorizes the policies, requirements, activities, and technical considerations relevant to the management of observational data and derived products by NOAA. The 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. The NOAA EDM Framework was an outgrowth of inter-agency work in which NOAA participated in 2011, namely the development of the *National Strategy for Civil Earth Observations* (US NSTC 2013); section 3.2 of that document introduced the concept of an EDM framework, which NOAA has elaborated with greater specificity.

The basic elements of the EDM Framework are illustrated in Figure 2 and include Principles, Governance, Resources, Standards, Architecture, and Assessment that apply broadly to many classes of data, and Data Lifecycles that may be specific to particular data collections. This Section discusses each of those elements.

Figure 2: GOES HERE

3.1. 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 3**Error! Reference source not found.**:

- **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).

Figure 3: GOES HERE

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. These dependencies are suggested by arrows in the diagram. 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.

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 greater detail in the NOAA EDM Framework (NOAA EDMC 2013).

3.2. Principles

The following basic principles, which were first enunciated in the *National Strategy for Civil Earth Observations* (US NSTC 2013), 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: Data should be made fully and openly available to all users promptly, in a non-discriminatory manner, and free of charge (or at no more than the cost of reproduction).

Long-Term Preservation: Data should be managed as an asset and preserved for future use.

Information Quality: Data should be well documented and of known quality.

Ease of Use: Raw 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 elaborated in the following subsections.

3.2.1. Full and Open Access

In general, data managed or paid for using public 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 law or policy.)

- **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.

3.2.2. 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.

3.2.3. 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.

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.

3.2.4. 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.

3.3. Governance

3.3.1. NOAA Governance

A systematic approach to data management requires agency-wide coordination. The Environmental Data Management Committee (EDMC)[i] is a nexus of EDM governance activities at NOAA. Established in

2010 by NOAA Administrative Order (NAO) 212-15 (NOAA 2010), EDMC reports to both the Chief Information Officers Council[ii] and the NOAA Observing Systems Council,[iii] and is a voting body with representatives from each division of NOAA. (The author of this paper is EDMC Chair.)

NOAA's National Centers for Environmental Information (NCEI)[iv] are responsible for long-term preservation and stewardship, and are crucial to governing and implementing appropriate data management practices at NOAA. NCEI was established in 2015 as an organization which groups the former National Climatic Data Center (NCDC), National Geophysical Data Center (NGDC), and National Oceanographic Data Center (NODC) under one management structure.

EDMC is chartered to develop and approve Procedural Directives (PDs); the following have been issued:

- *Data Management Planning Procedural Directive* (NOAA EDMC 2014)**:** Directs managers of all data production projects and systems to plan in advance for data management, and contains a DM Plan template with questions to be addressed by data production projects. This is version 2 of a directive originally issued in 2011.
- *Procedure for Scientific Records Appraisal and Archive Approval* (NOAA EDMC 2008): Defines the process used by NOAA data centers to approve archive submission requests.
- *Data Documentation Procedural Directive* (NOAA EDMC 2011): States that all NOAA data collections, products derived from these data, and services that provide NOAA data and products, shall be documented. It 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.
- *Data Sharing for NOAA Grants Procedural Directive* (NOAA EDMC 2012): 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 including 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. This directive is currently being revised in response to the White House Office of Science and Technology Policy (OSTP) memorandum *Increasing Access to the Results of Federally Funded Scientific Research* (US OSTP 2013).

- ***Data Access Procedural Directive*** (NOAA EDMC 2015): States that all NOAA environmental data shall be made discoverable and accessible via the Internet, except in limited circumstances, and requires that a formal waiver be sought to justify lack of public accessibility.
- ***Data Citation Procedural Directive*** (NOAA EDMC 2015): States requirements and procedure for datasets archived at NCEI to be assigned a persistent Digital Object Identifier (DOI), and provides guidance on appropriate level of granularity to do so for ongoing time series or complex data collections. Urges data users to cite datasets used in papers, decisions and other products, and recommends a citation format including the DOI.
- ***External Data Usage Recommended Practice*** (NOAA EDMC 2013): Provides a worksheet of potential issues to consider when using non-NOAA data for NOAA purposes.

