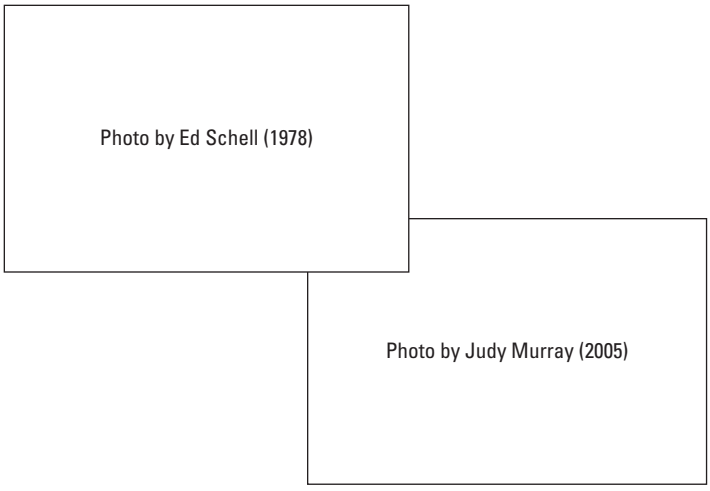
USGS
science for a changing world

NBII-SAIN Data Management Toolkit

Open-File Report 2009–1170

U.S. Department of the Interior
U.S. Geological Survey



Cover. A photopoint comparison of Round Bald on Roan Mountain showing encroachment of woody plants over a 20 year period. Well documented, geo-referenced, and archived data are critical for understanding such changes and associated ecosystem impacts. The value of such data grows over time. The left photo was taken by Mr. Ed Schell in 1978, and the right one was taken by Ms. Judy Murray at approximately the same location in 2005.

NBII-SAIN Data Management Toolkit

By Thomas E. Burley[1] and John D. Peine

[1]Formerly of the Institute for a Secure and Sustainable Environment,
University of Tennessee

Open-File Report 2009–1170

U.S. Department of the Interior
U.S. Geological Survey

U.S. Department of the Interior
KEN SALAZAR, Secretary

U.S. Geological Survey
Suzette M. Kimball, Acting Director

U.S. Geological Survey, Reston, Virginia: 2009

For more information on the USGS—the Federal source for science about the Earth, its natural and living resources, natural hazards, and the environment, visit http://www.usgs.gov or call 1-888-ASK-USGS

For an overview of USGS information products, including maps, imagery, and publications, visit http://www.usgs.gov/pubprod

To order this and other USGS information products, visit http://store.usgs.gov

Suggested citation:
Burley, T.E., and Peine, J.D., 2009, NBII-SAIN Data Management Toolkit, U.S. Geological Survey Open-File Report 2009–1170, 96 p.

DISCLAIMER

The purpose of this document is to present various techniques and methods for addressing the various aspects of data management that arise with projects, agencies, or organizations. Various geospatial and non-geospatial hardware and software programs and tools are presented and incorporated into specific workflows as well for illustration purposes only. The authors, the University of Tennessee, the U.S. Geological Survey, and the USGS National Biological Information Infrastructure do not endorse or promote the use of any of these specific (both public and private domain) programs or tools mentioned herein. Mention of any specific software package or equipment does not imply endorsement; use of any software or hardware mentioned herein is at the risk of the user and bears no responsibility on the authors, the University of Tennessee, the U.S. Geological Survey, and the USGS National Biological Information Infrastructure. These example tools and software packages simply serve as examples for the various areas of data management and are therefore not comprehensive, conclusive, endorsed, or promoted.

ACKNOWLEDGMENTS

The authors thank Brandon League with ESRI (formerly University of Tennessee); Franciel Azpurua-Linares, Shelaine Curd-Hetrick, John Rumble, and Fred Rascoe with Information International Associates; Jean Freeney, Mike Frame, and Viv Hutchison with USGS-NBII; Mark Roper with the U.S. Forest Service, San Juan National Forest; Anita Bailey with the U.S. Forest Service, Cherokee National Forest; Drew Selig with the U.S. Forest Service; Carolyn Wells with the U.S. Fish and Wildlife Service, Asheville, N.C., Field Office; Laura Pickens with the U.S. Fish and Wildlife Service, Ashville, N.C., Field Office; Judy Murray, Nora Schubert, Kristy Urquhart, and Bob Harvey with the Southern Appalachian Highlands Conservancy (SAHC); Terry Giles of the USGS Ft. Collins Science Center

Date of Version	Changes Made	Reason
25-Jan-07	--	--
19-Feb-07	Added Appendix D	Relevance
23-Apr-08	Major text edit and reformat of entire document for purpose of readability and presentation	Increase usability
18-June-08	Captions for front page pictures	Visual of value of Toolkit

Contents

Data Management Toolkit Part A (Sections 1–5)—Project Policy, Approach, and Planning Framework Overview....1
1 Data Management Toolkit Introduction: Background and Purpose....1
2 Planning: Project Management Fundamentals....9
2.1 Fundamental Project Implementation Tips....9
2.2 Project Planning....10
3 Planning: The Benefits of Well-Defined Standards....13
4 Planning: Strategic Data Management Principles and Guidelines....15
4.1 The Elements of Sound Data Management....15
4.2 Strategies for Data Management....15
5 Planning: Data Stewardship for Ensuring Data Longevity....17
5.1 Data Policy....17
5.2 Data Ownership....18
5.3 Data Custodianship....18
5.4 Data Storage and Archiving....19
5.5 Access and Security....20
5.5.1 Access....20
5.5.2 Security....21
5.6 Web Access Considerations....21
Data Management Toolkit Part B (Section 6)—Elements of Data Management Overview....25
6 Planning: Data Management Considerations for Meeting Goals and Objectives....25
6.1 Data Modeling and Database Design....25
6.2 Data Acquisition....26
6.2.1 The Data Clearinghouse Concept....26
6.3 Critical Elements of Data....27
6.3.1 Spatial Data Elements....27
6.3.2 Non-Spatial Data Elements....29
6.4 Metadata Documentation of Data Sets, Information, and Resources....30
6.5 Thematic Data Content Standards and Systems for Promoting Data Interoperability and Sharing....30
6.6 Quality Assurance and Quality Control....31
6.6.1 Quality Assurance....31
6.6.2 Quality Control....32
6.6.3 Aspects of Data Quality....32
6.6.4 Sources of Data Error....33
6.7 Cartographic Display and Maps....34
6.8 Demonstration of Example Project Data Sets—Areas of Data Management....35
6.9 Additional Data Management Information Sources....35
Data Management Toolkit Part C (Sections 7–12)—Example Approaches to Specific Elements of Data Management....37
7 Guidelines for Data Modeling and Design....37
7.1 Conceptual Model Design: Capturing the User's View....37

7.2 Logical Model Design 37
7.2.1 Establish Entities and Relations Among Entities 38
7.2.2 Entity Representation 38
7.2.3 Evaluation of the Model in Progress 39
7.3 Physical Model Design 39
7.3.1 Pilot/Prototype Test Model 39
7.3.2 Entities to Features and Objects 39
7.4 Data Dictionary 40
7.5 Pilot Project Test Database 40
8 Guidelines for Project Quality Assurance, Development of a QA Plan, and Quality Control 41
8.1 Quality Assurance and the QA Plan 41
8.1.1 Management 41
8.1.2 Design 42
8.1.3 Data Assessment 44
8.1.4 Reporting and Oversight 47
8.2 Quality Control 48
8.2.1 First-Round QC 48
8.2.2 Second-Round QC 51
8.3 Additional Quality Assurance and Quality Control Information Sources 53
9 Tools, Guidelines, and Work flows for Creation of Federal Geographic Data Committee-Compliant Metadata 55
9.1 Documentation Tools and Standards 55
9.1.1 Metadata Standards 55
9.1.2 Free Metadata Creation Tools 56
9.1.3 Metadata Training and Assistance 56
9.2 Process Steps for Documentation of Data Sets and Other Resources 56
9.3 Tips and Tricks for Creating Metadata with the FGDC-NBII Biological Data Profile 57
9.3.1 Example Work flows for Metadata Documentation 57
9.4 Instructions for Utilizing ITIS (Integrated Taxonomic Information System) for Documenting the Biological Dimensions of a Data Set 59
9.5 Tips for Documenting Legacy Data Sets 60
9.6 Additional Metadata Information Sources 60
10 Geospatial Data Acquisition Guidelines for Quality 61
10.1 Geo-Referencing With GPS Data Collection 61
10.1.1 GPS Unit Data Collection – Mapping Grade Unit Settings 61
10.1.2 GPS and Coordinate Systems 62
10.1.3 GPS Data Collection Guidelines for Accuracy 63
10.1.4 Datum Transformations: Differential Correction and the WGS 84 and NAD 83 Datum Transformation Issue 64
10.1.5 Getting GPS Data in a GIS-Compatible Format 65
10.1.6 GPS Accuracy Reporting and Recording Critical Information in FGDC Metadata 66
10.2 Scanning and Geo-Referencing Hardcopy Data and Maps 66
10.2.1 Scanning Hardcopy Maps and Other Source Data 66

10.2.2 Geo-Referencing 67
10.3 Digitizing Features from a Geo-Referenced Digital File 67
10.4 Working with Tabular Data 68
10.5 Additional Data Tools and Information Sources 68
11 Documentation Tool # 1—FGDC Bio-Profile Metadata Questions 71
12 Documentation Tool # 2 – Dublin Core Metadata Questions 73
Selected References 85
Appendix A—FGDC Bio-Profile Cross-walk 89
Appendix B—Quality Assurance Plan Template 91
Appendix C—The Top Ten Most Common Metadata Errors 95

Figures

1–2. Diagrams showing—
1. How adaptive management and monitoring functions all relate to and are dependent on effective information managment 12
2. How metadata helps prevent information entropy. 16

NBII-SAIN Data Management Toolkit

Data Management Toolkit Part A (Sections 1–5)– Project Policy, Approach, and Planning Framework Overview

1 Data Management Toolkit Introduction: Background and Purpose

Catawaba rhododendron, Rhododendrom catawbiense is the number one tourist attraction at Roan Mountain State Park. The old timers called impenetrable patches of woody shrubs like rhododendron and mountain laurel "laurel hell." Native shrubs such as these are invading the grassy balds. (Photograph from Southern Appalachian Highlands Conservancy)

Importance of Data Management

The Strategic Plan for the U.S. Geological Survey Biological Informatics Program (2005–2009) recognizes the need for effective data management:

> Though the Federal government invests more than $600 million per year in biological data collection, it is difficult to address these issues because of limited accessibility and lack of standards for data and information...variable quality, sources, methods, and formats (for example observations in the field, museum specimens, and satellite images) present additional challenges. This is further complicated by the fast-moving target of emerging and changing technologies such as GPS and GIS. Even though these technologies offer new solutions, they also create new informatics challenges (Ruggiero and others, 2005).

The USGS National Biological Information Infrastructure program, hereafter referred to as NBII, is charged with the mission to improve the way data and information are gathered, documented, stored, and accessed. The central objective of this project is a direct reflection of the purpose of NBII as described by John Mosesso, Program Manager of the U.S. Geological Survey-Biological Informatics Program-GAP Analysis:

> At the outset, the reason for bringing about NBII was that there were significant amounts of data and information scattered all over the U.S., not accessible, in incompatible formats, and that NBII was tasked with addressing this problem….NBII's focus is to pull data together that truly matters to someone or communities. Essentially, the core questions are: 1) what are the issues, 2) where is the data, and 3) how can we make it usable and accessible (John Mosesso, U.S. Geological Survey, oral commun., 2006).

Redundancy in data collection can be a major issue when multiple stakeholders are involved with a common effort. In 2001 the U.S. General Accounting Office (USGAO) estimated that about 50 percent of the Federal government's geospatial

Substantial cost savings and increased efficiency are direct results of a pro-active data management approach.

data at the time was redundant. In addition, approximately 80 percent of the cost of a spatial information system is associated with spatial data collection and management (U.S. General Accounting Office, 2003). These figures indicate that the resources (time, personnel, money) of many agencies and organizations could be used more efficiently and effectively. Dedicated and conscientious data management coordination and documentation is critical for reducing such redundancy. Substantial cost savings and increased efficiency are direct results of a pro-active data management approach. In addition, details of projects as well as data and information are frequently lost as a result of real-world occurrences such as the passing of time, job turnover, and equipment changes and failure. A standardized, well documented database allows resource managers to identify issues, analyze options, and ultimately make better decisions in the context of adaptive management (National Land and Water Resources Audit and the Australia New Zealand Land Information Council on behalf of the Australian National Government, 2003).

Many environmentally focused, scientific, or natural resource management organizations collect and create both spatial and non-spatial data in some form. Data management appropriate for those data will be contingent upon the project goal(s) and objectives and thus will vary on a case-by-case basis. This project and the resulting Data Management Toolkit, hereafter referred to as the Toolkit, is therefore not intended to be comprehensive in terms of addressing all of the data management needs of all projects that contain biological, geospatial, and other types of data. The Toolkit emphasizes the idea of connecting a project's data and the related management needs to the defined project goals and objectives from the outset. In that context, the Toolkit presents and describes the fundamental components of sound data and information management that are common to projects involving biological, geospatial, and other related data. These components include project planning, standards, data stewardship, data modeling, quality assurance/quality control (QA/QC), metadata, geospatial data acquisition, critical elements of data, and free tools and resources. Also, where possible, it provides guidelines for addressing those various components based on industry, Federal, and international best practices and standards.

Quality information can only be derived from quality data.

The effectiveness of planning and decision-making is closely related to the quality and completeness of available information. Quality information can only be derived from quality data. Global positioning systems (GPS) and geographic information systems (GIS) greatly contribute to improved resource management and decision making. However, efficiency and effectiveness in natural resource management and decision making are not direct results of using tools such as GPS and GIS. These tools must be implemented with adhered-to standard procedures and methodologies for data management and data documentation. Otherwise, management goals and objectives may never be fully realized and the effects from implemented decisions may never be fully quantifiable. The data management concepts presented herein are geared towards facilitating multi-agency efforts related to the adaptive management of Roan Mountain. Since these concepts are applicable to any project, however, the intention was to present them in such a way that the Toolkit can be applied and adapted to other scenarios. The ultimate goal is to allow for those engaged in a project to become better aware of fundamental data management issues that may be outside of their respective areas of expertise.

The greatest challenge of this project was helping natural resource managers, agency biologists and scientists, the non-government community (NGO), and the academic community to realize the importance of data and information management. Approaching this issue in a holistic and collaborative way will greatly enhance the value and utilization of data and information over the long term.

Roan Mountain—An Ideal Case Study Application of the Toolkit

The impetus for this project and the development of the Toolkit was the recognized longstanding need for better documentation and management of data and information related to adaptive management of the Roan Mountain Massif, hereafter referred to as "the Roan." The circumstances represent an ideal scenario to evaluate the Toolkit and demonstrate the extent of its value. Located in the southern Appalachian Highlands, the Roan contains one of the last remnants of high-elevation grassy balds in the United States and is home to many species of concern, including the Gray's lily, Golden Winged Warbler, and the Carolina northern flying squirrel. The exact origin of grassy balds is something of an unsolved mystery. However they evolved, today a major challenge to resource managers is monitoring and managing the Roan grassy balds with the long-term goal of restoring and maintaining this internationally significant habitat. It is also an ideal location to study the effects of climate change on vulnerable species which make up over 30 percent of the known 700 plus species (Jamie Donaldson, oral commun., 2006).

The utility of biological information starts with effective data management. Though the Roan is a unique landscape, the data management issues identified are not unique in the context of multi-agency, collaborative, adaptive management efforts. At the core of any on-the-ground natural resource management effort are data and information which help drive critical decision making. The value and utility of such resources, as in the case of the Roan, can become highly marginalized over time without properly defined and integrated data management and metadata documentation. For more than 30 years, researchers and resource managers from the Federal,

state, academic, and non-profit sectors have meticulously gathered data on the flora and fauna of the unique Roan ecosystem. The need for an overarching methodology for collecting, documenting, interpreting, storing, and using this treasure trove of data and information has become even more apparent given current management foci. The Roan Mountain Project addresses the critical need to integrate geo-referenced data with historical and ongoing resource-management activities. To that end, the Data Management Toolkit has been devised to make data and information available in a usable format to people in decision-making roles.

The long standing recognized need for data management was articulated in a January 2006 letter of support for the NBII-SAIN Roan Mountain project from Paul Bradley, former District Ranger of the U.S. Forest Service-Region 8 Appalachian Ranger District, who wrote:

> Improving documentation of and access to biological data collected on the Roan has long been a high priority of biologists and managers…In addition, the development of tools for assessing the effectiveness of past, current and future adaptive management of the grassy balds as it relates to biological inventory, monitoring and science associated with the Roan would be most valuable. I believe that if we are successful in accomplishing those positive relationships, there is potential for transferability of methodology to other high elevation biomes in the Appalachian Mountain Range. These needs have been recognized as critical to the long term success of management activities taking place within the extremely sensitive and internationally significant resources of the Roan Massif for more than twenty years (Paul Bradley, written commun., 2006).

The endangered Gray's lily, *Lilium grayi*, discovered on the Roan by Asa Gray in the 19th century, is found only in a few high elevation meadows primarily along the North Carolina, East Tennessee, and Virginia borders. (Photo by Southern Appalachian Highlands Conservancy.)

The ecosystems on the Roan are in the midst of a perfect storm of multiple stressors on biological resources. Roan Mountain, home to the southernmost extent of ranges of species more adapted to cooler climates, is in a critical geographic location. There is an increase in invasive species, pests and pathogens are increasing, and air pollution remains a major concern. The increase in environmental stressors coincides with a dramatic reduction in funds and expertise of Federal land management agencies attempting to cope with this mounting crisis in adaptive resource management for ecosystem sustainability. Addressing these emerging issues in the future will require interdisciplinary science and predictive modeling to ascertain indicators of cause and effect relationships. Finding and implementing adaptive management solutions will require interagency and academic partnerships and cooperation. To realize these goals the principles of efficient and effective data management practices must be applied. Adopting the procedures documented in the Toolkit offers a technical means to create a foundation from which interdisciplinary science and predictive modeling goals can be achieved.

> We simply cannot afford to continue gathering data without an overarching integrated science strategy and documentation methodology.

A second critical challenge is to convince scientists and resource managers to use the Toolkit guidelines. During the last 30 years, a large amount of data has been generated in various formats. Documents have been stored in government offices, in scientists' offices, and in the case of Roan Mountain, even in private homes and garages as some researchers feared the data might be lost, discarded, or damaged. Coding of species at different research sites has been uncoordinated. In some cases, no or inaccurate data on geospatial relationships were recorded. There is a need to efficiently and effectively evaluate the cost and biological sustainability of adaptive management practices. A thorough data management system along with a holistic yet feasible science-based monitoring strategy would provide a foundation for scientific inquiry.

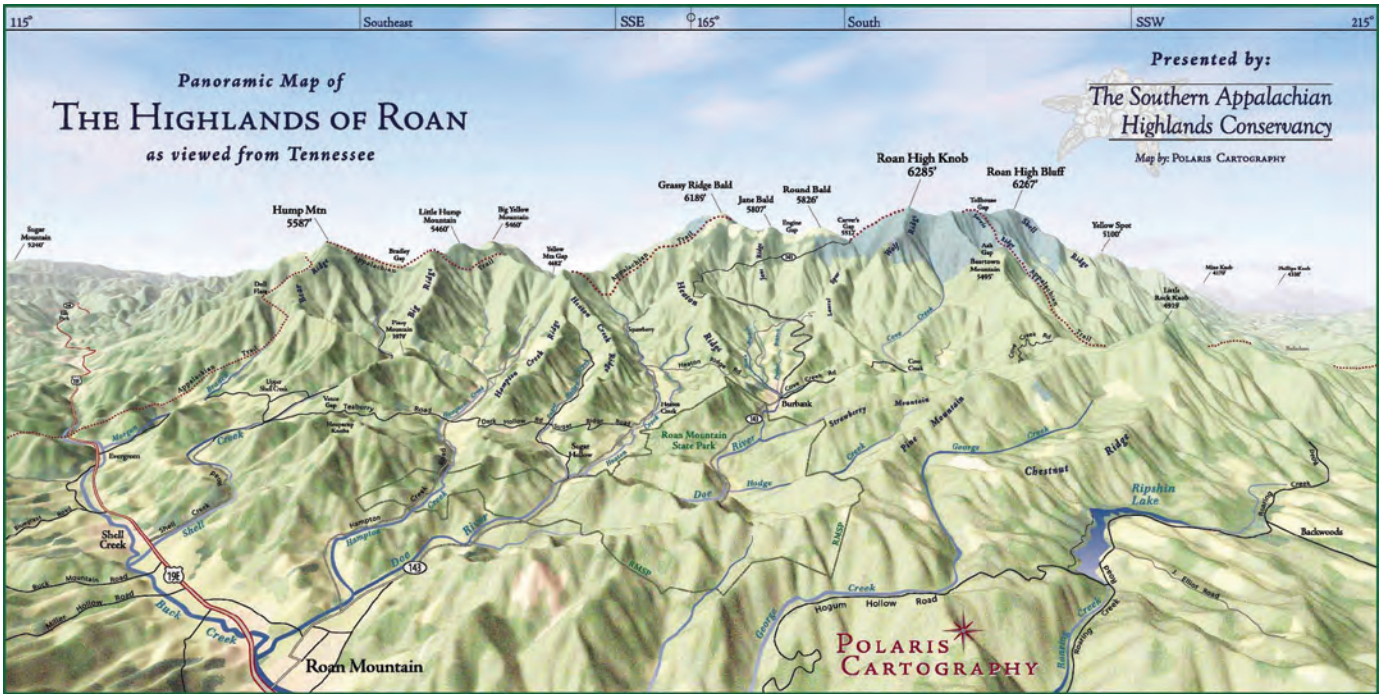
Perhaps of even greater importance, the Roan Mountain Project, with its multiple partners from the Federal,

state, university, and private sectors, can serve as a model of interagency cooperation. The Toolkit is ready to be put to use at Roan Mountain and other biomes in the Appalachian Highlands and ultimately at any other site where researchers, resource managers, governmental agencies, and the private sector are charged with monitoring and maintaining critical habitats. In a sense, the Toolkit provides a bridge from an earlier generation to future generations of scientists and managers. If used as intended, the Toolkit will serve as an impetus for more effective data analysis and productive resource management well into the 21st century.

Look to the Roan Mountain case study inserts to clarify the dimensions of the Toolkit and their importance to adaptive resource management applications.

ROAN COLLABORATORS

- U.S. Forest Service (Pisgah and Cherokee National Forests)
- National Park Service—Appalachian National Scenic Trail Park Office
- U.S. Fish and Wildlife Service—Ashville, NC field office
- U.S. Geological Survey—National Biological Information Infrastructure–Southern Appalachian Node
- The Nature Conservancy
- Southern Appalachian Highlands Conservancy
- North Carolina Natural Heritage Program
- Tennessee Natural Heritage Program
- Several regionally located universities
- Numerous individual citizen volunteers and contract biologists



The Highlands of Roan as seen from Tennessee (Image from Polaris Cartography, *http://www.polarismaps.com/prtpan_roan.shtml*)

User Guide to the NBII-SAIN Data Management Toolkit

Part A of the Toolkit is designed to introduce the critical components of data management. It can be considered a high level manager's overview that creates the context for the more detailed guidelines that follow. This is designed to help elucidate the concepts and their importance in the bigger picture of natural resource management.

Getting Started: Plan Your Work and Work Your Plan

The User Guide provides an overview of all the elements critical to effective data management included in the NBII-SAIN Data Management Toolkit. It can be used to focus on the dimensions of data management most relevant to the project. The User Guide helps frame the contents and approach of the Toolkit in the context of the NBII-SAIN project with which it is associated. In addition, the NBII-SAIN Roan Mountain FY05–06 Final Report (Burley and others, 2006) serves as an in-depth direct application of the data management concepts presented in the Toolkit to 12 example priority data sets resulting from years of collection, observation, and analysis at Roan Mountain. These are considered legacy data.

A brief introduction to representative topics included in the Toolkit is presented below with explanations of why they are relevant. The issues associated with the legacy data sets evaluated via the Toolkit are described in bulleted text relative to that section's context. Further details on these issues is provided in the NBII-SAIN Roan Mountain FY05–06 Final Report (Burley and others, 2006).

Project Feasibility and Pre-Planning

Ideally, before implementation, an assessment is done to determine whether or not a potential project will move forward. Specific research question(s), project goals, and objectives, as well as internal capacities and related issues are identified. The adequacy of resources (time, funds, and personnel) and upper management support in both the short and long term are evaluated. This is particularly important if the project is to have a long-term scope involving re-sampling or other work associated with management initiatives involving monitoring. These considerations help determine feasibility as well as implementation strategy.

Some identified issues with the example Roan legacy data sets in the context of project feasibility and planning include the following:

- The levels of resources needed to replicate the initial effort are not available.
- The data were not managed in a way that facilitates such legacy data uses as trend analysis.
- Much uncertainty existed regarding how legacy data sets related to a holistic management strategy.

The components of project feasibility and pre-planning are featured in

- **Toolkit Section 2 - Planning: Project Management Fundamentals** and
- **Toolkit Section 2.1 - Fundamental Project Implementation Tips**.

Planning for Project Implementation

If it is determined that a project will move forward, proper planning for data management helps ensure a project's success. This is particularly important for long-term initiatives with many partners involved. The various areas of the Toolkit below present an overview of the fundamental components of data management as related to the full lifecycle of a data set.

Implementing and(or) establishing data standards early in a project is an excellent way to ensure quality and data compatibility. This means taking the long-term investment approach to data and information. Standards can be applied to many areas of a project, including methods of data collection, equipment settings, data format, required data attributes, documentation methods, geo-referencing, and accuracy levels. Standards also support interoperability of data sets. Utilizing existing standards wherever possible also can help to increase efficiency so that researchers do not re-invent the data management wheel. Specifying all aspects of an anticipated long-term initiative from the outset can help to ensure quality and that any data or information will be consistent if the standards are followed consistently. This is especially important when multiple agencies and organizations are involved in a collaborative effort. Agreed-upon common methodologies help ensure the utility of data for all partners involved

and the consistency of data and information related to the geographic area of focus. Consistency also leads to the ability to conduct trend analyses for detecting change, which is a critical aspect of monitoring.

Some identified issues for the example Roan legacy data sets in the context of project implementation include the following:

- Data collection methods used with both spatial and non-spatial data were inconsistent.
- Some data have biological or adaptive management attributes; other data have spatial components but no attributes. In general, very little spatial data were available.
- No standard concise documentation format was available (for example, Federal Geographic Data Committee (FGDC) metadata) at the time the projects were designed.
- Species codes varied from year to year. Much uncertainty exists as to which species codes were used each year and which cover methodology, which determines the percentage of vegetative cover of plots each year, was applied.
- Considerable effort was required to reconcile differences between data sets.
- No experienced personnel involved with any collection or curation efforts remains available to assist with any reconciliation efforts beyond minimal quality assurance.

The important components of planning for project implementation are featured in

- **Toolkit Section 3—Planning: The Benefits of Well-Defined Standards**,
- **Toolkit Section 4—Planning: Strategic Data Management Principles and Guidelines**, and
- **Toolkit Section 6.5—Thematic Data Content Standards for Promoting Data Interoperability and Sharing**.

Planning: Data Stewardship

The life of data and information does not simply end after a project has been completed. Careful consideration of a project's data and information is needed for a long-term perspective of data management and physical state. When issues such as data policy and ownership, data custodianship, data and information archiving, and data accessibility are not addressed, numerous problems can arise that may require substantial time and effort to properly address. In addition, when multiple stakeholders are involved, relevant reports, data, and other project information can quickly become scattered and lost. This can be further complicated by personnel turnover and equipment change or failure. Proper data stewardship is needed even if the project is not meant to have a long-term scope of work so that the data and information can be accessible and usable for future unforeseen applications. Thorough consideration of data stewardship before a project starts can help to hedge against such problems.

Some identified issues with the example Roan legacy data sets in the context of data stewardship include the following:

- Data resources were physically scattered around the Southern Appalachian region and beyond, including some as far north as Massachusetts and as far south as Mississippi. Such locations included home garages, personal offices, and numerous Federal, state, university, and nongovernmental organization offices.
- Extensive time-consuming research was conducted in order to simply locate people who knew about the locations of documents and datasheets. The time and funds to conduct such research after-the-fact are resources that most organizations do not have today.
- Many former project partners noted that they had kept materials out of concern that they would be lost in the state and Federal agency domain.
- Duplication of effort was another serious issue. Many former partners were not aware of all the work and projects carried out, and in several specific cases, efforts were duplicated because materials were not accessible through a data stewardship management policy.

The important components of planning for data stewardship are featured in

- **Toolkit Section 5—Planning: Data Stewardship for Ensuring Data Longevity**.

Planning: Data Quality and Relevance

Defining the characteristics that determine a data set's relevance and utility is critical prior to beginning actual work on a project. Typically, careful consideration is paid to the purpose of the data to be collected and how it relates to the real-world system being studied. The development of a data model can be a valuable means of ensuring that all needed data and related attributes are represented and that all questions related to the project can be answered. A model also can allow for analysis and predictions based on how various data interact through relations that model the real world system being studied.

The data sources also are considered. Should new data be collected or can existing data be used? Understanding the various elements of data, spatial and non-spatial, and the manner in which the data can affect a project's outcome are important in the context of data acquisition. Such things as data representation and scale, resolution, currency, format, geographic datum and projection, data attribute field names, cartographic/graphical output specifications, and data collection methods all affect a data set's relevance and utility. Most importantly, ensuring the quality of a project's data is critical so that a project can achieve its intended outcome. Specifying many of the latter elements is a great first step for ensuring quality. Development of a quality assurance plan can be used to tie all of those elements together, in addition to quality control procedures for validating those elements.

Some identified issues with the example Roan legacy data sets in the context of data quality and relevance include the following:

- Various data needed for representing factors related to the real-world system being studied were not collected.
- Little or no quality spatial data are available, even after substantial improvements in Global Positioning System (GPS) technology and use.
- Many spatial and non-spatial data elements were not defined or were inconsistent.
- Overall uncertainty was common for crucial quality aspects of data such as accuracy, completeness, and consistency.

The necessary components of planning for data quality and relevance are featured in

- **Toolkit Sections 6.1, 6.2, 6.3, 6.6, 6.7—Planning: Data Management Considerations for Meeting Goals and Objectives**,
- **Toolkit Section 7—Suggested Guidelines for Database Modeling and Design**,
- **Toolkit Section 8—Suggested Guidelines for Project Quality Assurance (QA), Development of a QA Plan, and Quality Control**, and
- **Toolkit Section 10—Suggested Geospatial Data Acquisition Guidelines for Quality and Interoperability**.

Standardized Documentation

Standards apply to every area of a project; likewise, applying standards for data set documentation through metadata is necessary so that all of the critical characteristics of the data are captured. Metadata can be thought of as data about data. A standardized form of metadata that everyone is familiar with is the information presented in library card catalogs or even the standardized nutrition labels on packages of food. Standardized metadata help to ensure that the same information will be captured for any data set in a similar format that is not specific to one person's realm of expertise. Standard documentation are viewed as a part of the process of creating a data set, not as an after-thought once the project has been completed. Documentation that is standard, consistent, and accessible also can enhance the utility of a data set several years later, and standard documentation promotes efficiency, reduces data duplication, and prevents many of the problems encountered with legacy data sets.

Some identified issues with the example Roan legacy data sets in the context of documentation include the following:

- Extensive research was required to flesh out important data set details from current and former partners. This was further complicated by the scattered physical locations of both former and current partners, as well as relevant documents, reports, and hardcopy files.
- Details pertaining to the various elements of data mentioned above, as well as relevant project methodologies, were not documented consistently or at all.
- Many data set contributors could not recall undocumented information and details due to the passage of time. On several occasions, information obtained from one contributor contradicted information from others.
- Partners duplicated data sets, for example, through manual data entry or by other means. This likely occurred because they were unaware that the task being carried out had already been done (no documentation existed for those data sets) and because they had no access to those data.

The important components of planning for standardized documentation are featured in

- **Toolkit Section 6.4—Planning: Data Management Considerations for Meeting Goals and Objectives** and
- **Toolkit Section 9—Suggested Tools, Guidelines, and Work Flows for Creation of Federal Geographic Data Committee (FGDC)-Compliant Metadata**.

This neotropical migrant, the Golden-winged Warbler, *Vermivora chrysoptera*, is in rapid decline due to the loss of its historic preferred nesting habitat, open oak savanna, and the disappearance of its adapted habitat, shrubland. Considered a Federal Species of Special Concern, it summers on the Roan near the southernmost end of its nesting range. (Photo by C.S., Robbins, U.S. Geological Survey, Patuxent Wildlife Research Center, *http://www.mbr-pwrc.usgs.gov/bbs/htmsl/h6420pi.jpg*)

The Bottom Line

Exponential change in environmental conditions in the future may require managers to adopt a holistic perspective. Understanding the complexity of effects on the environment, natural and human, will require quality data for use in predictive models. Interdisciplinary science requires precise indicators of cumulative stressors. The foundation of that is data management.

