Regional Metadata Tool Recommendations

Authored by:
Steve Rentmeester
Environmental Data Services
Portland, OR
environmentaldataservices@gmail.com

Submitted to: Pacific Northwest Aquatic Monitoring Partnership Coordination Staff

May 2, 2011

Contract #: G110PX90146

Background

Implementing a regional-scale strategy for creating and making available metadata associated with regionally relevant monitoring data represents a significant challenge. Executive Order 12906 that established the Federal Geographic Data Committee from which early metadata standards originated and the U.S. Geological Survey's National Biological Information Infrastructure (NBII) that supports the development of biological metadata have been in existence since the early 1990s and yet, for many organizations, including federal agencies, metadata creation is not a consistent or standard business practice. Changing business practices around the creation and maintenance of metadata will require funding agencies to support and enforce metadata requirements as integral to data deliverables and will necessitate organizational level mandates, dedicated staffing, new tools, and changing community norms around metadata creation. While the challenge is significant, one simple recommendation clearly stands out and that this is to **get started now**.

The PNAMP Metadata Task Group has been working to promote metadata development within monitoring programs in the region and to advance regional implementation of standard metadata reporting as an integrated component of standard business practice. Metadata creation is a technical task that is detailed in nature and requires a dedicated time commitment. The challenge is compounded by variation in monitoring objectives, study designs, and methodologies. Some have proposed solving this dilemma by decreasing the required information content of existing metadata standards. However, leaders in information management have consistently recognized the need for greater detail in metadata to support data discovery, description, and proper use. This common perception has led to an adage among information managers and technical experts – "Minimal metadata is minimally useful". An alternative approach to reducing the work load in completing metadata is to develop tools that reuse and supplement information being stored in regional project proposal and contract management systems.

To promote the development and use of metadata, PNAMP contracted Environmental Data Services to facilitate a series of work sessions with the PNAMP Metadata Task Group with the goal of evaluating potential metadata tools. Two work sessions were held between December, 2010 and February, 2011. During the work sessions six proposed tools were reviewed and ideas for additional approaches were requested. Additional input was gathered though an on-line survey[1]; one-on-one conversations with technical experts from NBII, EPA, and Pacific States Marine Fisheries Commission (PSMFC); and through discussion with the PNAMP Steering Committee. The following document advances three recommendations and explains the rationale for other proposed tools considered but not advanced as recommendations.

Organizational Benefits of Metadata

Metadata provide significant benefits to both the organizations that collect data and to those who subsequently use the data. For organizations that collect data, metadata help enhance the quality, usability

[1] The results of the online survey are archived in the document PNAMP_Metadata Tool Survey Results.pdf at http://www.pnamp.org/document/3425.

and value of data for internal and external users. Additionally, metadata supports broader monitoring objectives and can:

- help avoid duplication of monitoring activities,
- foster sharing of data resources,
- help ensure data are interpreted and used appropriately,
- preserve institutional memory,
- publicize research, and
- reduce workload of compiling data for regional analyses.

Organizations are strongly encouraged to begin metadata documentation during the earliest stages of project planning and to view metadata creation as integral to their workflow and to their data products.

Metadata that describe regional monitoring activities (the locations, study designs, methodologies, and organizations) is critical information to support data analysis and is necessary to inform regional funding decisions. Given the multitude of monitoring programs and organizations within the Pacific Northwest, documenting their activities and associated data products is essential for advancing coordination efforts. This descriptive information documents the operation expenditures of regional funding entities. Each program collects and stores monitoring data in unique ways aimed at meeting program-specific objectives. Metadata can help track monitoring activities and describe data products across the multitude of monitoring programs and therefore metadata creation is a core business need for regional monitoring entities.

Metadata are "data about data"

Metadata are simply data used to describe other data. They are a description of the content, quality, lineage, condition, and other characteristics of data. For many people, the first exposure to metadata is with data in a Geographic Information System (GIS). However, metadata are critical for any dataset so that the data can be discovered, understood, used, and archived properly. Metadata records are similar in concept to library catalog records: details about a book such as title, author, and publisher are recorded in a standard way to ease the search for information. Like a library catalog, metadata are organized in a standardized format using a common set of terms. Each piece of information in a metadata record is referred to as a metadata element. Standardization facilitates searching and discovery of data.