3.3.2. Federal Governance

As a US agency, NOAA is naturally subject to federal policies and mandates regarding data management. Two documents of particular relevance were issued in 2013 which prompted specific action at NOAA:

- The *US Open Data Policy,* as expressed in Office of Management and Budget (OMB) Memorandum M-13-13 (US OMB 2013), 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 establish a public data inventory. This inventory was established as the NOAA Data Catalog described in Section 4.1.

- The OSTP memorandum *Increasing Access to the Results of Federally Funded Scientific Research* (US OSTP 2013) states that agencies with significant research budgets must develop and implement a plan to ensure that research data, whether created within the agency or by extramural funding recipients, were made publicly accessible. Further, the memo states that published papers written with federal funding must be made freely accessible after a one-year embargo period, and that such papers should appropriately cite their source data. This resulted in the development of the NOAA *Plan for Public Access to Research Results* (NOAA 2015) and also gave impetus to the data citation project described in Section 4.2.

Among other federal policies and documents of relevance, we also note the *Digital Government Strategy* (US EOP 2012), which 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, and the *25 Point Implementation Plan to Reform Federal IT* (US CIO 2010) which establishes a "Cloud-first" policy for acquisition of new computing capability (see also Section 4.5 on this point).

3.3.3. External Coordination

In order to maximize compatibility of NOAA observations with data from other sources 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. Relevant external bodies include, among others: World Meteorological Organization (WMO), Committee on Earth Observing Satellites (CEOS), Group on Earth

Observations (GEO), US Group on Earth Observations (USGEO) Data Management Working Group, Federal Geographic Data Committee (FGDC), Open Geospatial Consortium (OGC), and International Organization for Standardization (ISO) Technical Committee 211 for Geographic Information and Geomatics (ISO/TC211).

3.4. Resources

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.

Competent and motivated personnel are the key to proper management of environmental data. NOAA has many such individuals across the agency, and their work is more effective when they can exchange knowledge and work together. Such collaboration is supported in part by the Data Management Integration Team (DMIT), 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 teleconferences. The NOAA EDM Wiki[v] is a compendium of recommended practices and other guidance edited by DMIT members. Working groups focused on specific activities are sometimes spun off from the larger DMIT. NOAA also holds an annual Environmental Data Management Workshop which brings over 100 people together for three days of plenary and break-out sessions.

Significant improvements in 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. Activities such as creating and maintaining metadata, making data available to other users, or ensuring data are properly transmitted to an archival facility should be included among the regular duties of relevant personnel.

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. NOAA's *Data Management Planning Procedural Directive* (NOAA EDMC 2014) is intended in part to address this problem. It states that data-producing projects are required to consider how they will store, transmit, document and archive their data, and that Program managers are identified as ultimately responsible for proper management of data from their program.

With constrained budgets, organizations cannot improve everything at once. Therefore, the following approach is suggested for data system developers and data stewards:

- 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.

3.5. 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 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.

Specific recommendations are still under development at NOAA. The *Data Documentation* directive explicitly mandates the ISO 19115 family of standards. The *Data Access* directive includes recommendations to use protocols such as OPeNDAP and OGC Web Services.

3.6. Architecture

NOAA infrastructure involved in environmental data management includes the observing platforms and systems themselves, data collection and processing systems, the National Center for Environmental Information and its associated systems for data ingest, storage and stewardship, 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 and facilities. 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 will support and simplify information exchange among NOAA systems and between NOAA and external data providers. Costs may also be reduced by using Cloud services (shared, pay-as-you-go information technology resources such as storage, processing, or software that can be scaled up or down based on demand).

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 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: GOES HERE

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, its partners, and the private sector as needed. The Digital Government Strategy (US EOP 2012) 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. A federated systems approach, as illustrated in Figure 5, is therefore necessary to leverage and harmonize multiple legacy, modern, and future systems that 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.

