NOAA FISHERIES
NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION

Fisheries EDM: Vision and Strategy

Version: 2.1.5
Date published: March 14, 2013
Presented by: The Fisheries Information Management Advisory Committee

Contents

Executive Summary

In 2008, the Fisheries Information Management Advisory Committee (FIMAC) produced Version 1.7 of its vision and strategy for Enterprise Data Management (EDM). In February of 2013, the FIMAC Coordination Team convened in Miami to update its plan. The result is described here. The new vision and strategy is expanded in two ways. First, it begins with a higher level perspective, which makes the advantages of enterprise data management obvious to potential adopters. Second, it proceeds to describe a set of concrete best practices. These best practices help insure that the direction of data management at NOAA Fisheries is well defined. They are produced by combining data user experiences with established EDM principles.

In addition, the new strategy describes a key role for the data management plan; as a device for collaboration between data management professionals, data users, and management.

Fisheries EDM: A Pragmatic Goal Based Vision

Data management is nothing new. It has existed as long as records have been kept. What is new is the scale of data collections and their relevance to organizations. Organizations are increasingly data driven in nature and the scale of data collections has expanded dramatically. According to a 2012 EMC Corporation study there are now 2.8 zeta bytes (10^{21} bytes) stored on the planet. This amount is projected to grow exponentially between 2013 and 2020. And yet, some data driven organizations may have difficulty exploiting the data they collect. The EMC study estimates that less than 1% of collected data has been analyzed. Of that small portion, not all has been analyzed effectively.

These facts have led to a renewed interest in data management and a redeveloped data management discipline called *Enterprise Data Management* (EDM). It is motivated by the realization that the effectiveness of a modern organization is largely determined by the manner in which it manages data.

NOAA Fisheries' response to the need for data management was the formation of an EDM program, starting in January 2008, proposed by the Director of the Office of Science & Technology and endorsed by the Science Board. The Fisheries Information Management Advisory Committee (FIMAC) was charged with the EDM program. The position of Information Architect was also established at that time as chair of the FIMAC. FIMAC advises the Leadership Council.

A November 2008 survey of managers by the FIMAC found that NOAA Fisheries had no authoritative data inventory, lacked sufficient metadata, experienced data quality and consistency problems, could not effectively integrate its data to answer key questions of the day, had administrative systems that did not work well together, and had delays in data delivery to decision makers.

Since then, FIMAC has developed the Data and Information Management Policy Directive and the Data Documentation Procedural Directive; promoted data documentation at Science Centers, Regional Offices, and Headquarters; directed the development of the InPort metadata system; established Information Data Coordinators at Science Centers, Regional Offices, and Headquarters; supported the development of documentation plans throughout NOAA Fisheries; established an overall architecture for making data available through a variety of portals; and established the EDM Wiki for use throughout NOAA.

In February 2013 a new full time Information Architect was hired and the Coordination Team of FIMAC convened to begin the process of updating the existing EDM vision and strategy. An important FIMAC goal at that time was communication of EDM benefits to the greater NOAA Fisheries organization. This document is the result of that meeting.

Communicating the connection between organizational effectiveness and EDM can be challenging, especially when expressed in the unfamiliar language of EDM. Data users and data stewards might also be skeptical of anything that might be perceived as overhead. For these reasons, the current vision begins by enumerating universally accepted *effectiveness goals* and then connects those goals to corresponding EDM best practices for NOAA Fisheries. These goals are:

Effectiveness Goals

- **Capability**
- **Speed**
- **Efficiency**
- **Accuracy**
- **Cost reduction**
- **Confidence**

Capability refers to the ability to perform a particular analysis. The ability to perform such an analysis depends on data management practices and cannot be assumed. *Speed* is included because data requestors want or need results as fast as possible. *Efficiency* means doing a task with fewer resources, or doing more with the same resources. Fishery analyses have value only if they are *accurate*. *Cost reduction* is closely tied to efficiency but deserves special emphasis in times of shrinking budgets. NOAA, industry, and taxpayers must all have *confidence* in our analytical results.

Subsequent sections will describe the context of effectiveness goals, the scope of NOAA Fisheries EDM, the role of data management plans, directives, best practices, stakeholder roles, and related groups.

