Regional Guidance on Metadata for Environmental Data

Edited by Steve Rentmeester with assistance from
the Pacific Northwest Aquatic Monitoring Partnership Metadata Workgroup

January 4, 2010

Pacific Northwest
Aquatic Monitoring
Partnership

PNAMP Series 2010-001

[This page left intentionally blank]

Executive Summary

Simply stated, metadata are "data about data". Metadata describe the content, quality, condition, and other characteristics of data. Most commonly, metadata are used to enhance searching and discovery of data sets and to facilitate understanding of the meaning and proper use of datasets. Additionally, metadata can be used to automate workflows within organizations. This guidance document describes what metadata are, how metadata are used, and the benefits of creating and maintaining metadata. Four distinct metadata standards (Dublin Core, Content Standard for Digital Geospatial Metadata (CSDGM), North American Profile, and Ecological Metadata Language) and the intended use of each are compared in this document. The Pacific Northwest Aquatic Monitoring Partnership Metadata Workgroup has reviewed each standard in terms of appropriateness for describing ecological data and recommends the CSDGM with associated extension or Ecological Metadata Language documents for all datasets. Funding entities are encouraged to implement contracting language requiring metadata as an integral component of any data deliverable and to support metadata creation by providing funding for regional metadata stewards. Organizations are encouraged to create metadata for the purpose of protecting investments in data generation and to enhance the quality, usability, and value of data produced by the organization. Additionally, organizations are encouraged to institute mandates requiring metadata creation. In recognition of the long backlog of datasets with no metadata, organizations are encouraged to phase in metadata creation by starting with newly created datasets, then inventorying existing datasets, and then identifying priority datasets for additional stewardship.

Acknowledgements

Many individuals participate in the Pacific Northwest Aquatic Monitoring Partnership Metadata Workgroup and contributed to this document including Mike Banach, Kasey Bliesner, Rob Bochenek, Jen Carlino, Jan Conitz, Cedric Cooney, Henry Franzoni, Stan Frazier, Phil Herndon, Vivian Hutchison, Cathy Kellon, Dayv Lowry, Edna Mo, Michael Newsom, Tracy Olson, Eric Peterson, Sean Quigley, Steve Rentmeester, Jacque Schei, Bruce Schmidt, Russell Scranton, David Tetta, and Carol Volk. Thank you all for your valuable contributions.

Suggested Citation Format

Rentmeester, S., ed., 2010. Regional Guidance on Metadata for Environmental Data. PNAMP Series Report No. 2010-001. Cook, WA: Pacific Northwest Aquatic Monitoring Partnership.

Table of Contents

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 and can be organized in the following categories.

Metadata Categories

- Identification - basic information about data set origin, time span, and content
- Data quality - a general assessment of the quality of the data set
- Spatial data organization – details about the spatial location described by the data set
- Spatial reference – information about how spatial information is reported in the data set
- Entity and attribute - details about the information content of the data set
- Distribution - information about the distributor of and options for obtaining the data set
- Metadata reference - information on the currentness of the metadata information

Why use metadata?

Metadata provide significant benefits to both the organization that collect data and to those who subsequently use the data. While metadata creation can sometimes feel like additional work for someone else's benefit, metadata are also important to the organization that collect the data for:

- Protecting investment in data collection by:
  - improving tracking of data within the organization,
  - supporting long-term maintenance of data sets,
  - ensuring data are understood and properly used as people change jobs, and
  - supporting the organization's role in regional management and restoration efforts;
- Limiting liability associated with data sharing by:
  - documenting the data to avoid misuse or misrepresentation of the data, and
  - meeting funding entities' expectations that data will be shared;
- Improving organizational efficiency by:
  - minimizing staff time spent responding to questions about their data,
  - providing an inventory of data to enable rapid location of pertinent data, and
  - providing information needed to support computer automation.

For organizations that collect data, metadata help enhance the quality, usability and value of data for internal and external users. Organizations should view metadata creation as integral to their workflow and metadata as integral to datasets. Organizations are strongly encouraged to begin metadata documentation during the earliest stages of project planning and to use and maintain metadata at every stage of their workflow.