Figure 5: GOES HERE

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 organizations into old technologies or specific vendors.

3.7. Assessment

NOAA EDMC established a Data Management Dashboard to display metrics regarding metadata quality and compliance with policy directives. This Dashboard is only accessible to internal users and agency leadership, and currently displays 4 metrics:

- Time series graph showing cumulative number of Digital Object Identifiers (see Section 4.2).
- Time series graph showing number of metadata records and their overall completeness scores (see Section 4.4).
- Bar graph with results of baseline EDM assessment of key NOAA observing systems, indicating number of systems that have a data management plan, send data to the archive, have metadata, and have a public data access point.

- Number of data management plans submitted by each division of NOAA in compliance with the *Data Management Planning* directive.

EDMC is working to automatically compile and record additional metrics on the Dashboard.

4. Selected Implementation Activities

In addition to the NOAA policy directive work described in Section 3.3, specific implementation activities are underway in support of improved data management. The following is by no means a description of all NOAA activities, but rather focuses on those in which the present author is more directly involved.

4.1. NOAA Data Catalog

The NOAA Data Catalog[vi] was established in November 2013 in response to the US Open Data Policy (US OMB 2013) requirement that each agency have a public data inventory. The Catalog harvests metadata from data centers and other projects and provides a consolidated view. Users can search for data either through the web-based user interface or the OGC Catalog Service (CSW) query interface. Search results include summary information about each dataset as well as links to full metadata records and available data access points.

The Catalog was implemented using the open-source CKAN[vii] software and runs on the Amazon Federal GeoCloud. It harvests metadata in either ISO 19115(-2) or FGDC Content Standard for Digital Geospatial Metadata (CSDGM) format from web-accessible folders (WAFs) maintained by NOAA data centers or data producers. These WAFs are typically associated with subsidiary catalogs which provide domain-specific value beyond that of the comprehensive NOAA-level catalog. Over 60,000 datasets are currently advertised by the NOAA Data Catalog. As required by the Open Data Policy, a list of datasets is produced daily in Javascript Object Notation (JSON) format via an extension provided by the US data.gov project.

The Catalog project has had a number of benefits. Besides enabling NOAA to meet the inventory requirement and provide a single point of discovery, it has helped reveal areas where metadata and aggregation can be improved. Metadata improvements include better standardization of vocabularies and XML paths for key information, as well as disambiguation of related datasets that have identical titles but differ in time period or other attributes. Possible aggregation improvements include writing collection-level metadata records to represent groups of individual dataset granules.

The Catalog itself has a several shortcomings that we hope to address. Chief among these is the lack of a time filter or selection widget in the user interface; this is a serious omission for environmental data for which observation time is a fundamental attribute. We hope to add this feature as an enhancement to the open-source software.

4.2. Dataset Identifier Project

NOAA began in 2013 to assign Digital Object Identifiers (DOIs)[viii] to datasets archived at its national data centers. These DOIs are *resolvable*, which means that they can be used as a reference to the current location of the data, or more specifically the location of a landing page describing the data and its access mechanisms. They are also *persistent*, which means that once assigned they are never deleted or reassigned, and that if the landing page is moved then the target of the corresponding DOI is changed accordingly. Such identifiers, along with guidelines for data citation, allow datasets to be explicitly referenced in research papers or other work.

A working group was established at NOAA to govern the assignment of DOIs and to write the *Data Citation Procedural Directive* (NOAA EDMC 2015) mentioned previously (Section 3.3.1). Key decisions made include, among others:

- Use "opaque" DOIs (arbitrary alphanumeric strings) rather than building-in semantics (i.e., do not include names of programs or observing systems in the DOI).
- Only assign DOIs to data that have been accepted for long-term preservation at NOAA's National Center for Environmental Information (to ensure persistence).
- Use the data's ISO metadata record to automatically produce the human-readable landing page (using an XML stylesheet).
- For ongoing time-series data, assign a single DOI to the entire series rather than separate DOIs for individual segments (and guide users on how to cite a subset).