Effectiveness Goals are Only Meaningful in a Broad Context

The relevance of a particular data set typically transcends the boundaries between individuals, sections, and divisions. Therefore, when individuals or sections consider whether their practices advance some effectiveness goal, it should be judged with respect to the goals of the whole organization, that is, NOAA or NOAA Fisheries. It should not be measured only with respect to the individual, section, or division executing the practice. Positive effects that one group has upon another when executing good EDM practices must be taken into account. Too often data management practices, their effects, and credit given, are administered within inappropriately narrow contexts. Broadening the context in which goals are defined is especially important when managing data access.

Scope of NOAA Fisheries EDM

Figure 1 shows how data flows from its origins to useful results. It is clear that data collections should be included in the scope. It is equally important to include the observational process. This is because the manner in which data is collected usually determines what analyses can be performed on it. The analytic process should also be included, since not all analytic processes are equally accurate, fast, etc.

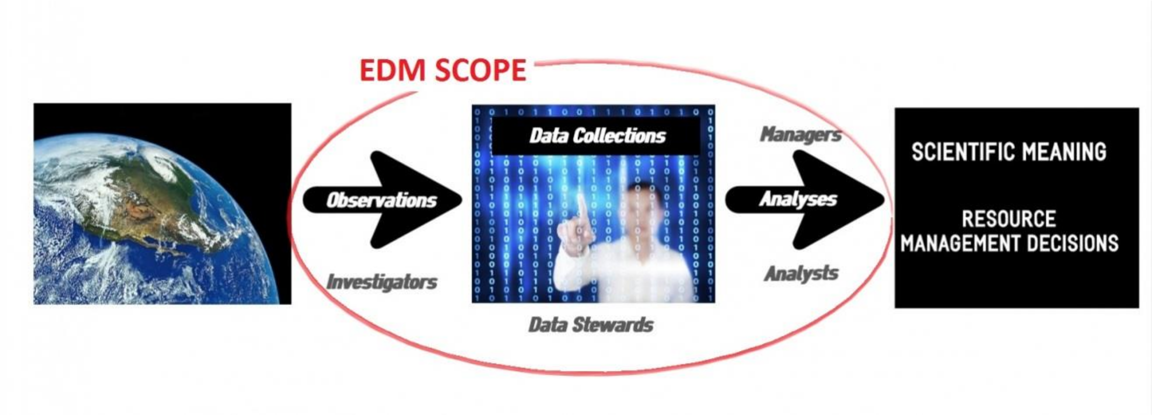

Figure 1. Fisheries data flow.

Data Management Plans

The primary vehicle for realizing EDM in an organization such as NOAA Fisheries is the data management plan (DMP). The DMP provides an opportunity for EDM advisors, such as the FIMAC, to recommend best practices in the form of example plans and templates. It also engages data users and stewards and affords them the opportunity to customize DMPs according to their special circumstances. It provides managers a framework for planning and increasing the effectiveness of their group. Finally, it defines metrics for measuring effective processes.

A DMP defines best practices in a number of data management areas and encourages users to measure the impact on effectiveness goals.

Directives

Guidance for preparing a DMP is provided by procedural directives. The procedural directives in turn define the implementation of policy directives. Procedural and policy directives exist at both the NOAA and NOAA Fisheries levels. Fisheries directives should build upon and be consistent with NOAA directives. White House Memos, circulars, etc. also provide guidance. The current status of these documents is shown below. For each directive, select topics, relevant to this document, are listed.

NOAA Policy Directive

- *NOA 212-15: Management of Environmental Data and Information*

Fisheries Policy Directive

- *NMFS Data and Information Management Policy Directive*

NOAA Procedural Directives

- *Data Management Planning Procedural Directive*
    - Documentation

    - Metadata
    - Quality assurance / Quality control
    - Data protection
    - Data archiving
    - Data Access and sharing
    - Data discovery
- *Data Documentation Procedural Directive*
    - Documentation
    - Metrics
    - Metadata
    - Discovery
    - Portals
- *Grants Data Sharing Procedural Directive*
    - Access and sharing

Fisheries Procedural Directives

- *NMFS Data Documentation Procedural Directive*
    - Documentation
- *NMFS Data Management Plan Procedural Directive* (under development)

White House Memos, Circulars, etc.