Metadata Tool Recommendations

During the past 10 years, several organizations in the Pacific Northwest (Science Applications International Corporation, Northwest Environmental Data Network, and PNAMP) have proposed data management strategies. Each of these strategies emphasized the importance of metadata and the need for shared technical infrastructure to support creation, discovery, and sharing of metadata. This call for shared infrastructure is driven by two core business needs that are common to all regional monitoring programs:

1. to dynamically report the who, when, why, and how of monitoring activities
2. to document monitoring data with FGDC compliant metadata

The ability to dynamically summarize current and historic monitoring activities is essential for planning future monitoring activities and for reporting activity to-date. Documenting monitoring data with FGDC compliant metadata helps ensure that data are understood and properly used in analysis. These two core business needs have consistently been identified by regional coordination programs. Developing regional-

scale approaches to metadata creation will alleviate a significant burden for data collection organizations, allowing those organizations to focus on ensuring data quality, deriving metrics, and sharing data. Current efforts within the region to develop data exchange templates could leverage benefits of a regional approach to metadata management by creating linkages between data shared in exchange template format and regional metadata systems.

While metadata affords many benefits to both the creators and users of monitoring data, the burden of metadata creation cannot be overlooked. Metadata creation is a tedious task that typically requires specialized training. Experts in the field (including members of the PNAMP Data Management Leadership Team) recommend that metadata creation be led by data stewards who have inter-disciplinary training in both biologic sciences and information management. Data stewards would provide significant support to metadata creation efforts within data collection entities.

Regional funding entities and monitoring organizations should anticipate and plan for the cost of metadata creation. Metadata creation and distribution will require specialize staff and training; regional metadata repositories for archiving and sharing metadata documents; and policy commitments at the organization-level. Fortunately, several good metadata creation tools already exist (Rentmeester, 2010), development of regional repositories is already underway, training opportunities are available, and a few organizations have begun implementing metadata policies. This document provides three key recommendations for advancing regional metadata creation and distribution efforts:

1. **Secure funding to hire data stewards to support metadata creation;**
2. **Implement a Pacific Northwest node within the NBII clearinghouse; and**
3. **Build a web-based Monitoring Activity Inventory Tool.**

Recommendation: Secure funding to hire data stewards to support metadata creation

Description:
In order for the region to have access to metadata about aquatic monitoring data, those metadata must be created. During meetings of PNAMP Metadata Workgroup, a consistent message has been re-iterated and affirmed by participants: The primary factor limiting metadata creation is availability of staff with time and appropriate skills. Regional funding entities have begun implementing contract requirements for metadata. These new requirements must be supported with funding for data stewards who can provide metadata training and support to data collection organizations. Regional data steward responsibilities should be modeled around the NBII Metadata Program. Data steward responsibilities should include:

- providing metadata training to natural resources staff,
- assisting natural resources staff in use of metadata creation tools,
- assisting organizations in coordinating metadata creation efforts,
- describing the benefits of metadata to organizational managers, and
- assisting organizations in identifying additional resources for metadata creation.

The goal of hiring data stewards should be to build capacity for metadata creation through program development within data collection organizations. Metadata creation is a long-term need that must be met through capacity building. Data stewards would distribute tools and training material, provide in-person training sessions, work one-on-one with biologists to create metadata records, and work with organizational managers to plan and implement metadata initiatives. The PNAMP Metadata Workgroup recommends a phased approach to implementing metadata creation (Rentmeester, 2010).

Phase 1: Create full metadata for future datasets
Phase 2: Create inventory-level metadata for existing datasets
Phase 3: Use inventories to prioritize existing datasets
Phase 4: Create full metadata for priority datasets

Pros:

- builds capacity within agencies
- decreases gap between need and capacity
- trains specialized staff with necessary skills to ensure quality and efficiency
- serves as pilot to increase understanding of the scope of need for metadata
- could build on existing Coordinated Assessments effort

Cons:

- PNAMP can provide recommendation, but may not have direct role in implementation
- does not streamline process through decreased detail or through increased automation

Requirements:

- An organization to create new positions and administer staff
- Funding for new positions
- PNAMP staff time to write National Spatial Data Infrastructure (NSDI) Cooperative Agreement Program (CAP) grant
- Well defined objectives and position descriptions

EDS Recommendation:

Secure funding to hire two metadata stewards through existing infrastructure at Pacific State Marine Fisheries Commission, StreamNet project. Define the objective for these new staff as advancing the existing Coordinated Assessment effort through training and technical support with the goal of incorporating data exchange templates (DET) into existing business practices at data collection agencies and supporting staff biologists in updating the associated metadata. Additionally, task the PNAMP Data Management Liaison with submitting a grant application for the NSDI Cooperative Agreement Program (CAP). The CAP grant proposal should request staff time to assist in providing training to the new metadata stewards (e.g. Train the Trainers course), providing and distributing existing training materials to internal agency staff, installing and configuring existing metadata creation tools on agency computers, defining agency strategies for metadata creation, and establishing crosswalks between DET and FGDC elements. PNAMP Data Management Liaison should review the "Submitting an NSDI CAP Proposal" document.
http://www.fgdc.gov/training/nsdi-training-program/materials/CAP_How2Submit_20101020.pdf

Recommendation: Implement a Pacific Northwest node within the NBII clearinghouse

Description:

This recommendation is to implement a Pacific Northwest Node through existing infrastructure at the USGS National Biological Information Infrastructure (NBII) Metadata Clearinghouse. A node (or portal) would be inexpensive to establish and would provide central storage and distribution of regional metadata. A PNW Metadata Node would provide a single web location to disseminate metadata and would ensure that those metadata are discoverable and searchable by the community. This node could also serve as a resource for distribution of existing tools and training materials.