NOAA DOIs are minted using the University of California Digital Library EZID[ix] service. So far all DOIs have been created manually through the human user interface rather than automatically through the application programming interface, but increased automation is contemplated as implementation experience is gained. Over 300 NOAA DOIs have been thus far been assigned to archival datasets.

4.3. Unified Access Framework

The Unified Access Framework (UAF)[x] project began in 2010 as an effort to improve the discoverability and accessibility of regularly gridded observations and model outputs. It has since expanded to support development of conventions for *in situ* observations. UAF is a gridded-data integration capability that leverages several de facto standards:

- netCDF, which provides the abstract data model, software libraries, and a persistent binary format;
- the Climate and Forecast (CF) metadata conventions;
- THREDDS XML catalogs[xi] which provide a directory-style listing of data available from distributed suppliers;

- the OPeNDAP protocol for web transport of data subsets;
- an OGC compatibility layer that provides access to the grids through OGC Web Map Service (WMS) and Web Coverage Service (WCS).

UAF established a NOAA-wide THREDDS catalog[xii] of CF-compliant datasets and works on improving their metadata to improve users' ability to discover and make use of these datasets. UAF also provides working examples of existing software that can access data through these standards.

4.4. Metadata Assessment

NOAA has well over 60,000 metadata records in either ISO 19115(-2) or FGDC CSDGM format. The NOAA Enterprise Metadata Management Architecture (EMMA)[xiii] project is a long-standing effort to assess and improve metadata completeness. EMMA uses an Extensible Stylesheet Language Template (XSLT) to test for the presence of some 40 fields in ISO metadata records, display scores of individual records in a visual "rubric" to enable metadata authors to assess and improve their individual scores, and record the results in a database to keep track of progress. The resulting time series shows an increase in both the number of metadata records at NOAA and their overall completeness scores (see figure at endnote xii).

EMMA also provides another XSLT to convert from FGDC to ISO format. Initially, converted records were scored as well, but later experience has shown that there exist some hand-crafted FGDC records of excellent quality that receive low scores because of assumptions in the transformation, so this is an area for further work.

EMMA also supports metadata creation through a database of reusable components. Elements that appear in many records (for example, the name and contact information of an individual metadata author) can be stored in the database and referenced by ID in the raw metadata record. When the record needs to be viewed or scored, these references are automatically resolved to create a complete record.

Finally, the NOAA Data Catalog (section 4.1) has helped to reveal metadata quality issues. Whereas EMMA tests for the presence of a field, it cannot determine whether the content of the field is valid. Exposing more metadata to public scrutiny through the Catalog helps to highlight typos or other usability issues.

4.5. NOAA Big Data Project

In April 2015 NOAA signed Cooperative Research and Development Agreements (CRADA) with five Cloud-based Infrastructure-as-a-Service (IaaS) providers: Amazon, Google, IBM, Microsoft, and Open Cloud Consortium. The "Cloud" here means scalable data storage and computing facilities operated by the private (or academic) sector and accessible by interested parties on a pay-as-you-go basis. The basic intent of this Big Data Project is to ascertain whether putting a copy of selected NOAA data in the Cloud, alongside scalable computing capabilities, will permit the creation of new value-added products and services and lead to a sustainable business model. We stress here that NOAA data will remain freely available, but that entrepreneurs may create, and possibly sell, new products based on those data.

Figure 6 illustrates the conceptual architecture. Starting at the bottom: NOAA produces and retains the master copy of environmental observations and model outputs. NOAA is continuing to establish and operate services to enable discovery and access of data and metadata in standard formats; these services operate within the NOAA network perimeter. Working copies of data are sent to the Cloud -- Software to perform data analysis and integration functions can be installed or developed to run in the Cloud directly on the data stored there. Entrepreneurs can create customized products and services tailored to particular industries or markets. The start-up cost to try creating new services is small because it is not necessary to first establish infrastructure and retrieve a copy of the needed data.