- *OMB Circular A-130*
    - Planning
    - Access
    - Integration
    - Documentation
    - Authorized access
    - Reduced cost
    - Increased quality (accuracy)
    - Speed
- *White House Memo Open Government Directive (M-10-06)*
    - Access
- *White House Memo Open Government Directive – Framework for the Quality of Federal Spending Information*
    - Access
    - Data quality

- *White House Memo Increasing Access to the Results of Federally Funded Scientific Research*
    - Sharing
    - Access
    - Data management planning
    - Preservation

Best Practices

These best practices, which were derived from data user experiences in Fisheries will contribute to implementing the guidance documents listed above. As more data user experiences are collected, the list will grow. Data management plans should specify best practices in the following areas:

Best Practices

- **Data access**
- **Documentation**
- **Integrated systems**
- **Ergonomic data collection**
- **Reduced reporting burden**
- **Quality assurance and quality control**
- **State of the art at sea communications**
- **Consistent definitions**
- **Streamlining**
- **Master data stores**
- **Change tracking**
- **Separation of obsolete and working data sets**
- **Standardized use of data types**
- **Clean table design and naming conventions**
- **Same site storage and computation services for bulk data**
- **Automation of repetitive analyses**

Data access

Goals affected: capability, speed, confidence, efficiency

Maximize Authorization

Inaccessible data has no value. Therefore, data management plans should specify what data access is available to authorized users. Authorization may occur at multiple levels. Access to trusted NOAA analysts and managers should be the default unless clear justification for denial can be provided. Denied users should have a NOAA Fisheries level appeal process available to them. Arbitrary denial of access greatly reduces the capability to perform needed analyses.

Simple solutions are available to make data available while also insuring security. Data can be protected from loss by providing read only access to certain users. Also, confidential data can be aggregated prior to being made public. Human error is always a possibility, but this alone is not a justification for making data inaccessible.

Discoverable data

Often considered separately from access, data discoverability is a data accessibility property that impacts speed goals. Data should be easy to find though the use of metadata, controlled vocabularies, master data stores, portals, and effective search engines. Metadata, or data that describes data, is the basis for efficient search. It includes documentation, keywords, and tags. Controlled vocabularies avoid the problem of searching for one of many possible synonyms that is not in the metadata. Thesauri can also be helpful with synonym problems. Master data stores and data portals can provide a single starting point for data searches, avoiding the need to explore many repositories.

Protection from loss

This is also often considered separate from access. However, lost data is not accessible, so the purpose of data protection is to provide continued access. Data should be protected by following established standards for secure system access and data archiving.

Documentation

Goals affected: capability, speed, accuracy, cost, efficiency, confidence

The role of documentation is to assign meaning to data. Undocumented data requires users to interview a number of potentially knowledgeable individuals regarding the meaning of a dataset. The results of such inquiries are usually varied, with different answers given to the same questions. This interview process costs time whenever a new data user encounters the dataset. Any time saved by not documenting is lost repeatedly during the life of the data. Undocumented data introduces uncertainty, reduced accuracy, the potential for misinterpretation and misuse. Additional costs are incurred in correcting erroneous analyses and the potential mismanagement of natural resources.

Guidance for documentation is provided in NOAA and NOAA Fisheries Procedural Directives. Documentation and metadata for NOAA fisheries data is now available in *InPort*, a joint effort of FIMAC, Fisheries Information Systems, and the Office of Science & Technology.

Integrated systems

Goals affected: capability, accuracy, speed, cost, efficiency, confidence

Comparison of un-integrated and integrated systems

Many important analyses depend upon the fusion of multiple data collections. For example, data from fishing trips is collected by multiple reporting systems, which populate multiple databases. A complete analysis of fishing trips requires the combination of these databases. Data from a single trip event is collected by a vessel reporting system, a dealer reporting system, a vessel monitoring system (VMS), and the NOAA Fisheries Observer Program and other assorted programs.

Numerous and costly problems arise if such systems are not integrated. In particular, if the systems each use a different identifier for a trip, then forming a complete description of the trip from all systems involves guesswork in the form of fuzzy matching. Such matching algorithms are expensive to program and prone to generate an unknown number of errors. In fact, there is no way to measure the error rate because there is no ground truth to compare against.