A multitude of monitoring programs and organizations throughout the Pacific Northwest collect research, monitoring, and evaluation data. Each program or organization collects, stores, and manages monitoring data in unique ways aimed at meeting program-specific objectives. Metadata facilitate the identification, discovery, assessment, storage, and use of data collected under widely varying objectives. No data set is complete without a metadata record. These standardized records describe such important features as why the data set was created, who created it, how accurate data are, what methodologies were used to develop it, and much more. Metadata support the use of data by multiple monitoring programs to meet broader objectives than just the original purpose. Metadata can:

- help avoid data duplication,
- foster sharing of data resources,
- help ensure that data are interpreted and used appropriately,
- preserve institutional memory,
- publicize research, and
- reduce workload required to compile data for regional analyses.

Metadata Standards

Standards form the core of nearly every activity aimed at managing biological information. Standards are critical in order to characterize data in a consistent manner, integrate data from multiple sources, and to make scientifically credible decisions based upon collected information. The more standardized the structure and content of information, the more effectively it can be used by both humans and machines. A metadata standard is simply a common set of terms and definitions that are presented in a structured format. While there are many national and international metadata standards, this document will focus on four common metadata standards: Dublin Core, Content Standard for Digital Geospatial Metadata, International Organization for Standardization 19115, and the Ecological Metadata Language.

1. The Dublin Core Metadata Standard is an internationally recognized metadata standard that supports a broad range of uses. It is useful for characterizing a variety of online and offline resources, including publications, tools, software, educational materials, grey literature, references, and other general resources. The Dublin Core contains 15 descriptive fields representing a core set of elements likely to be useful across a broad range of disciplines and business sectors. It is generally recognized that Dublin Core does not provide enough detail to support requirements for proper interpretation of biological or ecological data sets.

2. The Content Standard for Digital Geospatial Metadata (CSDGM) Version 2 (FGDC-STD-001-1998) is the metadata standard for all U.S. Federal agencies, as per Executive Order 12906. The standard is frequently referred to as the FGDC Metadata Standard. The objective of the standard is to provide a common set of terminology and definitions for the documentation of digital geospatial data. The standard was developed from the perspective of defining the information required to determine the availability**,** fitness for an intended use, means of accessing**,** and the means to transfer geospatial data sets. The standard does not

specify how the data are organized or presented in a computer system or how the data were collected.

A key feature of the CSDGM is the ability to customize or extend the standard through the development of profiles or extensions. Profiles are custom adaptations that may specify specific domain values for existing CSDGM elements and/or increase conditionality of a specific element. Extensions are a set of added elements that extend the standard to better serve the community or data type. Profiles may also include extensions and undergo an extensive review process. Profiles and extensions to the CSDGM standard currently exist for biological, remote sensing, shorelines, and wetland datasets. The Biological Data Profile, developed by the U.S. Geological Survey National Biological Information Infrastructure program (USGS-NBII), provides a common set of terminology and definitions for documentation of biological data. It allows biological information such as taxonomy, sampling methodology, and analytical tools to be added to a metadata record.

3. The International Organization for Standardization (ISO) has developed and approved an international metadata standard, ISO 19115. In December 2003, the American National Standards Institute (ANSI) adopted the international standard, ISO 19115 and the ANSI International Committee for Information Technology Standards signed an agreement with Canadian General Standards Board Committee on Geomatics to co-develop the North American Profile (NAP), a regional profile of ISO 19115 Geographic Information - Metadata. The North American Profile has been adopted by ANSI and is currently in final stages before release. National-level federal programs and private vendors are shouldering the initial responsibility for creating crosswalks from CSDGM to the ISO 19115 standard and are developing compliant metadata entry tools. The current suggestion from FGDC is continued use of the CSDGM with appropriate profiles and extensions until the NAP is fully implemented.