As of the time of this writing, initial transfer of some current and archival data is in progress but no specific outcomes have been reported. Results will be announced on the NOAA Data Alliance[xiv] web site.

Figure 6: GOES HERE

5. 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. A number of specific implementation activities have been undertaken to improve data discoverability, documentation, accessibility, and citability.

The following is a partial list of recommendations, extracted from the NOAA EDM Framework, that would advance the goals of improved environmental data management at any organization.

1. Write Data Management Plans and allocate an appropriate percentage of project funds to managing the resulting data.

2. Solicit feedback from users regarding the accessibility, usability and quality of data, and make improvements if appropriate.
3. Use existing domestic and international data, metadata, and protocol standards wherever suitable in preference to *ad hoc* or proprietary methods.
4. 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.
5. Ensure that personnel responsible for environmental data understand the need for data management and are trained in good EDM practices.

END NOTES

References

NOAA EDMC. "Data Access Procedural Directive." *US National Oceanic and Atmospheric Administration.* 2015. https://www.nosc.noaa.gov/EDMC/PD.DA.php.

—. "Data Citation Procedural Directive." *US National Oceanic and Atmospheric Administration.* 2015. https://www.nosc.noaa.gov/EDMC/PD.DC.php.

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

—. "Data Management Planning Procedural Directive." *US National Oceanic and Atmospheric Administration.* 2014. https://www.nosc.noaa.gov/EDMC/PD.all.php.

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

—. "External Data Usage Recommended Practice." *US National Oceanic and Atmospheric Administration.* 2013. https://www.nosc.noaa.gov/EDMC/external.data.php.

—. "NOAA Environmental Data Management Framework." *US National Oceanic and Atmospheric Administration.* 2013. https://www.nosc.noaa.gov/EDMC/framework.php.

—. "Procedure for Scientific Records Appraisal and Archive Approval." *US National Oceanic and Atmospheric Administration.* 2008. https://www.nosc.noaa.gov/EDMC/documents/NOAA_Procedure_document_final_12-16-1.pdf.

NOAA. "NAO 212-15: Management of Environmental Data and Information." *US National Oceanic and Atmospheric Administration.* 11 04, 2010. http://www.corporateservices.noaa.gov/ames/administrative_orders/chapter_212/212-15.html.

US CIO. "25 Point Implementation Plan to Reform Federal Information Technology Management." *US Chief Information Officer.* 2010. http://www.cio.gov/documents/25-Point-Implementation-Plan-to-Reform-Federal%20IT.pdf.

US EOP. *Digital Government: Building a 21st Century Platform to Better Serve the American People.* Washington DC: US Executive Office of the President, 2012.

US NSTC. "National Strategy for Civil Earth Observations." *Executive Office of the President.* 2013. https://www.whitehouse.gov/sites/default/files/microsites/ostp/nstc_2013_earthobsstrategy.pdf.

US OMB. "Memorandum M-13-13: Open Data Policy -- Managing Information as an Asset." *US Office of Management and Budget.* 2013. http://www.whitehouse.gov/sites/default/files/omb/memoranda/2013/m-13-13.pdf.

US OSTP. "Increasing Access to the Results of Federally Funded Scientific Research." *White House Office of Science and Technology Policy.* 2013. http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf.

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

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

[iii] https://www.nosc.noaa.gov/

[iv] http://ncei.noaa.gov/

[v] https://geo-ide.noaa.gov/wiki/

[vi] https://data.noaa.gov/

[vii] http://ckan.org

[viii] http://doi.org/

[ix] http://ezid.cdlib.org/

[x] https://geo-ide.noaa.gov/wiki/index.php?title=Overview_and_benefits_of_the_GEO-IDE_UAF_Grid_Project

[xi] http://www.unidata.ucar.edu/software/thredds/current/tds/catalog/

[xii] http://ferret.pmel.noaa.gov/geoide/geoIDECleanCatalog.html

[xiii] http://www.ngdc.noaa.gov/metadata/emma/

[xiv] https://data-alliance.noaa.gov/

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.