Another problem is the use of trip identifiers at different levels. For example, if a trip identifier in one system is actually a sub-trip identifier in another system, matching is again costly and error prone.

A third problem with un-integrated systems is manual and redundant data entry. Suppose that the vessel system and the dealer system both report vessel permit number, dealer permit number, and trip identifier. Each entry takes time and is an opportunity for error. Each error then propagates further errors in the form of mismatches.

An integrated reporting system addresses all these problems. In such a system, all data associated with a single trip goes to a single electronic report. This reduces the risk of data entry error, reduces redundant effort, takes less time, and eliminates matching processes since the data is matched upon entry. Such systems can also be designed to auto-populate certain variable fields, further reducing errors and workload.

Design order

Data systems are often organized in a hierarchy of super-systems and subsystems. In the above example, the vessel, dealer, VMS, and observer systems are all subsystems within a larger super-system. There are two possible approaches to design. One is to design and implement the subsystems first and the other is to design the super-system first.

In the first approach, the resulting subsystems are created with no regard of how they will interact. The result is the un-integrated super-system as in the above example, with all the associated drawbacks. Once this situation develops it is costly to redesign and re-implement. By designing the super-system first, even if not actually implementing it, such costs can be avoided. It is sometimes thought that the design of super-systems is costly. However, designing need not be costly, but lack of planning almost always is.

Controlled vocabularies

Joining related data tables requires the use of consistent terminology and codes across tables and databases.

Ergonomic data entry

Goals affected: speed, efficiency, accuracy, cost, confidence

The design of systems that collect human entered data has a major impact. Keeping invalid data from entering a data system is relatively simple compared to the task of detecting and correcting errors once invalid data propagates throughout a system. Variables typically have well defined valid formats and ranges. When data is entered these constraints can be checked by the data entry system and users can be provided real time feedback and the opportunity to enter a corrected value. Data entries can also be compared with other entries on the same form for consistency.

Providing choices when possible is superior to free form data entry. Paper data entry should be avoided if at all possible. Paper data entry introduces the problem of legibility and also introduces an additional opportunity for errors during transcription.

Some data stewards would rather accept faulty data than reject errors. The best way to accommodate this approach is to 1) Provide feedback as described above. 2) If the user is persistent in entering an invalid value, accept the value but flag it as invalid or separate it into a quarantine table for analysis by quality control (QC) personnel.

The user interface in any data entry system should be intuitive. Ideally it does not require any documentation in order to operate, although such documentation should still be provided.

Reduced reporting burden

Goals affected: efficiency, accuracy, cost, confidence

Industry partners have a finite ability to report during fishing activity. In order to improve the quality of reported data, care should be taken in choosing what reporting is required. Two areas where reporting can be reduced are variables that are not used by analysts and redundant reporting.

Quality assurance and quality control

Goals affected: speed, efficiency, accuracy, cost, confidence

Quality assurance (QA) is the discipline concerned with error prevention. Examples have already been provided in sections on integration and ergonomic data entry. Quality control (QC) is the discipline of detecting and correcting existing errors. The error prevention approach is preferred over error correction. Nevertheless, errors will inevitably enter a data system making QC necessary.

QC can be carried out in one of two ways. It can be accomplished in an improvised manner, as errors are identified by alert analysts, or it can be done systematically. The systematic approach is far superior. In the systematic approach, all data in the system are checked automatically for two primary error types, invalid data and mismatched data. Invalid data can be detected on the basis of format, out of range, or being null. Mismatched data appear as inconsistencies or as orphans, where a record appears to have no match in a related data set. It is most important to detect errors in source data because source data errors propagate into derived data sets.

The systematic approach is also superior because it provides an opportunity to detect the major causes of errors and enables prioritization of QC efforts. For example, if errors involving fishing trip orphans are aggregated according to (vessel, dealer) pairs, and then ranked according to the number of errors, it will be easy to see which (vessel, dealer) pairs have the greatest difficulty working together. By addressing high priority errors first, and identifying their causes, great efficiencies can be gained and future errors prevented.

Data management plans should specify that all detected errors are to be fixed if at all possible.