4. Ecological Metadata Language (EML) was developed in conjunction with the National Center for Ecological Analysis and Synthesis and was based on prior work done by the Ecological Society of America. It was developed by and for the ecology discipline. EML is implemented as a series of XML (Extensible Markup Language) document types that can be used in a modular and extensible manner. EML provides a flexible metadata standard for use in data analysis and archiving that will allow automated machine processing, searching, and retrieval. EML is more detailed and descriptive than CSDGM. EML documents can be converted to the CSDGM standard using existing tools. Several national-level metadata clearinghouses support EML, including USGS-NBII, NOAA's Metadata Enterprise Resource Management Aid (MERMAid), and the Mercury clearinghouses at the Oak Ridge National Laboratory.

Selecting an Appropriate Metadata Standard

Selecting an appropriate metadata standard is partially dependent on how a user community anticipates using the metadata to support their objectives. Metadata can be used for a range of purposes including inventorying existing datasets, searching for and discovering datasets, proper interpretation and use of data, and automation of data validation or analysis. Each of these uses

requires increasing levels of detail in metadata (Table 1). Inventorying existing datasets may only require a minimal set of metadata elements (e.g. a subset of CSDGM elements). The CSDGM metadata standard supports search, discovery and distribution of datasets and supports proper use of the data. This represents the most common use of metadata. Supporting the proper interpretation and use of data requires description of the entities and attributes in the dataset. These metadata elements are included in the Biological Data Profile. In addition to supporting those activities, metadata can also be used to automate workflows by facilitating creation of data entry applications, automation of data validation, exchange of data between data management systems, and automation of metric creation. Supporting the automation of workflows can only be accomplished through the use of detailed, machine-readable metadata. Data practitioners at the National Center for Ecological Analysis and Synthesis (NCEAS) and the Long Term Ecological Research Network (LTER) created the Ecological Metadata Language (EML) and have demonstrated the potential for automation through the use of detailed, machine-readable metadata.

Table 1. Functional abilities of the major metadata standards.

Standard	Strength	Intended Use
Dublin Core	Simple and broadly recognized	Software, tools, website, publications
CSDGM	The standard for U.S. Federal agencies. Supports data discovery and retrieval. Sufficient detail to use data properly.	Both spatially explicit and non-spatially explicit data sets; biological and physical data sets
ISO 19115 North American Profile	International metadata standard	Metadata clearinghouses will convert CSDGM metadata to North American Profile format
EML	Modular and flexible. Machine readable and supports automation of data validation and analysis	Both spatially explicit and non-spatially explicit data sets; biological and physical data sets

Organizations must evaluate their objectives for creating metadata and select metadata standards with the appropriate level of detail to support those objectives. Given limited resources and tools for metadata creation and varying objectives for how organizations use metadata, it is recommended that organizations select metadata standards based on criteria including:

- relevance to organizational and regional monitoring objectives,
- duration of time since data creation, and
- desired level of workflow automation.

Most organizations in the Pacific Northwest have large backlogs of data with limited or non-existing metadata. Completing an inventory of historic datasets using a metadata standard of limited detail will help organizations prioritize historic datasets for further stewardship. For selected datasets, metadata elements documented during the inventory stage can be supplemented with additional elements to meet requirements of a more detailed standard.

Metadata Decision Tree

Existing Datasets

The Metadata Decision Tree (Figure 1) aims to assist organizations in selecting a metadata standard that is appropriate to the dataset and intended use of the metadata. The Metadata Decision Tree begins by recommending a distinct pathway for existing datasets. To facilitate the scoping, planning, and implementation of any metadata documentation effort, the PNAMP Metadata Workgroup recommends that organizations begin by documenting all existing data resources with an inventory-level metadata standard. The inventory-level metadata standard is composed of a subset of elements from the full CSDGM metadata standard. The inventory will allow an organization to asses the utility of datasets and prioritize datasets for further stewardship. Organizations will need to define utility criteria based on each dataset's relevance to organizational reporting requirements, relevance to regional-level high priority indicators, and the ability to capture metadata and assess the quality of the data. Datasets that meet an organization's utility criteria should be advanced along the decision tree and follow the pathway for future datasets. For these datasets, the PNAMP Metadata Workgroup recommends that inventory-level metadata be supplemented to create a full metadata record.