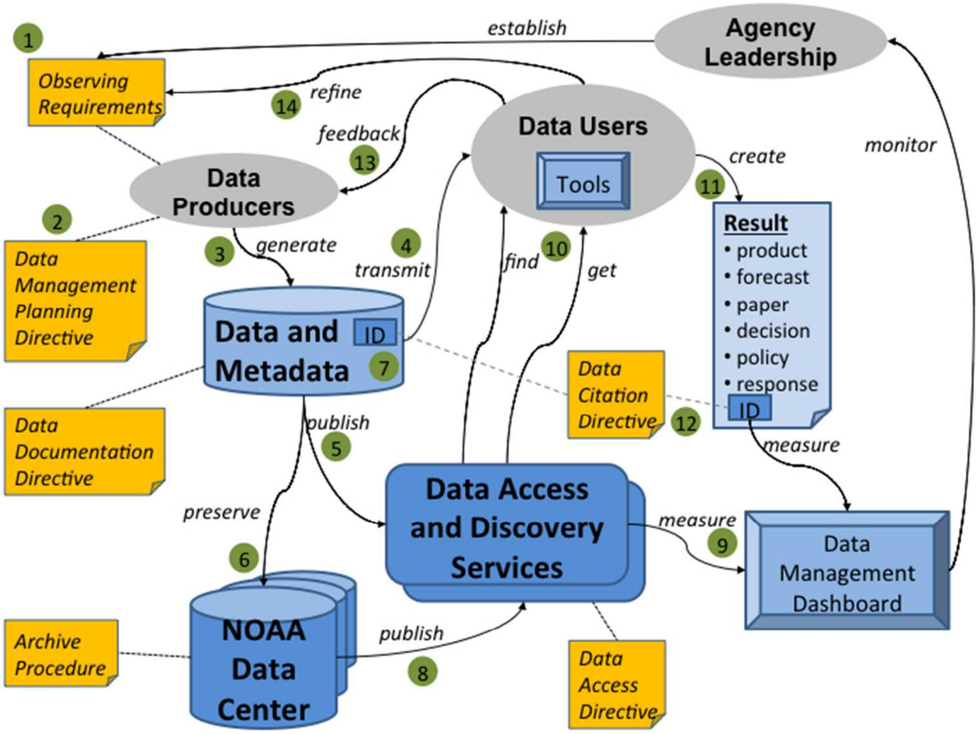
Figure 2: The Environmental Data Management Framework includes Principles, Governance, Resources, Standards, Architecture, and Assessment that apply broadly to many classes of data, and individual Data Lifecycles for particular data collections.

Figure 3: Activities in the Data Lifecycle, including some of the downstream support or feedback loops that connect various activities.

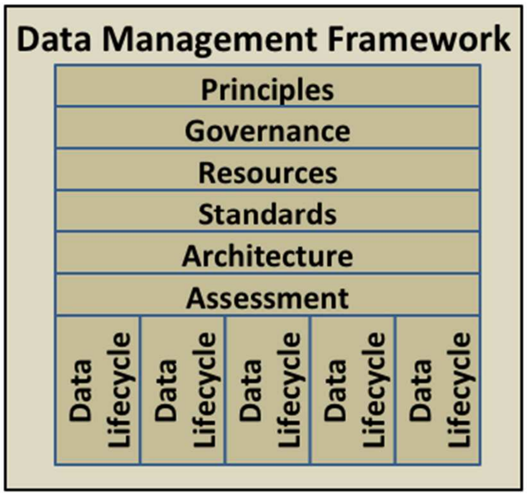
Figure 4: Schematic of shared-services architecture. Rather than explicitly linking individual data producers to specific customer applications, data management services and tools are generalized and decoupled as much as possible. Shared services can be established at an agency level (e.g., for data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

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.