State of the art at sea communications

Goals affected: capability, speed

High bandwidth at sea communications coupled with web based reporting can serve as the basis for real time reporting. Currently, at sea bandwidth is very low for some systems and alternative communication streams have reporting lags that occupy weeks. Since fishing quotas can be exceeded in the course of days, real time reporting provides superior capability for resource management.

Consistent definitions

Goals affected: efficiency, accuracy, confidence

The concept of a fishing trip seems simple and intuitive at first glance. Legally, it is a vessel's journey from castoff to re-docking. However, from a VMS perspective it is the journey from an offshore line of demarcation crossing to the re-crossing. Further complicating the situation is the fact that vessels may weave across a line of demarcation repeatedly during a trip.

Trip definition is sometimes associated with unloading of catch. In the ideal case, a vessel unloads at the end of a trip. In reality, some vessels may attempt to perform multiple offloads during a trip, returning to sea between offloads. Is this one trip or many?

Additional complications involve vessels which return to dock or cross demarcation lines due to weather or mechanical problems.

Current attempts to reconcile these different trip versions are sub-optimal. In the case of vessels returning to port more than once, additional complex reporting requirements are used. In the case of multiple demarcation crossings, heuristic algorithms are sometimes used to combine the many trips generated by VMS systems. Such heuristics may generate errors as well as fix them.

Since many regulations refer to a trip, a simple and consistent definition would be of great benefit to law enforcement and analysts.

Streamlining

Goals affected: confidence, accuracy

Over time, data processing streams accumulate incremental changes. Typically these are not the result of any overall plan. As a result, the system can become unnecessarily complex and convoluted. So complicated are some of these systems that no one person knows or understands them. Errors are difficult to detect and correct. Such data flows should be reengineered to provide the simplest possible data flow which achieves the desired result.

Master data stores

Goals affected: confidence, efficiency

A master data store is a single trusted source of data. Data in the store is carefully QCed, has appropriate metadata and documentation. It also has streamlined and mapped processing paths for its derived data. Master data stores provide a number of advantages. They make search easier by providing a single starting point for search. They also provide consistency and confidence for data consumers. The alternative is multiple collections of similar data. These collections tend to diverge based on differences in QC methods and other processing algorithms. When analytic results emerge from two or more such sources they will not be consistent, raising questions and reducing confidence among data consumers.

Change tracking

Goals affected: confidence

Data is often updated to reflect QC activity or in light of new information. As in the case of multiple data stores, data consumers will want explanations. For this reason, all changes to data records should be made easy to track.

Separation of obsolete and working data sets

Goals affected: speed, accuracy, efficiency, cost, confidence

Obsolete and unmaintained data tables should not be stored with working data tables. Obsolete tables should be archived separately so that analysts do not mistakenly use them, producing inaccurate results. Obsolete tables also waste analyst's time as they attempt to determine which table is reliable. Also, obsolete columns in tables should be removed from working tables. Any such changes should be documented and managed by a change management board.

Standardized use of data types

Goals affected: speed, accuracy, efficiency, cost, confidence

There are many ways to represent a dataset. Dates provide a good example. There are numeric formats, alpha numeric formats, and all sorts of orderings of day, month and year. Hours, minutes and seconds may or may not be included. A date may be stored in multiple columns, one for day, one for month, and one for year. These forms do not all react the same to processing and data queries. Results can be difficult to predict. Analysts often spend considerable time determining the effects of these differences and errors arise. Therefore, preferred date formats should be agreed upon whenever practical and exceptions justified.

Serial numbers also see some degree of representational variance. Sometimes they are stored as numbers and sometimes as strings.

Clean table design

Goals affected: speed, efficiency, accuracy, cost, confidence

Database table columns should serve a single purpose and store a single variable. Table designers occasionally attempt to store more than n things in n columns. For example, suppose a table contains a column for *start date* and another for *end date*. So far these columns represent two variables. The designer may then decide to encode some meaning in a reversal of the start and end dates, where the interval ends before it begins. Such an encoding, while clever, is likely to introduce errors when analysts assume the natural order of the time interval. Kludgy encodings are unnecessary and should be avoided.

Tables and columns should have consistent and easy to understand names.