Substantial amounts of money and labor, nearly $70,000 and 3,600 hours of labor, were expended to recover data and put the 12 legacy data sets into perspective for the Roan Mountain Project alone (Burley and others, 2006). The primary funding support for this effort was provided by the USGS National Biological Information Infrastructure – Southern Appalachian Information Node (NBII-SAIN).

The intent of this retrospective analysis is to provide a blueprint for stakeholders to plan and implement future projects in a consistent manner. In the future, building on this assessment, and with proper planning, research efforts and management decisions on the Roan will be more streamlined, efficient, and transparent to multiple agencies involved.

2 Planning: Project Management Fundamentals

Photo ©Dmitriy Shironosov

Successful project management is more an art than a learned science as no project is the same. There are fundamental concepts, however, that can be applied to nearly all projects to help ensure success. According to David Hamil, a certified project management professional with considerable experience in the geospatial industry, project failure can frequently be attributed to four challenging areas (Hamil, 2006)—planning, management, support, and inclusiveness.

Planning and Project Management

Lack of thorough planning can be the downfall of any project with a budget and a finite timeframe. Agencies and organizations cannot afford to expend resources on projects that have little chance of realizing the goals and objectives simply because they were not planned well.

Even with an adequate plan and adequate resources, if not properly managed a project may not meet its intended goals and objectives. Project planning and management of all resources need to (Hamil, 2006):

- Be flexible so that any scenario can be dealt with efficiently and appropriately.
- Be comprehensive so that all areas and activities are addressed.
- Be simple and straightforward. Overly complex plans and processes will become a bottleneck creating misinterpretations and confusion.
- Be agreed upon by all project stakeholders so that support is 100 percent.
- Be enforced. Consistency and completeness are critical because project plans and guidelines will not work otherwise.

Management Support

Without adequate support or sponsorship from high-level management, a project may be doomed from the beginning. Full support is needed from inception to completion to ensure that the resources needed for the project will remain available (funds, staffing, time) (Tomlinson, 2003). As Hamil stated, "A project succeeds only when senior leadership makes it a top priority and broadly communicates their sponsorship across the organization. Organizations respond when leadership emphatically communicates their commitment to a project" (Hamil, 2006).

Long-Term Consideration

All too often the life cycle costs of a project, data set, or effort are underestimated, unaccounted for, and(or) not considered at all during a cost/benefit analysis. This can lead to failed projects, unanticipated costs, and unsuccessful efforts.

Inclusiveness and Participation

Ideally, all stakeholders are engaged in every aspect of a project, including pre-planning, the implementation process, and assessment of results completion (Tomlinson, 2003). Without communication and dialogue throughout, misconceptions may arise on either end, unanticipated considerations may not be addressed, and ultimately inclusiveness may not be achieved.

2.1 Fundamental Project Implementation Tips

The following useful tips are modified from Hamil (2006), National Land and Resources Management Audit and the Australia New Zealand Land Information Council on behalf of the Australian National Government (2003), and Tomlinson (2003).

- Understand the problem, issue, or need before jumping to a solution.
- Define project requirements clearly.
- Always include key stakeholders or anticipated data users in the feasibility process so that all expectations are included.

- Maintain open communication among all project participants.
- Ensure that there are no misconceptions or unrealistic expectations about the data or the project between the customers and providers, if they are two different groups.
- Carefully assess internal capabilities, resources (time, funding), and skills of personnel in determining a project's feasibility.
- If the project is to have a long-term scope, ensure to the extent practicable that it will be sustainable in the context of work and internal capabilities, skills, resources, and high-level management support.
- Resolve any political issues so that they do not become a bottleneck later on.

2.2 Project Planning

The planning strategy described in this section provides a background for technical data management components of planning described in Section 6.

Proper planning is crucial to the success of any project. Planning, however, does not only occur once a decision has been made to move forward with a project. Pre-planning through a feasibility study/risk assessment can be extremely useful for assessing the validity of a project's approach and tentative methodology. Though a pre-implementation planning study may be initially viewed as time consuming and unproductive, the long-term benefits of such an approach will greatly increase the chance a project will succeed by identifying potential risks. Success criteria may include final product quality, adhering to the project timeframe, and staying within the project budget. The following considerations are crucial in making the decision on whether or not to proceed (Hamil, 2006):

Understand the Nature of the Project and Its Goals and Objectives

The first step, the feasibility study, serves to ensure that all project stakeholders are on the same page with their understanding of the project's goals, objectives, scope, problem statement, constraints, and so on (Hamil, 2006; Tomlinson, 2003). The questions for everyone involved are: What is this project all about? Are the goals clear and concise? Any individual differences and discrepancies of interpretations by anyone involved (management, project managers, technicians, field personnel, or end-users) indicate initial misconceptions that ideally are addressed before further planning occurs.

Consider All Alternatives and Determine Whether or Not It Is Appropriate for a Project to Move Forward

The second step is to look at all possible scenarios and alternatives for proceeding, including the possibility of not proceeding (Hamil, 2006). At this point, problems and opportunities for proceeding with the project are identified and weighted, including individual egos and(or) politics. Getting these out in the open early on can help prevent negative factors from overly influencing the decision to move forward with a potential project when in reality it is not feasible to proceed beyond this step (Hamil, 2006).

Once all problems and opportunities have been identified, consensus can be reached on the criteria for an acceptable solution. Factors affecting a consensus may include political, economical, technical, and organizational factors (ESRI, 2006b). All alternatives are analyzed and ultimately the best chosen. Hamil (2006) gives some warning signs that may indicate an unfeasible project:

- Persisting political issues,
- Insufficient staff experience and training,
- Unfavorable cost/benefit ratio,
- High risk and rewards that do not warrant taking on such potentially high risks,
- Continued disagreement regarding the problem, issue, or need among project stakeholders,
- Lack of higher up/management support,
- Insufficient funds and other key resources,
- Insufficient time available to complete the project, and
- Key project stakeholders not participating or unwilling to provide input to the project.

Develop a Broad Implementation Outline Addressing How to Proceed

If the decision is made to move forward, the third step is to develop a broad project implementation plan (Hamil, 2006; Tomlinson, 2003). This includes developing estimates of resources needed, the scale of the project, and major milestones so as to begin giving the project a time framework. This can help keep momentum going by getting the project team thinking about critical implementation issues (Hamil, 2006).

Once the key elements presented above have been defined it is appropriate to move forward with project implementation planning. Much of the feasibility study content is used as

the foundation for the actual project plan. The project plan, composed of a single document or a collection of documents, serves as the "game plan" for management and execution of a whole project (Hamil, 2006). It is helpful for a project plan to be flexible enough to adjust for unexpected events and issues. A project plan, at a minimum, addresses the following elements (Hamil, 2006):

- Project description and overview of goals and objectives;
- Any necessary descriptions of the project management approach or strategy;
- Scope statement including full project schedule, project deliverables, major milestones and associated target dates as well as any other timeframe oriented work;
- Detailed breakdown of work;
- Full budget that addresses all necessary project areas and takes into consideration all the areas of sound data management as identified in Section 6 and as described in this document;
- Documentation of specifications for all areas of sound data management as identified in Section 6 and as described in detail in this document;
- Any relevant performance measurement baselines for schedule and cost; and
- Sufficient time available to complete the project.

Other more specific management plans may be relevant depending on the scope of the project. At a minimum, Quality Assurance and Quality Control procedures are needed. The various other types of potentially relevant management plans can include the following from Tomlinson (2003), ESRI (2006b), ESRI (2006c), and Hamil (2006):

- Risk management plans including key potential risks, any constraints and assumptions, and planned responses or Problem and Resolution (PAR) scenarios;
- Schedule management plan;
- Cost management plan;
- Communications plan; and
- Quality Assurance/Quality Control Plan (QA/QC; an extremely important component of any data-driven project. A QA/QC overview as well as a suggested QA plan approach, an example QA plan template, and suggested example QC procedures are presented in this Toolkit).

Consistent communication among all project stakeholders is of paramount importance. If a project is to fully meet its goals and objectives, all stakeholders are fully engaged in every aspect of a project lifecycle through open dialogue using conference calls, regular updates, and face-to-face meetings. This serves to keep everyone on the same page and helps to avoid project roadblocks such as misconceptions and misunderstandings, which ultimately result in wasted resources and inefficiency.

The Carolina northern flying squirrel, *Glaucomys sabrinus coloratus*, is extremely rare and listed as endangered. Its range has shrunk since the last ice age, and only a few specimens have been documented in its southernmost range in isolated, high-elevation habitat in East Tennessee and Western North Carolina. (Photo from U.S. Fish and Wildlife Service, *http://www.fws.gov/cookeville/docs/endspec/flsqrl.html*)

The Bottom Line

The theory described above is very nice and tidy. In the real world of adaptive management of the grassy balds on the Roan, decision making has a long history of collaboration by consensus among partners on the Roan Mountain Stewardship Committee. The committee includes representatives from two national forests; Appalachian Trail affiliates, including the National Park Service, the U.S. Fish and Wildlife Service, the U.S. Geological Survey, two land conservancies, and two state heritage programs; along with several universities, numerous volunteers, and many contract biologists.

The arduous work of controlling woody plants on the balds is conducted mostly by dedicated volunteers. Federal funding for the Roan has always been limited. Given the circumstances, it is truly remarkable what has been consistently accomplished over 30+ years. Each agency has its own policy, planning, and adaptive management practices; yet collaboration is accomplished in an informal way. Given this cultural precedent, the central challenge in utilizing the Toolkit is to encourage standardization in data management practices, to the extent practicable, among all these partners. And this partnership collaboration will happen when planning for, and implementation of, adaptive management practices become more specific in terms of implementing goals, objectives, and science based on QA/QC evaluation procedures. As shown in figure 1, data management is a central component of project planning and implementation.

It is time to take the universally expressed need for holistic data management to an operational level, particularly during this time of limited resources.

Additional Reference (Burley and others, 2006)

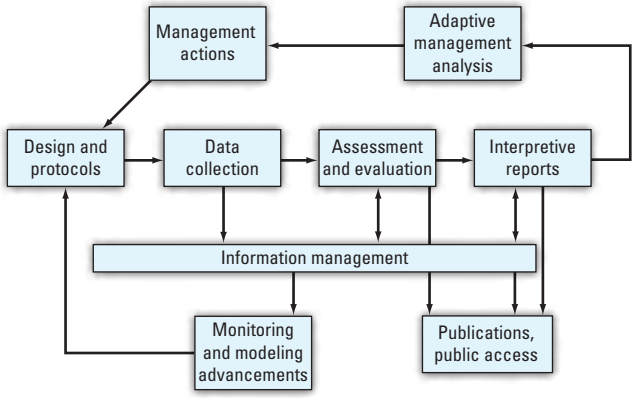


Figure 1. How adaptive management and monitoring functions all relate to and are dependent on effective information managment (Trexler and Busch, 2003).

3 Planning: The Benefits of Well-Defined Standards

Photo by National Park Service

Standards are relevant to all aspects of a project including data collection, data processing protocols, and even data stewardship. Standards, when established at the outset of a project and adhered to throughout a project, can help to ensure a consistent level of detail and data quality by giving confidence and credibility to results (Tomlinson, 2003). Inconsistent data collection, whether within a project or over time by an agency, can result in wasted time as well as increased efforts and cost associated with having to reconcile data sets (if even possible) so that the data can be used in an integrated database.

See Section 9.1 and 9.2 for tools, guidelines, and work flows for creating standardized metadata documentation

Documented and adhered-to standards also help promote consistency in cases of real-world challenges such as project staff turnover. In addition, data collected inconsistently is only as accurate as the lowest resolution format to be included in the database. For example, if 10-meter resolution imagery and 30-meter resolution imagery are defined collectively as an imagery data set, the resolution of that data set is 30 meters. The higher resolution associated with the 10-meter imagery is lost when combined with the lower 30-meter imagery because accuracy cannot be created where it does not previously exist.

Wherever possible, standards are adopted from existing, reliable sources. Experts in many different disciplines have created standards for specific subject and thematic areas. These established standards typically have a baseline minimum of required specified data, attributes, and characteristics, so as to allow integration and interoperability with other thematically similar data and information based on those required minimums. Besides the baseline minimum required elements, there are usually other optional or more flexible elements.

It is generally best to refrain from modifying an existing standard's minimum required elements and specifications to fit a need, unless there are adequate reasons to do so. If a standard is modified, it is no longer standard! It is now a new standard, but it might only be supported by the organization who modified it. This now reintroduces the challenges that standards are meant to address such as interoperability, long-term data set availability, and so forth. However, standards can usually be expanded beyond the minimum required elements that are intended to serve as a "baseline" of common information. The long-term benefit of standards is to allow data to be more easily integrated with and(or) compared with other similar data sets and reused in applications that were not anticipated at the time the data set was created. The older the well-documented data, the more valuable they becomes for trend analysis.

There are multiple benefits of standards for data management (Burley and others, 2006; National Land and Water Resources Audit and the Australia New Zealand Land Information Council on behalf of the Australian National Government, 2003). Standards allow a better understanding of data and, in turn, better information. Better information leads to improved decision making. Adherence to standards also leads to

high-quality data; improved accuracy, consistency, and completeness of data; and reduced data redundancy. As a result, there is increased potential for data integration and interoperability, increased potential for sharing data, improved control over data updating and consistency among different versions, and improved data security and access.

There are a number of common issues and bottlenecks that can arise when standardized methodology is not used or followed. (Burley and others, 2006; National Land and Water Resources Audit and the Australia New Zealand Land Information Council on behalf of the Australian National Government, 2003). If data become scattered within and among organizations, for example, it is difficult to know what data are available and where they are stored.

When data are collected and stored according to different standards, the chances increase for incompatibility among data sets of similar themes and subject matter within and among organizations. In addition, multiple large fragmented databases have limited usefulness in support of the complex science challenges facing our Nation. Difficulties also arise in integrating data from other locations or with other themes. Lack of standards also leads to data that are not compatible with some computer technology.

In addition, if data are collected for a single purpose, without regard to the potential for future uses, the result may be a fragmented database with gaps in coverage or overlapping coverage. This often results in increased costs and lost opportunity and investment.

If data are poorly documented and publicized, they may be of limited use or unavailable to potential users. Moreover, poor documentation of existing data may lead to duplication of efforts in data collection. Uncertainty about the usefulness of a data set in another context may arise, and technical, acquisition, and project details are forgotten. Poor documentation and application of standards also potentially leads to data quality issues in terms of questioning the validity of the study, approach, and methods.

Lack of a standardized methodology also makes it difficult to set up policies for providing access and use of data by others. The concept of "ownership" of data by multiple users and a "silo" mentality are not conducive to partnership or sharing. In short, data and information need to be treated as a long-term investment of resources, much like financial or human resources.

Documentation through Toolkit protocols of several legacy databases collected for the Roan poignantly illustrates the value lost from a lack of standardization of data documentation procedures. Project goals, objectives, hypotheses tested, and data collection procedures were not consistently documented (Burley and others, 2006). Locating the data collection sites was often difficult or impossible. Stakes and landmarks were inconsistently located. Geo-referencing procedures were often not accurate enough to locate data collection sites. Data storage locations were often not adequately documented nor secured; in some cases, data were scattered around the country in part because of concerns by investigators about a lack of standardized data archiving procedures, controlled access, and secured storage areas.

A classic example of lack of standardized archiving occurred when a large number of photographs from plot locations were thrown together uncatalogued in a box. This shortcoming is not limited to data. Research reports and planning and adaptive management policy documents are not archived in a systematic way among agency partners. The concern about archiving has been partially addressed through this project by the re-establishment of the Southern Appalachian Highlands Conservancy Collection for Roan materials at the Archives of Appalachia at East Tennessee State University. As was expressed in the introduction of this document, lack of standardization ultimately dramatically reduces the value of data collected that could instead increase exponentially over time as environmental conditions are anticipated to change dramatically. At a location where more than one-third of the species are considered at risk, the need for standardization of data management is imperative.

For More Information:

- General standards information, information on the benefits of standards is available at *http://www.fgdc.gov/standards*
- Documents and information pertaining to biodiversity informatics and biodiversity data management standards are available at *http://digitaltaxonomy.infobio.net/index.php?Documents*

The NBII-SAIN Roan Mountain Project FY05–06 Final Report provides real examples of the issues that can arise when various aspects of data management are not fully or properly addressed (Burley and others, 2006). The sections of this report listed below reiterate the importance of standards for all aspects of data management through an in-depth evaluation of 12 priority legacy data sets related to natural resource management.

- Summary of the Example Priority Legacy Data Set Issues and the Value Added by Utilization of the Data Management Toolkit, p. 15–21
- Detailed Evaluation of Example Priority Legacy Data Sets by Application of the Toolkit, p. 22–45

4 Planning: Strategic Data Management Principles and Guidelines

4.1 The Elements of Sound Data Management

Project strategic planning ideally includes the rationale for data collection, analysis, and management from the very beginning. Documentation of information sources is key to establishing credibility in the formulation of appropriate project goals and objectives as well as a foundation for selecting appropriate indicators to evaluate environmental conditions and proposed environmental restoration strategies. Key elements that make up a comprehensive data management plan include the following (Burley and others, 2006; National Land and Water Resources Audit and the Australia New Zealand Land Information Council on behalf of the Australian National Government, 2003; Tomlinson, 2003):

- Data supporting rationale for project goals, objectives, and analysis;
- Data modeling and database design;
- Data quality assurance and quality control;
- Data documentation and metadata compilation;
- Data standardization through pre-defined procedures and methodology;
- Data policy and ownership;
- Data storage and archiving;
- Data custodianship;
- Data access; and
- Data security.

See Section 6 in the Toolkit on data management considerations for meeting goals and objectives.

4.2 Strategies for Data Management

Embracing data management as a long-term investment helps to maximize the value of data both during and after the project for which it was collected. Strategic long-term organizational goals and a well-defined data policy help support data management over time. With careful management, the utility and longevity of data can be increased and extended so as to achieve cost economies of scale.

By using established standards and principles, managers can avoid re-inventing the data and information management wheel. This promotes cost savings, efficiency, and interoperability. In addition, taking advantage of existing infrastructure and resources and using existing systems, facilities, and data whenever possible results in significant savings of time and money.

It is to everyone's advantage that all institutional partners support all aspects of data management and stewardship. Robust organizations with the broadest span of interest are often the most appropriate custodians of high-value general use information. A university or college library that maintains archives of data and information of relevance to its region, or a larger-scale national database for a thematically similar type of data, could be an ideal repository for this valuable information. Should the organization overseeing a project fold or lose key personnel, the legacy will survive for the next generation of researchers. The data do not have to cease to exist once the parent organization ceases to exist!

The Bottom Line

Looking to the future, the partners on the Roan would do a great service by devising a holistic strategy for the conduct of science, monitoring, and collaborative adaptive resource management for the grassy balds. It is in the best interest of all to design the data management strategy in this context. The dynamics of species diversity, richness and distribution, and degree and nature of risk is of concern, but a cost-effective strategy to monitor these dynamics with an acceptable level of accuracy has not been devised. A holistic data management strategy will insure that the collective value of data will increase over the long term.

The NBII-SAIN Roan Mountain Project FY05–FY06 Final Report provides real examples of the issues that can arise when various aspects of data management are not fully or properly addressed (Burley and others, 2006).

This first diagram (fig. 2), adapted from Michener and others (1997), provides an excellent visual representation of the highly probable consequences associated with the absence of standard documentation using metadata as an integral part of a project from the beginning.

The second diagram (fig. 2) shows how metadata, if documented properly, can help prevent "information entropy," which is so common in scientific and natural resource management, work by promoting information sustainability.

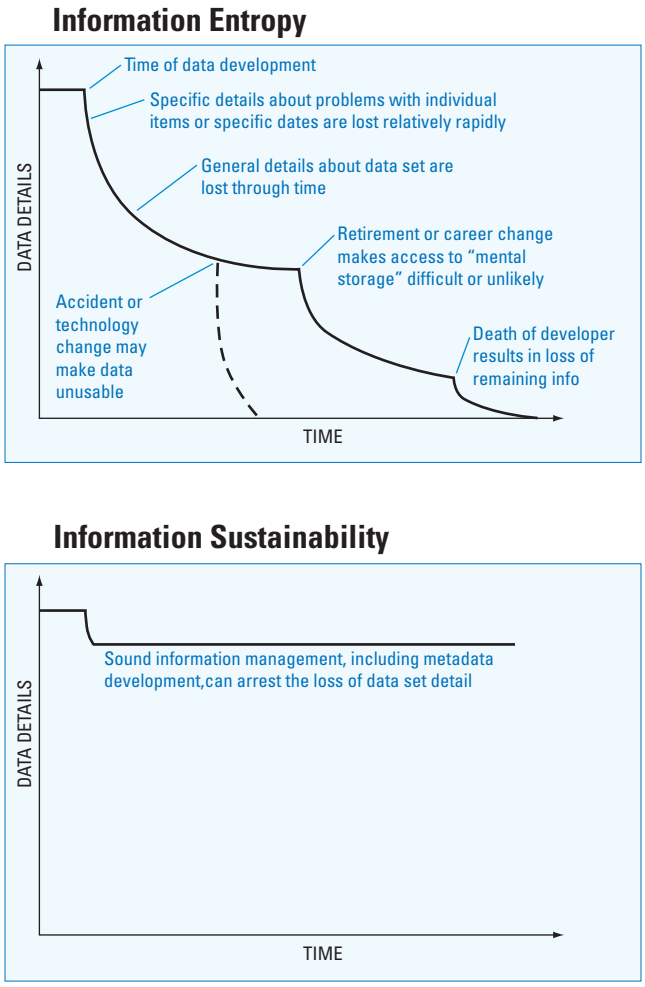


Figure 2. How metadata helps prevent information entropy. (Adapted from Michener and others, 1997)

5 Planning: Data Stewardship for Ensuring Data Longevity

Volunteers collecting data. (Photo by National Park Service)

The life of a data set does not simply end after a project has been completed. The various elements of data stewardship need to be carefully considered and properly addressed before a project has been initiated and typically before budgets have been finalized. This is particularly important if these elements are not defined at a higher organizational level, as it may be necessary to purchase equipment and hardware for infrastructure (such as servers).

All elements pertaining to data stewardship in this section interact with each other and can be addressed in an integrated fashion so that conflicts and compatibility issues do not arise. By addressing these elements, the foundation is built with which the data can have a properly managed home, be made properly accessible, and have proper security characteristics established for maximizing utility in the long term.

Multi-agency partnerships can be a great way to address these elements of data management. Resources, skills, and knowledge can be pooled for the purpose of developing a sound data infrastructure and management plan. By integrating a wider range of resources and expertise by way of appropriate public, private, and academic/educational groups, these data management elements can be addressed in a manner that supports all involved partners in their use of the data contained therein (National Land and Water Resources Audit and the Australia New Zealand Land Information Council on behalf of the Australian National Government, 2003). Depending on the circumstances, these elements also could be integrated into data policy as discussed below.

> Consider your data to be the object of a treasure hunt. Leave as many clues as possible so it can be found and understood, even after you are pushing up daisies.

5.1 Data Policy

Development of sound data policy is the first step towards addressing strategic long-term goals for data management. Data policy serves to establish a broad strategic framework of operating principles, goals, and objectives that can encompass many of the elements included in this Toolkit (National Land and Water Resources Audit and the Australia New Zealand Land Information Council on behalf of the Australian National Government, 2003). Depending on the scope and breadth of coverage, data policy also can encompass other strategic issues such as relevant legal matters, data stewardship issues and custodial duties, data acquisition, and other such issues addressed in this Toolkit (National Land and Water Resources Audit and the Australia New Zealand Land Information Council on behalf of the Australian National Government, 2003). As a high-level framework, data policy is most useful when it is flexible and dynamic. Such an approach will allow data policy to address relevant data management issues in a way that can be readily adapted for unanticipated challenges, different types of projects, and potentially opportunistic partnerships while still maintaining its guiding strategic focus.

> Organizational data policies that are too rigid may serve the immediate needs of their originating organization but can prove to be an inflexible hindrance in other areas.

Organizational data policies that are too rigid may serve the immediate needs of their originating organization but can prove to be an inflexible hindrance in other areas. Data policy can be designed to embrace new data management challenges or opportunistic partnerships. This can aid in the prevention of a silo mentality that can develop and persist within individual agency cultures, a real consideration with multi-agency collaborative efforts. In addition, a flexible structure can be an advantage when addressing unforeseen issues, challenges, and new technology so that people do not get bogged down with "the way things have been done for years."

One possible approach for keeping data policy flexible is to initially design it as minimally as possible so as to serve as a framework that can be applied to many different data management scenarios. That framework can potentially be used as a template for developing more specific project or partnership data policies. Depending on the nature of an organization's work, two or three data policy and subsequent management scenarios could be developed from that framework based on experience and perceived needs. This approach would allow for options as needed that still integrate and adhere to a common overarching framework.

Another consideration when developing data policy is whether there are any other agencies currently involved in similar work at the project or even broad agency level.

Establishing a dialogue with organizations already involved in similar types of work can be beneficial and can help avoid re-inventing the data and information management wheel. Such communication and dialogue could also give way to project partnerships and collaboration for addressing the various elements of data management. One organization may have more experience and resources related to one element of data management and infrastructure, whereas another may have more experience in another area. Strategic relationships built on this premise can benefit everyone involved because organizations can learn from each other and even establish strong synergistic ties for addressing common data management issues.

See Section 9.2 for suggested process steps for documentation of data sets and other resources.

5.2 Data Ownership

Data ownership is an integral aspect of any project that needs to be clearly addressed before data collection has begun. This aspect can and probably should be incorporated into any data policy applicable to a project, agency, or multi-agency effort. The responsibilities and rights associated with data custodianship (addressed below this section) do not necessarily imply data ownership (Biodiversity Conservation Information System, 2000). These elements of data stewardship, therefore, need to be carefully considered and defined with regard to the various partners. Data collection on Federal lands usually requires a permit and provides the opportunity to encourage use of the principles described in this section. Data ownership and the way it is defined for a data set can be contingent on many things, such as (but not limited to) the following from Biodiversity Conservation Information System (2000) and Burley and others (2006).

- Are data to be created through a contractual agreement with another individual(s) or organization(s)?
- Are the data in the private or public domain?
- If multiple individuals or organizations are involved, is a data set to be subject to the respective data policies and ownership standards of those involved with the creation of that data set? Will a separate project or a policy specific to a data set and subsequent ownership criteria be needed?
- Are all individuals or agencies involved with creating a data set to have full legal ownership rights or are full ownership rights to be limited to a subset of those involved?
- Are the data to consist of new data, merged data sets, or modified/value-added data? For the latter two, are there any applicable data policies or ownership rights to be considered in regard to the originating data sources?
- How will those with ownership rights use the data after the work is complete?
- Will those designated as data owners need to address any other data management lifecycle aspects such as maintenance, updating, archiving, security, and release and dissemination (essentially those associated with data custodianship)? Will the designated data owners be different from the designated data custodians? Are all parties aware of the differences (if any) between these two distinctions?
- How long will the data be maintained?
- Are there any copyright or intellectual property rights that need to be considered?
- Are there any liability issues to be considered?
- Will the data be sold in various forms for profit by anyone with ownership? If so, how will profits and(or) applicable royalties be handled?

5.3 Data Custodianship

Data custodianship and the concept of a data custodian are associated with the person, organization, or agency designated with the formal responsibility of overseeing custodianship responsibilities. These considerations may include adherence to data ownership criteria and the archiving and storage, access, security, and quality-assurance procedures associated with data stewardship after a data set has been finalized. Custodians are established to ensure that "important data sets are developed, maintained, and are accessible" within their defined specifications (Biodiversity Conservation Information System, 2000).

Unfortunately, data sets do not proactively manage themselves.

Unfortunately, data sets do not proactively manage themselves. Designating a person or agency to be in charge of overseeing these aspects of data management helps to ensure that data sets do not become compromised. Ideally, these aspects are managed in accordance with the defined data policy, as well as any other applicable data stewardship specifications. Some typical responsibilities of a data custodian may include the following (Biodiversity Conservation Information System, 2000)

- Maintaining adherence of a data set to appropriate and relevant data policy and data ownership regulations.
- Ensuring accessibility to appropriate users through rules, as defined in any relevant documentation such as data policy, data ownership, data security, and data-use restrictions.
- Maintaining adequate levels of data set security, as appropriate. This can vary greatly on the basis of the

data set, usage, and long-term applicability of the data set. For example, the National Institute of Standards and various other Federal agencies publish best practices for securing scientific data.

- Practicing fundamental data set maintenance, including but not limited to data storage and archiving (oversight of maintenance, updating, quality).
- Practicing data set documentation through metadata and any subsequent documentation updating.
- Assuring quality and validating of any additions to a data set through defined relevant data set standards and processes. This may include periodic data audits to spot-check for issues and to ensure quality.

Custodianship probably is best handled by a single agency or organization that is most familiar with the content of the data set and associated management criteria. If it is necessary that more than one agency or organization be in charge of custodianship for a data set, custodianship duties can be broken down and assigned in conjunction with the capacities of the various custodians in terms of practicality, technical resources, and ability, as well as relevant experience. If the data are electronic/digital and to be housed in an existing database, the database system administrator may already be tasked with general duties that overlap with those of a data custodian. Efforts can be made to prevent duplicate work on either end of the scenario where a database administrator and a data custodian are two different people. Some general criteria for a data custodian include the following topics from Biodiversity Conservation Information System (2000) and Burley and others (2006):

- Assigned responsibility for management
- Vested interest and subsequent need for data set oversight
- Adequacy of technical and(or) physical resources
- Adequacy of long-term financial resources
- Data custodianship experience and competency
- No conflicts of interest
- Consensus from other stakeholders as to the appropriateness of a designated data custodian

For the purposes of management and custodianship feasibility in terms of resources (time, funding, hardware/software), it may be appropriate to develop different levels of custodianship service. Different levels of service may be associated with, for example, various levels of importance associated with data sets. Consideration needs to be given as to whether or not any data sets should be designated as core data sets. Core data sets are those that are vital to an organization's or agency's routine operations and projects such that their quality and integrity is of high importance (Biodiversity Conservation Information System, 2000). If any data sets are to be designated at a high priority, custodianship details need to be well defined for those data sets first, before addressing others.

5.4 Data Storage and Archiving

Storing data in digital form is preferred for analysis, transferability, and long-term archiving.

Data storage and archiving address where data will be housed after a project has ended. Data dispersed within and among multiple locations and organizations will ultimately lead to difficulty in accessing the data. This element includes considerations for digital/electronic data and information, as well as relevant hardcopy data and information. Without careful planning for storage and archiving, many problems arise that result in the data becoming out of date and possibly unusable as a result of not being properly managed and stored. Multiple versions of data sets among various partners may result in uncertainty as to which is the most current. Existing metadata documentation also can become out-of-date. In addition, data and related information may be scattered across various geographic locations, updating may become impossible, portions of the data set may be lost, and in general, a data set's utility may quickly diminish. Other factors such as retirement of personnel and(or) job changes, as well as the simple passage of time also affect data utility in this context.

Some physical data set storage and archiving considerations for electronic/digital data include hardware/software, data set format and maintenance, backup and recovery, and procedures.

Hardware and Software

For digital data, what type of database will be needed for storing the digital data? Will any physical system infrastructure need to be set up or is the infrastructure already in place? Will a major database software package be needed along with a geographic information systems (GIS) interface product? What software will be needed to meet the data or system requirements? Will this system be utilized for other projects and data? Who will oversee the administration of this system?

Size and Format of Data Sets

The size of a data set can be estimated so that storage space can be properly estimated. It is typically better to overestimate than under-estimate, particularly where data might be added or updated and the database used for other projects. The data types and formats also can be identified so that no "surprises" in the form of database capabilities and compatibility will arise.

Data Set Maintenance and Updating

Updating procedures for a data set ideally are carefully defined. If a data set is live or ongoing, the procedure may

include such elements as additions, modifications, deletions, and frequency of updates. Tracking multiple versions of a data set is extremely important in a multi-user environment. The number of potential users and simultaneous users can be estimated. If it is anticipated that a data set will undergo major future updates and revisions, it may be helpful to have an "official" version available while the updating process is occurring. One approach is to have a "production" database for the official data set version, a "development" database for the updating, and a "test" database for testing, as well as QA/QC of a newly updated data set.