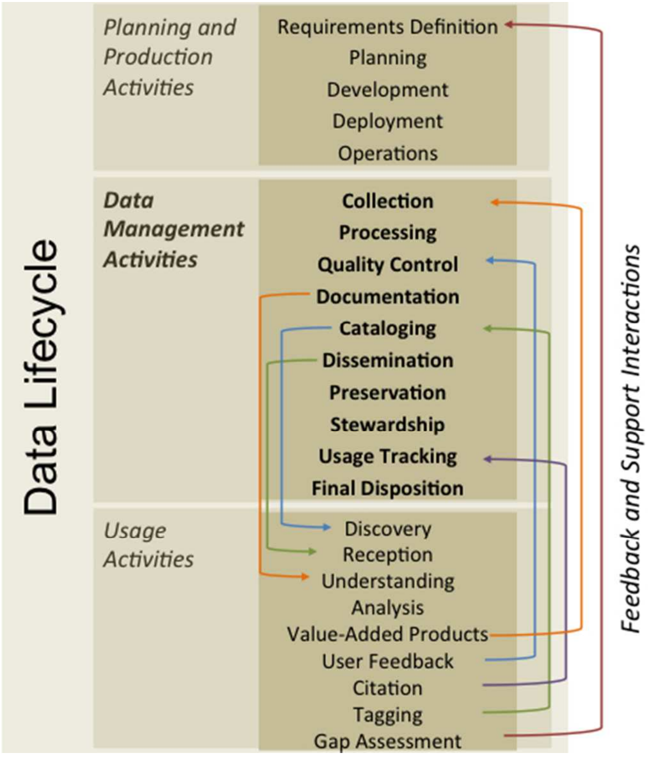
Figure 6: Conceptual architecture of the NOAA Big Data Partnership.



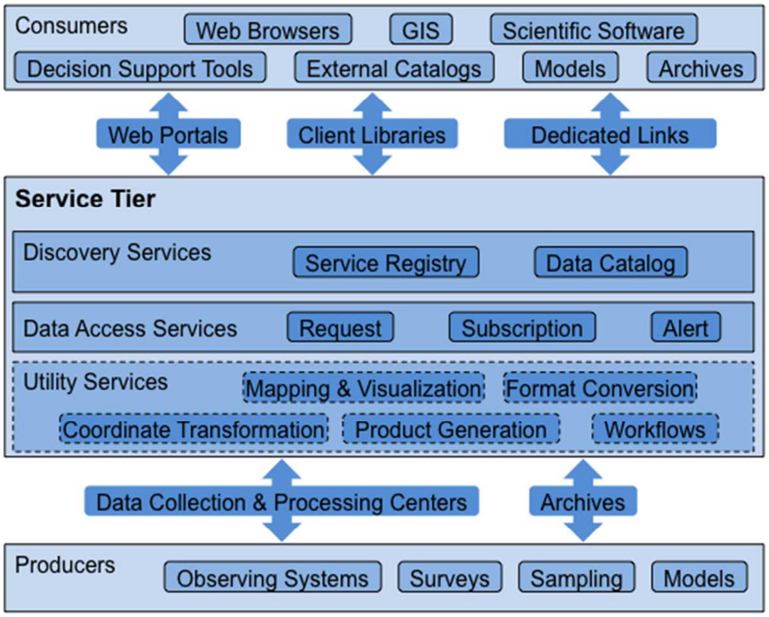
Conceptual overview of the desired target state of NOAA data management activities. Not all activities are illustrated. The numbers correspond to steps in the walk-through below.
254x190mm (72 x 72 DPI)