The PNW Metadata Clearinghouse could be built as a node within existing infrastructure at the NBII. Aquatic monitoring datasets could be discovered through the web-based clearinghouse using submitted metadata. Users would be able to search based on geography, time frame, keywords or full text, protocol category, fish population, watershed, or a known location. Search results would provide access to the full metadata record for a given dataset. Instructions for accessing the full dataset would be included as a field within the metadata record. If the dataset were available electronically, then the metadata record would include a hyper link to the dataset itself.

In a recent survey of PNAMP participants two-thirds indicated that this tool had moderate or high potential benefit to the region and to their individual organization.

The Northwest Environmental Data Network (NED) portal development effort was a pilot program that developed and tested a regional data discovery portal (http://gis.bpa.gov/NPCC/default.htm). The PNW Metadata Clearinghouse proposal should build on lessons learned from the NED Portal effort and would leverage resources at the NBII (NED, 2006).

Pros:

- Relatively inexpensive
- Utilizes an existing metadata repository
- Provides central storage and accessibility for regional metadata
- Can be used along with other metadata creation tools
- Does not require a new centralized managing entity

Cons:

- Does not address metadata creation
- Useful only if metadata files have been created through use of other tools, by biologists, or by metadata stewards
- Does not support dynamic combination of metadata records or a subset of elements for compiled datasets
- Does not provide for centralization of metadata creation within the region
- Care must be taken to prevent proliferation of different version of metadata
- Dependent upon the business processes of each data collecting entity

Requirements:

- A clearly articulated mission statement must be written to define scope of the node
- Development time from NBII Staff

- PNAMP staff time to coordinate and manage project
- Well established practice of metadata creation within data collection agencies

EDS Recommendation:
While using a PNW node on the NBII clearinghouse is a relatively inexpensive task and has long been discussed as a need, EDS cautions against implementing the node prematurely. The caution is driven based on the risk of creating unrealized expectations. If decision makers or analyst visit the node and there is limited content available, there is risk they will walk away frustrated and may not return. EDS recommends prioritizing metadata creation and establishing it as standard business practice before developing a highly public interface for displaying that content.

Recommendation: Build a web-based Monitoring Activity Inventory Tool

Description:
The Monitoring Activity Inventory Tool would be a web-enabled GIS that supports users in recording the spatial location of monitoring activities or the area of inference for data analysis activities. Additionally, the tool would support the tracking and reporting of monitoring activities. Location information would be stored as a latitude and longitude and also as a location associated with a stream network. Location information could be imported from a text file or could be drawn on a map in a web browser. The spatial information would be managed in a central database and could be shared or exported as a set of metadata elements. Metadata elements exported from this tool could be supplemented with additional metadata to form a complete metadata record. This tool would include linkages with MonitoringMethods.org and with existing contract management systems. The tool could be built to support a component of the Generalized Random Tessellation Stratified (GRTS) site evaluation process (http://www.pnamp.org/project/3263). The Monitoring Activity Inventory Tool would be managed as regional database system.

Tracking the history of monitoring activities will support a range of business needs. The ability to dynamically summarize current and historic monitoring activities is essential for coordination of future monitoring activities and for reporting activity to-date. Similarly, selection of sites for inclusion in a monitoring program may be dependent on prior monitoring activities at the site. Compilation of data to address specific monitoring questions requires an understanding of monitoring history across the basin. This tool will support these business needs by providing a single regional system of record for tracking the history of monitoring activities at individual sites throughout the region.

Aquatic monitoring data is collected for a location on the surface of the Earth and as such, is inherently spatial. Analysis of monitoring data to support reporting on status, trends, or mechanic relationships of environmental resource requires understanding the spatial context of these data. While some analyses can be completed using a coarse-level description of location (sub-basin or watershed), many analyses require explicit latitude and longitude coordinates for each monitoring event. This tool will also support data analysis efforts by providing a single regional repository for capture and storage of monitoring locations and area of inference polygons and to support reporting of those locations in a variety of formats and scales of resolution.

In a recent survey of PNAMP participants 100% indicated that this tool had high potential benefit to the region and over 70% indicated this tool had high potential benefit to their individual organization. In 2006, StreamNet developed an Aquatic Monitoring Activity Inventory database to support PNAMP in answering question about regional monitoring. This proposed tool would build on lessons learned from the StreamNet effort (Storch, 2006).