Database Backup and Recovery

To ensure the longevity of a data set(s), the requirements for backing up a database in case of user error, recovery from software or media failure, and disaster recovery need to be clearly defined and agreed upon. This critical aspect of data management will likely be already defined at an organizational level, but project-specific or data-set specific requirements may be necessary depending on the importance of the data, as well as producer, user, and(or) customer discretion. Besides the mechanisms, schedules, and frequency for backups and any appropriate recovery plans, the types of backups need to be specified and planned. This can include types of storage media (for example, tape, CD, DVD) for onsite backups and even whether or not off-site backing up is necessary. Because backup requirements may differ for data sets, depending on frequency of use, priority, and any other relevant characteristics, it may be beneficial to conduct an analysis on data sets to determine the anticipated changes—"rate of change of information in a database and the acceptable level of loss of that information will strongly influence the backup plan" (Morris, 2005).

Considerations for hardcopy materials include the above elements but in a different context. The physical space that hardcopy materials take up is an important consideration as well as the actual site where the materials will be stored. Implementation of a filing system to maintain organization also is a consideration, along with ensuring that someone with the appropriate expertise is assigned custodial responsibility for maintaining and managing that filing system. Previously organized file drawers can quickly turn into a nightmare if care is not taken to ensure the management and organization of materials over time. Tracking of materials to ensure that they do not become incomplete when someone "borrows" or checks them out also is an important consideration.

Standard Operating Procedures (SOPs)

Procedures for carrying out tasks related to hardware and software maintenance as well as for addressing incident and change management are critical for sustaining data over time. Procedures are not unlike standards for data collection or standards for data processing. The concept of SOPs can be applied to all of the aforementioned aspects of data storage and archiving, as well as data custodianship in general. These SOPs need to be well documented in a readily accessible format so that people are not discouraged from taking the time to understand and use them. Thorough SOP documentation also helps to manage and buffer against staff turnover and similar occurrences that can disrupt a system.

Configuration Management

Good documentation of a system's configuration is useful for sustaining data and information over time and includes such details as hardware specifications, software specifications, versions and updates/patches applied, locations of software (for example, which server does application X live on), network addresses and physical locations of hardware. Placing the information in a readily accessible format facilitates the use of the information for verifying system security relative to software patches installed, or assessing the potential implications of software upgrades relative to other software packages (for example, software version compatibility). Documentation also is useful when staff turnover occurs.

5.5 Access and Security

5.5.1 Access

If current or future partners cannot access or do not know about a data set, does it really exist?

The elements of data access and security are affected by the storage and archiving element described above. Data and information need to be readily accessible to those who need it or those who are given permission to access it (for example, those with data ownership rights, the general public). Without proper access, the utility of data is essentially null.

Some issues to address concerning access to data include:

- Relevant data policy and data ownership matters regarding access and use of data
- The needs of those who will require access to the data
- Various types and differentiated levels of access needed and as deemed appropriate (for example, read only access versus full system administration read/write/delete permissions)
- The cost of actually providing data as opposed to the cost of providing access to data (often as applied to digital data)
- Format appropriate for end-users
- System design considerations, including any data that require restricted access to a subset of users

- Matters of private and public domain in the context of the data being collected
- Liability issues. These need to be included in the metadata in terms of accuracy, recommended use, or use restrictions. Depending on how widely the data are made available, a carefully worded disclaimer statement can be included in the metadata so as to free the provider, data collector, or anyone associated with the data set of any legal responsibility for misuse or inaccuracies in the data.
- The need for single-access or multi-user access and related issues associated with multi-user access systems

If relevant, a Memorandum of Understanding (MOU) or a License Agreement can be created and provided to users if data access needs to be more tightly controlled (National Land and Water Resources Audit and the Australia New Zealand Land Information Council on behalf of the Australian National Government, 2003). Upon signing and returning the agreements, for example to the data custodian or a data owner, the user is provided access to the data but is required to adhere to the conditions of use outlined in the MOU or License Agreement. This approach can be used with both hardcopy materials and digital/electronic data.

5.5.2 Security

Database security, particularly for electronic digital data, is an extremely important consideration today (Morris, 2005; Tomlinson, 2003). The damage that viruses do often makes headlines, and there are numerous examples of organizations and corporations that have had data and systems compromised because of vulnerabilities. Hackers also are a very real threat for any data stored in a database and made available over the web.

A common threat for any web-enabled system is automated software designed to exploit system resources for other purposes through vulnerabilities in operating systems, server services, or applications (Morris, 2005). Actual physical equipment theft is another important consideration. By staying current on new threats from hackers, an administrator can employ available technologies so that a database and its data are not put at risk. Many hackers also target systems simply because they are there and because they serve as an easy exercise in hacking, even if they are not interested in the data on the system (Morris, 2005). Taking control of a system's hardware is a major issue in today's computing environment. Appropriate measures and safeguards can be put in place for any feasible threats.

If necessary, a threat analysis, sometimes referred to as a risk analysis, can serve as an effective means for determining the appropriate allocation of resources (time, funds) for safeguarding a system. A threat analysis is a "comprehensive review and ranking of the risks associated with computer infrastructure and electronic data" (Morris, 2005). Morris quotes Michael Wojcik as saying "…a weighted threat model…can address threats in an appropriate order. The exact metric is debatable, but it should probably combine attack probability, likely degree of damage, and at a lesser weight the effort of implementing defense. And, of course, where it is trivial to protect against a threat, it is worth adding that protection even if the threat is unlikely" (Morris, 2005). Such an analysis might be helpful in situations where several projects and numerous data sets will be stored on a system. The nature of the data also determines what measures to take to ensure security. Sensitive information such as threatened and endangered (T&E) species data and other similarly sensitive and proprietary holdings may require a thorough analysis.

The consensus is that security needs to be implemented in layers and rely on more than one method (Morris, 2005). Several items and methods are available for security, including an uninterruptible power supply (UPS); backups, as suggested in the storage and archiving section above; backup integrity testing and ability to restore from backup; physical access control; limited administrative access to resources on the network (read/write/delete access types); any appropriate technology such as a firewall, sensitive traffic encrypting, maintaining up-to-date software security patches; incident response capabilities; and a full recovery plan (Morris, 2005). Where possible, any implemented security features need to be tested to determine their effectiveness. The ideal scenario for ensuring data security and longevity is a proper computer infrastructure/facility. Such a scenario would include such monitoring devices as previously mentioned, raised floor, 24x7x365 staffing and problem notifications. Though it may not feasible to implement all of these, the security methods used ideally coincide with the sensitivity of the resources at stake, the risks as determined by a threat/risk analysis, and the resources available for safeguarding against identified risks.

The article by Paul J. Morris titled "Relational Database Design and Implementation for Biodiversity Informatics" from which the content of this section was drawn expands on the issue of data security and gives many good tips and considerations (Morris, 2005).

5.6 Web Access Considerations

Many different options are available for making data accessible from the web. The types of data, as well as end-user needs, determine which option is most appropriate. Other key considerations are the level of expertise of the end user, the sensitivity of the information, and the complexity of the data. Another consideration is whether or not the database will be open or secure. An open database is accessible by anyone with access to the Internet and is appropriate for information that is not considered proprietary or that is already in the public domain. A secure database requires a login password or implements similar security. This allows the information to be confined to a pre-approved group of users, for example all members of a research team. Frequently data are provided in a

"read only" format where updates are not allowed or are only permitted through a secure login access method.

A number of integrated international efforts, or thematic "communities", exist for facilitating access to various types of data. One effort, the Global Biodiversity Information Facility (GBIF), is a community of providers that creates a standardized network through which people can make their biodiversity data available to others (Global Biodiversity Information Facility, 2006). GBIF coordinates and integrates thematic data accessibility through a series of National Country Nodes (for example, in the United States the NBII is the U.S. node) with use of its standard communication and metadata protocols. These protocols are designed to help transcend the political and institutional barriers of organizational-specific data. Implementation of these standard protocols for access allows data to be accessed by a large standardized community of participants. Anyone can make their data accessible through GBIF by installing a free downloadable "data provider" package that implements the GBIF standards. Once a data provider package is set up, data to be shared are linked to the provider, and the provider is registered with the community.

A Web Mapping Service (WMS), based on the Open Geospatial Consortium (OGC) Open GIS Standards, can be implemented to make geospatial data accessible by sharing map layers through OGC-compliant services (Open Geospatial Consrtium, 2008). However, a map made available this way is not the data itself but generally a pictorial format. Another format such as a Web Feature Service (WFS) provides the actual data that conveys coordinates and geometry of geographic features. Another possible option is that end users can be permitted to access data through GIS software using an open or secure database connection. This will often require a login password to access the database but may make data oversight more difficult because users have direct access to the database.

- **Distributed Generic Information Retrieval (DiGIR)**
  GBIF uses the Distributed Generic Information Retrieval system as its standard communication protocol. DiGIR is a standardized way of sharing data and information. DiGIR functions as a client/server protocol by allowing users to search across heterogeneous databases from one access point. It implements a common interface that hides the details of the underlying databases by mapping them to a standard representation of the data thereby providing a single, "virtual" collection of data (Specify Biodiversity Collection Software, 2006). This interface, called a DiGIR Portal, functions by keeping the digital addresses of participating providers so that their databases can be searched. Once implemented, a DiGIR Portal provides access to a community of similar data sources, such as biodiversity data through GBIF.

  DiGIR is free and can be an extremely useful way to make biological and species occurrence data accessible from the web. More information on GBIF and DiGIR can be found at the web links below.

  How to make biodiversity data available through GBIF:
  http://www.gbif.org/DataProviders/HowTo

  GBIF standards and protocols:
  http://gbif.nbii.gov/portal/server.pt?open=512&objID=761&&PageID=2142&mode=2&in_hi_userid=2&cached=true

  Information on DiGIR:
  http://digir.net/ and *http://www.specifysoftware.org/Specify/specify/Specify%20DiGIR/index_html#1*

- **Open Geospatial Consortium (OGC)**
  A useful consideration for web-accessible geospatial data is the Open Geospatial Consortium. OGC is an international voluntary consensus industry consortium of hundreds of companies, government agencies, and universities. The goal of OGC is to promote standards and interoperability with web-accessible geospatial data and content through its OGC-compliant specifications (Open Geospatial Consortium). The idea behind OGC compliance is that geospatial data in accordance with OGC compliant standards can be accessed or "consumed" by heterogeneous applications, thus helping to promote sharing, integration, and interoperability. OGC continuously develops and refines its specifications to ensure that they are in accordance with the latest technology and industry advancements.

  More information on OGC can be found at these web links:
  http://www.opengeospatial.org/ and http://www.opengeospatial.org/specs/?page=abstract

- **NBII Geospatial Interoperability Framework (NBII-GIF)**
  The NBII GIF supports the integration of distributed map layers through the establishment of the various standards, protocols, and processes. This effort, described in detail at *http://geospatial.nbii.gov/portal/server.pt*, was established in 2004 to aid NBII partners and collaborators with integrated geospatial data layers, mapping applications and content. The framework consists of a bounding box component, place-name gazetteer, geospatial portal, training modules, and various toolkits to aid in the implementation of the GIF.

The NBII-SAIN Roan Mountain Project FY05–06 Final Report provides real examples of the issues that can arise when various aspects of data management are not fully or properly addressed (Burley and others, 2006). The sections of this report listed below demonstrate the importance of the components of data stewardship through an in-depth evaluation of 12 priority legacy data sets related to natural resource management.

- Summary of the Example Priority Legacy Data set Issues and the Value-Added by Utilization of the Toolkit, p. 15–21.

- Detailed Evaluation of Example Priority Legacy Data sets using Application of the Toolkit

This topic is arguably one of the most important dimensions of data management, while at the same time the most neglected. As expressed in the opening paragraph of this report, there is an enormous redundancy of biological data and, at the same time, incomplete documentation and stewardship of it (Ruggiero and others, 2005; John Mosesso, oral commun, 2006; U.S. General Accounting Office, 2003). Unintentional lack of stewardship also occurs for documents related to policy, planning, research, and resource management practices. For example a report on the Roan regarding the grassy balds management policy for the Pisgah National Forest was marked "not for public distribution" even though it was published in 1987 and amended several times through 1994. After some inquiry, it was determined to make the document available to the public on the World Wide Web, and a copy was scanned and placed on the USGS-NBII Portal in the AT Community—Roan project site, available to all. Several other reports followed.

Providing access to hardcopy and digital materials through an official repository is a viable option for long-term preservation and access. As part of this project, an archive was re-established for documents and data related to Roan Mountain called the Southern Appalachian Highlands Conservancy Collection at the Archives of Appalachia at East Tennessee State University. Several key policy and resource management documents are stored there now, and there are plans to archive many others. Data sets can be stored there as well. In each case, specific arrangements are made as to accessibility. The older the data, the more valuable it becomes, particularly for the Roan populated by several species at risk at the southernmost extent of their range. The Roan is an ideal site for detecting the early biological effects of climate change.

For More Information

The National Park Service Northeast Temperate Network Inventory and Monitoring Program Data Management Page contains example NPS reference documents for such things as computer and system back-up methodology, example cooperative agreement language, and document coordination and preparation among other references:

- *http://science.nature.nps.gov/im/units/NETN/downloads/Plan/NETN_DataManagementPlan.pdf*

—— THIS PAGE INTENTIONALLY LEFT BLANK ——

Data Management Toolkit Part B (Section 6)—Elements of Data Management Overview

6 Planning: Data Management Considerations for Meeting Goals and Objectives

6.1 Data Modeling and Database Design

The utilization of a data model is ideal for any project involving data that is used to model a real-world system (Zeiler, 1999). A data model is a representation of a real-world system (for example, components, linkages) on which simulations and analysis can be run and from which inputs and predictions are made (Breman and Zeiler, 2005). Integrating a model into a GIS by way of a geodatabase can be beneficial in that a geodatabase model can help to bring a physical data model closer to the logical data model that it is based on (Tomlinson, 2003). Through thematic layers and their respective relations in a spatial database, data can be categorized to match the user's view of the world, thereby allowing for robust spatial analysis. In addition, QA/QC can be implemented through a robust pool of existing data "behavior" when using the geodatabase model approach (Danielle Hopkins, ESRI, written commun., 2006; Zeiler, 1999).

Development of a data model is an essential consideration in the planning process; use of a data model helps to ensure that all aspects of the real-world system are represented. It also helps ensure that the goals and objectives will be met by cross-referencing the elements of the real world system with elements of a database. By establishing this framework at the outset, problems associated with data inadequacy that could possibly derail a project can be avoided down the road. Some risks of not properly addressing this step include poorly organized and structured data; incomplete data that do not meet the needs and goals of the project; duplicate, missing, or unnecessary data; inadequate representation of data; and lack of proper data management implementation relative to the data associated with the project (Zeiler, 1999; Burley and others, 2006).

There are three main types of data models; Tomlinson (2003) includes the following:

- **Relational Data Models**
  In this type of model, the data (both spatial and non-spatial) are stored in tables which are logically associated with shared attributes (primary and foreign keys). Individual records are stored in the rows, and the attributes of the records are represented by columns. Use of software such as, for example, ESRI ArcSDE allows for geospatial data to be integrated with non-spatial tabular data because they both can exist in a relational database with this type of data framework.

- **Object-Oriented Data Models**
  The object-oriented data model allows for real-world entities to be more accurately represented than does the relational data model. Objects can be assigned behavior that models their behavior in the real world. Behaviors are methods the objects can perform that mimic real-world functions. With this type of model, objects store information about themselves (attributes and behavior) within themselves rather than in related tables. Objects also communicate with one another by messages that invoke a receiving object's behavior.

  An object-oriented model allows for objects with similar attributes and behavior to be organized into classes, creating a hierarchical nesting structure. Relations also can be used to describe how objects associate with each other within and between classes. A primary feature of the object-oriented model is that data attributes and behavior are encapsulated. Encapsulation means that the data within an object can only be accessed through means that coincide with the object's behavior.

- **Object-Relational Data Model**
  Object-relational data models combine object-oriented data model capabilities with relational database functions. This model essentially meets in the middle of the latter two models with its design and capabilities. This model extends the relational data model by adding an abstract data type that combines the alphanumeric data types used in the relational model. There are advantages to using this model, but there are also limitations because it is a compromise between the two models. One of the main disadvantages is that data are not encapsulated as with the object-oriented data model.

Roger Tomlinson's book "Thinking About GIS" explains these model types further (Tomlinson, 2003). The type of model chosen will be contingent on the needs of the particular project. A useful way to manage the modeling process is through the use of Unified Modeling Language (UML) diagrams. UML is a standardized diagrammatic notation methodology for documenting object models and serves as an excellent standard way to develop and document data models (Object Management Group, 2006;Tomlinson, 2003; Zeiler, 1999). This can improve the modeling process by portraying the model in a standard visual manner, which may allow deficiencies or other problems to be identified that would be

more difficult to catch otherwise. The following web sites have more information on UML, as well as introductions and tutorials for software that can be used for the model design process: *http://support.esri.com/index.cfm?fa=knowledgebase.documentation.viewDoc&PID=43&MetaID=658* and *http://www.uml.org/*

6.2 Data Acquisition

Several methods are available for obtaining data that can be used for projects. These include collecting and creating new data as well as downloading, purchasing, and deriving existing data. Whichever route is taken, duplication of data is to be avoided, especially when the initial thought is to collect or create new data. A key question in this context to be addressed comes from the National Land and Water Resources Audit and the Australia New Zealand Land Information Council on behalf of the Australian National Government (2003):

- Do similar data already exist in a potentially suitable format that could be utilized for the project?

There are fundamental considerations when utilizing pre-existing data. If a potentially useful data set is discovered, several key points are considered. Good metadata documentation (if available) adequately addresses all of the following issues (National Land and Water Resources Audit and the Australia New Zealand Land Information Council on behalf of the Australian National Government, 2003; Tomlinson, 2003):

- The potential data set in question can be evaluated in terms of its fitness for the purpose; this involves assessment of criteria such as scale, resolution, datum and projection, accuracy, and data content. (Section 6.3 of this Toolkit identifies these elements.) Proper metadata documentation provides information about these criteria. The metadata associated with the files and(or) any other relevant documentation may contain information that can be used to determine whether the database is useful to the project.

- Where possible, clarification is obtained from the owner on whether or not the data set is the most up-to-date version. Issues that arise when using outdated or unofficial data sets can lead to confusion, major problems during analysis, and a lack of interoperability and integration with other data sets. Proper metadata documentation and(or) other documentation can provide information regarding these issues.

- It is essential to consider the format of source data. It is helpful to consider whether or not it is worth the time and effort to digitize hardcopy data. Proper metadata documentation and(or) other documentation will provide information regarding these details.

Whether or not new data are to be collected or an existing data set is to be used or modified, the following are some questions from the National Land and Water Resources Audit and the Australia New Zealand Land Information Council on behalf of the Australian National Government (2003) and Tomlinson (2003) that address project data acquisition.

- Has metadata creation/documentation been included/budgeted for?

- What are the anticipated applications or end products of the data?

- Who are the targeted or anticipated end users of the data? Where relevant, have those users been consulted about their expectations of the data and its applicability to the issues to be addressed?

- Have all of the specific data types relevant to the project objectives and project mission been defined?

- Have all of the expectations of the data been defined?

- How do those expectations and the associated data to be collected fit into the overarching project mission?

- What data scale is appropriate for the project?

- Which type of spatial representations (point, line, polygon, raster) will best suit the spatial aspects of the data?

- Have data-capturing specifications (equipment, software, settings, storage) been made? What level of data accuracy is appropriate to achieve the project objectives?

- Have data acceptance criteria and sampling strategies been defined?

- Could data collected for this project be used or modified slightly to satisfy this project as well as other anticipated needs? Whenever possible, economies of scale need to be achieved with data collection.

6.2.1 The Data Clearinghouse Concept

Data and metadata clearinghouses can be a great way to search for existing data. The concept behind a metadata clearinghouse is to create a centrally accessible location on the Internet with access to metadata, data sets, and web-mapping services that can be queried by anyone. One example of a metadata clearinghouse is the NBII metadata clearinghouse. Another example, the Geospatial One-Stop, part of the Federal E-Gov Initiative, allows for data search and discovery through basic search questions such as “What” is being searched for or “Where” is the area of interest (the geographic location). Queries also can be run for special interests or by predefined data categories. In addition to metadata, Geospatial One-Stop provides map services that can be pulled into GIS applications.

For more information about the NBII metadata clearinghouse, follow these links:

- Primary NBII Metadata Information:
  http://metadata.nbii.gov/
- Primary NBII Metadata Search Interface:
  http://mercury.ornl.gov/nbii/
- For more information about Geospatial One Stop, follow this link:
  http://gos2.geodata.gov/wps/portal/gos

Numerous other examples of data clearinghouses can be found on the Geospatial One Stop web site in the predefined data categories.

6.3 Critical Elements of Data

This section presents data elements that need to be considered with regard to proper management of data. Defining these elements (where relevant) contributes to a project's consistency, accuracy, and precision so that a project's objectives can be met. In addition, many of these elements improve the representation of real world systems so as to allow for more accurate and realistic analysis. These elements ideally are taken into consideration during the project planning period and are specifically defined and agreed upon by all parties involved. Establishment of these criteria can contribute to the likelihood of data acceptance during QA/QC processes, thus giving confidence and credibility to data from the outset. The agreed-upon specifications for these types of elements need to be specified in the acceptance criteria of the project QA plan.

6.3.1 Spatial Data Elements

- **Geographic Datums and Map Projections**
  It is essential to specify a geographic datum and a map projection at the outset of a project (Tomlinson, 2003). If project data are to be combined with already existing data in a geodatabase, the coordinate system of the new data needs to be the same as that of the existing data or geodatabase. Careful attention is paid to data that are developed or digitized so that they match relevant existing data. For a project recording new data, this needs to be specified in the QA plan (Danielle Hopkins, ESRI, written commun., 2006).
- **Data Representation and Scale**
  The appropriate spatial data representation can be determined so that the data will adequately represent the types of features/entities being represented (Tomlinson, 2003). Consistent symbology and appearance can be defined for representations of features in a project. This is heavily affected by the scale determined appropriate for the project.

GIS spatial data representations include vector data, raster data, and triangulated data.

- **Vector data**
  Vector data work well for discrete, fine-detail features that require a high level of precision and a precise shape and position, as well as data that requires topological relations. Vector data can be represented by points, lines, and polygons (for representing a "discrete" area).
- **Raster data**
  Raster data work well for modeling classifications or continuous phenomena, both of which can be thought of as thematic data (Tomlinson, 2003). These are either discrete (such as vegetation types across a mountain range) or continuous (such the concentrations of a chemical in an area), respectively. Raster data also can be image data such as imagery (satellite and aerial) that has been scanned and scanned maps, as well as pictures that are well represented by raster features.
- **Triangulated data**
  Triangulated data or TINs work well for modeling and analysis of earth surfaces.

Also, though it is not technically a "spatial" data type, features that do not need to be represented by a spatial data type but are an important aspect of that type, such as owner(s) of a building(s), can be represented as "objects" in tables and joined with a primary key.

The appropriate scale for a project can be determined at the outset so that the data meet the expectations of the users. As an example, for a project with a working scale that can be described as "small" (one that shows moderate detail, such as 1:50,000 scale) it would be difficult to represent monitoring plot locations with polygons. In this case, points would be more appropriate. Both data representations and the scale aspect of a project need to be specified in the project QA plan so that they coincide with any analysis and work to be done (Danielle Hopkins, ESRI, written commun., 2006).

- **Data Resolution**
  Resolution is defined as "the size of the smallest features that can be mapped or sampled at a given scale" (Tomlinson, 2003). Information on resolution is essential for data derived from source data such as aerial and satellite imagery. This is largely contingent on the appropriate scale for a project (see above).
- **Topological Relationships**
  Topological relations between features can play a vital part in the integrity and quality of geospatial data (Zeiler, 1999). Topology addresses how points, lines, and polygons precisely interact and relate to each other in terms of geometric configuration (Zeiler 1999). A lot parcel database, a road network, or a stream/hydrologi-

cal network are all examples of systems that require topological configurations for correct representation. Data integrity rules, topological relation queries and navigation, complex editing, and feature construction are all components of topology that can add richness to data so as to more accurately represent real-world features (ESRI, 2006e). In the context of a GIS, topology is considered if components of the system being modeled are best represented by precise geometric relations. Below are some important aspects of a topology to carefully consider and define when creating a topology in a geodatabase (ESRI, 2006c; Danielle Hopkins, ESRI, written commun., 2006):

- **Participating feature classes**
  Any feature classes participating in a topology must be contained in the same feature data set so as to ensure a common spatial reference.

- **Topological rules**
  These determine the valid relations between features in a topology so as to better constrain modeled features to real-world parameters (Zeiler, 1999). Topological rules can apply to features in the same feature class, features in two different feature classes, or between subtypes of features. Rules exist for points, lines, and polygons. An example might be "lines must not self overlap." Following is a link that describes the various topological roles applicable for a geodatabase: *http://webhelp.esri.com/arcgisdesktop/9.1/index.cfm?id=1680&pid=1677&topicname=Topology%20rules*

- **Feature rank**
  The ranking of features in a topology dictates how they will be affected if a topology is validated. As an example, features ranked the highest (for example, one) will not move at all. On the other hand, features ranked the lowest (for example, five) may snap to other features depending on the rules and the cluster tolerance.

- **Cluster tolerance**
  Cluster tolerance refers to "the minimum tolerated distance between vertices in a topology" (ESRI, 2006e). If a topology is to be validated during QC, the cluster tolerance of the topology will snap together any vertices within the cluster tolerance proximity (for example, 0.5 or half a foot) of each other. Care is taken to ensure that any cluster tolerance used during data production is the same as that specified when creating a topology in a GIS. In addition, any cluster tolerance used is larger than the precision and less than the accuracy of the most accurate feature class in a topology.

- **Annotation and Labeling**
  Annotation is text or labeling graphics on a map used to provide information for the user on places, features, and so forth. Annotation may or may not be relevant for a project, but if there are to be end products of a project geared towards presentations, displays, and other means of conveying information about the project from data, annotation needs to be considered and specified in the project QA plan (Tomlinson, 2003). This helps to ensure the quality of the "look and feel" regarding what is to be labeled. Annotation specifics can consist of Font Type, Font Size, placement-horizontal and vertical orientation, and other similar types of formatting. This helps the viewer or user gain the most information from the visual portrayals of the data.

- **Spatial File Formats**
  The most appropriate spatial file format type for a project is decided and specified at the outset. The ESRI ArcGIS Geodatabase format can provide a very robust and widely supported framework as it provides robust capabilities for accurate representation of landscape features (Tomlinson, 2003; Zeiler, 1999). If working with data already in a format other than the ArcGIS geodatabase format, data can be converted from one representation or format to another either through available software tools or through manual digital enabling using some of the options described in Section 10 of this document.

Utilization of the geodatabase data model can provide a number of benefits, including these from Zeiler (1999).

- High-quality cartographic output and data analysis.
- Robust feature geometry for representing shapes of features being modeled. Features are not limited to generic representation or behavior.
- All data can be stored in an industry-accepted database format.
- A geodatabase can enforce the integrity of attributes through domain and validation rules, which provide for built-in Quality Assurance by attribute domains, validation rules, and subtypes.
- Relationships between objects (non-spatial entities such as tabular data) and features (spatial entities) can be defined.
- The ESRI geodatabase format provides advantages over the ESRI Shapefile and ArcINFO Coverage format besides increased feature modeling capabilities. The latter two formats are not ideal for storage and management of complex relational data that requires such things as data normalization or, for example, one-to-many relationships.

If it is anticipated that a project's data will have concurrent editing needs as well as a need for scalability, the ESRI ArcSDE multi-user data extension to ArcGIS, when used with leading relational databases, is a viable solution for a spatially enabled database. Some advantages of a multi-user geodatabase include the following from Tomlinson (2003) and Zeiler (1999).

- Flexibility for scaling.
- Very large sets of geographic data and tabular data can be managed and large numbers of viewers and editors can be served.
- Data can be served to other applications such as ArcIMS, CAD, and ArcView applications as well.
- Data can be centrally stored and administered.
- Many options for web-enabling data access, as well as other accessibility options.
- Open GIS Consortium (OGC) compliant applications can be built. In fact, a number of such compliant viewers are readily available on the internet.
- Structured Query Language (SQL) can be used to access the tables and rows in the geodatabase.

6.3.2 Non-Spatial Data Elements

- **Metadata**
  Metadata are concise documentation about the critical elements and details of data. Metadata are to be viewed as "living" documentation that is developed throughout the life of a project and data set. It is recommended that the FGDC standard along with the NBII Biological Extension Profile be used for proper documentation of data sets and databases. Section 9 of this Toolkit discusses the FGDC-NBII standard and provides work flow suggestions for metadata creation as well as tips and tricks for working with various available metadata tools. Suggestions for legacy data set documentation are included. This Toolkit also contains information extraction questionnaire-style tools in Sections 11 and 12 to facilitate metadata documentation during the lifecycle of a project.

  For more information on the FGDC-NBII Biological Profile standard, follow this link:

  http://www.fgdc.gov/metadata/geospatial-metadata-standards or http://metadata.nbii.gov/portal/server.pt?open=512&objID=255&mode=2&in_hi_userid=2&cached=true

  For documentation of web resources, documents, and reports, the Dublin Core metadata standard can serve as a potential solution. The Dublin Core standard is a common metadata standard for information resource description (Dublin Core Metadata Initiative, 2009). Dublin Core is a standardized method for documenting such resources that includes fundamental details such as Title, Creator, and Subject, as well as a suggested means of describing those details. Section 12 of this Toolkit provides a questionnaire-style documentation tool and examples of how these details can be documented.

- **Data Currency**
  Data currency addresses the time period of data. For example, a project analyzing the current conditions of a site would likely require the most current data available. If temporal-change analysis is an aspect of a project, the time period of the data is particularly important depending on the timeframe of the work being done. It would not make sense to use data from 1986 if the scope of the project is focused on change in a geographic location from 1988 to 2005. Metadata or the data originator most likely will be able to provide information about a data set's currency.

- **Data Capture Methods**
  If new data are to be derived, created, or collected, the data capture methods need to be specified in terms of horizontal and vertical accuracy, as well as needed data precision, standard classification systems, or methodologies. REMEMBER: data are only as accurate as the least accurate part of it. Section 6.6 of this Toolkit contains information to be considered regarding error in the context of data derived from source data, as well as aspects of data quality for both derived data and collected data.

  REMEMBER: Data are only as accurate as the least accurate part.

  Consideration of equipment and software to be used and the specific settings for that equipment and software are essential aspects to be defined and addressed. Some suggested standard spatial data acquisition guidelines for accuracy and interoperability based on industry and Federal guidelines are outlined in Section 10 of this Toolkit, but as with any project this aspect needs to be properly evaluated in the context of the project goals and objectives. All of these elements need to be addressed and defined in the project QA plan (Danielle Hopkins, ESRI, written commun., 2006).

  Federal accuracy guidelines for maps are defined by the National Map Accuracy Standard (NMAS) based on map scale (U.S. Geological Survey, 1947). The following link provides a USGS web site with a downloadable PDF for the NMAS standard: *http://rockyweb.cr.usgs.gov/nmpstds/nmas.html*

- **Individual File and Data Attribute Naming Conventions**
  File and attribute names can be created in such a manner that they convey the nature of the data. It is helpful to avoid spacebar characters in filenames. However, use of an underscore between words is a generally acceptable way of making filenames more readable. Careful attention is be paid to data attribute naming conventions so that attribute field names conform to a program's rules and restrictions regarding them.

- **Data Processing and Analysis Standards**
  Data processing standards include such things as the software environments and equipment used for data processing and manipulation. Data analysis refers to the computing and analysis methodology used with the specified software and equipment on a data set if an information product is to be developed from the data set. Similar to the data capture methods, the processing and analysis parameters are specifically defined in the context of the project objectives and agreed upon by all parties in the initial QA plan (Danielle Hopkins, ESRI, written commun., 2006). A QA plan framework is discussed in Section 8.1 of this Toolkit.

6.4 Metadata Documentation of Data Sets, Information, and Resources

Although just as crucial as carrying out a project and generating information and data, the documentation of the processes and details associated with the creation of the data and information is often overlooked or disregarded. Metadata ideally are developed throughout the life of a data set from planning through digitizing to analysis and through publication. Such documentation can be invaluable both internally and externally, for example, when trying to explain obscure data field names or when trying to determine the data collection methodology used. It can be too easy to assume that the people involved with a project will always be in that job position and that their memories are infallible. It is in the best interest of organizations to proactively take steps to institutionalize metadata production and maintenance so that metadata is a key component of their data development and management process.

The common excuse for not doing it is that the cost of metadata production is too high or the creation process itself is too difficult. All too often, the costs (time, money, and efficiency) associated with NOT creating metadata—including loss of information with staff changes, data redundancy, accuracy issues, data conflicts due to incompatibility, liability, misapplications, and decisions based on poorly documented data—are not acknowledged or realized until a problem presents itself (Burley and others, 2006). Unfortunately, these realizations often occur after significant time and money have been invested in an effort. It is worth noting that Presidential Executive Order 12906 was intended to address this issue by requiring utilization of Federal metadata standards for any geospatial data collected by Federal agencies.