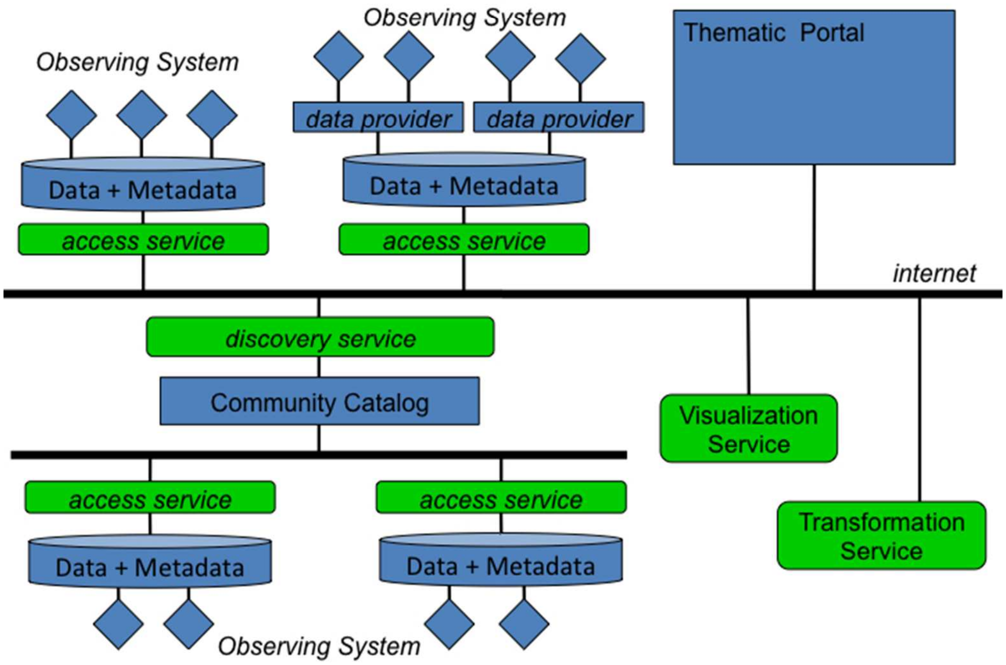
The Environmental Data Management Framework includes Principles, Governance, Resources, Standards, Architecture, and Assessment that apply broadly to many classes of data, and individual Data Lifecycles for particular data collections.
254x190mm (72 x 72 DPI)



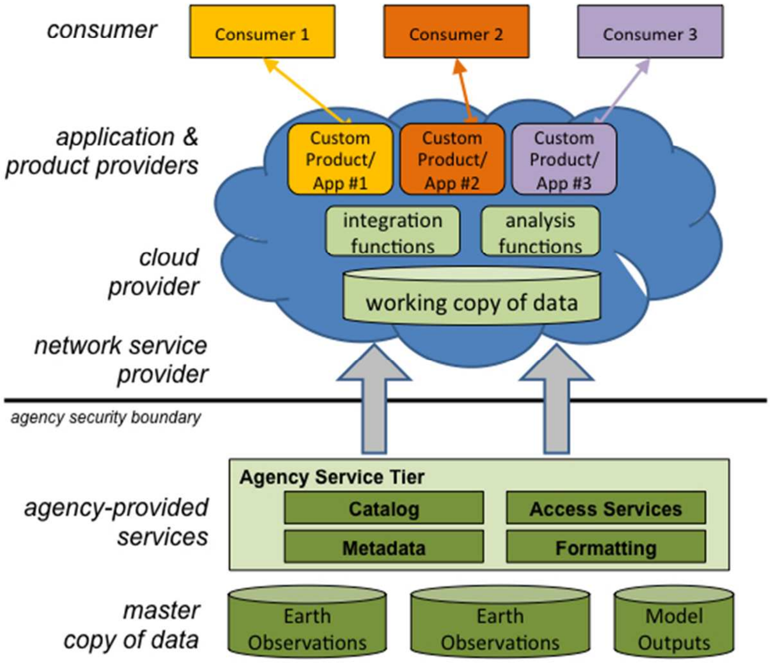
Activities in the Data Lifecycle, including some of the downstream support or feedback loops that connect various activities.
254x190mm (72 x 72 DPI)



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.
254x190mm (72 x 72 DPI)



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.
254x190mm (72 x 72 DPI)



Conceptual architecture of the NOAA Big Data Partnership
254x190mm (72 x 72 DPI)