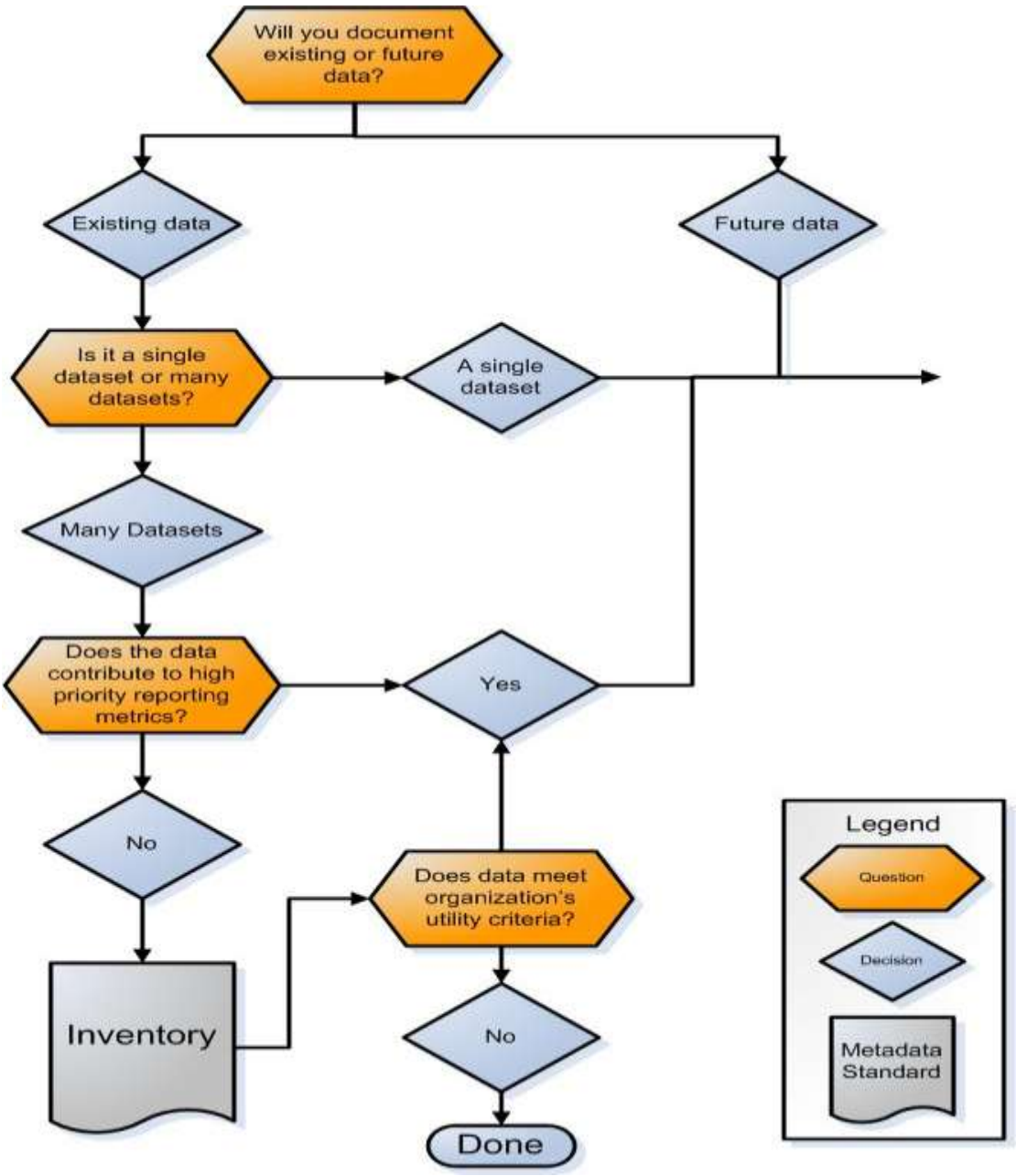

Figure 1. Metadata Decision Tree – Existing Datasets Section

Future Datasets

The PNAMP Metadata Workgroup recommends that all newly created datasets be documented with the CSDGM compliant metadata and any appropriate extensions to the standard (Figure 2). All geographic data, including habitat and watershed condition datasets, should be documented using the CSDGM metadata standard. Remote sensing data should be documented using the CSDGM metadata standard with the remote sensing extension. Biological data, including all data about fish or other wildlife, should be documented using the CSDGM metadata standard with the Biological Data Profile. While it is likely too early to recommend adoption of EML by all monitoring programs in the Pacific Northwest, work at NCEAS and LTER has demonstrated the potential of EML to support long-term data management efforts in the region. The PNAMP Metadata Workgroup recommends that demonstration data management and monitoring programs should adopt the use of EML as their metadata standard. Metadata described with EML can easily be exported as a CSDGM compliant metadata document.