Metadata are essential for promoting efficiency as well as for avoiding duplication of data. Storing metadata in a reputable clearinghouse (see Section 6.2.1 of this Toolkit) makes a data set "discoverable" for other people. By searching a clearinghouse, anyone may become aware of an existing data set through its metadata that meet their project needs. Proper metadata also allow people to understand the important details of a data set so that they may determine whether or not it might be applicable to their project. This creates the potential for project cost savings, as well as possible collaboration opportunities with other researchers involved in similar work.

6.5 Thematic Data Content Standards and Systems for Promoting Data Interoperability and Sharing

Wherever possible, thematic data content standards are adopted to avoid re-inventing the data and information management wheel. Thematic data content standards promote an established and accepted set of data elements or attributes for a thematic data type or subject that work to hedge against future difficulties of integrating data from multi-source heterogeneous data sets. Thematic content standards typically specify the minimum information to be collected for promoting interoperability, in addition to optional suggested attributes. Users are not confined to just those elements when the project requires other types of data. It is acceptable to add more attributes to the specified minimum. By taking this approach, a project's data are more likely to be compatible with existing data sets of similar theme and subject matter. This can be advantageous for further leveraging of existing resources, as well as for building collaboration among other projects or organizations with the same goals locally or nationally.

An example of such a standard is the North American Weed Management Association's (NAWMA) weed mapping standard for non-native plant species. This is a collectively agreed upon standard developed by a diverse number of Federal and state agencies in the context of a common goal and theme (Stohlgren and others, 2003). As a result, data developed to this standard can be shared and transferred across state and Federal agency boundaries. For more detailed information about the NAWMA standard , what it entails, and what agencies promote it, follow this link: *http://www.nawma.org/*

Another example standard is the Federal Geographic Data Committee (FGDC) Vegetation Classification and Information Standard (Ecological Society of America, 2006a). Signed into effect in October 1997, this is the standard vegetation classification for Federal agencies and their cooperators (Federal Geographic Data Committee Vegetation Subcommittee, 2006). The membership subcommittee includes representatives from 14 Federal agencies, as well as The Nature Conservancy and the Ecological Society of America. Besides addressing core data, it also addresses issues such as data collection, metadata,

scale, management and reporting of vegetation information and includes background information on the standard and other related information (Ecological Society of America, 2006b). For more information, follow this link: *http://www.fgdc.gov/standards/projects/FGDC-standards-projects/vegetation/index_html*

In relation to the latter Vegetation Classification and Information Standard, the VegBank initiative builds on the FGDC Vegetation Classification and Information Standard. The goal of VegBank is to bring together data and information on the same subject matter from a diverse group of researchers and scientists (Ecological Society of America, 2006b). The VegBank initiative consists of a vegetation plot database operated by the Ecological Society of America's Panel on Vegetation Classification (Ecological Society of America, 2006b). As a national effort, its goal is to create and maintain an integrated vegetation plot database based on "three linked databases that contain (1) the actual plot records, (2) vegetation types recognized in the U.S. National Vegetation Classification and other vegetation types submitted by users, and (3) all plant taxa recognized by ITIS/USDA (Integrated Taxonomic Information System/U.S. Department of Agriculture) as well as all other plant taxa recorded in plot records" (Ecological Society of America, 2006b). Its standards and best practices are embodied by the U.S. National Vegetation Classification (*http://www.vegbank.org/vegdocs/panel/standards.html*), though the database is designed to be flexible so that legacy data can be incorporated (Ecological Society of America, 2006b). The system is set up to require registration and establishment of a certified user profile for people wanting to contribute their plot data. This is done so that others who are interested in data know the background and credentials of the contributor and the data. This information can help determine whether the data fit the needs (Ecological Society of America, 2006b).

To faciliate the contribution of legacy data to VegBank, a free downloadable program called VegBranch is available from the web site. VegBranch is a Microsoft Access database interface designed "to import legacy vegetation plot data, modify those data to conform to VegBank standards, migrate the data to the VegBank format, and then prepare the data to be sent to the VegBank server" (Ecological Society of America, 2006c). An interface that users can readily download and use at their own personal computer with widely available software and an internet connection creates a strong advantage for a collaborative program by making use easy. For more information, follow this link: *http://www.vegbank.org/vegdocs/vegbranch/vbr-overview.html*

The advantage of using an established, supported system like VegBank is that all of the data are permanently archived and documented so that longevity is ensured. Data and information contributed can be searched, viewed, annotated, revised, interpreted, downloaded, and cited by others (Ecological Society of America, 2006b). This allows anyone to search for data that might meet a need of a project. For more information, follow this link: *http://www.vegbank.org/vegdocs/history/development.html*

6.6 Quality Assurance and Quality Control

QA/QC are some of the most important aspects of any project. QA/QC can be thought of as a proactive approach for ensuring the usefulness and longevity of a data set throughout the lifecycle of the data set (ESRI, 2006c; Danielle Hopkins, ESRI, written commun., 2006; Tomlinson, 2003). QA/QC also can be helpful with the project scoping process—to determine goals and objectives and how they will be achieved through a project's data. Because of this, a QA Plan is introduced at the outset of a project during the scoping process because it applies to all aspects of a project. The QA plan is very much a "living" document, however, evolving during the project planning and execution stages (Danielle Hopkins, ESRI, written commun., 2006).

6.6.1 Quality Assurance

Quality Assurance is an "all-encompassing management approach combining technical aspects of quality, quantitative and qualitative evaluation methods, and human resources in a system designed to meet the customer's expectations" (ESRI, 2006c). In other words, it is the technical game-plan for the way everything will be addressed in a project. With the approaches and technical aspects detailed, data set error can be evaluated and quantified using the aspects of data quality described below. The evaluation is conducted by comparing the data details to established acceptable levels of error and(or) acceptance criteria identified in the QA plan. The process of analyzing or comparing to the acceptance criteria is done using the QC process methods established in the QA plan.

Establishing a proper QA plan at the beginning of a project, before any work has begun, is necessary to ensure precision, accuracy, consistency, and completeness (Danielle Hopkins, ESRI, written commun., 2006). Simply developing a QA plan is not enough though. QA, like proper metadata documentation, is an ongoing process that must be conducted throughout the whole project. Development of work flows for different areas of the project allows the QA process to parallel the rest of the work being done. Section 8 of this Toolkit further addresses this aspect of QA/QC.

Consistent communication among stakeholders is crucial during the development of a QA plan (ESRI, 2006c; Danielle Hopkins, ESRI, written commun., 2006; Tomlinson, 2003). In some cases the end users may not be sure what expectations are appropriate for certain data, or they may have misconceptions about what type of information they will be able to gain from the data. This can occur if proper analysis and understanding of the problem or issue at hand has not been accomplished. In such cases, the end users may be unsure of their objectives and goals as well. Establishing the goals and objectives of the project and cross-referencing the needed quality aspects of the data with the goals and objectives to ensure that they are met though the use of the QA plan can help to resolve

such uncertainties, if they exist. A QA plan helps to shape the project and can help hedge against problems that may pop up later if changes in methods are made or if initially vague goals and objectives suddenly change or take a definite shape after work has begun.

Some of the important issues to be addressed outright through the QA plan include several other elements laid out in Section 6 of this Toolkit. By specifying how these will be addressed, a thorough QA plan helps to ensure the long-term investment approach with regard to data. The considerations in Section 4 of this Toolkit can be used to guide the elements in Section 6 in the context of defined goals and objectives. Acceptance criteria are another important aspect of a project that a QA plan addresses. Quality Control (QC) also needs to be addressed in the QA plan. In general, any aspect of a project that can be addressed in the QA plan probably should be addressed there. This in turn leads to an understanding of the way quality issues will be handled so that errors can be measured and tracked. As a result, the final product of a project is more likely to meet the identified expectations and requirements.

6.6.2 Quality Control

Quality Control refers to the "individual task or set of tasks performed at a given level of the production process that is measured and aimed at ensuring integrity of a product, output, or action" (ESRI, 2006c). For example with GIS data, QC can entail performing visual checks on data, as well as using a specified sampling technique to test a subset of features in a data set. Quality control is important because it evaluates the product against the specified quality acceptance criteria in the QA plan.

The process of QC typically involves a two-pronged manner for discovering error (ESRI, 2006c)—exploratory and confirmatory approaches.

- **Exploratory approach**
  The exploratory approach occurs when data are examined for anomalies or things that do not follow specified guidelines. An example is a query of data or of a particular data attribute for values that do not fall in an acceptable value range. The query will select values that are outside the specified value domain, essentially exploring and discovering those anomalies.

- **Confirmatory approach**
  The confirmatory approach for discovering error entails validating data against specified rules and criteria in the QA plan to see which things do conform to the established guidelines. A query searching for data within an acceptable value range will select only the data that fall within that domain but not values that are outside the specified value domain. This differentiates between those that do conform and those that do not.

These two approaches can be used with a number of methods including visual checks, statistical methods, automated checks, and metadata review. Simple visual checks can be an effective means of identifying error with the exploratory approach. With spatial data, examples of error that can turn up during an exploratory visual search are incomplete features or polygons that are not closed. With tabular data, red flags during an exploratory visual search might indicate incomplete records or missing attributes. An example of the confirmatory approach can involve automated checks on a data set such as a query to find the attributes with values that fall within a specified value domain. Other examples include verifying project data adherence to the database schema or data model for the project and using an automated procedure to search for polygon features that have less than a specified amount of area (for example, polygon slivers in a subdivision parcel data set).

6.6.3 Aspects of Data Quality

Aspects of data quality, like many of the elements of data management need to be addressed with projects. ESRI (2006c) lists the following aspects.

- **Completeness**
  Completeness refers to whether or not any features or data are missing, based on specifications and source materials.

- **Consistency**
  Consistency refers to the level of uniformity of detail throughout data and the validity of attribute values. An example scenario that would result in inconsistent data is the mixing of different types of GPS receivers in a project. Use of default values, validation rules, and value domains are examples of ways to enforce and evaluate data consistency.

Logical consistency and physical consistency describe the two kinds of consistency with data. For example, physical consistency can pertain to making sure that all roads in a city connect to an interstate with an on-ramp in a GIS roads layer. In the real world, a city road cannot directly connect with the interstate. Logical consistency can refer to whether or not all polygon boundaries close in a GIS layer.

- **Accuracy**
  Accuracy refers to the quality of being near to an actual value, for example, a measure of correctness.

- **Precision**
  Precision refers to the quality of being reproducible or the exactness of a measurement. In the example case of a GPS unit, it refers to the performance of the unit and how well it records a location. With data values, precision refers to the number of significant digits associated with the value.

In addition, data resolution and currency can contribute to overall data quality depending on the scope of a project. Resolution is a fundamental consideration for imagery data along with scale. Currency refers to the time period of the data. Both of these are mentioned in Section 6.3 of this Toolkit.

6.6.4 Sources of Data Error

Error can result from several sources in a project. Some of the most frequent sources include source data used, data collection, data entry, as well as data analysis. Two kinds of error are typical: systemic error and random error. Systemic error can be thought of as consistent error that repeats in a similar way that, for example, can be a result of a faulty aspect of a data production work flow. Random error, on the other hand, is just that: random and unpredictable. The source of random error can be much harder to pinpoint.

Below are some examples of error types that can occur from the respective error sources. The examples are from ESRI (2006c) and Danielle Hopkins, ESRI, written commun., 2006).

- **Source Data**
  Source data can result in error in a project because any existing error in the source data will be brought into the new project if used as is or propagated into data derived from the source data. These types of error typically are related to error associated with the various aspects of data quality described above. When using source data, it is useful to acquire metadata for that data, as metadata will have valuable details that can be important during the QC process. Maintaining a copy of the source data is also critical if any visual QC is to be done on derived data.

  Spatial and temporal error from this source can occur relative to the time frames of various source data and differences in the scales of the sources when map scales are mixed. The date of a "borrowed" data set and any related source material for it most likely are available from metadata and, because the date may be needed, it is noted in new metadata created for any derived data. Scale is critical in that generalization is common with small-scale data. Whenever possible, compile data from large-scale sources as compared to small. Attempting to "blow up" small-scale data for a larger scale map can cause many quality problems. Also, if small-scale data are blown up for use with larger-scale data in a new data set, the different levels of generalization in the data become obscured; consequently, the validity of the data could become compromised in the context of its intended purpose. If various map scales are used to contribute data to a data set, the data set's scale is only as accurate as the smallest of the source data scales.

  For data generated from a source (for example, digitized), indistinct or overly generalized boundaries of features in the source can result in error in the newly derived data, including the misrepresentation of the feature's area or extent. If the source data are not carefully collected and created, an error present in the source data will be propagated into the newly derived data. The mixing of geographic datums and projections also can result in error. For an example of the issues associated with two different datums, see Section 10.1 of this Toolkit.

- **Data Collection**
  Errors in data collection can result from human error and(or) such things as equipment malfunction or improper equipment calibration. For example, if a team of researchers is collecting plot data using strictly ocular estimates, accuracy and precision can be difficult to quantify. Completeness, in that context, could refer to whether all the intended plots were actually sampled. Consistency could be attributed to how well they followed an agreed-upon classification standard, as well as whether all of the researchers were using the same type of GPS unit to collect spatial data.

- **Data Entry**
  Error from this source is generally a result of human error. For example, it can occur when digitizing features from source data or when digitally entering plot sampling data that were recorded on a field sheet.

- **Data Analysis**
  An error in data analysis likely stems from the first three error sources which, in turn, are all affected by the identified aspects of data quality. Any error resulting from the above sources will be carried over into the analysis, then to any resulting output. The aspects of data quality synchronize output from data analysis with the data quality aspects of the data acquisition sources and methods. It is not possible for the output of analysis to be more accurate or precise than the accuracy level or precision of the input data. For example, if an analysis output value that has seven significant digits is generated from two input values of two significant digits each, the output value with such a level of precision, in this context, is not correct in that precision cannot be created where it does not exist. The general rule of thumb is that the result is only as good as the least accurate, complete, consistent, and(or) concise data set.

  Precision cannot be created where it does not exist.

6.7 Cartographic Display and Maps

A map can be thought of as "a graphic representation of all aspects of the physical and cultural environment". A map serves as "a graphic device for storing and communicating information about the earth's surface, ranging in scale from a land parcel to the entire world" (Will Fontanez, University of Tennessee, written commun., 2006). Maps can be an extremely efficient means of conveying vital information derived from data to an audience. However, if not done properly, maps can be inefficient and problematic when data are not represented well and if poor technique is used.

In order to efficiently convey information, various elements of map design are taken into account such as the following from Will Fontanez (University of Tennessee, written commun., 2006).

- **Clarity and Legibility**
  Items on a map should be clear and easy to read.
- **Contrast**
  It is vital for map objects to be distinguishable. Various items can be clearly distinguished from one another. For example, interval representations of population can be shown by graduated symbology. The symbols can be sized so that they can be easily differentiated one from another.
- **Figure/Ground**
  The information content of the map stands out from the background.
- **Balance**
  This refers to the positioning of components in a logical fashion.

Factors that influence a map design, or controls of map design, include the following from Will Fontanez (University of Tennessee, written commun., 2006).

- **Objective**
  The objective drives the purpose for which the map is being made.
- **Audience**
  The audience is the target group for which the map is being made. This aspect is particularly important in that, for example, overly technical thematic maps would likely be information overload for the general public.
- **Reality**
  Reality is the real world factor, or the physical and cultural dimensions of the earth that cannot be changed.
- **Scale**
  Scale represents the earth to map relationship, (for example, 1:24,000 is the standard scale for USGS topographic maps).
- **Technical Limits**
  Technical limits refer to the techniques, equipment, and requirements available for producing maps.

Maps contain a minimum number of items that help the audience better understand the information being conveyed, such as:

- **Title**
  A map needs a title describing the geographic area, topic, date.
- **Legend**
  A legend is needed so that audiences know what the symbology of the map means.
- **Source**
  A map includes a statement of the source of the data (for example, the project, data set, agency/organization).
- **Directional Arrow**
  A map needs a directional arrow, as the North direction is not always intuitively obvious and should not be assumed.
- **Scale**
  A scale bar is extremely helpful, particularly with the prominence of dynamic geospatial data that can be used to produce maps.
- **Date**
  The date that the map was produced is included, as well as the date of the data portrayed in the map.

Scale is essential to consider in cartographic output because scale affects the level of detail portrayed in the map. An example of a bad map in this context would be a medium-scale map of 1:75,000 that displays locations of meter by meter botany monitoring plots. In such cases as this, generalization would have to be used. Generalization with maps is "the process of adjusting map content in order to provide as useful and recognizable a representation of the real world as is feasible within the map's limits and scales" (Will Fontanez, University of Tennessee, written commun., 2006). Section 6.6 of this Toolkit addresses aspects of data quality and sources of data error that are applicable to cartographic output.

Generalization techniques for data in maps of a smaller scale might entail the following considerations from Will Fontanez (University of Tennessee, written commun., 2006):

- **Selection**
  Deciding which map features should be displayed
- **Simplification**
  Smoothing or generalizing linear features
- **Classification**
  Putting map items in categories to reduce the complexity of the data being portrayed

- **Symbolization**
  Using quantitative and qualitative symbols to represent map items

- **Displacement**
  Slightly shifting or exaggerating features in order to make them legible

Efforts can be made to specify during the planning period the type of symbology and the look/feel that the maps should have so as to utilize a standardized approach. Criteria like these can obviously be adjusted and modified as needed, but by establishing specifications at the outset, the quality of information will be increased in terms of completeness and consistency. Also, the ability of audiences to get the information they need from the maps will be greatly enhanced.

6.8 Demonstration of Example Project Data Sets—Areas of Data Management

The NBII-SAIN Roan Mountain Project FY05 Final Report provides real examples of the issues that can arise when various aspects of data management are not fully or properly addressed (Burley and others, 2006). The sections of this report listed below reiterate the importance of the areas of data management identified above through an in-depth evaluation of two priority legacy data sets related to natural resource management.

- Summary of the Example Priority Legacy Dataset Issues and the Value-Added by Utilization of the Data Management Toolkit

- Detailed Evaluation of Example Priority Legacy Datasets using the Toolkit

6.9 Additional Data Management Information Sources

- The National Park Service (NPS) Northeast Temperate Network Inventory and Monitoring Program Data Management Page. This site contains example NPS reference documents for such things as basic database specifications and best practices, database strategies and naming recommendations, spatial data specifications, and data mining among other references, available at *http://science.nature.nps.gov/im/units/NETN/downloads/Plan/NETN_DataManagementPlan.pdf*

—— THIS PAGE INTENTIONALLY LEFT BLANK ——

Data Management Toolkit Part C (Sections 7–12)—Example Approaches to Specific Elements of Data Management

7 Guidelines for Data Modeling and Design

Unless otherwise cited, this section draws its conceptual, logical, and physical modeling approach from Breman and Zeiler (2005) and Zeiler (1999).

The process of developing a data model can be extremely valuable in establishing an understanding of the project's data framework. A data model can help to determine the types of data and associated data attributes needed for meeting a project's requirements. This section attempts to present a general approach to modeling that can be used for projects with spatial and non-spatial data. As with any project, though, more detail or more specific techniques might be needed for various data identified as relevant. The approach presented below incorporates aspects of relational database models, as well as capabilities associated with object-oriented models. The aspect and features of the ESRI ArcGIS geodatabase data model, an object-relational data model (see Section 6.1), are integrated into this approach so that spatial data can be represented (Clint Brown, Environmental Systems Research Institute, Inc., oral commun., 2005; Cunningham and Silverstrand, 2005; Tomlinson 2003; Zeiler and Arctur, 2004; Zeiler, 1999).

This section is divided into five parts, and within those parts are various steps. Elements and notes about this process probably should be documented with metadata and other appropriate methods so that they can be referenced for future work.

7.1 Conceptual Model Design: Capturing the User's View

The beginning step in the modeling process results from a foundation built on extensive communication between the data producers and the end users or customers. The focus is on keeping the over-all model and data structure as simple as possible, while still adequately addressing project participants' business rules and project goals and objectives. The main objective is to identify the major data entities and the relations among data at a high level (Data Warehousing and Business Intelligence, 2006).

Project goals and objectives need to be identified and understood from the outset. The types of data and information outputs associated with these goals and objectives need to be clearly identified. The relevant functions and rules of the organization/business also are identified. A general broad example of such a business function could be "to maintain the grassy balds of Roan Mountain from encroachment of woody plants" (Burley and others, 2006). For those relevant functions, relevant activities associated with each function are identified and described. An example of such an activity might be "coordinating volunteer briar-cutting efforts" (Burley and others, 2006). In addition, whether a particular element or function creates or requires data needs to be analyzed. An example of this might be "area of bald cut by volunteers" which would produce data (Burley and others, 2006). This process can also help to differentiate between data of interest and background data; data of interest will typically receive more detail in the modeling process than background data.

For each function, all of the types of data needed to support that function are identified in order to make that particular function operational. Modeling of key data flows can be improved by identifying providers and consumers of data and information. The functions can be examined while keeping the following question in mind: With whom or what does this function interact and what is the nature of this reaction? The likely sources for each type of data also are identified. Data that flow in are often the responsibility of another function but may be received from an external entity or another type of source. During this process, it is also important to address instances of data duplication so that they do not become problematic later on (known as data normalization). Normalization helps promote a sound physical database design. Difficult instances of duplication are flagged for later review if they cannot be addressed immediately.

Once all of the anticipated types of data needed to support all aspects of the identified functions and objectives are identified and described, the data types can be grouped into general thematic layers or groupings (for example, roads). In addition, spatial data elements such as the appropriate coordinate system, as well as other Critical Elements of Data items identified in Section 6.3 above, are addressed for the entities grouped together

7.2 Logical Model Design

The next part of the modeling process is implementation of a logical design based on the conceptual design. This involves identifying the specific "entities" that will make up the logical data set groupings, as well as identifying the appropriate representations and relationships between them (Data Warehousing and Business Intelligence, 2006). These "logical" elements in an object-relational model consist of entities

(features and objects), classes of entities, and attributes that can further define the entities (Data Warehousing and Business Intelligence, 2006; Tomlinson, 2003; Zeiler, 1999). Entities in a data model serve to represent real-world entities such as, for example, individual briar patches on a mountain. These may include, for example, geometric representations with attributes or alphanumeric/tabular information only. The data should be described in as much detail as possible during this process. Attributes or characteristics of those example objects/entities could represent, in this case, the density or height of those patches. A class or a similar group of these objects might be "Roan Mountain-Round Grassy Bald briar patches."

Establishing the data structure through the relations between entities is accomplished during this stage, including the relations between spatial and non-spatial (tabular) data. Relevant entity behavior should also needs to be developed.

7.2.1 Establish Entities and Relations Among Entities

This step involves identifying and describing entities and describing the relationships that exist among them. Unified Modeling Language (UML) diagrams can be used during this process to organize and develop the model (National Land and Water Resources Audit and the Australia New Zealand Land Information Council on behalf of the Australian National Government, 2003; Zeiler and Arctur, 2004). Ideally, all parties contribute during the modeling process, as their input is invaluable for the definition process as well as for validation of a model (Object Management Group, 2006; Zeiler, 1999).

The goal of this process is to identify all the relevant entities that make up the detailed logical groupings identified in the Conceptual Design phase. All of the relevant entities that will make up each of the groupings are now identified. One technique for establishing entity behavior is capturing and articulating behavior through a series of statements or sentences. Nouns typically will represent entities while verbs demonstrate relations between entities. Some examples of these types of statements include:

- Statements that describe a sub-classification of entities. An example might be "A driveway is a type of road."
- Statements that describe an entity. An example might be "An on-ramp eases merging vehicles onto the interstate."
- Statements that describe structural relations between entities. An example might be "An on-ramp connects to two types of roads."
- Statements that describe the aggregation of entities into a new entity. An example might be "An interstate is composed of lanes."

Once a list of statements has been established that are believed to be nearly complete, the various types of relations between entities can be developed from the statements. Attributes for individual entities also can be specified at this stage. Specifiying primary and foreign keys for entities helps to establish the relations among entities. Entitiy behavior, validation rules, and relation types also are essential for model design. Behaviors are the "methods, or the operations, that an entity can perform" (Tomlinson, 2003). Behavior for attributes within the geodatabase data model can include valid attribute domains (constraints on attributes, such as valid value ranges), subtypes of features, simple and composite relations, and validation rules for feature and attribute integrity (Cunningham and Silvertrand, 2005; Gillgrass and McGrath, 2005; Zeiler, 1999). Validation rules can entail constraints on the cardinality of relations between origin and destination components, as well as connectivity rules for geometric network features. Composite relations can be used to ensure the integrity of data associated with dependent relations. When an entity is deleted from an origin class, the related entity in the destination class will be deleted as well so that "orphans" are not left behind. If necessary, custom behavior can be implemented through programming.

Also, topological relations among entities can be useful to model how features share geometry. This can be done with planar topology and geometric networks. A planar topology shows the precise relation among features, for example, parcels in a subdivision interfacing with each other correctly. A geometric network is useful for modeling transportation networks or utility networks, for example, and aids in the capture and implementation of the functional logic behind the manner in which various parts of those types of networks connect and interact in the context of the full system being modeled.

The arrows below demonstrate how the concepts of a standard database and a geodatabase relate to aspects of entity relationships and data behavior (Cunningham and Silvertrand, 2005)

Database Management system		Geodatabase
Check Constraint	→	Domains/Subtypes
Primary/Foreign Key	→	Relationship class

UML diagrams can be very helpful with the logical design process because diagrams allow for visualization of entities and the relations between them.

7.2.2 Entity Representation

Classifying entities by their appropriate representation is the next step in developing a model. Classifications associate with the types of feature classes (spatial) or object classes (non-spatial) that will be created in a geodatabase (Gillgrass and McGrath, 2005; Zeiler, 1999). Some entities may be mappable (such as a building), whereas others may consist of tabular data (the respective owners of the buildings). Some considerations that can flesh out the aspects of scale and appropriate spatial/non-spatial data representation and elements are discussed in Section 6.3 of this Toolkit. The questions below can be used to determine whether or not an entitiy is modeled.

- Can or should the feature be mapped?
- Is the feature discrete (example, a building) or continuous (example,: terrain data)?
- What is the appropriate feature/geometric shape?
- What scale is appropriate for display and(or) what scale is appropriate for the work to be done?
- Will the feature require any labels or text-related attributes (annotation)?

Features that cannot be represented geometrically are specified as objects for inclusion in an object/tabular class table (such as owner(s) of a building).

To reiterate, if constant communication is maintained throughout this process among all anticipated users of the data, it is likely that nothing will be overlooked.

7.2.3 Evaluation of the Model in Progress

Constructing data models is an iterative process, so some evaluation of the model is done at this point. The end-user's business rules or functions are used to validate the logical model, as well as the relevant user's requirements for updating, accessing, and entering data (Morris, 2005; Zeiler, 1999).

Also, checking the database in the context of normalization is necessary. Normalization is generally composed of three minimum phases or "forms", although the first two phases typically are thought of as the most important for achieving a functioning database (Morris, 2005).

- **First Normal Form**
  Individual fields in a row contain only one piece of information or "concept," and no information or "concepts" repeat in a row.
- **Second Normal Form**
  Where a Primary Key is to be something other than a unique index of numbers (such as a catalog number), a Primary Key cannot contain redundant information. As a result, multiple rows cannot contain information about the same entity because the Primary Key catalog number associated with that entity would have to appear in multiple rows. This repetition of the Primary Key would, in the context of a database, prevent that Primary Key from being unique because it would be required in more than one row.
- **Third Normal Form**
  A database in third normal form contains no repeated information in any field except for foreign keys.

Some further key criteria to evaluate a logical design against include:

- Does the logical data model represent all data without duplication?
- Does the logical data model support relevant organizational business rules?
- Does the logical data model accommodate different views of data for distinct groups of known or anticipated users?
- Are project goals, objectives, and organizational requirements fully addressed?
- Are geographic features appropriately represented and organized?
- Are tabular data appropriately formatted and organized?

7.3 Physical Model Design

The logical data model, during this step of the modeling process, is now converted into database elements during the physical design stage. In the case of a geodatabase, the model is further developed by determining how entities will be organized as feature classes and object classes (explained below). If a pilot or test database model is developed, it can be used to evaluate the effectiveness of the model (Breman and Zeiler, 2005). The "logical" elements identified above, such as objects, attributes, and classes, are cross-referenced to the physical database elements, such as rows, columns and individual fields, and tables (respectively) in the database schema.

7.3.1 Pilot/Prototype Test Model

Implementation of a prototype or pilot database to test, review, and calibrate the model can be useful in ensuring that the model will serve its intended purpose. In addition, work flows can be developed during this process for the building and maintaining of each data layer for updating purposes later on. Though the process of testing a pilot database may be viewed as time consuming, in the long run it is a check to ensure that the model is sound and that all needed data are represented appropriately and adequately. The risk and subsequent unforeseeable problems that could result from not fully carrying out this phase of the modeling process could cost more time and resources than actually taking the time to go through the process correctly the first time.

7.3.2 Entities to Features and Objects

When working with geographic data in a geodatabase, entities are assigned to feature classes and(or) subtypes and object classes (for tabular data). A feature class consists of geographic features with the same geometry type, the same attributes, and the same spatial reference (ESRI, 2006e). An object class consists of non-spatial data of the same type or class (ESRI, 2006e). Groups of related feature classes can be

organized into feature data sets; the grouping of feature classes is an organizing method that enforces a common coordinate system for the features contained in the feature data set. The common coordinate system can be used to improve the organization of entities that have been cross-referenced to features in the context of the data model. Topological roles are also a consideration when grouping related features as any topologically related features must be in the same feature data set. If features are to be part of a geometric network or a planar topology, they need to be in the same feature data set.

Geodatabase performance can be affected by the decision of whether or not to map an entitiy to a feature class subtype or an entire feature class. Classing related entities as subtypes in one feature class helps improve overall performance. However, all entities in the same feature class will have the same attributes, so all subtypes in a feature class will be constrained by this. For example, a Roads feature class might be made up to five subtypes of roads. All of these subtypes will have the same attributes. Considerations for when to create separate feature classes include the following (from Breman and Zeiler, 2005; Cunninham and Silverstrand, 2005; Zeiler, 1999).

- If varied access privileges for groups of features are needed, the same access privileges are granted to the entire contents of a feature class.
- In a multi-user geodatabase, if a subset of some specific features are to be accessed through versions, all features in a feature class are accessed the same way.
- If the feature attributes are to be different from other related features, all features in a feature class must have the same types of attributes.
- If a feature is to have distinct customized behavior, all features of a feature class must have the same behavior.

7.4 Data Dictionary

It is very helpful to develop a data dictionary once the model has been developed so as to thoroughly document the types of data included in the model as well as the associated data attributes and behavior. A data dictionary is a "catalog or table containing information about the data sets stored in a database. In a GIS, a data dictionary might contain the full names of attributes, meanings of codes, scale of source data, accuracy of locations, and map projections use" (ESRI, 2006e). It is also very useful to incorporate a data dictionary into a data set's metadata as well. An excellent free tool called Geodatabase Designer that can assist with this is discussed below.

7.5 Pilot Project Test Database

Once a data model has been established, it can be beneficial to test the model (depending on the size, perhaps small portions of the whole model or simply an adequately representative subset of it) to see if it functions as intended. A pilot project test database also can be a good way to evaluate the effectiveness of a draft QA plan.

Sample data models are available for a variety of industry-specific areas including biodiversity, forestry, groundwater, and base-mapping (Zeiler and Arctur, 2004). Through collaboration and input from respective sector leaders, including the academic sector, the template data models available for download can be used to promote the concept of standardization for interoperability in the GIS and spatial data realm. Design templates and application case studies also are available. These are good examples of models. The sample models, along with the design templates and case studies, are helpful with various aspects of a project during the modeling process. They also provide insight into what a UML diagram or a data dictionary might include. (Available at *http://support.esri.com/datamodels*)

More standard methods and best practices for working with geodatabases and data models can be found at this link: *http://support.esri.com/index.cfm?fa=knowledgebase.documentation.viewDoc&PID=43&MetaID=571*

This site provides overview information on UML design software for modeling: *http://support.esri.com/index.cfm?fa=knowledgebase.documentation.viewDoc&PID=43&MetaID=658*

An extremely useful tool, Geodatabase Designer, is available for download for free from ESRI ArcScripts that can be of assistance with analyzing a geodatabase and data model during the pilot/test phase; it can be used to ensure that all aspects and configuration are as they should be. Geodatabase Designer, interfaces with ESRI ArcCatalog and allows a whole geodatabase schema (all classes, attributes) to be documented and analyzed. Individual elements such as Domains, Object Classes, Relationship Classes, Geometric Networks, and Topologies can be examined by way of an HTML report, XML, or a Notepad-style Log file. Geodatabase Designer also allows the importing of schemas in XML format from other sources where a Data Model Template (either as-is or modified) can be used. In addition, if Microsoft Visio is available, a data model can be designed through all aspects of the Logical and Physical phases in standard UML and loaded directly into ArcCatalog using Microsoft Visio. This will physically create the geodatabase data model based on the UML diagram. This, in turn, can be used to ensure the accuracy of a model's UML documentation.