Same site storage and computation services for bulk data

Goals affected: capability, efficiency, confidence

The size of some data sets can present challenges. For example, acoustic data is very voluminous and is not readily transferred over networks. Data are often stored on media in investigators offices. This limits opportunities for collaboration between investigators and puts the data at risk of loss due to non-standard archiving. One solution is to store the data on a server that also hosts processing capability. In this way, multiple investigators can process and collaborate using a shared storage/processing service.

Automation of repetitive analyses

Goals affected: efficiency, cost, accuracy

Analysts' time is expensive, so it should not be spent on repetitive tasks that can be automated. A good data management plan budgets time and resources for automation. Such an investment returns time savings each time the automated process runs. Accuracy improvements are also expected since a debugged automated process is less error prone than a manual process.

Stakeholder Roles

Users and Data Stewards

Users and data stewards (hereafter called users) are the primary drivers of the EDM process. They are the ones who know the limitations of existing systems and often conceive of necessary improvements. A special subset of data users is the *Catalysts*. These individuals are compelled to view data systems not as they are but as they should be. They are the FIMAC's primary source of ideas.

Another special group is the Information Management Coordinators (IMC). These individuals are the contact points between users and the FIMAC and they serve as FIMAC members. Each regional office, science center, and headquarters office is represented by an IMC.

FIMAC

The FIMAC has several roles. Firstly, it advises management in the form of the Leadership Council and the Fisheries Science Board. It develops policy in the form of procedural directives. It supports the user community and their management in a consulting role, in the same way that professional consulting firms help businesses become more effective. The FIMAC identifies Catalysts and catalogs their ideas. The FIMAC distributes ideas from Catalysts as well as proven solutions to other regions, maximizing the impact of innovations. The FIMAC coordinates EDM efforts in EDM related organizations (See Related Groups below). Finally, FIMAC matches identified Fisheries data management needs to corresponding EDM practices and solutions.

In the consulting role, FIMAC assists users and managers to refine best practices, incorporate them into DMPs, and measure the effects upon Fisheries effectiveness goals.

The FIMAC is composed of individuals from across Fisheries and it collaborates using its Wiki and other web based collaboration tools.

Fisheries Information System Program (FIS)

FIS implements and develops EDM systems consistent with policies recommended by FIMAC. Major examples include the InPort metadata system and Fisheries One Stop Shop (FOSS). FIMAC and FIS members populate one another's many committees and working groups.

Management

Collaboration with management at all levels is critical for realizing the benefits of EDM. Managers who wish to improve the data related operation of their organizations can call upon

FIMAC for support. Implementing the abundance of EDM practices will take time and managers should set priorities with regards to which goals to pursue first. Setting EDM related expectations in performance plans is also an important role for managers.

Related Groups

Table 1 lists a number of EDM related groups within NOAA.

Table 1

Group	Mission	Relation to FIMAC
Environmental Data Management Committee (EDMC)	Develops EDM policy NOAA-wide	FIMAC develops policy consistent with EDMC. FIMAC provides EDMC with Fisheries performance metric data. Fisheries Information Architect serves on the EDMC.
NOAA Observing Systems Council (NOSC)	Advises Under Secretary of Commerce for Oceans and Atmosphere (NOAA Administrator) on matters of Earth observation systems.	FIMAC and NOSC collaborate in defining Observing Systems of Record
National Oceanographic Data Center (NODC)	Provides scientific stewardship of marine data.	FIMAC will help organize the archiving of critical Fisheries data sets at NODC.
Integrated Ocean Observing System (IOOS)	Delivers ocean related data to decision makers to improve safety, enhance the economy, and protect the environment.	FIMAC will help IOOS gain access to Fisheries data collections.

Appendix I: Glossary of Terms

Controlled Vocabulary – A set of preferred terms. Controlled vocabularies are useful in system integration and search.

DMP – Data Management Plan

EDM – Enterprise Data Management, a set of modern data management practices

FIS – Fisheries Information Systems

IMC – Information Management Coordinator. Represents the FIMAC at each regional office, science center and at headquarters.

Master Data Store – A single trusted source of data for a given domain.

Metadata – Data about data, such as documentation, keywords, flow diagrams, etc.

QA - Quality Assurance. The discipline of error prevention.

QC – Quality Control. The discipline of error correction.

VMS – Vessel Monitoring System