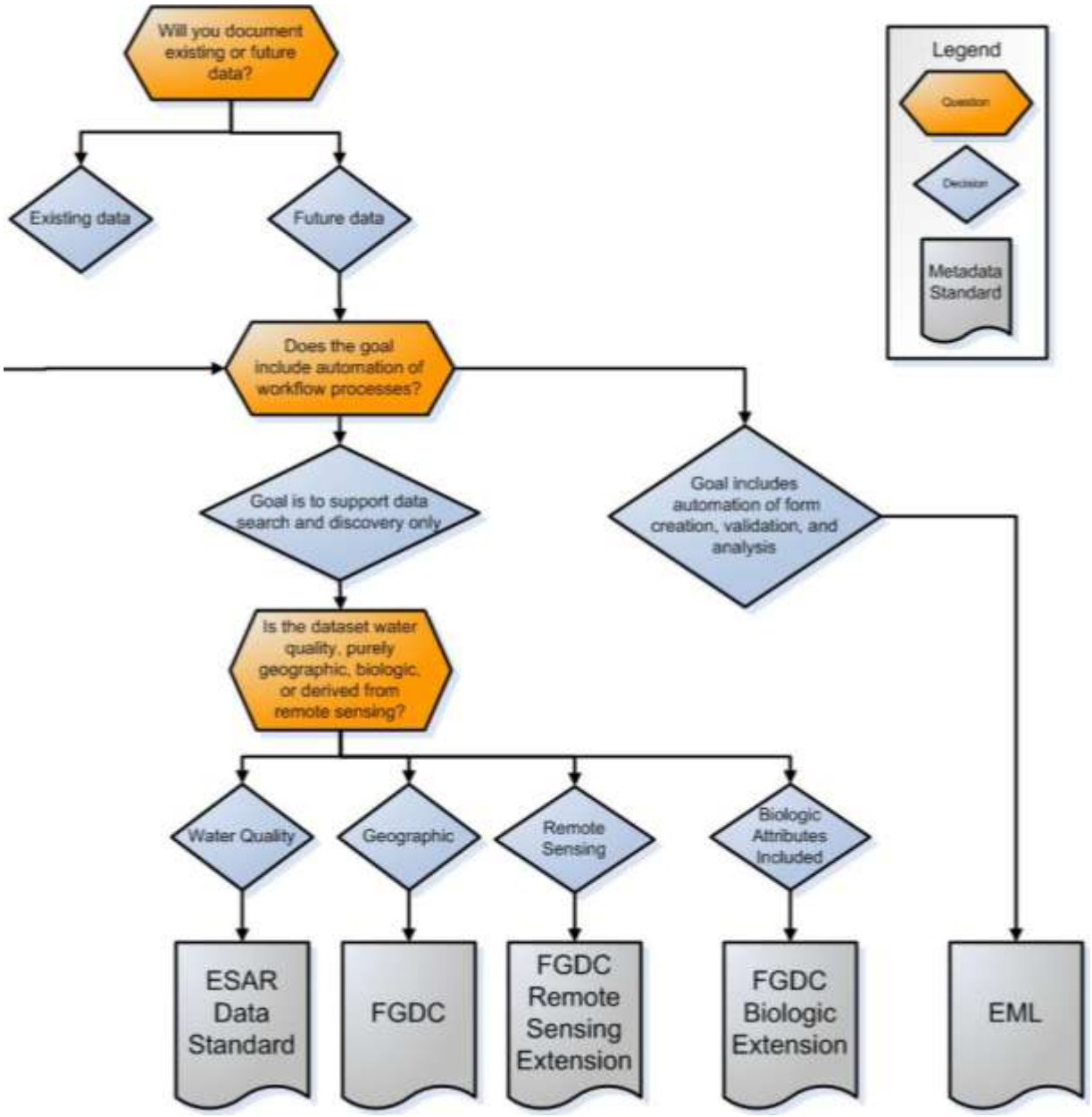

Figure 2. Metadata Decision Tree – Future Datasets Section

Recommendations and Next Steps

Implementing a regional-scale strategy for documenting metadata associated with all regionally relevant monitoring data represents a significant challenge. Executive Order 12906 and the NBII 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 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**.