The tool is available at *http://arcscripts.esri.com/details.asp?dbid=13484*

Another handy tool available for use with ESRI ArcCatalog is the Spatial Domain Analyzer. This tool allows for analysis of a proposed spatial domain (X, Y, Z domain) of a geodatabase and(or) a feature data set; it allows the spatial domain of vector data to be loaded into it. The Spatial Domain Analyzer is available for free from ESRI ArcScripts: *http://arcscripts.esri.com/details.asp?dbid=13916*

8 Guidelines for Project Quality Assurance, Development of a QA Plan, and Quality Control

Unless otherwise cited, this section draws its Quality Assurance/Quality Control approach from ESRI Educational Services Course Exercises (2006b), ESRI Educational Services Course Lectures (2006c), and Danielle Hopkins (ESRI, oral commun., 2006).

The QA plan and QC procedures, as with the elements of any project, should fit the needs of that project and, consequently, may vary from project to project.

8.1 Quality Assurance and the QA Plan

The QA plan is a critical part of a project that is designed prior to any actual production work. The QA Plan serves to bring together and document all aspects of a project that require specification and consideration at the outset in the context of the project's goals and objectives (Danielle Hopkins, ESRI, oral commun., 2006). Because of this, consistency and completeness of data quality are major components of, and are affected by, any specifications laid out in a QA plan. If all aspects of a project are specified and documented, but the specifications are adhered to only 80 percent of the time, data consistency and completeness will be affected accordingly.

QA serves to establish the expectations for all stakeholders, defines who will be doing what and the details of those functions, establishes the QC processes, and documents the standards and acceptance criteria to which the project will adhere (ESRI, 2006c; Danielle Hopkins, ESRI, oral commun., 2006). The nature of the QA plan, when properly defined and agreed upon, will facilitate communication between all project stakeholders so that everyone understands how the project will be carried out. It is impossible to draft a good QA plan, and consequently, difficult to carry out a project well, without constant communication among all project participants.

An effective QA plan will reference and not repeat information from other relelvant project documents. If the specifications in the original documents are modified during the planning process, updates will be needed in multiple locations, which could potentially cause confusion. This concept is similar to that of the relational database design principle of not repeating information anywhere in the database. Instead, the modified documents are referenced by name and, where possible, referenced down to the level of specificity (such as section names).

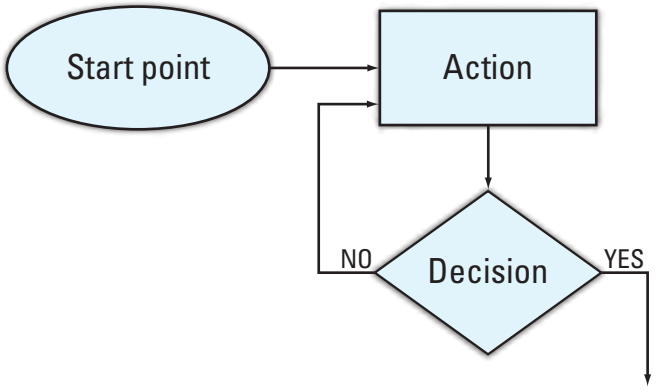
A general example framework for the content of a QA plan can be broken down into four conceptual components: Management, Design, Data Assessment, and Reporting and Oversight (ESRI, 2006b; ESRI, 2006c; Danielle Hopkins, ESRI, oral commun., 2006). An example QA plan template (ESRI, 2006b) is included in this report in Appendix B.

8.1.1 Management

The management portion of the QA plan shapes the QA plan's structure and its purpose in the context of the project's goals and objectives (ESRI, 2006c; Danielle Hopkins, ESRI, oral commun., 2006). The management portion also identifies the roles and responsibilities for the project QA/QC functions and lays out an acceptance work flow for the development of the QA plan itself. The QA plan is a living document that will likely have several revisions during a project as the collaboration among project participants evolves. A history of revisions is useful to document the ways the plan has changed over time and the changes made.

Careful management of the QA plan is necessary so that the implications of planned and unplanned changes and revisions are fully evaluated. It is possible that the QA plan may require unplanned revision at later stages in the project. A formalized process for implementing change control can ensure that any perceived improvements are fully evaluated. A revision work flow can be used to define this formal process.

The creation of a QA plan acceptance work flow can be helpful in managing and facilitating the revision and development process. Visual diagrams often can help with such work flows. A standard way of representing elements of a work flow consists of rectangles or squares representing actions and diamonds representing decisions with possible decision scenarios indicated by flow lines exiting the diamond (the decision). Other symbols can be incorporated as well, such as oval shapes for start and end points (Cornell University, Central Technical Services, 2003).



The QA plan acceptance work flow can include the application of the draft QA plan for the assessment of a pilot project test database. A pilot project test database and a QA plan acceptance work flow can be used together to identify and resolve problems associated with a project's database schema or data model (Breman and Zeiler, 2005; ESRI, 2006b; Danielle Hopkins, ESRI, oral commun., 2006). Such a "test run" can also be helpful in identifying any QA/QC work flow problems that cannot be realized until they are actually carried out. It is usually much more cost effective to address any unexpected problems during the planning phase and(or) pilot test phase than after actual work and production have begun. The pilot project test database is not necessarily a huge undertaking; it simply represents a solid portion of the full project that adequately represents the major elements to be addressed by the QA plan.

To increase efficiency, the pilot project test database could be integrated into the actual project as the first phase or portion of several phases to be addressed during the normal production work flow. When doing this, however, if any adjustments and revisions are made to the QA plan, the pilot project database and any associated data will not be valid for use. This is a decision that is evaluated on a per-project basis. It may not be worth the time and effort to make the pilot project a first phase of the full project unless re-doing the whole first phase is not a problem.

An example QA plan acceptance work flow might consist of the following steps (which can be efficiently diagrammed using the above suggested method) (ESRI, 2006c).

1. Submit a draft QA plan.
2. Review the draft.
3. Determine whether or not revisions should be made. If yes—make revisions and return to the step 2; if no—proceed to step 4.
4. Test the application of the QA plan against a pilot project test database. This will allow the QA plan to actually be test-driven on a test subset of the project, thus helping to determine whether more revisions on the QA plan are needed.
5. Review the results of the QA plan application to the pilot project database.
6. Determine if more revisions to the QA plan should be made. If yes—make the required revisions and return to step 2; if no—proceed to step 7.
7. Accept the QA plan, and finalize it.
8. Move on to full project implementation and production schedule.

8.1.2 Design

The design portion of the QA plan serves to establish the details of how the data will be handled for the project (ESRI, 2006c; Danielle Hopkins, ESRI, oral commun., 2006). This is where various data elements such as those described in Section 6 can be specified, including all of the critical elements of data, any thematic data standards that are to be used, metadata specifications, and specifics (if applicable) related to Data Stewardship, along with any other aspects of a project that can benefit from initial consideration and definition. Alternatively, documents containing these specifications can be referenced here. In addition, project deliverables and schedules can be clearly defined and agreed upon by all parties involved with the project. Again, if other documents are available that address these data management elements, the QA plan can simply reference those so as to avoid duplication of the specifications therein.

- **Data Acceptance Criteria**
  The design section is the first step in addressing the QC aspect of a project by incorporating the acceptance criteria and sampling strategies for the project data. Acceptance criteria are the "maximum level of error allowed for data to be acceptable in a project or by a client" (ESRI, 2006c). Any acceptance criteria specified for a project's data need to be specific, attainable, and measurable (ESRI, 2006c). Other essential aspects of the data development process (timelines for delivery, QC tasks and any rework, reporting timelines) can be addressed through the QA plan as well.

  Often the acceptance criteria are already set in the project contract, database design or definition documentation, or other similar project documents. Acceptance criteria are specified for each appropriate project area (for example, conversion, migration), as well as for each data type because these criteria will have their own characteristics and nuances. Vague data error acceptance statements cast for the whole project will not work well for maintaining quality. If the project is to be successful, all project participants will need to agree on the acceptance criteria for the different project areas.

  Typically the various categories of data acceptance consist of acceptance or rejection on the basis of comparison of the QC results to the defined acceptance criteria. Any QC tasks can be managed effectively by acceptance or rejection work flows similar to the example QA plan acceptance work flow. In the case of unexpected problems with data or other parts of the project, a flexible Problem and Resolution (PAR) work flow might be needed and is one way to address such issues should they arise (ESRI, 2006c). A PAR work flow serves as a general guideline for handling issues that cannot be anticipated. Because these potential

problems are unknown, simply outlining a general framework for how to deal with them can expedite resolution.

Depending on the project and data, it may be appropriate to include an optional "accepted with rework" category that could be applicable to situations where, for example, most of the data meet the criteria and the nonconformities have been identified and isolated (ESRI, 2006c). Such a scenario could occur, for example, with a series of systemic errors resulting in an identifiable problem in the production work flow. Addressing this category is necessary if it is to be allowed as an option in the acceptance work flow so that it does not turn in to a gray area of misinterpretation.

- **Data Acceptance Evaluation Strategies**
  The establishment of standards and policies for data management and the data elements identified in this Toolkit contribute to the likelihood of data acceptance while also giving confidence and credibility to the results (Danielle Hopkins, ESRI, oral commun., 2006). Data acceptance can be organized by the various types of project areas to be evaluated for example, database design, attribute accuracy, completeness. For management purposes, these areas can be grouped by their evaluation category type for example, pass/fail, where no error is tolerated; quantifiable, where acceptable error might be specified as a percentage of the total; or subjective, such as the "look and feel" of cartographic label placement.

  Pass/fail group items are often subject to automated QC checks where 100 percent of the elements are checked. An example of this could be a database design. Quantifiable group items may be less straightforward and may be subject to a sample check or an evaluation of a subset of the data or features if the cost of a pass/fail evaluation is prohibitive. If sampling is used, samples should be representative of the full type of that particular data element. Subjective items such as label placement can be a bit trickier with pass/fail than with the other two options in that one person's evaluation and interpretation of the item can differ from another's. The best way to hedge against items that could fall into the subjective category is to specifically define, using the acceptance criteria, how such potentially subjective elements should be presented or formatted.

  Sample sizes and types to be used are a vital aspect of the QC process to be determined by all project participants. Confidence in data can affect sampling size in that as confidence in data increases, sampling size may decrease. Because of this, confidence may shape, in part, the sampling size and the sampling methods used, depending on how rigorously the project adheres to the specified data elements and data management protocols. Depending on the data, such aspects as time and the quality of source data also can influence confidence. For example, if a project uses existing data or derived data, metadata for the source data sets also can, in part, determine the necessary sampling methods and sample sizes.

- **Data Sampling Methods**
  Sampling is typically done for evaluation categories other than pass/fail where no error is tolerated. The purpose of using samples when performing QC checks is to examine a "group of features or a data subset representative of the entire database" (ESRI, 2006c). The three primary methods for sampling data are random, systematic, and stratified.

  - Random
    With this method, the selection process is essentially left up to chance. Every part of the data set has an equal chance of being sampled.

  - Systematic
    This method performs the selection process in a regular way such that, for example, every third record in a database might be sampled or a spatial feature every X feet might be sampled.

  - Stratified
    This method performs the selection process based on a variable or grouping approach. For spatial data, this might be done by overlaying a grid on the features to be sampled whereby X numbers of features in each grid cell are included in the sample. A cluster method also can be used with the grid. Stratified sampling on tabular data would typically occur when project participants choose a predetermined number of record types based on attributes.

Sample size can be determined by several means. It can be a percentage of features or records or a specific number of features or records; it can be based on a set standard, based on a sampling formula, or determined by other statistical means. The appropriateness of a sampling method is usually determined by project strategy, the importance of that particular feature set or data type, or the total number of features in that data set (ESRI, 2006c).

- **Error Calculation**
  One of the most important yet difficult aspects of the QC process is determining what value is acceptable for each type of identified data evaluation category. Depending on the project areas to be evaluated, as well as the type of evaluation category to be applied, the QA plan needs to be as specific as possible in stating what is unacceptable (ESRI, 2006c). For a data set and its relative importance and composition, it might be necessary to specify this for each attribute. For example, the QA plan might specify what would constitute

as an error per individual records or features and their associated individual attributes.

For QA and QC to function as intended, it is necessary to apply rules and specifications consistently. The best way to insure consistency is to be as specific as possible for all data acceptance criteria and evaluation categories in the QA plan. This is particularly true for the subjective evaluation category since one person's view of, for example, an annotation for feature labeling might differ from another's. Automated and visual QC checks can also be detailed along with the specifics (type, methods, tools, settings, layers to be performed on) needed for carrying out those tasks (ESRI, 2006c). Relevant documents that contain information regarding such things as database design or other aspects of a project also can be identified and referenced.

There are various ways in which error can be calculated for data sets. The appropriateness of one way over another typically is shaped by project strategy and the importance of that particular feature set or data type, in addition to some of the other factors (data criteria, evaluation categories) described above that can influence QA and QC. Common methods for calculating error include the following from ESRI (2006c):

- Per Feature or Record
  When error is calculated on a per record or feature basis, whether or not that instance of error is actually one or more errors is not specified. In other words, one feature specified as an instance of error may incorporate several incorrect attributes, but it is still counted as one collective instance of error. This can be considered "feature-centric" whereby the total number of errors will be less than or equal to the total number of database records or features. Such a method might be more appropriate for simple databases.

- Per Attribute
  This method of error calculation is more specific in that it recognizes the potential for more than one error being associated with a record or feature. For example, if one database record has five incorrect attributes, five instances of error will be recorded. With this method, the total number of errors will be less than or equal to the total number of features times the number of evaluated attributes. This method may be more appropriate for larger, more complex databases since it gives all attributes equal importance.

- Weighted Attribute
  This method assigns determined weights or "importance" to attributes. Specific types of attributes are evaluated and their respective level of error is calculated against an assigned weight. This equates to the sum of the errors for a category of attributes multiplied by the weight (1 x the weight) equals the "score" for that attribute category. This method can be useful when attributes are of varying importance. This method can be more appropriate for larger, more complex databases.

To calculate error rate, take the number of errors divided by the total number of features reviewed times 100. Percent correct is equal to 100 minus the error rate (100 being equal to 100 percent of the acceptance criteria).

8.1.3 Data Assessment

The data assessment portion of the QA plan is where the QC work flows are described and documented (ESRI, 2006c; Danielle Hopkins, ESRI, oral commun., 2006). Any QC should have appropriate actions and methods, QC reports, and decisions noted in the work flow. QC reports are a great way to concisely summarize the results of a QC test. The content of the QC summary information in the QC reports should be clear so that the decision process can go smoothly. The person(s) responsible for each of the steps of each respective QC work flow are noted, and the order in which the QC work flow actions, reports, and decisions occur are numbered as well (ESRI, 2006c). It is probably beneficial to have the QC for a data set performed by someone other than the person responsible for the production of the data.

Depending on the nature of the QC deemed necessary for a project, it may be helpful to organize the various types of individual project data sets and project elements to be evaluated into similar groups. Those groups can be organized into "inspection tables" in a spreadsheet or database based on their respective evaluation category type as mentioned above (pass/fail, quantifiable, subjective). The rows of the "inspection tables" can be associated with individual project data sets or areas. The columns or attributes can be associated with such things as the criteria categories (for example, completeness, accuracy), total feature or record counts for each data set, the sample size, and percent correct or pass/fail. A predetermined ranking system (1, 2, 3...A, B, C…) with related data-set inspection sampling percentages that directly correlate with the importance and(or) size of the project area being evaluated can be helpful. Any other pertinent information also can be handled through this approach so as to streamline and automate the QC process (ESRI, 2006c; Danielle Hopkins, ESRI, oral commun., 2006).

- **Work flow Diagrams and Checklists**
  Work flow diagrams and checklists can be an efficient means of documenting and managing a project's QC assessments (ESRI, 2006c; Danielle Hopkins, ESRI, oral commun., 2006). These can help the participants understand the work to be done by providing a visual display of the work flows in incremental steps. The visual diagrammatic technique described with the QA

plan acceptance work flow in the QA plan management section (Section 8.1.1 of this Toolkit) can work well for this. Checklists developed to complement a diagrammatic work flow can be helpful by presenting the tasks in a straightforward and concise manner. As a result, the steps of a QC work flow can be checked off as they are completed.

If desirable, checklists could be made even more specific by further breaking down the steps in the QC work flow to show each component of the QC to be done (for example, confirm feature or database record counts, look for badly formed geometry) with each requiring a check. This can further ensure consistency and completeness in addition to being an excellent form of documentation. It could also be required that each individual step be initialed and dated by the person carrying out that step to help ensure documented accountability. If people know that they are documented as being responsible for a step in a work flow, they are much less likely to make assumptions or to skim over important details. Work flow documentation such as this also can make it very easy for someone else to pick up where another left off. If the person originally assigned to the work flow or task is unavailable, the remaining tasks can be completed without confusion.

The use of work flows for addressing unexpected data and QA/QC problems also can be an effective and efficient management technique. Because of the nature of unexpected problems, a flexible, agreed upon system is one that can be adapted accordingly. A Problem and Resolution (PAR) work flow can be an effective means for documenting and addressing a problem with all appropriate project partners. This approach can ensure that everyone is in agreement with the circumstances surrounding the issue, the implications of the issue, and how the issue should be addressed. Depending on the project, a simple spreadsheet table or database could be set up to track such problems, along with any relevant information so as to ensure that they are resolved.

An example QC work flow, similar to the QA Plan Acceptance work flow above, might be the following five steps (ESRI, 2006c; Danielle Hopkins, ESRI, oral commun., 2006):

1. Prepare the data,
2. Perform QC,
3. Create QC report,
4. Determine whether or not the data met the acceptance criteria, and

5a. If yes, accept the data and create an acceptance report;

5b. If no, reject the data.

In this example of a QC work flow, the 5b step results in the data going to a re-work stage and back to step 1. Fully addressing how rejection scenarios are to be handled for the QA plan is critical so that errors do not become problematic. This is particularly important if the data set or project element is subject to sampling during the QC process. If the data or project area under review is rejected, a new sample of the whole reworked data set similar to the first sample is used so as to properly re-analyze the data or element (Danielle Hopkins, ESRI, oral commun., 2006). Documenting all of this in the QA plan helps ensure repeatability.

The above work flow could also be expanded to include multiple phases of QC to be done by different people. For example, the first phase could include quick automated checks done by the project data manager. If the data passed that phase, it could move onto a data technician for more time consuming QC such as visual analysis and sampling. Relevant source materials might also be used for QC during this second phase, or an end-user might perform their own QC by testing the data in a particular application (ESRI, 2006c).

Tips for QC and for Developing Successful QC Work flows.

- Share QC tools and procedures.
- Ensure consistent and constant communication through meetings or conference calls among all project partners, so that everyone is on the same page.
- Include all aspects of QC and related work flows in the QA plan.
- Ensure that rejection scenarios are properly addressed.
- Use work flow diagrams and checklists.
- Document all decisions and processes.
- Most importantly, make no assumptions.

- **Error Lifecycle and Tracking**
  Documenting and tracking error during the QA/QC process is an essential proactive management technique for ensuring that errors are addressed and resolved (ESRI, 2006c; Danielle Hopkins, ESRI, oral commun., 2006). Such documentation can also be vital for evaluating the effectiveness of a project's QA/QC processes for the current project as well as for any future work. Similar to identifying random or systemic error in data, random or systemic problems associated with a project's QA/QC procedures can be equally as bad. Documenting and tracking error also can help to

ensure accountability in a project so that responsibilities can be delegated accordingly. Built-in accountability helps to guard against the bottleneck questions of "Who should be handling it?" or "Who should have handled it?"

There are essentially three basic phases of the error resolution process. These consist of error review, error correction, and error verification (ESRI, 2006c; Danielle Hopkins, ESRI, oral commun., 2006). Review entails analysis of the discovered error. Correction involves taking the appropriate steps to correct the error depending on the type of error and the type of evaluation category (pass/fail, an quantifiable, subjective). A rejection or re-work work flow can be created to assist with this. Verification involves ensuring that the error has been resolved and, if necessary, resubmitted to the appropriate step of the QC work flow. How the verification phase is handled may be contingent upon whether "accepted with re-work" was an option as well as the type of evaluation category (pass/fail, quantifiable, subjective) used for that project element. Delegating these phases to different people can be advantageous for ensuring quality and that all errors are discovered and addressed.

Below are some key questions associated with the error lifecycle that can flesh out the pertinent details regarding the nature of the error and the accountability aspect.

- What is the condition of the problem, error, and(or) data?
- Where and when in the data production process were the data compiled?
- Who is responsible for reviewing the data?
- How and when were the data reviewed?
- Who is responsible for correcting the data?
- How and when were the data corrected?
- Who is responsible for verifying the corrected data?
- How and when were the data verified?

A great way of tracking error is the use of a GIS feature class, a database, or a spreadsheet with pre-defined fields for relevant error attributes (ESRI, 2006c; Danielle Hopkins, ESRI, oral commun., 2006). Such an approach can be very useful for tracking both spatial and non-spatial data error as well as any error associated with database design. Each individual row in the table will constitute an instance of error. This aspect will be influenced by the error calculation method used as described in Section 8.1.2 of this Toolkit. If multiple error calculation methods will be used, it might be appropriate to have individual error tracking tables associated with each type of error calculation used. The three phases of the error resolution process can be used to guide the appropriate attributes to be included with an instance or error. An initial set of fields can also be included to give the location of the error. The remaining fields can be associated with the three phases of the error resolution process. Below is a general example of an error tracking table. The example attributes represent the attribute column names in the table, and individual error records would make up the rows in the table (ESRI, 2006c; Danielle Hopkins, ESRI, oral commun., 2006).

Entity/Table Name: "Error Tracking"
Example Attributes:

- Unique Identifier: A unique identifying number
- X Coordinate: (applicable for spatial data – be cognizant of consistency with decimal degrees, degrees-minutes-seconds, or projected coordinates)
- Y Coordinate: (applicable for spatial data – be cognizant of consistency with decimal degrees, degrees-minutes-seconds, or projected coordinates)
- Employee ID: The name or some other form of identification for the person that discovered the error or that performed the QC
- QC Check Type: The QC method that discovered the error
- Layer: The feature class or object class/table where the error is located
- Object Type: For spatial data: the spatial representation type. For tabular: the object unique ID. For an attribute: the unique ID and name of the feature or record that it is associated with
- Workspace: The location of the data set or project element and error, (for example, a file path or location on a computer or server, a geodatabase file path)
- Review Date: The date when the error was found
- Review Status: The correction that needs to be made (for example-correct attribute value)
- Review Description: A concise, detailed description about what should be corrected, where it is located
- Correction date: Date the error was corrected
- Employee Correcting ID: Identification of the person that corrected the error
- Correction Status: Describes what was corrected

- Correction Description: Gives a concise, detailed description about what changed and corrected
- Verification Date: Date the error was verified
- Employee Verifying ID: Identification of the person who verified the correction
- Verification Status: States if the error correction has been verified
- Verification Description: Describes in concise detail how the error was verified

To give the error correction process additional QA/QC and standardization, the above example error tracking table can be set up with predefined standard domains and subtypes for attributes to help enforce data integrity. For example, a domain for the QC Check Type attribute would allow the error technician to choose only from a predefined list of the relevant QC methods used for the project. Such a tracking table could easily be established as a template and reused for multiple projects with only the predefined domains and subtypes requiring project-specific modification. Another advantage of using such an error tracking table with predefined domains and subtypes is that analysis can be performed on the different error attributes, which can reveal patterns or systemic error. For example, a query could be run for the different types of QC check types. If one type of QC check is the dominant type associated with a high percentage of the project error, it is likely that there is a production problem associated with the method used to create the data or project element associated with that QC check type (ESRI, 2006c).

Any type of table or spreadsheet can work well for organizing and managing the error tracking process for both tabular and spatial data. A GIS error tracking point feature class, however, can be particularly helpful for tracking spatial error in spatial data sets (ESRI, 2006c; Danielle Hopkins, ESRI, written commun., 2006) using ESRI ArcGIS software. During the QC process, point features can be edited into the error feature class to precisely document the locations of discovered error during QC. The error tracking point feature class can have the above attributes set up with predefined project-specific geodatabase domains and subtypes. As error is discovered, a point is created for each spatial error (again-depending on the error calculation method used) and the associated attributes are edited into that error point feature during an ArcMap editing session. Depending on the spatial representations of error as well as the type of QC being performed, features also can be copied and pasted in to the error tracking feature class. This method of error tracking is further addressed in the example QC section of this document below in Section 8.2.

The creation of a separate error tracking geodatabase can be a useful way of managing and implementing all of this. This could simply be a validated copy of the project geodatabase schema that has the appropriate error related domains and subtypes in addition to the project data validation rules. Such an approach would allow quality control measures for the actual project QC to be implemented without having to add them to the actual project database. Data could be loaded into the error tracking geodatabase and project-data QC performed. If any error is discovered, the project's error discovery and resolution process is carried out. This could help to keep the error tracking and resolution process separate from the actual project database.

8.1.4 Reporting and Oversight

Reporting is a critical part of the QC work flow process for providing summary information that helps determine the next steps to be taken. The reporting section of the QA plan defines the various reporting mechanisms for a project (ESRI, 2006c; Danielle Hopkins, ESRI, oral commun., 2006). QC reports are important in that they help to convey information about the QC performed such as calculations, notes, and results. This information helps to determine whether or not the data can be accepted or rejected. Acceptance and rejection reports generally occur at the end of a QC work flow and can represent a formal acceptance or rejection of the data. PAR reports are useful in the case of unexpected problems because the reports ensure that those problems are documented and addressed accordingly (ESRI, 2006c; Danielle Hopkins, ESRI, oral commun., 2006).

Development of an error summary table that presents the issues and errors associated with a data set can be a good way to concisely summarize the results of QC in addition to a report. Error summary tables can be structured to present information on one data set per table by having the specific QC inspection method(s) used as the records or rows in the table. An "error count" attribute column with fields for each method used would record the errors associated with each check. Other attribute columns can be included as needed. Another approach would be to have the respective QC inspection methods each constitute a table while having the data sets that were evaluated represented as the records or rows in the table. An error count attribute column with fields for each data set would be appropriate in addition to any other necessary QC attributes.

- **The QC Report**
  There are several pieces of information that should be included in any QC report. Basic identification information such as the project title, who the report is

addressed to and who it is coming from, the purpose and scope of the report, the date, and the person(s) responsible for the QC can be included. Other information pertaining to the actual QC could include a simple summary of the QC results, any relevant methods and calculations, the acceptance status (accepted or rejected), and any relevant notes pertaining to error and other issues. If a data set is rejected or accepted with rework, specifying a resubmition time window helps to keep things moving.

- **The PAR Report**
  The PAR report serves as a way to document unexpected problems or larger issues that arise during a project (ESRI, 2006c; Danielle Hopkins, ESRI, oral commun., 2006). Basic identification information such as that mentioned with the QC report can be included. Descriptions of the problem and the circumstances around it are important as well as information on proposed and accepted resolutions.

8.2 Quality Control

The QC portion of a project is the phase used to evaluate data and project areas against predefined acceptance criteria. Consistency and completeness are both aspects of data quality that are necessary for any QC performed on a project and its data. The type of QC to be used can be predefined in the QA plan in the Data Assessment section and ideally needs to agree with the various areas of the QA plan previously mentioned. Appropriate work flows and checklists to help document and manage the QC process, in addition to helping establish project accountability, can assist with this.

One possible approach for QC is to divide it into first-round and second-round checks. With this approach, the first group is associated with automated methods, whereas the second group consists of more intensive and time consuming methods (ESRI, 2006c; Danielle Hopkins, ESRI, oral commun., 2006).

8.2.1 First-Round QC

First-Round QC consists primarily of automated checks that analyze 100 percent of a data set or project element. These typically fall into the pass/fail evaluation category (ESRI, 2006c; Danielle Hopkins, ESRI, oral commun., 2006). First-Round QC is typically less time consuming and, as a result, less expensive than other methods such as visual QC and QC that requires sampling. Below are some possible automated approaches that incorporate ESRI ArcGIS Desktop functionality, as well as other methods for first-round QC:

- **Database Schema**
  Evaluating a project's database/data model schema and its properties helps to ensure that all aspects are correct. Such things as the actual structure (schema), any relationships and relationship classes, primary and foreign keys, database properties (domains), spatial reference (for geodatabases and any feature data sets, feature classes), as well as field properties such as data type, null values, and subtypes need to be checked (ESRI, 2006c). These elements help to determine the quality and integrity of the project's data as well as whether or not the database or model will serve to meet the project's goals and objectives. Depending on how the database design/modeling process is approached, Section 7 of this Toolkit might provide other aspects to check.

  An excellent free tool for use with ESRI ArcGIS that can be used to verify the latter aspects of a geodatabase is the Geodatabase Designer tool. For more information about this tool, see Section 7.5 of this Toolkit. Geodatabase Designer is available for download from the ESRI ArcScript web site at this link: *http://arcscripts.esri.com/details.asp?dbid=13484*

  An example database checklist might

  - Verify database schema,
  - Check database properties,
  - Verify spatial reference, and(or)
  - Verify field properties.

  Any discovered errors can be recorded in an error tracking table. Once the geodatabase schema has been validated, a copy of the empty corrected geodatabase can be used as an error tracking geodatabase as described in the Error Lifecycle and Tracking portion of Section 8.1.3 of this Toolkit.

- **File Formats, File and Attribute Naming Conventions**
  Per Section 6.3 of this Toolkit, file formats as well as file and attribute naming conventions are aspects of data that can be verified using their specifications in the QA plan. Any discovered error can be recorded in an error tracking table.

  Performing QC on these typically consists of visual verification, such as

  - Verifying that the data are in the correct specified format (ex: ArcGIS 9.x as opposed to ArcView 3.2) and(or)
  - Verifying that file naming conventions are consistent for all data delivered (for example, named appropriately, no spaces in the file name) and that both file and attribute field names conform to the software environment's rules and restrictions for them.

- **Validation Rule Conformance**
  Tools readily available in the ESRI ArcGIS ArcToolbox and the Editor toolbar can be used to validate data loaded into a project geodatabase target feature class to ensure that it matches with any specified domains, field properties, and(or) geometric network rules (ESRI, 2006c; Danielle Hopkins, ESRI, oral commun., 2006). This is important because it is possible to load data into an ArcGIS geodatabase that is not consistent with established validation rules such as, for example, attribute domains of valid values.

  One example is the Object Loader tool, which works in conjunction with an open edit session through the Editor toolbar. For the Object Loader tool to work, the Editor toolbar target feature class must be set to the project geodatabase feature class designed for the source data being loaded so that both have the same geometry. Also, multiple data sources must have the same schema for this tool to work properly. It is helpful to use an error tracking geodatabase (a copy of the approved project geodatabase schema) for this QC procedure.

  First, be sure the Editor Toolbar is visible on ArcMap (version 9.x). To add the Object Loader tool, go up to Tools-Customize, click the Commands tab, select Data Converters from the list. Drag the Object Loader tool on to the Editor Toolbar.

  To use this tool, first begin an edit session in an error tracking project geodatabase (the copy of the project's geodatabase configured for QC).

  1. Be sure that the editing "Target" on the toolbar is set to the appropriate data feature class or object class in the error tracking project geodatabase designed to receive the source data.
  2. Click the Object Loader tool and browse to the location of the data to be evaluated. Click add, click Next.
  3. On the next screen, check to make sure that the attribute field names match between the data being loaded and the error tracking geodatabase. If these match, proceed to step 4.
  4. Click next again, and select "Load all of the source data".
  5. Click next again and select "No" for the question regarding feature snapping based on the current snapping environment. Select "Yes" regarding the statement "validate the new features and create a selection of all invalid features."
  6. Click next and review the latter options to make sure they are as listed above; and
  7. Click finish.

  Once the data are loaded, ArcMap will automatically select any features that do not contain valid attribute values per the predefined domains and field properties of the geodatabase. Visually inspect the map and attribute table for any selected features as they should be recorded as errors. Note: Be careful not to mouse-click anywhere inside the map as this will deselect any selected features!

  If any features are selected, they can now be recorded in an error tracking table. If they are simply point features, they can be copied and pasted into the error tracking point feature class as mentioned in the Error Lifecycle and Tracking portion of Section 8.1.3 of this Toolkit. This can be done by changing the Target on the Editor toolbar to the error tracking feature class and performing a copy/paste function.