Pros:
- Streamlines metadata creation process
- Increased automation will help alleviate current workload
- Subset of FGDC elements that directly support monitoring coordination and funding decisions
- Improve the resolution, accuracy, and availability of location information for monitoring activities and data analysis efforts
- Assist funding agencies in tracking where monitoring resources are being allocated

- Integration with other regional data systems including MonitoringMethods.org
- Support data collection agencies with reporting monitoring locations
- Support automation of data analysis through improved spatial location information
- Support implementation of NED recommendation for reporting spatial location

Cons:

- Requires policy enforcement to ensure use by data collectors
- Requires ability to import location information from user provided files
- Requires regional organization to manage the system
- Does not include all FGDC metadata elements

Requirements:

The system should support users in creating location information by either uploading files to the system or by using a web-based mapping tool to manually draw points, lines, or polygons. Uploading files would allow users to upload text files or standard GIS files. Data collectors who use GPS units to gather coordinate information during monitoring activities need to be able to upload text files or comma separated values (csv) files to the system. Users who have GIS capacity within their organization need the ability to upload point, line, or polygon files to the system. The system should provide the ability to manually draw location information on a web-based map. The system should allow users to select a location type (point, line, or polygon), zoom to a location on the map, and then use the mouse to click on the map. To support manual drawing, the web-based maps need to include multiple background layers including aerial/satellite imagery, a standard stream layer, roads layer, national wetlands layer, and a topographic layer. This use-case represents the most advanced aspect of web-based GIS capabilities. The data system would need to support create, update, delete and edit functions for point, line, and polygon features. Standard fields should be included for all new features to allow user to attribute newly created features.

Upon completion of monitoring activities for a given season, crew leaders would be responsible to record location, protocol information, organization, and date of monitoring activities for each site visited during the field season. For sites that were visited, but not sampled, the reason for not sampling should also be recorded to the data system. This end-of-season reporting should be funded and required under the data collection contract from the funding organization.

The Monitoring Activity Inventory Tool must support dynamic reporting of monitoring activities at a range of spatial and temporal scales. The system must support searches based on geography, time frame, keywords or full text, protocol category, fish population, watershed, or a known location. Search results may be a single visit, multiple visits for a single site, most recent visit for a collection of sites, or a collection of sites and visits. The system must support several different outputs including coordinates for a bounding box of sites, distinct list of sites, distinct list of protocols, and distinct list of monitoring data storage systems or data contacts. Additionally, a metadata report containing a subset of FGDC elements could be produced that reported the spatial bounding box for the set of all sites in the search results. This report would include all appropriate Spatial Domain and Spatial Reference metadata elements from the FGDC standard. Regional data stewards could supplement this report with additional metadata elements to create a full metadata document.

EDS Recommendation:
PNAMP should secure a funding source and release a request for proposals from qualified vendors for development of this tool.

Other Tools Suggested, But Not Recommended

Metadata Builder

This proposed tool would be comprised of a set of scripts that could pull together metadata from a variety of sources. This approach will only work if source metadata or source data systems exist, are well established, and have data holdings. Likely, this approach is premature.

Rational for Not Advancing:

- Requires unique scripts for each contract management system
- Difficulty in coordination at regional scale
- Institutional boundaries around enterprise data systems

Database Documentation Tool

This proposed tool would be comprised of code that scans Microsoft Access or SQL Server databases and generates entity attribute info for FGDC.

Rational for Not Advancing:

- Requires that data is stored in a database
- Would only support groups that are already technologically advance

Develop Excel Worksheet with Subset of FGDC Elements

This proposed tool would be to develop a Microsoft Excel worksheet that supports documenting a subset of FGDC Elements and include worksheet with field data.

Rational for Not Advancing:

- Minimal metadata is minimally useful
- This distributed approach requires significant staff time for data compilation
- Limited capacity in Excel to lock data format and enforce data standards
- Targets concerns of highest paid biologic staff, when in practice, metadata maintenance is typically assigned to more junior staff.

References

Rentmeester, S. 2010. Regional Guidance on Metadata for Environmental Data. Pacific Northwest Aquatic Monitoring Partnership. Series 2010-001.

Storch, A. 2006. Pacific Northwest Aquatic Monitoring Partnership Inventory of Aquatic Monitoring Activities. Pacific States Marine Fisheries Commission. http://www.pnamp.org/sites/default/files/2004-002-00_28571_FinalReport.pdf

NED. 2006. Northwest Environmental Data Network Portal Channels and Data Steward Roles and Responsibilities. Version 4 (July 2006). http://www.nwcouncil.org/ned/ChannelSteward.pdf