Regional funding entities can support metadata creation by implementing contract language that requires all newly created datasets be delivered with a CSDGM and appropriate profile compliant metadata or an EML metadata document, as recommended in the metadata decision tree (Figures 1,2). However, implementing contract requirements for metadata must be supported with funding for regional metadata stewards who can provide training and support to regional data collection organizations. Regional metadata steward responsibilities should be modeled around the NBII Metadata Program. Metadata steward responsibilities should include:

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

Creation of metadata is a shared responsibility and will require staffing commitments from both regional funding entities and data collection organizations. Organizations benefit from creation and maintenance of metadata (see "Why Use Metadata?"). Organizational leads should appoint metadata coordinators within each division of the organization to acquire additional training, coordinate internal metadata creation efforts, and provide guidance to other staff. While federal agencies are required to create metadata under Executive Order 12906, the PNAMP Metadata Workgroup recommends that state, tribal, and local organizations implement similar mandates. Mandates should phase metadata creation in gradually.

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

Once a few datasets have been described it becomes much easier to create metadata for additional datasets. Most metadata editors support creation of templates where the majority of descriptive fields are predefined for the specific organization, division, or program. These predefined fields can be copied over from existing metadata.

The PNAMP Metadata Workgroup has created a webpage with links to metadata standards, metadata creation tools, and other resources (http://www.pnamp.org/metadata). Additionally, assistance with creating and maintaining metadata can be obtained through the NBII Metadata Program (http://www.nbii.gov/portal/community/Communities/Toolkit/Metadata/).

Creation of metadata requires tools and fortunately several good tools exist (Table 2). Several of these tools support the use of templates. Metadata stewards can help organizations or programs create templates with predefined values for many descriptive fields. Alternatively, database programmers may be able to create metadata templates using data stored within an organization's contract management application (e.g. Pisces at Bonneville Power Administration). This approach has been demonstrated at a USGS field office and likely has broad utility.

Table 2. Commonly used metadata creation tools.

Tool	CSDGM	Description	Links
ArcCatalog 9.x	Yes	Editor with style sheets for various formats. Automated creation and update of spatial metadata elements and list of attributes. Stored as XML flat file. Can create custom editor.	www.esri.com
Metavist	Yes	Editor includes the Biological Data Profile and has recently improved usability for the Entity and Attribute section	http://metavist.djames.net/Default.aspx
Metascribe	Yes	Significantly reduces effort to produce metadata. It is template driven. Once a template is created, the user can create multiple records quickly and easily.	www.csc.noaa.gov/metadata/metascribe/
EPA Metadata Editor	Yes	Developed to simplify and standardize geospatial metadata development across the U.S. Environmental Protection Agency (EPA).	www.epa.gov/geospatial/eme.html
Morpho	Yes	Supports creation and editing of metadata along with viewing and querying data. Primarily used for creating EML.	http://knb.ecoinformatics.org/morphoportal.jsp

Ultimately, changing business practices around metadata creation will require new perspectives about the value of metadata creation. Typically, metadata are created as a final step in publishing results and data. When metadata are created at the end of the data creation workflow, there are limited opportunities for data creation organizations to benefit from metadata. Creating metadata at the earliest stages of a data collection project can support project planning, data entry, data validation, and data analysis. These workflow efficiencies can only be realized if metadata are created prior to data collection. Tools being created at NCEAS and within the Integrated Status and Effectiveness Monitoring Program are demonstrating the benefits of creating metadata prior to data collection.

Appendices

Appendix A. Metadata Decision Tree - All Sections