  If the selected features are not points, they must be processed before they can be added to the tracking feature class. The ArcToolbox tool Feature to Point can be used to create points based on the centroid locations of selected features. These new point features can be appended to the error tracking feature class using the ArcToolbox Append tool found by clicking the Data Management Toolbox—General—Append. Note: Be sure to select "NO_TEST" on the schema type option on the Append tool (ESRI, 2006b). Once this processing is done, the error tracking attributes can now be manually populated for each of those features with appropriate error attribute information.

  Another ArcGIS tool, the Editor toolbar Validate Features tool, can be used to check for broken validation rules after the data have been loaded (ESRI, 2006c). Similar to the Object Loader tool, the Validate Features tool only works during an open edit session. This tool supports analysis of subtypes and domains, network connectivity, and relationships (ESRI, 2006c). The Validate Features tool can be found by clicking on the drop-down Editor button on the Editor toolbar. The tool appears grayed out until a feature(s) is selected. Simply right-click on a layer in the table of contents, click Selection and click Select All. Navigate to the Validate Features tool and click on it. The tool will automatically run and will display whether or not the features selected are valid. Any discovered error can be recorded in an error tracking table or denoted with a new point in the error tracking point feature class.

- **Missing and Unique Attribute Values**
  Attributes of features and layers need to be evaluated

to ensure that the column names and field values are correct. Inspecting attribute columns to make sure they are all there and named properly is a good first step. The use of SQL queries on attributes in standard databases and in the ArcGIS Desktop Select by Attributes dialog box can help to discover unique and(or) missing values (ESRI, 2006c). Performing frequency or alphabetical sorts on the data also can help to ensure that there are no misspellings and that the data meet the standard specifications. Note: the Frequency tool or sort table are available only at the ESRI ArcGIS ArcInfo license level (ESRI, 2006c). Any discovered attribute error can be recorded in an error tracking table or the error can be related to the feature and denoted with a point in the error tracking point feature class.

- **Duplicate Attribute Values**
  Data that contain attributes which require a unique identifier (for example, Monitoring Plot ID) can be analyzed to ensure that duplicate values do not exist. This is particularly important if the database schema contains relations between data based on these unique fields or if joins and relates will be used. The ArcGIS ArcToolbox Frequency tool can be used to ensure that unique fields do not contain any duplicate values (ESRI, 2006c).

  1. In ArcToolbox, click the Analysis Tools toolboxà Statistics-Frequency tool. Select the appropriate feature attribute table, and select the appropriate field that should contain unique values to be analyzed. Click OK. Note: if an attribute table has more than one unique field, this process can be run individually for each of those respective unique attribute types.

  2. The Frequency tool creates a new table that is automatically added to ArcMap that assigns a frequency or occurrence number to each value in the previously selected attribute field. To view this table, click the source tab in the ArcMap Table of Contents; right click the new Frequency table; select Open. Unique values will show a frequency of one. If any values occur more than once, the frequency number will equal the number of times they repeat.

  3. To locate the features associated with any repeating attribute values that are unique, an ArcMap relate can be created between the Frequency table and the feature attribute table based on the attribute field name being analyzed.

  4. Create a Relate by right clicking on the feature Frequency table, select Joins and Relates, select Relate. Create the Relate to the table containing the data being evaluated.

  5. Open the two respective attribute tables. On the Frequency attribute table, right-click the Frequency column and select Sort Descending. Any values with a Frequency greater than one will be sorted to the top. Select the gray box to the left of those records to select those records.

  6. Click Options at the bottom of that table, and select Related Tables. Select the table related in steps 3 and 4 above.

  7. Click the Selected Button on the related feature attribute table to display the selected features.

  Depending on the number of features selected and their spatial relationship with each other, they may or may not have the correct respective unique attribute value. It can be difficult to tell if just one value shows a frequency of two, and consequently very difficult to tell which attribute is correct if several values show up with a frequency greater than one. Because of this, it is recommended that all attribute values with a frequency greater than one be considered an error and recorded in an error tracking table, or the associated features can be denoted with a point in the error tracking point feature class until each is evaluated (ESRI, 2006b).

- **Feature and Tabular Data Record Counts**
  Feature and tabular data record counts can be checked to ensure that their totals are correct for the different respective types of data. Comparisons can be done for both feature classes as well as standard object/tabular entity tables against the source data or any data collection notes/metadata. Differences may indicate problems resulting from the conversion or migration process, data loading and schema conflicts, and the spatial domain of the geodatabase and(or) feature data sets (X,Y, and Z domains) (ESRI, 2006c). If the spatial extent of a geodatabase or feature data set is not large enough to include the spatial extent of the project data, features or records may be left out as a result of being outside the specified bounds.
  One great free tool available for analyzing the spatial domain of feature data sets is the Spatial Domain Analyzer tool. This tool is available from the ESRI ArcScript web site at: *http://arcscripts.esri.com/details.asp?dbid=13916*

  Any discovered error can be recorded in an error tracking table.

- **Feature Geometry**
  Data geometry can be checked to ensure that badly formed feature geometry do not make their way into the final product. The ArcGIS ArcToolbox "Check Geometry" tool can be used for this function (ESRI, 2006c). This tool can be found in the Data Management Toolbox under Features (ESRI, 2006c). Discovered errors can be denoted with a point in the error tracking point feature class.

- **Annotation**
  Annotation is an important aspect of cartographic display that helps convey information to the user. The concepts of Confirmatory and Exploratory QC presented in Section 6.6.2 of this Toolkit can be used with ArcGIS to evaluate map annotation.

  Feature-linked annotation also can be evaluated for format consistency. This is done by using the Select by Attributes tool available under "Selection" at the top left of the screen (ESRI, 2006c). Attribute queries can be created to select those that do conform and those that do not on the basis of the different specifications for the annotation. For example, a query can be created to select any annotation that does not have a font size of 10, or a query can be created to search for annotation that meets all of the specified annotation parameters such as font type, font size, orientation angle, and horizontal and vertical alignment. Once these are selected, the selection can be switched so as to select any that do not meet all of the specifications. Any discovered errors can be recorded in an error tracking table or denoted with a point in the error tracking point feature class.

- **Topology**
  To evaluate the topology of a data set, a topology validation can be performed in ESRI ArcGIS so as to ensure a project's data set topological integrity (ESRI, 2006c). The basic components of topology that can be evaluated include participating feature classes, cluster tolerance, rules, and ranks. These are described (participating feature classes, topological rules, ranks, and cluster tolerance) in Section 6.3 of this Toolkit.

  The validation of a topology can be performed in ESRI ArcMap. By loading data into the topological feature classes, validation can be performed so as to flesh out any errors. Use of the error tracking geodatabase as described in this section to identify, evaluate, and fix errors is preferred. Corrections can be applied to the project geodatabase.

  1. Be sure the Topology and Editing toolbars are added to ArcMap.
  2. Be sure that an editing session has been started. Most of the tools on the Topology toolbar should no longer be grayed out if an editing session is open.
  3. The Topology toolbar gives three options for validating a geometry: validate a specified area, validate in the current ArcMap visible extent, or validate the entire topology. For consistency, validation of the entire topology is preferred. Once this is selected on the toolbar, the choice will be re-confirmed.
  4. After processing, any errors will be symbolized on the map. Any errors that appear can be investigated using the Error Inspector on the Topology toolbar. The Error Inspector has sort functions with different display options such as viewing all topology errors at once or viewing the errors per rule.

  By right-clicking on identified errors in the Error Inspector, individual errors can be zoomed in on so as to inspect them and their context. Discovered errors can be recorded in an error tracking table or denoted with a point in the error tracking point feature class. It is essential to be cognizant of the types of errors and whether or not there seem to be patterns (for example systemic) that could indicate a problem with a data production work flow resulting from human error or a flaw with the work flow itself. For example, a series of similar errors on a series of features that were digitized from hardcopy maps could be the result of the editing session snapping environments that were not turned on appropriately during digitizing. A topology error report can be generated through the topology Layer Properties "Errors" tab which can be helpful for the reporting aspect of the topology QC work flow.

8.2.2 Second-Round QC

Second-Round QC is typically more intensive and time consuming than First-Round methods (ESRI, 2006c). Evaluation categories associated with Second-Round QC might include the subjective type or the quantifiable/sampling type. Sampling and visual QC methods, both of which are not as cut and dry as automated pass/fail checks, are two primary methods of Second-Round QC as presented in this Toolkit:

- **Visual QC**
  Many errors and discrepancies can frequently be identified by those familiar with the project by simply examining the data using a GIS application, digital images, and hardcopy review. Below are four criteria for performing visual QC (ESRI, 2006c):

- **Extensive knowledge of the project and data**
  This is crucial so that the person responsible for the QC knows what to look for and what the data ought to look like.

- **Understanding of the QA plan and data specifications**
  Visual QC can be more subjective so it is helpful for those that are performing it are aware of any relevant defined project data specifications. A project manager will be well aware of specifications applicable to visual QC and, consequently, might be a good candidate for carrying out any visual QC.

- **Organization and Planning**
  Access to any source data used to derive project data is crucial for performing visual QC. In addition, metadata documentation for the source data are helpful as well.

Loading and examining data, particularly with a GIS application, can reveal many errors with such elements as labeling, symbology, and feature representation. The use of basic zoom tools and other display properties can reveal discrepancies. If data are derived from source data (for example, feature digitized from a source map or imagery), the source data probably are readily available and can be loaded into the GIS so that comparisons can be made (ESRI, 2006c). The source data would be the bottom layer, with the derived data overlaid for visual comparison. Source data metadata can give valuable information relating to source data quality and subsequent derived data quality.

- **Visual QC - Symbolization and Labeling of Features and Attributes**
  Display of the data to look for errors such as misspellings, spacing issues, and mislabeled features can be helpful for annotation and labeling. The Symbology Tab found in the layer properties for each respective layer in ESRI ArcGIS ArcMap also can be used to display attributes for features directly on the data. The Symbology Tab is accessed by double-clicking directly on a layer name in the Table of Contents in ArcMap and by selecting the Symbology Tab. The advantage of this technique is that the attributes are displayed in a visual context with their associated features which allows for validation on the basis of project knowledge. This can be particularly useful when done by someone familiar with the data or in cases where there are only a few acceptable choices for that particular attribute.

Using the various functions in the Symbology area of the Layer Properties also can be helpful with feature and attribute validation. Spatial features can be validated with their various classifications or with their various attributes (ESRI, 2006c; Danielle Hopkins, ESRI, oral commun., 2006):

- **Visual QC – Points**
  To validate attributes associated with point features, size and color of the point features can be varied accordingly with different attributes or to coincide with the different classifications of point features so as to ensure they are correctly represented. The Symbology Tab in the layer properties in ArcMap is used for this.

- **Visual – Lines**
  Varying the color and(or) thickness associated with line features (for example, different types of roads) against a classification or attribute can be helpful for contrast against source data for accuracy. The Symbology Tab in the layer properties in ArcMap is used for this. Other aspects such as direction flow (in the case of waterways) and endpoint display also can be used to validate line data.

- **Visual QC – Polygons**
  Setting transparency and hollow fills can be helpful for validating polygon data against any source data. These are done by double-clicking directly on a layer name and selecting the Display tab in ArcMap and by double clicking directly on the color of a layer below the layer name. In this case, the source data (such as an aerial photograph) would be loaded into ArcGIS ArcMap along with the polygons being validated. The source data layer would be placed below the polygons. Solid fills also can be used with attribute displays for validating attributes.

Other tools in ArcGIS can be helpful for performing visual QC. The Overview window and the Magnifier window (found on the top toolbar of ArcMap under "Window") are great tools for visual QC review.

The creation of a grid using a polygon feature class that overlays the full extent of a project's geographic area can be helpful for data with many features (ESRI, 2006c). The grid can serve to track areas that have been subject to visual QC for the purpose of completeness. Similar to an error tracking point feature class, a visual QC grid polygon feature class can be created using the drawing tools in ArcMap with the Editing toolbar. Ideally, each square or rectangle of the grid will be approximately the same size (they do not need to be exactly the same), and each will be a separate polygon. Snapping environments can be set when creating such a grid feature class so that the grid squares all connect consistently.

As each polygon is created, they should also be sequentially numbered in the grid polygon feature class attributes. Or, the automatic ArcMap OBJECTID could be used for this. The polygons are labeled with the numbers using the labels tab in the ArcMap layer properties to ensure systematic coverage of the whole area. The polygons can be given a hollow fill so that just the outlines are visible to allow for the project data to be visible for validation. Once an area has been visually examined and validated, the polygon covering that area should be deleted so as to denote completion.

If errors are discovered during the visual QC process, an error tracking spreadsheet or table can be used to manage them. The error tracking point feature class can be extremely helpful with visual QC for pinpointing the precise location and documenting the details associated with any identified errors. ArcMap spatial bookmarks also can be helpful so that specific error locations can be relocated quickly. These can be created by clicking on View on the top ArcMap toolbar-Bookmarks-Create.

Creation of a separate ArcMap map document (.mxd file) is a great way to preserve data visualization settings specific for visual QC. The data are not stored in the .mxd file; the file simply references the data from a specified location for display and analysis purposes. ArcMap layer files (.lyr file) function in the same way by storing cartographic display settings for individual data layers. ArcMap layer files and .mxd can be used to store specific layer properties for a QC work flow. They also can be used for optimal display settings of approved finalized data (ESRI, 2006c).

- **Sampling of Data**
  Sampling is often done when data sets are so large that it is burdensome to inspect 100 percent of the data. Sampling is used to extract what is considered a representative subset of the total number of record or features. QC is then performed on that subset. An established acceptance criteria for that type of QC and the data attribute that was evaluated (for example, 95 percent of a particular type of attribute must be correct) can then be used. The Data Sampling portion of Section 8.1.2 of this Toolkit addresses many of the considerations that arise when devising a sampling methodology. The use of a grid as described in the Visual QC portion of Section 8.2.2 above also can be helpful. Once an appropriate sampling methodology has been specified, many of the various types of example QC methods identified in both the First Round and Second-Round sections of the QC section can be applied to a sampled subset of data. Results of those tests can then be evaluated against the established acceptance criteria for that type of QC and the data attribute checked.

8.3 Additional Quality Assurance and Quality Control Information Sources

There are several sources available for additional Quality Assurance and Quality Control techniques and procedures. The above examples represent one framework used to perform QA/QC. Here are many others from Danielle Hopkins (written commun., 2006).

- Geospatial and QA/QC-Managing Error: *http://www.colorado.edu/geography/gcraft/notes/manerror/manerror_f.html*
- ANSI/ASQC Z1.4-1993: Sampling Procedures and Tables for Inspection by Attributes - E-Standard: *http://qualitypress.asq.org/perl/catalog.cgi?item=T51E*
- Other ANSI/ASQC/ISO QA/QC Standard Methodologies: *http://qualitypress.asq.org/perl/catalog.cgi?category=Standards*
- Communication and Acceptance Criteria – the Gift that Keeps on Giving: *http://www.allpm.com/modules.php?op=modload&name=News&file=article&sid=1388*
- Acceptance Criteria Part I- The True Measure of Task and Project Success: *http://www.allpm.com/modules.php?op=modload&name=News&file=article&sid=1354&newlang=eng*
- Acceptance Criteria Part II, The process of acceptance: *http://www.allpm.com/modules.php?op=modload&name=News&file=article&sid=1366*
- Various QA/QC papers: *http://www.laurelhillgis.com/papers.htm*
- EPA Quality Management Tools and Quality Assurance Plans: *http://www.epa.gov/quality/qapps.html*
- Intergovernmental Panel on Climate Change – Chapter 8: Quality Assurance and Quality Control: *http://www.ipcc-nggip.iges.or.jp/public/gp/english/8_QA-QC.pdf*
- The National Park Service Northeast Temperate Network Inventory and Monitoring Program Data Management Page. This site contains example NPS reference documents for Quality Assurance/Quality Control among other references: *http://science.nature.nps.gov/im/units/NETN/downloads/Plan/NETN_DataManagementPlan.pdf*

—— THIS PAGE INTENTIONALLY LEFT BLANK ——

9 Tools, Guidelines, and Work flows for Creation of Federal Geographic Data Committee-Compliant Metadata

Unless otherwise cited, this section draws from the primary author's personal knowledge and experience with metadata. This section provides an overview of how to approach the process of documenting data and information resources with several free tools and example work flows. It also gives an overview of useful documentation standard formats and web-based defining tools for data elements such as taxonomy. As with any new experience, typically there is an initial learning curve with the first few tries. However, there are many options included below that help to automate the process of metadata creation for efficiency so that anyone can become competent in metadata creation. Many resources are available by way of people and expertise, so help is almost always available.

9.1 Documentation Tools and Standards

9.1.1 Metadata Standards

For geo-referenced data with a spatial component and for tabular research data, the Content Standard for Digital Geospatial Metadata (CSDGM) Version 2 (FGDC-STD-001-1998, also known as the Federal Geographic Data Committee (FGDC) standard) and the Biological Data Profile (BDP) is appropriate. This standard addresses:

- Date(s) in which data were initially collected (year, month and day if known),
- Stated or inferred purpose(s) of data collection,
- Location(s) of data collection,
- Person(s) collecting data,
- Detailed description of data collection methods,
- Person(s) currently in possession of data,
- Location(s) where data are currently stored,
- Format(s) in which data are currently available (for example, electronic spreadsheets, electronic word processing files, paper copies) including any known publications of the data,
- Agency(s) or organization(s) funding data collection,
- Point(s) of contact within these agency(s) or organizations,
- Spatial and non-spatial elements of the data,
- Definitions of geospatial attributes for tabular data to describe fields and values of the data set, and
- Taxonomic elements of biological data.

For more-detailed information on the FGDC-NBII Biological Extension standard, follow these links:

- *http://metadata.nbii.gov/portal/server.pt?open=512&objID=255&mode=2&in_hi_userid=2&cached=true*
- *http://www.fgdc.gov/metadata/geospatial-metadata-standards*

This link provides a graphical representation of the standard:

- *http://www.nbii.gov/images/uploaded/151871_1166715705793_NBII_bdp_graphical.doc*

For documenting reports and other documents, the Dublin Core Metadata standard can be used to address such resources. This standard addresses:

- Resource title;
- Creators, contributors, and publishers;
- Controlled and uncontrolled subject matter keywords;
- Resource description;
- Resource format;
- Resource data issued, such as when the resource was made available;
- Sources from which the resource was derived; and
- Geographic details (if relevant).

For more detailed information on the Dublin Core Standard, follow these links:

- *http://metadata.nbii.gov/portal/server.pt?open=512&objID=718&mode=2&in_hi_userid=2&cached=true*
- *http://dublincore.org/*

For standardized biological/taxonomic information to be included with the FGDC-NBII Biological Extension metadata, the ITIS (Integrated Taxonomic Information System) standard should be used:

- *http://www.itis.gov/taxmatch_ftp.html*

NBII supported standard metadata keyword thesauri and guides for promoting standardization and consistency of resource descriptors. Utilization of a standardized vocabulary for descriptors can help search engines and metadata clearinghouses work more efficiently.

- *http://www.nbii.gov/portal/server.pt?open=512&objID=578&&PageID=1798&mode=2&in_hi_userid=2&cached=true*

9.1.2 Free Metadata Creation Tools

Free Metadata Software Tools Used for Developing Formatted FGDC-NBII Metadata:

- Metavist: *http://ncrs.fs.fed.us/pubs/viewpub.asp?key=2737*
- MetaScribe (online metadata creation tool provided by NOAA): *http://www.csc.noaa.gov/metadata/metascribe/*
- NPS Metadata Tools and Editor (Version 1.1) available as an extension for ESRI ArcCatalog or as a stand-alone program: *http://science.nature.nps.gov/nrdata/tools/*

  When downloading and installing this tool, both the stand-alone and ESRI ArcCatalog extension are installed.
- MetaParser (mp) for validating and quality control of a completed metadata record: *http://geology.usgs.gov/tools/metadata/*
- MetaParser (mp) Batch provides a Windows interface for using the above MetaParser tool: *http://support.intergraph.com/Geospatial/Downloads/Tools.asp?ID=48&SORT=Title*
- Online web-enabled version of MetaParser for metadata QC and validation: *http://geo-nsdi.er.usgs.gov/validate.php*
- Others tools available at: *http://sco.wisc.edu/wisclinc/metatool/ http://geology.usgs.gov/tools/metadata/ http://metadata.nbii.gov/portal/server.pt?open=512&objID=255&&PageID=338&mode=2&in_hi_userid=2&cached=true*

9.1.3 Metadata Training and Assistance

Metadata Expertise, Help, and Training Resources

- *http://geology.usgs.gov/tools/metadata/*
- *http://www.fgdc.gov/participation/individual/trainers/*
- *http://www.fgdc.gov/training*
- *http://metadata.nbii.gov/portal/server.pt?open=512&objID=255&&PageID=339&mode=2&in_hi_userid=2&cached=true*
- *http://www.fgdc.gov/metadata*

9.2 Process Steps for Documentation of Data Sets and Other Resources

In order to capture the most detail about a project or data set, metadata development ideally begins at the start of a project and continues through publication. Appendix D of this Toolkit provides an excellent visual portrayal of how critical details and even general information pertaining to a data set are lost over time and how metadata helps prevent such "information entropy." The following steps are meant to help guide documentation in the context of this Toolkit.

1. What type of data or resource is being documented?
   a. For spatial data sets-utilize Documentation Tool #1 in Section 11 of this Toolkit along with the FGDC cross-walk presented in Appendix A of this Toolkit.
   b. For data sets in tabular format containing or not containing a spatial component-utilize Documentation Tool #1 along with the FGDC cross-walk presented in Appendix A of this Toolkit.
   c. For documents and reports-utilize Documentation Tool #2 found in Section 12 of this Toolkit.
   d. For a web-enabled resource-utilize Documentation Tool #2
2. Utilize the appropriate documentation tool for the type of data or information to be documented by filling out the tool's questions throughout the project lifecycle.
3. If Documentation Tool #1 was used for a data set containing spatial data, use one of the tools listed near the beginning of this section. Section 9.3.1 below describes a user-friendly metadata work flow with some of the available free metadata tools noted in this document.
4. If Documentation Tool #2 was used, a text document can be compiled containing the itemized information.
5. To make information about data available to others once documentation is complete, FGDC metadata should be submitted to a metadata clearinghouse such as the NBII Clearinghouse: *http://159.189.176.5/portal/server.pt?space=CommunityPage&cached=true&parentname=CommunityPage&parentid=0&in_hi_userid=2&control=SetCommunity&CommunityID=410&PageID=606*

9.3 Tips and Tricks for Creating Metadata with the FGDC-NBII Biological Data Profile

When working with the FGDC-NBII content standard, there are several workarounds and shortcuts available as a result of the variety of free tools available on the internet. A particularly useful tool is provided by the National Park Service (NPS) Natural Resources Program. It is a free extension for ESRI's ArcCatalog (version 8.3 and 9.x) software that increases the metadata management and editing functionality. This allows for automated extraction of the spatial reference information (Question 18 in Documentation Tool #1 of this document) and the fields and values of the data (Question 17 in Documentation Tool #1-entity and attribute information), as well as the Bounding Coordinates of Question #1 in Documentation Tool #1. Using this Extension with ESRI ArcCatalog, in addition to one of the free tools available such as the U.S. Forest Service Metavist Program, provides a robust way to create FGDC-compliant metadata.

It is advantageous to implement metadata record templates whenever possible for projects or even whole agencies to streamline the process of metadata creation (Giles and Kutner, 2005). Templates are prepared text that can be copied and pasted into areas of a metadata record. The use of templates creates consistency in a project, between projects, or across an organization. For example, if several projects focus on one geographic area, or if all of the contact information and access/use constraints are the same for a series of data sets or agencies, a predefined template can speed up the metadata process because those areas are completed with "boilerplate" text. This can be done by developing the reusable text in Documentation Tool #1 in this document, or by developing a metadata record template with an XML record with one of the metadata software tools in Section 9.1.2 of this Toolkit.

XML (extensible markup language) is a general purpose language that is widely used for describing various types of structured data and information. Developed by the World Wide Web consortium, XML is designed to facilitate data sharing across different systems through its standardized format (World Wide Web Consortium, 2006). Many of the free tools identified above use XML as one of the accepted formats for metadata record importing and exporting. For example, parts of a metadata record can be created in ESRI ArcCatalog, exported as an XML file, and opened in another metadata program such as Metavist. For more information on XML, follow these links: *http://www.xml.com/* or *http://www.w3.org/XML/*

For more information supporting the benefits of metadata, an FGDC "quick-guide" to the standard with further explanations of the elements, as well as methods for making the creation process easier, follow this link: *http://www.fgdc.gov/metadata/metadata-publications-list*

9.3.1 Example Work flows for Metadata Documentation

For geospatial data with attributes resulting from project work, the spatial and entity/attribute elements can be extracted using ESRI ArcGIS ArcCatalog and incorporated into the free U.S. Forest Service metadata tool called Metavist. The free NPS ESRI ArcCatalog Metadata Extension tool also is included in this work flow along with the Free metadata QA/QC tool called MetaParser. Ideally, the ArcCatalog work flow steps are done prior to adding the information collected through Documentation Tool #1 of this Toolkit.

If ESRI ArcCatalog software is not available, a metadata record can still be created using any of the free tools identified above, including the U.S. Forest Service Metavist tool, the NPS Metadata Tool, and the MetaParser tool. Using ArcCatalog simply helps to automate documentation of the elements associated with Questions 1, 17, and 18 of Documentation Tool #1 of this Toolkit.

For further reference, Appendix C of this Toolkit contains a useful FGDC document titled "Top Ten Most Common Metadata Errors" (FGDC, 2006).

For Spatial Data:

1. Be sure that copies of Metavist and ESRI ArcCatalog (8.3 or 9.x) are available on the computer being used.

2. Download and install the free NPS ArcCatalog Metadata Extension available from this link: *http://science.nature.nps.gov/nrgis/tools/tools.cfm*

3. Open ArcCatalog and add the NPS Extension Toolbar to ArcCatalog if it is not already visible. Click the NPS Metadata Editor tab and set the NPS stylesheet to FGDC BioProfile. Add the ArcCatalog Metadata toolbar to ArcCatalog if it is not already displayed. Click the Metadata tab and set the ArcCatalog metadata style sheet to FGDC-NBII or FGDC BioProfile. Select the data to be documented in the ArcCatalog Catalog Tree on the left. On the right side window, make sure the "Metadata" tab is the active tab.

4. If the data set contains biological components (Question 11 in Documentation Tool #1), utilize the Integrated Taxonomic Information System (ITIS) as described in Section 9.4 below to generate the text file containing the taxonomic information and to incorporate it into the metadata record. Otherwise, skip this step.

5. Export the metadata from ArcCatalog by clicking the Export button on the ArcCatalog metadata toolbar. Be sure to specify a known location on the computer, and be sure to specify the format as XML.

6. Verify that the questions in Documentation Tool #1 of this document are complete to the extent practicable for the data set. If the questions in Documentation Tool #1 have not been filled out for the data set, fill them out now. After steps 1–3 above, questions 17 and 18 in Documentation Tool #1 should already be complete in the exported XML, along with question 11 which addresses any biological components (if applicable) and the Bounding Coordinates part of question 1. However, the geographic extent (for example counties or other geographic indicators) description still must be completed.

7. Open Metavist and open the XML record just created in ArcCatalog with Metavist. Verify that the spatial elements, entity/attribute elements, bounding coordinates, and ITIS taxonomic information extracted through steps 3–5 above are correct for the data set. If using a metadata template with boilerplate language for specific sections as per Section 9.3, copy/paste that information into the appropriate sections. For any metadata keywords, use the NBII Keyword Thesaurus Catalog at: *http://www.nbii.gov/portal/server.pt?open=512&objID=578&&PageID=1798&mode=2&in_hi_userid=2&cached=true*

8. Incorporate the information from Documentation Tool #1 into the FGDC-NBII Biological Extension Standard format through Metavist by using the "Cross Walk of FGDC Interview Questions to the FGDC-Biological Data Profile Metadata Standard" provided in Appendix A of this Toolkit. This serves to cross walk the information associated with the questions to the various sections of the FGDC-NBII metadata profile. Simply copy/paste information from Documentation Tool #1 into the appropriate FGDC elements in referenced in Appendix A with the computer mouse if/where possible.

9. Save the XML file to a known location on the computer. Validate the metadata record to ensure its compliance with the FGDC format and the required fields for FGDC Metadata. This can be done quite easily with an online version of the FGDC metadata validation tool called MetaParser at this Web site: *http://geo-nsdi.er.usgs.gov/validate.php*

   MetaParser is a Quality Control tool for metadata that checks for the required elements of the FGDC-NBII Biological Extension Standard. After running, MetaParser generates error reports showing areas of the record that need to be filled out or corrected. The online version above is the easiest way to use MetaParser for metadata validation as it requires just uploading the XML file in progress.

   Also, MetaParser can be used by downloading the free MetaParser program to the computer. MetaParser also is available with a Windows interface from the links at the beginning of this section. MetaParser has the capabilities to generate text files and HTML files, which are more readable than the standard working XML format.

10. Once the record has been validated, submit it to the NBII metadata clearinghouse so that this valuable data set can be maintained and discoverable by others. If any other pertinent information is needed for the metadata, the submitter will be contacted by phone or email.

    To submit metadata, go to this web site and follow the instructions: *http://159.189.176.5/portal/server.pt?space=CommunityPage&cached=true&parentname=CommunityPage&parentid=0&in_hi_userid=2&control=SetCommunity&CommunityID=410&PageID=606*

For Non-Spatial Data in Tabular Format:

1. Be sure that copies of Metavist and ESRI ArcCatalog (8.3 or 9.x) are available on the computer being used.

2. Save the data to be documented in a Microsoft Access or a Microsoft Excel table saved in dBase IV .dbf format.

3. Download and install the free NPS ArcCatalog Metadata Extension available at the above link.

4. Open ArcCatalog and add the NPS Extension Toolbar to ArcCatalog if it is not already visible. Click the NPS Metadata Editor tab and set the NPS stylesheet to FGDC BioProfile. Add the ArcCatalog Metadata toolbar to ArcCatalog if it is not already displayed. Click the metadata tab and set the ArcCatalog metadata stylesheet to FGDC-NBII or FGDC BioProfile. Select the data to be documented in the ArcCatalog Catalog Tree on the left. On the right side window, make sure the metadata tab is the active tab.

5. If the data set contains biological components (Question 11 in Documentation Tool #1), utilize the Integrated Taxonomic Information System (ITIS) as described in Section 9.4 below to generate the text file containing the taxonomic information and to incorporate it into the metadata record. Otherwise, skip this step.

6. Export the metadata from ArcCatalog by clicking the Export button on the ArcCatalog Metadata toolbar. Be sure to specify a known location on the computer, and be sure to specify the format as XML.

7. Verify that the questions in Documentation Tool #1 in Appendix A of this document are complete to the extent practicable for the data set. If the questions in Documentation Tool #1 have not yet been filled out, do so. Be sure to fill out all applicable questions. Questions 17 should already be complete in the exported XML file during steps 1–6 of this work flow. Question 18 is not applicable since the data do not have a spatial component. However, a description of the geographic extent (for example, counties or other geographic indicators) of the data must be given.

8. Open Metavist and open the XML record just created in ArcCatalog with Metavist. Verify that the spatial elements, entity/attribute elements, bounding coordinates, and ITIS taxonomic information extracted during steps 3–5 above are correct for the data set. If using a metadata record template with boilerplate language for specific sections as per Section 9.3, copy/paste that information into the appropriate sections. For any metadata keywords, use the NBII Keyword Thesaurus Catalog at: *http://www.nbii.gov/portal/server.pt?open=512&objID=578&&PageID=1798&mode=2&in_hi_userid=2&cached=true*

9. Follow steps 8–10 directly above this work flow.

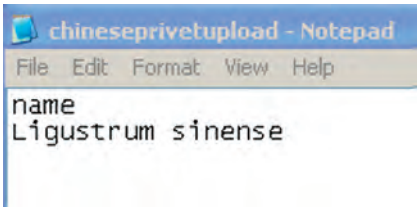
9.4 Instructions for Utilizing ITIS (Integrated Taxonomic Information System) for Documenting the Biological Dimensions of a Data Set

The Integrated Taxonomic Information System (ITIS) is a tool developed through a partnership of several Federal and non-governmental agencies (*http://www.itis.gov* and *http://www.itis.gov/taxmatch_ftp.html*) in response to the need for scientifically credible taxonomic information (Integrated Taxonomic Information System, 2004). ITIS provides an easily accessible database with reliable information on taxonomic names and classifications. The database is reviewed periodically to ensure high quality with valid classifications, revisions, and additions of newly described species (Integrated Taxonomic Information System, 2004). ITIS includes documented taxonomic information of flora and fauna from both aquatic and terrestrial habitats.

ITIS provides a function to generate a downloadable text file containing taxonomic classification elements that can be incorporated into an XML-formatted metadata record. Note: This function of ITIS only works with Scientific Names, and can only search within one Kingdom at a time. Below are the steps for creating an example file using the ITIS tool.

1. Create a text Notepad file containing the scientific names of a species and a header line stating "name" and save it to a known location on the computer. As an example, in the screen capture below is the scientific name for Chinese Privet. Note: The search tool is case sensitive—Genus species.

   An example text file should look like this:

   All species included in a text file must be of the same Kingdom. Multiple species may be included in one file.

2. Open up *http://www.itis.gov/taxmatchftp.html* in an internet browser. Using the Upload tool in the center, browse to the file created in step one (saved on the computer) and use the Upload button to send it to ITIS.

3. Select the Kingdom for the species in the list in step 3. Select "Scientific Name (FGDC Biological Profile Report)" in step 4. Select both "View Matches" and "View non-matches" in Step 5.

4. Click the Taxonomy Compare button to run. This goes to the Report summary page. This table lists all the matches and non-matches between the species uploaded and the ITIS database. If any are listed under "Non-matches from Input Data", check the spelling in the file, or if necessary use one of the ITIS search functions (*http://www.itis.usda.gov/*) to see what the ITIS-endorsed species name is. To get to the necessary SGML file click the "Generate FGDC Biological Profile SGML" button.

5. Click "Download SGML data." The file should open directly in the web browser. Go to "FILE"- "SAVE AS" in the web browser to save as text file (.txt). Be sure to specify Unicode (UTF-8) encoding. Save it to an appropriate location on the computer.

6. There are two ways to incorporate the taxonomic information from ITIS into a metadata record. The first is to utilize the previously mentioned ArcCatalog NPS Metadata extension. (This can also be done with the stand-alone NPS Metadata Tool.) With ArcCatalog open and the NPS Metadata Editor activated, open the in-progress metadata file through ArcCatalog, and click the "Import ITIS" button on the "NPS Metadata" dropdown tab/button. Browse to the location of the text file that was downloaded in step 5. Note: The NPS tool will warn if the record already contains the Taxonomic Classification section. The existing taxonomic classification will not be deleted but appended to with the taxonomic information from the ITIS text file.

   The second method of incorporating the taxonomic information from ITIS is a slightly more manual way done with the actual XML metadata file. For this method, create a metadata record using Metavist (mentioned at the beginning of this metadata section) and(or) by following steps 1–8 above. Save the in-progress metadata record as an XML file. Open the XML file in Notepad, and open the ITIS download using Notepad as well. Copy and Paste the text from the ITIS file into the XML just below the </keywtax> tag in the Metavist XML metadata file, and save with Unicode (UTF-8) encoding. Open the file again and the taxonomic information should be complete.

9.5 Tips for Documenting Legacy Data Sets

A common issue when creating metadata is how to properly create documentation for legacy data sets. Properly documenting data from the outset helps to avoid wasted time and money that must be sunk into documenting data sets 1, 2 or even 10 years after the fact. At best, metadata are not viewed as something done at the end of the data development process but are recorded through the life of a data set so that vital details do not become lost over time. Nevertheless, legacy metadata are better than no metadata at all.

When documenting a legacy data set, it can be a frustrating effort to uncover fundamental details. Often, uncertainties or holes will remain in the documentation even after completion. Depending on the data set, it might be difficult or impossible to ascertain whether or not the data set has been fully documented. Typically, it is necessary to use several different sources (hardcopy documents/reports/field notes, databases in older file formats, and(or) personal interviews with people that were involved with the project) must be thoroughly searched in order to gather as much information as possible. The unfortunate reality is that it is unlikely that the data set can be documented as well as it could have been had the metadata process been an integral part of the project from the beginning.

When creating legacy metadata, it may be helpful to include a statement such as: "This metadata record documents legacy data to the extent practical, as required by Executive Order 12906, 'Coordinating Geographic Data Acquisition and Access: The National Spatial Data Infrastructure,' dated April 11, 1994." This can serve as a disclaimer so that the user is made aware of the situation they are working with (Terry Giles, U.S. Geological Survey, written commun., 2005). This could be added at the top of the Supplemental Information section (Question #20 in Documentation Tool #1 of this Toolkit).

The legacy metadata record can be enhanced by citing the sources (people and documents) used for the metadata (Terry Giles, written commun., 2005). It is best to include some information about where the content of the metadata was obtained, or at least to include a list of sources that were used when creating the metadata. For example, one might include a reference to "maps found in the XYZ refuge folder, in Region X's land management office's filing cabinet" and include a means of contacting that office. Or in the case of someone who was interviewed, obtain permission to include the name and a means of contact. Some type of caveat could be included in the record stating that some details might be missing, but given the resources available, the metadata record is as complete as possible within a reasonable amount of time and work (Terry Giles, written commun., 2005).

A good place to include such disclaimer and reference information would be in the Supplemental Information section (in Documentation Tool #1 of this document, Question #20). Depending on the information, it could be noted in the particular section in which it is recorded during the metadata creation process.

9.6 Additional Metadata Information Sources

- The National Park Service Northeast Temperate Network Inventory and Monitoring Program Data Management Page. This site contains example NPS reference documents for GPS metadata, project metadata, and metadata tools among other information: *http://science.nature.nps.gov/im/units/NETN/downloads/Plan/NETN_DataManagementPlan.pdf*

- For useful FGDC metadata publications about the benefits of metadata and the most common misconceptions about metadata: *http://www.fgdc.gov/metadata/metadata-publications-list*

10 Geospatial Data Acquisition Guidelines for Quality

A Geographic Information System (GIS) allows different types of data (for example, spatial and non-spatial) to be integrated and associated in ways that would otherwise be difficult. This section addresses common issues associated with acquiring new data and converting it into a GIS-usable format. Because nearly 80 percent of all data have a spatial component, the primary focus is on issues associated with spatial data collection. Conformance to standard methodologies when working with spatial data is crucial to allow for interoperability between data sets and consistency over time. A consistent level of detail is vital for data quality and credibility.

10.1 Geo-Referencing With GPS Data Collection

This section includes GPS data collection, format, and processing protocols based on industry-accepted and Federal guidelines. The guidelines below describe recommended settings primarily for highly accurate mapping grade GPS receivers. This section also addresses other critical issues to be aware of when collecting GPS data that could affect accuracy and consequently the expected outcome of a project. This section is not intended to be comprehensive but is primarily intended to address the prominent issues associated with GPS data collection, to provide example protocols and equipment settings to be standardized within projects, and to serve as a supplement to other guides and resources.

The references in this section, in addition to the web links cited, provide additional information on this and related topics. The objective is to address critical GPS data collection issues in a way that is not brand-specific and that assumes access to popular GIS software, such as ESRI ArcGIS and possibly ESRI ArcPad GPS software. As with any aspect of any project, the desired accuracy and means by which to obtain it (for example, equipment and settings) can be evaluated and specified in the QA plan.

If a mapping grade GPS unit is used for field work, it will allow for a consistent level of quality positional accuracy (National Park Service, 2004). A mapping grade GPS unit is useful for maintaining the accuracy and integrity of a project GIS. As with any information system, the quality of information coming out depends on the accuracy of the data going in. To gain a better understanding of the critical differences between recreational and mapping grade GPS units, see the link below to the ESRI Online Article-Recreational vs. Professional GPS: What's the Difference? *http://www.esri.com/news/arcuser/0104/rec-gps.html*

Another helpful resource is the document, GPS to GIS Procedural Handbook and Reference Guide by Mark Roper of the San Juan National Forest, found at the link below. This Forest Service guide discusses Forest Service accuracy guidelines and similar specifications as well several other GPS related issues. It also addresses issues such as converting between the North American Datum of 1927 (NAD 27) and other datums and includes information regarding mapping accuracy standards, GPS/GIS operations, and other common issues that may be useful to anyone engaged in GPS data collection. This document can be found at: *http://www.fs.fed.us/database/gps/gpsusfs.htm*

Also, the Draft GPS Data Accuracy Standards document from the U.S. Forest Service (U.S. Forest Service, 2003) is available at: *http://www.fs.fed.us/database/gps/gps_standards/GPS_Data_Standard.pdf*

10.1.1 GPS Unit Data Collection – Mapping Grade Unit Settings

The following discussions are based on Busch and Morrison (2001a), Busch and Morrison (2001b), ESRI (2006a), Forest Science Research Network (2002), National Geodetic Survey (2006), National Park Service (2006), Roper (2005), Roper (2006a), National Park Service (2005), U.S. Forest Service (2004), and U.S. Forest Service (2003).

- **Mission planning**
  Software or GPS unit capabilities can be used to ascertain satellite availability, positioning, and Position Dilution of Precision (PDOP) for the project location for tentatively selected field work dates. This allows for the best day(s) and time of day for data collection quality and maximized productivity to be determined.

  The GPS unit or GPS software package usually allows for the approximate coordinates to be entered (nearest degree is sufficient) for the work location. Before doing this, be sure to obtain an up-to-date satellite almanac file by going outside and turning the GPS unit on for at least 15 minutes. A GPS satellite almanac file provides the orbits, GPS clock correction information, and atmospheric delay parameters for the GPS satellite constellation. Almanac files are globally applicable, so physical location relative to the field work location does not matter when doing this. The current almanac file provides the approximate PDOP and satellite availability for the intended field dates at the field work location.

  Trimble, a company that provides GPS and other positioning technologies, has a free downloadable stand-alone software tool that allows for proper mission planning so one can ascertain visibility for GPS,

GLONASS, IGSO and geostationary satellites for a project location. This can be found at: *http://www.trimble.com/planningsoftware.html*

The Trimble web site also provides an up-to-date almanac/ephemeris file that can be obtained at the link below to be used with the software: *http://www.trimble.com/planningsoftware_ts.asp?Nav=Collection-8425*

- **Satellite Availability**
  A minimum of four satellites should be available when collecting data so that a 3D position can be obtained. If using real-time differential correction (addressed below) by way of Wide Area Augmentation System (WAAS) or another method, five or six satellites may be needed to achieve proper real-time correction functioning.

- **Differential Correction**
  To increase the accuracy and precision of GPS data, ensure that at least WAAS real-time correction is turned on if another preferable method of correction (external beacon for real-time correction or differential post-data correction) is not available. Such correction systems help to account for and remove error that may be introduced during the data collection process. Such error can be a result of GPS satellite clock errors, orbit/ephemeris errors, multi-path errors, and(or) atmospheric effects.

  WAAS should not be relied upon solely as the primary means of differential correction for Differential Global Positioning System (DGPS). WAAS was originally designed by the FAA for use by aircraft to approximately 200 feet above the earth's surface. Because of this, WAAS may not be reliable under a vegetation canopy or when the southern sky is obstructed (National Park Service, 2004).

- **Position Dilution of Precision (PDOP)**
  The lower the PDOP, the better the data quality. The receiver should at least be configured to stop collecting data when the PDOP is more than 6. If the receiver can be configured to do so, it is useful to log PDOP values.

- **Elevation Mask Angle**
  If user configurable, it is useful to set the elevation mask angle of the GPS unit to 15 degrees or greater, depending on surrounding vegetation and topography.

- **Signal-to-Noise Ratio (SNR)**
  Ideally, set the SNR to less than or equal to 4.

- **ESRI ArcPad GPS Software Parameter Settings**
  ArcPad parameters depend largely on the type of GPS receiver being used. The GPS to GIS Procedural Handbook and Reference Guide addresses common ArcPad parameters, depending on the type of unit being used (Roper, 2005). Any applicable GPS unit settings ideally will match the settings in ArcPad (for example, the Geographic Datum being used). World Geodetic System 1984 (WGS 84) is typically the default datum used for GPS units, but it should never be assumed that the unit's display datum is set to WGS 84 (always verify!).

10.1.2 GPS and Coordinate Systems

The following discussions are based on Bisio (2005), ESRI (2006e), ESRI (2006f), National Geodetic Survey (2006), Roper (2005), Roper (2006a), Mark Roper, written commun., 2006, Roper (2006b), National Park Service, (2004), and from primary author's personal knowledge.

If not using ESRI ArcPad, these options are considered.

- The standard datum for GPS units is typically WGS 84, so data collected will be in WGS 84. (It is necessary to always confirm this with the unit in use.) Depending on the type of GPS unit being used, desktop PC software may be available for exporting the data to GIS format. Most resource grade units offer this type of software. For a list of options and suggestions, the GPS to GIS Procedural Handbook and Reference Guide provides good information and links to discussions about how this may be handled (Roper, 2005).

If using ESRI ArcPad GPS software, these options are considered.

- What is the datum and projection of the GIS layers that are to be used with the collected data? If the collected data will be added to an existing GIS geodatabase, the datum and projection in the ArcPad map on the unit need to match the datum and projection used in the geodatabase. Do not intermix NAD 27 data, North American Datum of 1983 (NAD 83) data, and WGS 84 data because substantial positional errors could result.

- If performing additions or updates to GIS data, a projection file (.prj) needs to accompany any GIS files loaded into ArcPad from the source data for use in the field. The shapefile to which ArcPad is adding GPS features needs a .prj (projection) file so as to allow the map in ArcPad to correctly display the data. If the shapefile does not have a .prj file, ArcPad will assume a default datum of WGS 84 and the GIS data may not display correctly.

- Please make this agree with the title of 10.1.4.

10.1.3 GPS Data Collection Guidelines for Accuracy

The following discussions are based on Busch and Morrison (2001a), Busch and Morrison (2001b), National Park Service (2006), National Park Service (2005), Roper (2005), Mark Roper (written commun., 2006), Subcommittee for Base Cartographic Data (1998), U.S. Forest Service (2003), U.S. Geological Survey (1947), and from primary author's personal knowledge.

The following data collection specifications serve as an example of the types of details that can be standardized in a project. Level of accuracy can be dependent on several things (terrain and surrounding relief, canopy, time of day, GPS settings, and(or) user's knowledge of the equipment being used) and consequently is contingent upon, but not limited to, the items addressed herein. As with any aspect of data management, the logging interval specifications below ideally would be addressed at the outset of a project.

- **Federal Geospatial Accuracy Standards**
  National accuracy level standards include the National Map Accuracy Standard (NMAS) and the National Standard for Spatial Data Accuracy (NSSDA). The NMAS standard, drafted in 1947, was intended for published graphical maps (U.S. Geological Survey, 1947). Because accuracy is dictated by map scale, the NSSDA standard was developed to account for the dynamic nature of digital geospatial data and GIS applications (Subcommittee for Base Cartographic Data, 1998). The NSSDA standard is specified by the Federal Geography Data Committee for use as a means to "evaluate and report the positional accuracy of geospatial data produced, revised, or disseminated by or for the Federal Government" per Executive Order 12906 (U.S. Forest Service, 2003). "NSSDA was developed to report accuracy of digital geospatial data that is not constrained by scale" (U.S. Forest Service, 2003).

  The NSSDA standard (Part 3 of the Geospatial Positioning Accuracy Standards) can be found at: *http://www.fgdc.gov/standards/standards_publications/*

  The NMAS standard can be found at: *http://nationalmap.gov/gio/standards/#overview*

  The USGS GPS Data Accuracy Standard also addresses accuracy reporting and describes it in the context of the National Map Accuracy Standard. The GPS to GIS Procedural Handbook and Reference Guide also addresses the NSSDA standard in conjunction with the NMAS standard, in terms of accuracy and distances. They are both available at the following link: *http://www.fs.fed.us/database/gps/gpsusfs.htm*

- **Point Data Collection**
  A logging interval set to one position per second is optimal for accuracy. For point data locations, the positions taken at that point can be averaged. Ideally, movement is minimized while collecting/averaging point data as this can result in substantial error. A good GPS point location can generally be obtained by taking 120 positions at a one second interval. Depending on the type of unit, as well as the project needs, more or fewer positions may be appropriate.

  If data are being collected in poor GPS signal areas with steep surrounding elevation and heavy vegetation/canopy, small movements of the receiver may help with signal acquisition if there is difficulty in getting a signal.

- **Line Data Collection**
  The logging interval set to one position every five seconds for linear features, is optimal. Averaging of vertices of lines or polygons also may help increase accuracy.

- **Polygon Data Collection**
  A logging interval set to one position every five seconds for polygon features is optimal.

Below are some tips to ensure accuracy in data collection.

- Holding the GPS unit at approximately eye level with minimal movement is helpful for accuracy.

- A GPS data dictionary can be predefined prior to data collection so as to simplify data collection and to standardize the data that are collected. A GPS data dictionary lists standard feature attributes (characteristics of data) to be collected with data and the default values and valid values and domains for the attributes. This can be developed in conjunction with any applicable project data model, or, if the data are to be added to an existing GIS, the GPS can be set up so that the data are recorded with attributes that match the existing GIS layers for compatibility.

- The use of GPS data dictionaries can help to speed up field work and also can be a form of Quality Control by reducing user/human error.

Many resource grade mapping units offer desktop PC software that allows for the development of data dictionaries for use in the field when collecting data. If using ESRI ArcPad software, ArcPad Application Builder can be used to develop custom forms for attributes. This also serves as a form of Quality Control for field spatial data by helping control what attributes and the type of attributes that are entered.

- ESRI ArcPad Tip for Capturing Latitude, Longitude, and Altitude values as Attributes

  For automatically capturing the X, Y, and Z values in the appropriate respective fields in ESRI ArcPad

while collecting data, an ESRI ArcScript is available that can be easily installed to collect this data. ArcPad does not automatically record these attributes when collecting data. Use of this script only in conjunction with real-time differential correction or autonomous (non-corrected) GPS is preferred. Post-processing differential correction might result in discrepancies between the X, Y, and Z attribute values and the GIS corrected locations of the data. This script is available at *http://arcscripts.esri.com/details.asp?dbid=12850*

10.1.4 Datum Transformations: Differential Correction and the WGS 84 and NAD 83 Datum Transformation Issue

The following discussion is based on Bisio (2005), ESRI (2006a), National Geodetic Survey (2006), Roper (2005), Roper (2006a), Mark Roper (written commun., 2006), Roper (2006b), Gary Thompson (North Carolina Geodetic Survey. oral commun., 2006), and Trimble Navigation Limited (2005).

This section addresses the problem associated with the difference between the geographic datum WGS 84 and the geographic datum NAD 83. This is essential to consider when attempting to achieve sub-meter GPS accuracy for a project. This problem can be avoided altogether if all project data are kept in the WGS 84 geographic datum. WGS 84 is the primary geographic datum for most GPS units and differential correction sources.

A proper transformation needs to be made to reconcile data collected with data in an existing GIS if they are calibrated to different geographic datums (WGS 84 and NAD 83). Previously, NAD 83 and WGS 84 were nearly identical, but with the evolution of datums over time, a small difference has developed between the two and thus they are no longer assumed to be the same (Bisio, 2005). As a result, error of up to a meter can be introduced into data when a transformation is not properly made between the two (Bisio, 2005). For example, if data collected in the field need to be compatible with data in a project GIS in NAD 83, the appropriate transformation needs to be made. Error resulting from this datum evolution can be introduced as a result of any differential correction performed on the data. The geographic datum of any differential correction source used (both real-time and post-processing) needs to be considered before data collection is carried out.

When differentially correcting GPS data (real-time or post-processed), the corrected data will be in the geographic datum of the correction source used. If using WAAS (Wide Area Augmentation System) real-time correction, a National Geodetic Survey CORS base station (*http://www.ngs.noaa.gov/CORS/cors-data.html*), or any other station that lists its survey coordinates with a WGS 84 or ITRF00 position (the datums ITRF00 and WGS 84 are nearly the same), the corrected data will be in WGS 84 or ITRF00. The default datum for nearly all GPS units is WGS 84, so GPS data will be in WGS 84 if no corrections have been made. With this scenario, a proper transformation needs to be made so that error is not introduced if the data are to be used with existing GIS data in NAD 83. This would entail using a conversion with NAD 83 (CORS96) to correctly carry it out.

On the other hand, if using a Coast Guard Beacon for real-time correction or if working with another source that provides their data in NAD 83, the corrected data would already be in NAD 83. In this case, NAD 83 (Conus) would need to be used for the transformation which assumes WGS 84=NAD 83 so the coordinate values do not change.

Here is a link to an ESRI technical article that addresses transformations between geographic datums, as well downloadable documents that show the various datum transformation methods that are supported by different versions of ESRI ArcGIS (ESRI, 2006g): *http://support.esri.com/index.cfm?fa=knowledgebase.techarticles.articleShow&d=21327*

Below are guidelines for converting data to NAD 83 using both ESRI ArcGIS and ESRI ArcPad. If using ESRI ArcPad, the appropriate transformation ID to use in ArcPad will depend on the version of ArcPad being used.

Working with a WGS 84 (or ITRF00) base station correction source in real-time or post-processing, or using autonomous (non-corrected) GPS

The following are modified from Roper (2005) Mark Roper (written commun., 2006), and Roper (2006b).

- **Performing the transformation from WGS 84 to NAD 83 using ESRI ArcPad:**
  Datum Transformation #8494 is used in ArcPad 6, whereas in ArcPad 7 Datum Transformation #1515 is used.

- **Performing the transformation from WGS 84 to NAD 83 using ArcGIS 9.x:**
  Use the ArcToolbox Data Management Tool: Projections and Transformations: Feature: Project Tool to reproject the data. This can be done for GIS data collected using ArcPad, or from data exported to GIS file formats from GPS unit-specific software, or from GPS data manually entered into tables and converted into Feature Classes (see Section 10.1.5 below). When prompted for the option of specifying a Geographic Transformation, specify Transformation #1515 – NAD_1983_To_WGS_1984_5.

  Note: Datum transformations work in either direction in ArcGIS. For example, this transformation listed as NAD_1983_to_WGS_1984_5 will transform NAD 1983 to WGS 1984 or WGS 1984 to NAD 1983.

Working with a NAD 83 base station correction source in real-time or post-processing

Discussion is modified from Roper (2005), Mark Roper (written commun., 2006), and Roper (2006b).

- **Performing the transformation to NAD 83 using ESRI ArcPad:**
  Datum Transformation #8088 is used in ArcPad 6, whereas in ArcPad 7, Datum Transformation #1188 is used.

- **Performing the transformation to NAD 83 using ArcGIS 9.x:**
  Datum Transformation #1188 is used. Use the ArcToolbox Data Management Tool: Projections and Transformations: Feature: Project Tool to re-project the data. This can be done for GIS data collected using ArcPad, or from data exported to GIS file formats from GPS unit-specific software, or from GPS data manually entered into tables and converted into Feature Classes (see Section 10.1.5 below). When prompted for the option of specifying a Geographic Transformation, specify Transformation #1188 – NAD_1983_To_WGS_1984_1.

When taking data from a GIS back into the field for updating or for use in navigating to a previously mapped entity

- Care is needed when taking data back into the field with a GPS field device. It is best if the data are converted into the datum of the differential correction source method to be used (Bisio, 2005). Depending on the respective datums, this can be done with ESRI ArcGIS as described above. Doing so is critical in that it helps maintain the integrity of the data and ensures that features are accurately identified and updated. Data creep can result if many check-ins/check-outs are performed without proper datum adjustment.

10.1.5 Getting GPS Data in a GIS-Compatible Format

The follow discussion is based on ESRI (2006d) and , primary author's personal knowledge)

There are a few options for GIS-enabling data captured from a GPS unit for use in map-making and analysis in conjunction with other data:

- **Using ESRI ArcPad:**
  The advantage of using ArcPad when collecting data is that it can create shapefiles while in the field, in addition to recording the desired attributes associated with the data. ESRI ArcPad also allows for better accuracy in that points, lines, and polygons can be specified for the type of GIS data being collected. This allows for minimal processing once back in the office.

- **Using ESRI ArcGIS or GPS unit specific software**
  ESRI ArcGIS can be used to format the data being collected for use with a GIS. Also, many GPS units come with computer software that allows data to be exported in a GIS format. Some web links to other options are presented at the end of Section 10.

  Though it is the least-desirable option because of accuracy issues, GPS data collected with a recreational grade unit can be formatted for use with a GIS. Data that were manually recorded on field sheets can be entered into a Microsoft Access table or a Microsoft Excel table (saved as a DBF IV dBase file) and can be used to generate a point feature class or shapefile. This can be done with ESRI ArcCatalog. The ArcGIS Desktop Help has instructions on how to do this. These can be found by searching the index for "x,y, creating geographic data from."

  The following are some things to keep in mind when doing this.

  1. Columns have to be explicitly formatted (for example numeric, specify width, and number of decimals).

  2. Field names are to be 10 characters or less, with no spaces and beginning with a character, not a number (underscores are an acceptable spacing character).

  3. Blank rows are not permitted anywhere in the table.

When adding x,y data using this approach, coordinates need to be recorded in the table in decimal degrees. Be sure to keep track of the datum of the collected data because the datum needs to be specified when creating a feature class or shapefile in ArcCatalog since that datum is what the coordinates were captured in. Once the feature class or shapefile is created, it will need to be displayed in the appropriate projection. Be certain to read Section 10.1.4 of this document. It addresses transformation issues between WGS 84 and NAD 83. The metadata for these data need to also reflect this methodology and accuracy so that users can understand how these particular data were captured and GIS-enabled.

For converting GPS data from latitude/longitude into Decimal Degrees and vice versa, here is a link to a web site designed by the Federal Communications Commission (FCC) for automatic conversion: *http://www.fcc.gov/mb/audio/bickel/DDDMMSS-decimal.html*

10.1.6 GPS Accuracy Reporting and Recording Critical Information in FGDC Metadata

The following discussion is based on Roper (2005), Mark Roper (written commun., 2006), Subcommittee for Base Cartographic Data (1998), U.S. Forest Service (2003), and from the primary author's personal knowledge.

- It is necessary to report GPS accuracy in accordance with Executive Order 12906 for all Federal Agencies. The Geospatial Positioning Accuracy Standards Part 3, National Standard for Spatial Data Accuracy (NSSDA), FGDC-STD-007.3-1998 provides guidelines on this standard. This requires a 95 percent confidence level for all geospatial data acquired in the context of the accuracy level deemed appropriate for the project or application. The accuracy reporting should be done by comparing the data set with an independent data set of higher accuracy. It may be possible to compare GPS positions with other existing source data of a known accuracy or preferably with imagery of the project area in the correct projection and of a known accuracy.

- The NSSDA standard is meant to be used as the means for reporting accuracy levels for newly collected data. It is helpful to include an explanation regarding the accuracy of the horizontal coordinate measurements and vertical coordinate measurements (if applicable) and how the accuracy confidence level was obtained. In this Toolkit, Question 16 of Documentation Tool #1 is the appropriate section for that information. The Geospatial Positioning Accuracy Standards PDF document presented here describes acceptable accuracy testing methods and proper format, such as "Tested or Compiled to meet": *http://www.fgdc.gov/standards/projects/FGDC-standards-projects/accuracy/part3/chapter3*

The USGS GPS Data Accuracy Standard also addresses accuracy reporting and describes it in the context of the National Map Accuracy Standard. The GPS to GIS Procedural Handbook and Reference Guide also addresses the NSSDA standard in conjunction with the NMAS standard. They are both available at the following link: *http://www.fs.fed.us/database/gps/gpsusfs.htm*

In addition to the accuracy statement, it is helpful to describe the techniques used, including:

- **Attribute Descriptions:** This can be gathered from the project GPS data dictionary, including features, attributes, and attribute values. This can be addressed with the metadata documentation work flows in Section 9.3.1 if the data are in a table or GIS format.

- **Source:** GPS receiver type, antenna type, receiver settings (SNR, PDOP, logging rates, or Elevation Mask), number of positions per point feature, correction method, and any other relevant field procedures or equipment details. These elements can be included in Question #16 of Documentation Tool #1 in this Toolkit.

This critical information enables users to easily evaluate how the data fit their applications requirements.

10.2 Scanning and Geo-Referencing Hardcopy Data and Maps

This section addresses issues that may occur when digitizing and spatially enabling hardcopy maps and data for use in a GIS. If working with aerial imagery, it is crucial that the desired resolution, the needed level of accuracy and the end use of the derived product be evaluated and determined so that appropriate methods and software are used.

10.2.1 Scanning Hardcopy Maps and Other Source Data

This discussion is based on Will Fontanez (University of Tennessee. written commun., 2006).

- **Scanner size:** Generally a scanner the size of 11x17 inches or 12x18 inches is good for most scanning applications.

- **DPI (dots per inch) resolution:** This determines the visible detail or quality of the item that is being scanned. DPI also affects the file size of the scan (for example, the higher the DPI, the larger the file size). The physical size of the source data also affects the file size. (for example, a full-size USGS topographic map scanned at 600 DPI would be an extremely large file.

If the scan is to be used as a background in a GIS, a lower resolution DPI is generally sufficient. If it is to be part of a print, a higher resolution DPI can be used. A minimum threshold of 150-200 DPI gives a good display on a computer screen. A minimum DPI of 300 is a good quality threshold if the scan is to be printed. A DPI of 600 is generally good for source data and items with a high level of detail. However, depending on the final product needs, a color scheme such as a greyscale can be used to cut the file size down by one-third. The concept of greyscale does not mean that the scan will be grey in color, but that it will be one uniform color. This can cut the file size down by a third, and typically the uniform color can be specified.

- **File format:** Generally, the TIFF file format works well for scanning output because it preserves detail fairly well.

- It is also advisable to have imaging software, so that the scans can be cleaned up (for example, any unwanted text or features removed).

10.2.2 Geo-Referencing

This discussion is based on ESRI (2006d) and the primary author's personal knowledge.

This section addresses basic georeferencing of raster data sets using ESRI ArcGIS software.

1. Be sure that the source (already referenced) data set is in the projection determined appropriate.
2. Enable the Georeferencing Toolbar in ArcMap.
3. Add to the map window the referenced data set and the data set to be geo-referenced.
4. Use the Rotate and Shift Georeferencing Tools in ArcMap to get the target layer initially as close as possible to the aligned data set.
5. A minimum of three link/control points should be used.
   - If possible, the links are spread over the entire raster data set rather than concentrating them in one area. Typically, having at least one link near each corner of the raster data set and a few throughout the interior produces the best results.
   - Use well-known and identifiable features such as road intersections, waterways, identifiable land features, building corners, or other objects that can be clearly identified in the referenced and target data set.
   - If multiple maps for the same geographic area will be geo-referenced, use the original referenced data each time (the original) so that error from one newly geo-referenced data set is not propagated to others.
   - For detailed specifics and functionality on using ESRI ArcGIS for geo-referencing data, consult the Help Files included with the software or consult the ESRI online help files at: *http://webhelp.esri.com/arcgisdesktop/9.1/index.cfm?TopicName=welcome*
6. Depending on the end use of the rectified data set, the data set can be permanently rectified/transformed to create a new data set that can be saved either in GRID, TIFF, or Erdas IMAGINE format. If analysis will be performed on it or if it will be used in another software package, it might be best to permanently transform the data set. The ESRI ArcGIS help files can provide more information on formats and their appropriate applications.

10.3 Digitizing Features from a Geo-Referenced Digital File

This discussion is based on ESRI (2006d), ESRI (2006e), and the primary author's personal knowledge.

This section addresses digitizing vector features from a source image or data set for use in a GIS by digitizing on screen, also known as "heads-up digitizing."

Using ESRI ArcGIS software:

- Be sure that the source electronic data to be digitized are projected in the correct coordinate system. The data frame properties in ArcMap can be set to the preferred coordinate system for the data before adding the data to be digitized. This helps ensure that any layers added to the map are projected on the fly to the preferred coordinate system.
- Create an empty feature class in a personal geodatabase using ESRI ArcCatalog. Name it appropriately, select the feature type, and be sure to specify the correct spatial reference (the geometry). Once the feature class is created, be sure to add the attribute fields to be included with the data. Here the attribute data types, as well as the attribute field "precision," are specified. In ArcGIS, characters such as dashes, spaces, and brackets are not supported. Underscores can be used to denote a space between characters, however.
- If a unique Object Identifier is to be used with the data other than the one automatically created by ArcGIS, be sure to create an attribute field for this.
- Once any attribute fields have been edited, add the empty feature class to ArcMap along with the source data. Be sure to set appropriate quality assurance guidelines before digitizing any features. These might include the following:
  - **Snapping Environments:** Snapping refers to "An automatic editing operation in which points or features within a specified distance (tolerance) of other points or features are moved to match or coincide exactly with each others' coordinates" (ESRI, 2006e). Snapping environments can be very helpful for ensuring data quality, for example, by making features connect precisely with other features when digitizing land parcels from a residential subdivision plat.

    Setting the snapping environments during an editing session can be configured with the following specifications: snapping priorities which specify

layers and what parts of the features in those layers (vertex, edge, endpoint) are subject to snapping, as well as layer snapping priority ranking, and snapping tolerance, which refers to the proximity or distance within which a feature will snap to another.

- **Selectable Layers:** Selectable layers refer to the layers that have select permissions turned on. Layers with this setting turned on can have their features selected with the various ArcMap selection tools. Turning off the selection option for a layer can help, for example, during an editing session and subsequent digitizing process so that layers are not unintentionally selected. Selectable layers can be configured by clicking on the Selection tab on the ArcMap table of contents. There the various data layers can have their selectable state turned on or off.

- Be sure the Editing toolbar is added to ArcMap and that an editing session has been started. Verify that the "Target" on the Editing toolbar is the appropriate feature class to be edited to contain the new digitized features. Use the editor toolbar to digitize/trace the source data by editing the new feature class with the new digitized features.

- For detailed specifics and functionality on the ESRI ArcGIS Editor toolbar, consult the Help Files included with the software or consult the ESRI online help files at: *http://webhelp.esri.com/arcgisdesktop/9.1/index.cfm?TopicName=welcome*

10.4 Working with Tabular Data

This discussion is based on ESRI (2006d), Morris (2005), and the primary author's personal knowledge.

This section addresses some basic considerations when incorporating tabular research data into a GIS so that it may be analyzed in a spatial context. One of the primary advantages of utilizing a GIS is that various types of data and information that would normally be difficult to associate can be integrated and analyzed.

The tabular research data should be formatted and entered in a table format such as a Microsoft Access table or a Microsoft Excel table saved as a dBase IV .dbf file with appropriate unique Field names for columns. If a unique primary key identifier other than the one automatically created by ArcGIS is to be used (for example, a monitoring plot identification number), create a column for such an identifier in the table. It is advantageous if the custom unique object identifier reflects an identifying element of the data; in the case of botany monitoring plots for example, a Plot Identifier is used for each row associated with an individual monitoring plot feature.

When entering tabular data into a table, it is necessary to structure the attributes so that only one instance of a single concept is placed in an attribute field. Also, it is necessary to avoid lists of items in an attribute field because that typically results in more than one concept being assigned to that attribute field. A single element such as a text string represents a single piece of information.

It is necessary to reduce redundant information so that multiple instances of the same value or concept are not entered. Normalization is a key aspect of database design which helps to ensure a usable database that can be managed as a long-term asset. Proper normalization of tabular data allows for a design where certain attributes can be linked, where necessary, to existing rows rather than to create a second instance or duplicate of data already entered. Section 7.2.3 of this Toolkit briefly touches on normalization and the concepts of "normal forms."

A journal article titled "Relational Database Design and Implementation for Biodiversity Informatics" by Paul J. Morris (Morris, 2005) is also an informative reference document for an in-depth understanding of database development with tabular data. It is also a great resource for gaining a better understanding of issues such as normalization and other issues associated with biological informatics. This article can be located at: *http://systbio.org/files/phyloinformatics/7.pdf* This link works only when copied and pasted into a browser.

Also, to learn how data cross-walk from an input table to ArcGIS, the ESRI Help files included with the software or available online can show table data types and the equivalent ArcGIS data type supported (ESRI, 2006d).

10.5 Additional Data Tools and Information Sources

- Example NPS GPS and GIS diagrammatic work flow: *http://www.nps.gov/gis/gps/gps4gis/*

- A large compilation of GIS Educational Information, Lessons, and Online Tutorials: *http://spatialnews.geocomm.com/education/tutorials/*

- National Geodetic Survey general information on Continuously Operating Reference System GPS stations for differential GPS correction: *http://www.ngs.noaa.gov/CORS/cors-data.html*

- NPS GIS links and other information: *http://www.nps.gov/gis/data_info/links.html*

- Colorado University GIS web site: *http://www.colorado.edu/geography/gcraft/notes/notes.html*

- NASA Goddard Space Flight Center Remote Sensing Tutorial: *http://rst.gsfc.nasa.gov/*

- ESRI GIS Data Dictionary: *http://support.esri.com/index.cfm?fa=knowledgebase.gisDictionary.gateway*

- GPS Utility- A Free GPS application for managing, manipulating, and mapping GPS data: *http://www.gpsu.co.uk*
- Nationwide Differential Global Positioning System (DGPS): *http://www.tfhrc.gov/its/ndgps/index.htm*
- GPS-CORPSCON-a freely downloadable coordinate converter program developed by the U.S. Army: *http://crunch.tec.army.mil/software/corpscon/corpscon.htm*l
- ESRI GIS User Forums where information can be queried, the ArcGIS help files can be accessed, and specific questions can be posted for other GIS users to respond to: *http://support.esri.com/index.cfm?fa=forums.gateway*
- Sources of base data for projects: *http://nationalmap.gov/* & *http://seamless.usgs.gov/*
- A Useful Overview of Aerial Imagery: *http://gis.ucsc.edu/Projects/aerial/guidelines.htm*
- Federal Geographic Data Committee Accuracy Requirements and Specifications: *http://www.fgdc.gov/standards/standards_publications/*
- National Spatial Data Transfer Standard (SDTS) guides and publications: *http://mcmcweb.er.usgs.gov/sdts/training.html*
- Federal Aviation Administration GPS web site (provides information on FAA's WAAS differential correction system): *http://gps.faa.gov*

—— THIS PAGE INTENTIONALLY LEFT BLANK ——

11 Documentation Tool # 1—FGDC Bio-Profile Metadata Questions

FGDC Biological Data Profile (BDP) Metadata Collection: Information Collection Tool for Novice Metadata Creators

Use: Any project-related data, both spatial and non-spatial

This set of questions can be used to extract the relevant elements of a data set for creating FGDC-NBII compliant metadata. For a cross-walk of the questions below to the FGDC-Biological Data Profile standard, see Appendix A at the end of this document. This tool is to be used in conjunction with Section 9.3 of this Toolkit.

Follow this link to see a graphical representation of the FGDC Bio-Profile Standard: *http://www.nbii.gov/images/uploaded/151871_1166715705793_NBII_bdp_graphical.doc*

FGDC Biological Data Profile Interview Style Questions:

Bold text below denotes the FGDC Profile section with which it is associated.

Project & Task Title:

Data set:

1. What is the approximate location of the project? Please provide general descriptions (for example, county names, NPS unit, lat-longs). Note: Section 9.3 of this Toolkit contains tips for addressing the bounding coordinates part of this element. (**Description of Geographic Extent, Bounding Coordinates**)
2. Who is the originator/owner of the data set? (Include address and telephone number.)
    a. Someone besides the originator/owner also can be listed in the metadata to answer questions about the data. List a name, address, and telephone number. (**Contact**)
    b. Are there other organizations or individuals who deserve credit for support, funding, or data collection and analysis? (**Data Set Credit**)
3. Include a description of the data set. (**Abstract**)
    a. What format are the data stored in—spreadsheet, database, ArcInfo coverage, ArcInfo Geodatabase, text file, other (please identify). If known, also include the software type and version. (**Native Data Set Environment**)
    b. Are there legal restrictions on who may obtain/use the data (T&E species, classified, incomplete, agency-specific proprietary)? (Access / Use Constraints)
    c. Where are the formats in subsection (a) above physically located? Please provide relevant names of individuals, organizations, and contact information.
4. Include any restrictions associated with the data in terms of the sensitivity of the data or any applicable access or use constraints. (**Access / Use Constraints**)
5. Why were the data collected (How will they be used in the scheme of the project)? (**Purpose**)
6. Please provide a citation for any publications pertaining to this data set, or update any listed as "in press" that have been published. (**Cross Reference**)
7. What is the time period represented by the data? (What are the beginning and ending dates of data collection?) (**Time Period**)
8. Were the data developed primarily through (**Currentness Reference**)
    a. Field visits,
    b. Remote instrumentation (for example, temperature recorders),
    c. Existing data sources?
9. What is the status of the data being documented—complete, in progress, planned (**Status**)?
    a. Will the data be updated? (**Maintenance & Update Frequency**)
    b. If so, how frequently (for example, weekly, monthly, annually, irregularly, or as needed)?
10. Please list any keywords associated with this project (minimum of one thematic keyword required). A keyword thesaurus such as the NBII Biocomplexity Thesaurus can be used for needed keywords so as to enable better search capabilities in metadata clearinghouses through standardized terminology. *http://thesaurus.nbii.gov/portal/server.pt?open=512&objID=578&&PageID=1657&mode=2&in_hi_userid=2&cached=true*
(**Keywords: Theme, Place, Stratum, Temporal, Taxonomy**).
11. Do the data contain taxonomic information? If yes, ITIS (Integrated Taxonomic Information System: *http://www.*

itis.gov/taxmatch_ftp.html) can be used for completing taxonomy information for documenting consistency purposes. The ITIS is the result of a partnership of Federal agencies formed to satisfy their mutual needs for scientifically credible taxonomic information. Note: Section 9.4 of this Toolkit contains tips for documenting a data set's taxonomic elements. If no, skip to question 12.

   a. What important species or communities were examined or are documented in the data? Please provide the genus/species names? (Taxonomy)

   b. Was a taxonomic authority or field guide for identification used?

      i. If so, what is the reference?

      ii. Describe any modifications, if any, to the classification.

      iii. Please describe any taxonomic procedures used (for example, specimen processing, comparison with museum materials, keys, genetic analysis).

12. Briefly summarize any field, lab, or analysis methods (cut and paste from other documents when possible). If standard or published protocols/methods were used, simply put the complete citation for the reference in 12a below. (**Methodology, Methodology Keywords**)

   a. If existing protocols or methods were used, list the references. (**Methodology Citation**)

   b. List any information sources for the data set, and provide a citation as well as a short description of the information contributed by each. (**Source Information**)

13. Were data processed with a model or other analytical tool? Examples include DISTANCE, Program Mark, SAS, the Century Model. If no, skip to question 14. (**Analytical Tools**)

   a. Please provide a brief description of the model or tool.

   b. Is the tool or model readily available? If so, please include a URL and(or) a contact address.

14. What measures, if any, were taken to make certain that the data were as nearly correct as possible? How "correct" are the attribute values? (Some examples of measures include QA/QC, instrument calibrations, spot-checking data, spreadsheet macros for outliers, and accuracy assessment matrices.) (**Attribute Accuracy Report**)

15. Were there any things excluded from the data collection (for example, stems less that a certain diameter, streams without surface flow)? If the data are from multiple years or sites, are all years/sites represented? What kind of data or information, if any, may be missing? (**Completeness Report**)

16. Please provide an assessment as to the accuracy of horizontal and vertical locations of features. This may include equipment settings (for example, with GPS, the maximum allowable PDOP), field checks, survey quality, and cross-checks with other locational references. (**Positional Accuracy**)

17. Are values in the data set restricted to a data dictionary or code book? If so, the information about those will need to be included. Describe and define what each attribute field means (include units of measure if applicable). (**Entity & Attribute**)

18. Do the data contain spatial information such as geographic datums and map projects or latitude/longitude? Section 9.3 of this Toolkit contains tips for addressing this element. If no, skip to question 19. (**Spatial Reference**)

   a. What are the geographic datum and spatial map projection parameters?

19. List the processing steps used to create the data set, including the approximate date of processing. (**Methodology; Process Steps**).

   a. List any source data and information used (**Source**). For each source list:

   b. Source name, originator and publication date,

   c. Source time period and scale,

   d. Source presentation form and media type,

   e. Contribution of the source to any analysis.

20. Is the data set available to other researchers? If no, please give a short explanation as to why and proceed to question 20. (**Distribution**)

   a. Who is the data distributor? How can the data be obtained? If available on the web, please include the URL. Please provide name, address, and phone number of the person/agency who is distributing the data.

   b. Should any advice be provided to potential users of the data set?

21. If there is any other pertinent information that should be captured in the metadata, please enclose here as well. The best section for this is the (**Supplemental Information Section**).

12 Documentation Tool # 2 – Dublin Core Metadata Questions

Dublin Core Metadata Collection: Information Collection Tool for Novice Metadata Creators

Use: Reports, Documents, Web-Enabled Resources

The below matrices were adapted from Zolly (2005). This set of questions can be used to extract the relevant elements of a data set for creating Dublin Core compliant metadata. The Dublin Core Metadata Element Matrices associated with each question give the guidelines that should be used to shape the answer to each question for consistency and standardization purposes.

1. What is the Title of the resource? (**Title**)

DC.Element Name	Title
Definition	A name given to a resource or resource component.
Purpose	Provides descriptive context for a resource; major point of entry from search engine results.
Obligation	Mandatory
Format	Title: Free text entry
Existing Standards	
Guidelines	→ Drop the articles "a," "an," or "the" at the beginning of a title → Place the most relevant title information at the beginning of the title for better intellectual access; drop corporate or institutional names from the title itself; these can go in Creator or Publisher elements (numbers 2 and 4 below, respectively)
Examples	"The GLOBE Program" should be documented as: **Title: GLOBE Program** "The National Wildlife Federation Backyard Habitat" should be titled as: **Title: Backyard Habitat** This enables the key terms to appear first in search returns list. National Wildlife Federation can be placed in the Publisher element. A resource from the University of XYZ having the title "University of XYZ College of Agriculture" and addressing the impacts of the varroa mite on honeybee populations in California should be renamed with a more descriptive title, such as: **Title: Varroa Mite Impacts on Honeybee Populations in California**. "University of XYZ College of Agriculture" will appear in the Publisher element.

2. What is the name of the person(s) or organization primarily responsible for the intellectual content of the resource? (**Creator**)

DC.Element Name	Creator
Definition	The name of the person(s) or organization primarily responsible for the intellectual content of the resource.
Purpose	This element allows users to search for a resource by the creator's name.
Obligation	Mandatory
Format	Creator: Free text entry
Existing Standards	
Guidelines	→ For personal names: FirstName LastName. → Multiple personal names should be separated by a semi-colon followed by a space. → For agency, organizational, or institutional names, use full agency name with no abbreviations or acronyms → To indicate a subdivision of the agency or institution, separate entities with a comma → To indicate multiple agencies, separate agencies with a semi-colon → If no author is specifically identified, the Publisher is the default Creator: enter identical data for Creator and Publisher.
Examples	• Creator: Rachel Carson • Creator: Rachel Carson; Stephen Jay Gould; Edmund O. Wilson • Creator: United States Geological Survey • Creator: United States Fish and Wildlife Service, Division of Migratory Bird Management • Creator: United States Geological Survey, Pacific Island Ecosystems Science Center • Creator: Houston Advanced Research Center • Creator: Conservation Management Institute, Virginia Tech • Creator: National Museum of Natural History, Smithsonian Institution • Creator: United States Geological Survey; United States Fish and Wildlife Agency, Endangered Species Program; United States Department of Defense, Office of Environmental Information Technology Management

3. Please list any contributors to the content of the resource. (**Contributor**)

DC.Element Name	Contributor
Definition	An entity responsible for making contributions to the content of the resource.
Purpose	This element gives credit to a minor author or an illustrator, photographer, editor, data provider, programmer, etc. A Contributor may be a person or an agency/organization. The cataloguer may optionally choose to delineate the role of the contributor, using parenthetical comments.
Obligation	Optional
Format	Contributor: Free text entry
Existing Standards	
Guidelines	See Guidelines under Creator.
Examples	• Contributor: John Mosesso, Jr. (photographer) • Contributor: Environmental Protection Agency (data set) • Contributor: Ron Sepic (editor) • Contributor: Deanne DiPietro; Allan Hollander

4. Who is responsible for making the resource available? (**Publisher**)

DC.Element Name	Publisher
Definition	The name of the entity responsible for making the resource available.
Purpose	The element establishes responsibility for a resource's publication. It can serve as a contact point for additional information about a resource, and to direct users to related information on a topic or issue. Frequently the Publisher is an agency, organization, corporation, institution, or non-profit; however, it also may be an individual.
Obligation	Mandatory
Format	Publisher: Free text entry
Existing Standards	
Guidelines	→ For Internet resources, the Publisher is often identifiable from the domain name; this is the entity responsible for making the resource available on the Web. → For agency, organizational, or institutional names, use full agency name with no abbreviations or acronyms → To indicate a subdivision of the agency or institution, separate parent entity from child entity with a comma → For personal names: FirstName LastName
Examples	• Publisher: Rachel Carson • Publisher: Washington Post • Publisher: United States Geological Survey • Publisher: Information Center for the Environment, University of California, Davis • Publisher: United States Fish and Wildlife Service, Division of Migratory Bird Management • Publisher: United States Geological Survey, Pacific Island Ecosystems Science Center • Publisher: John Wiley and Sons, Inc. • Publisher: Ecological Society of America • Publisher: National Museum of Natural History, Smithsonian Institution

5. Please provide some keywords about the resource. For search capability optimization, these should be selected from a controlled source such as the NBII/CSA Thesaurus (*http://thesaurus.nbii.gov/portal/server.pt?open=512&objID=578&&PageID=1657&mode=2&in_hi_userid=2&cached=true*). (**Subject-Controlled**)

DC.Element Name	Subject.Controlled
Definition	Keyword terms from a controlled list or thesaurus which accurately describe the subject and the specificity of a resource.
Purpose	This element enables users to locate relevant resources from both a general and a fielded search, and allows them to narrow or expand their search to the specificity of their information needs. Use of a controlled list or thesaurus ensures that all cataloguers are using a common "language" to describe resources, greatly enhancing the consistency and the relevance of documents retrieved by the user.
Obligation	Mandatory
Format	Subject: Free text entries
Existing Standards	• NBII/CSA Thesaurus is the recommended standard used for standardization and consistency purposes
Guidelines	→ Catalogue to the specificity of the document; if a narrow term is used, there is no need to include the approved broader term. For example, if the term "forest management" specifically applies to the resource being catalogued, there is no need to also include approved broader terms such as "ecosystem management" or "resource management." → Capitalize the first word of each term, as well as all proper nouns → Terms in this field are comma-delimited, with a space after each comma. → Do not include non-preferred terms in this field → To include terms not in the thesaurus, use the "Uncontrolled Subject" field. Only preferred terms occurring in the thesaurus should be included in the Subject field.
Examples	• Subject: Amphibians, Malformations, Contaminants • Subject: Urban environments, Watersheds, Contaminants, • Subject: Climatic change, Kyoto Protocol

6. Provide some uncontrolled subject keywords for the resource. These should be terms that were not included in #5 above. (**Subject-Uncontrolled**)

DC.Element Name	Subject.Uncontrolled
Definition	Keyword terms not appearing in the CSA/NBII Biocomplexity Thesaurus which are nevertheless key to describing the subject and the specificity of a resource.
Purpose	This element enables users to locate relevant resources from a general search, even though it does not appear in the Biocomplexity Thesaurus.
Obligation	Optional
Format	Subject: Free text entries
Existing Standards	
Guidelines	→ Terms in this field are comma-delimited, with a space after each comma → Capitalize the first word of each term, as well as all proper nouns Do not include non-preferred terms in this field
Examples	• Uncontrolled Subject: Protocols, Citizen science • Uncontrolled Subject: Convention on Biodiversity

7. Please provide an abstract for a description of the content, scope, and purpose of the resource. (**Description**)

DC.Element Name	Description
Definition	An abstract or textual description of the content, scope, and(or) purpose of the resource.
Purpose	This element provides the user with crucial context in deciding the relevance of specific documents to an information need. It is also a searchable field, and can aid users who are using generic free text searching, or whose information needs are vaguely understood.
Obligation	Mandatory
Format	Description: Free text entry
Existing Standards	
Guidelines	An abstract of the resource may include: → Topic or focus → Intended audience or use → Summary of findings → Limitations of use, or conditions of use
Examples	• Description: Global temperature fluctuations, and the resulting alterations to precipitation rates, appear to play a direct role in the declines in some amphibian populations. Research teams have linked low water levels in amphibian wetlands habitats to debilitating outbreaks of stress-induced diseases in frogs. (from FrogWeb) • Description: LMS is an evolving application designed to assist in analysis and planning of forest ecosystems by automating the tasks of stand projection, graphical and tabular summarization, stand visualization, and landscape visualization within a cohesive system. (from PNWIN) • Description: The Harold L. Lyon Arboretum coordinates, facilitates, and executes research, instruction, and service activities that utilize its collections and resources. Its major emphases are tropical plants, native Hawaiian plants, conservation biology, and Hawaiian ethnobotany.

8. Please provide a specific reference such as a weblink, ISBN, or ISSN number so that a user may find the resource. (**Resource Identifier**)

DC.Element Name	Resource Identifier
Definition	An unambiguous reference to the resource within a given context.
Purpose	A finding mechanism to provide the user with physical access to the resource. Recommended best practice is to identify the resource by means of a string or number conforming to a formal identification system. For most electronic resources, this is the Uniform Resource Identifier (a URL, an FTP address, or DOI). For print resources, the identifier may be an ISBN or ISSN. For some resources there is more than one identification value – for instance, a report available both online and in print.
Obligation	Mandatory
Format	Limited to the following qualifiers and free text values: Resource Identifier: [a URI] (free text value) Resource Identifier: ISBN (free text value) Resource Identifier: ISSN (free text value) Resource Identifier: Other (free text value)
Existing Standards	URI – Uniform Resource Identifier (http:// or ftp://) for web resources ISBN – International Standard Book Number ISSN – International Standard Serial Number
Guidelines	Cataloguers should be allowed to choose more than one qualifier.
Examples	Resource Identifier: *http://www.issg.org/database/* Resource Identifier: *ftp://cbi.usgs.gov* Resource Identifier: *http://dx/doi.org/10.4046/j.1523-1739.1999.98075.x* Resource Identifier: ISSN: 0027-9633 Resource Identifier: ISBN: 0-471-89736-2 Resource Identifier: Other: Report# T79138-1

9. What type of resource is it? (**Resource Type**)

DC.Element Name	Resource Type
Definition	The category or genre of the resource.
Purpose	This element allows users to restrict a search to resource of a specific kind. It also provides valuable information regarding the context, scope, and purpose of a resource.
Obligation	Mandatory
Format	
Existing Standards	
Guidelines	
Examples	• Resource Type: Proceedings • Resource Type: Case Studies, Management Plans and Reports • Resource Type: Internet Map Services, Data sets

10. What type of media format is the resource in (ex. Microsoft Word format-.DOC, Zip file-.ZIP, etc.)? (**Resource Format**)

DC.Element Name	Resource Format
Definition	The physical or digital manifestation of the resource.
Purpose	Typically, Resource Format may include the media-type or dimensions of the resource. Format may be used to determine the software, hardware or other equipment needed to display or operate the resource. Providing this information to users prior to their physically accessing the resource enables them to determine quickly if they may utilize the resource – partially or completely – from their current hardware and software. Cataloguers may choose more than one Resource Format, if applicable.
Obligation	Mandatory
Format	
Existing Standards	
Guidelines	Use multiple Formats only if they characterize major, independent components of the resource that may require individual equipment to access them. For instance, if the resource is a general Web page which includes two illustrative images, do not characterize the Format as URL, GIF; there is no need to include the image format GIF, as Web browsers can interpret and display the image. Alternatively, if a mapping application allows for export of data to an Access database, a Resource Format of Mapping Service, MDB would be appropriate.
Examples	Resource Format: URL Resource Format: DOC Resource Format: XLS Resource Format: ESRI E00 Resource Format: MDB Resource Format: ZIP

11. Who holds the intellectual property rights to the resource? Are there any restrictions on the accessibility of the resource? (**Rights**)

DC.Element Name	Rights
Definition	A statement about the rights management of the resource.
Purpose	This element indicates intellectual property rights to the resource, as well as any restrictions to accessibility of the resource.
Obligation	Mandatory
Format	Restricted checklist. Default value example is “Copyright held by Publisher”; multiple values may be selected. Rights: Copyright held by Publisher Rights: Copyright held by Creator Rights: Copyright held by Source Rights: Public Domain
Existing Standards	
Guidelines	Default value is “Copyright held by Publisher”.
Examples	See Format section above.

12. What language is the content of the resource in? (**Language**)

DC. Element Name	Language
Definition	The language(s) of the intellectual content of the resource.
Purpose	While not a primary access point for searching, this field allows a search to be further restricted to resources in a specific language. It also provides information to the user regarding potential intellectual access to the resource.
Obligation	Mandatory
Format	Language: selected from picklist using Existing Standard below. Selection of more than one language should be permitted; multiple language codes should be comma-delimited.
Existing Standards	ISO [RFC1766] at *http://dublincore.org/documents/dces/#rfc1766*
Guidelines	
Examples	Language: English Language: English, French Language: English, French, Spanish

13. What is the date of the resource? The format should be YYYY-MM-DD (**Date Issued**)

DC.Element Name	Date.Issued
Definition	A date associated with an event in the life cycle of the resource.
Purpose	Typically, Date is associated with the creation or availability of the resource.
Obligation	Optional
Format	Field entry • YYYY-MM-DD If no date is entered, the element should either be omitted in the end-user display, or should contain a null value.
Existing Standards	ISO 8601 [W3CDTF] at *http://dublincore.org/documents/dces/#w3cdtf*
Guidelines	→ Recommended best practice for encoding the date value is defined in a profile of and follows the YYYY-MM-DD format. → Use "date issued" only for fixed resources such as print or digital versions of news releases, articles, journals, fact sheets, guidelines, standards, and other documents. → Leave blank if the above resources have no issue date. → If no day is specified, format YYYY-MM may be used. → If neither day nor month is specified, YYYY may be used. → The Date field should not be confused with the Coverage.Temporal field. The Date field indicates a specific/fixed moment in the life cycle of a resource; the date that the version in question was issued. [If multiple versions continue to be available within the resource catalogue, the Relation element should also be employed.] Coverage.Temporal addresses the relation of a time period to the content itself: a range or span of time the content covers.
Examples	• 2002-06-28 • 2002 • 2002-06

14. Is this resource part or derived from a Source resource? If so, please provide a reference to that resource. (**Source**)

DC.Element Name	Source
Definition	A reference to a resource from which the present resource is derived.
Purpose	The present resource may be derived from the Source resource whole or in part.
Obligation	Optional
Format	Source: (free text value of a Resource Identifier for the source document, plus additional bibliographic information, if appropriate.)
Existing Standards	Use Existing Standards for the Resource Identifier.
Guidelines	Recommended best practice is to reference the resource by means of a string or number conforming to a formal identification system – for example, a Resource Identifier.
Examples	• The referenced resource, the Dublin Core Metadata Element Set 1.1, at <*http://dublincore.org/documents/dces/*>, is a component of a larger project, the Dublin Core Metadata Initiative. Source: Dublin Core Metadata Initiative <*http://dublincore.org*> • The referenced resource, "Environmental Contaminants: Factors Affecting Natural Resources" at <*http://biology.usgs.gov/s+t/SNT/noframe/idx-co.htm*>, is a component of the print and digital resource, Status and Trends of the Nation's Biological Resources. Source: Status and Trends of the Nation's Biological Resources, *http://biology.usgs.gov/s+t/SNT/index.htm* GPO stock # 024-001-03603-7

15. Are there any resources that are related to this one that would be of value to the user to note? (**Relation**)

DC. Element Name	Relation
Definition	A reference to a related resource.
Purpose	Identification of other sources that are related to the current resource, and the type of relationship.
Obligation	Optional
Format	Relation: is part of (Resource Identifier) Relation: is version of (Resource Identifier) Relation: is replaced by (Resource Identifier) Relation: replaces (Resource Identifier) Relation: Other (explain)
Existing Standards	See Existing Standards for Resource Identifier element.
Guidelines	Recommended best practice is to reference the resource by means of a string or number conforming to a formal identification system – for example, a Resource Identifier.
Examples	• Replaces *http://www.website.org/000123.pdf*

16. What is the temporal range of the resource? Please provide if applicable (ex. From YYYY-MM-DD to YYYY-MM-DD). (**Coverage-Temporal**)

DC.Element Name	Coverage.Temporal
Definition	The temporal scope of a resource.
Purpose	Provides temporal context to the scope of a resource, delineated by a date range. This is especially useful when documenting a data set, to denote the interval during which data was collected. For completed data sets, both a Beginning and an Ending Date are required; for a data set whose data collection efforts are ongoing, only a Beginning date is required.
Obligation	Mandatory if applicable to the resource
Format	Coverage.Temporal.Beginning: YYYY-MM-DD Coverage.Temporal.Ending: YYYY-MM-DD
Existing Standards	For date ranges, use Existing Standard ISO 8601 [W3CDTF] at *http://dublincore.org/documents/dces/#w3cdtf*
Guidelines	Do not confuse Coverage.Temporal with the Date element. Coverage.Temporal should be used only when the resource's intellectual content is characterized by a given interval of time.
Examples	Example 1: A transect was monitored for amphibians from 1 July 1990 to 1 July 1991: Coverage.Temporal.Beginning: 1990-07-01 Coverage.Temporal.Ending: 1995-12-31 Example 2: Project FeederWatch, an annual volunteer bird monitoring program, has been collecting observational data on birds present at feeders around the country: Coverage.Temporal.Beginning: 1999-01 [Note that there is no Coverage.Temporal.Ending Date, as data collection is ongoing.]

17. What is the geographic extent or scope of the content of the resource? Please provide if applicable. Please use one of the approved place name authorities given in the Appendix. (**Coverage-Geographic**)

DC.Element Name	Coverage.Geographic
Definition	The geographic extent or scope of the content of the resource. Coverage typically includes spatial location: place name(s) or geographic coordinates.
Purpose	Use of this element enables geospatial referencing and searching on resources, using place names or georeferenced coordinates.
Obligation	Mandatory, if applicable.
Format	Picklist from approved sources Coverage.Geographic.Hemisphere: Coverage.Geographic.Continent: (one or multiple) Coverage.Geographic.Country: (one or multiple) Coverage.Geographic.State: (one or multiple) Coverage.Geographic.County: (one or multiple) Coverage.Geographic.Coordinates: Coverage.Geographic:
Existing Standards	Recommended place name authorities for standardization and consistency: • The Geographic Names Information System [GNIS] from USGS • Federal Information Names Processing Codes [FIPS] from US Census • GeoNet Names Server [GNPS] from NIMA
Guidelines	• Coverage.Geographic metadata is specific to the scope of the resource's geographic area of study or concern; it should not be used to characterize the origin of the resource itself. For example, a resource published by the University of Texas on lynx populations in Canada would have a geographic coverage metadata for the regions of Canada, and no geographic coverage metadata related to Texas. Therefore, Texas would not be considered part of the resource's geographic coverage.
Examples	Coverage.Geographic.Hemisphere: Northern Hemisphere Coverage.Geographic.Continent: North America Coverage.Geographic.Country: Canada Coverage.Geographic.State: British Columbia, Alberta

18. What is the jurisdictional coverage area? Is any aspect of jurisdiction applicable to the resource being documented? (**Coverage-Jurisdictional**)

DC.Element Name	Coverage.Jurisdictional
Definition	The entity or entities with legal standing or obligation to interpret or apply the law, agreement, or binding commitment articulated in the resource.
Purpose	Use of this element enables geospatial referencing and searching on resources, using place names or geo-referenced coordinates.
Obligation	Mandatory if applicable
Format	Free text
Existing Standards	Recommended place name authorities for standardization and consistency: • The Geographic Names Information System [GNIS] from USGS • Federal Information Names Processing Codes [FIPS] from US Census • GeoNet Names Server [GNPS] from NIMA
Guidelines	→ Coverage.Geographic metadata is specific to the scope of the resource's geographic area of study or concern; it should not be used to characterize the origin of the resource itself. For instance, a resource published by the University of Texas on lynx populations in Canada would have a geographic coverage metadata for the regions of Canada, and no geographic coverage metadata related to Texas.
Examples	North American Free Trade Agreement: Coverage.Jurisdictional.Entity: Canada; Mexico; United States Canada-United States Air Quality Agreement: Coverage.Jurisdictional.Entity: Canada; United States Chesapeake Bay Program: Coverage.Jurisdictional.Entity: Virginia; Maryland; District of Columbia; Pennsylvania Directive on the Conservation of Wild Birds: Coverage.Jurisdictional.Entity: European Union

19. Does the resource contain any species-specific references? If so, please provide the Genus and Species based on the Integrated Taxonomic Information System (ITIS) (*http://www.itis.gov/*). (**Scientific Name**)

DC.Element Name	Scientific Name
Definition	The unequivocal scientific name by which an organism is universally recognized.
Purpose	Establishes a standard name observed by all researchers to describe an organism.
Obligation	Mandatory if the resource is species-specific
Format	Scientific Name: Genus species
Existing Standards	The Integrated Taxonomic Information System (ITIS) (*http://www.itis.gov/*) should be used for consistency and standardization. If ITIS does not provide coverage for a particular species, an appropriate alternative may be used. The name of the resource used should be specified
Guidelines	• Genus name should be capitalized; species name should be lower case. • Multiple species are comma delimited, with a space after the comma. • If the resource deals with a multiple species within a given genus, use the genus name, followed by spp. (for example Rana spp.) • If the resource does not address species-level data, this should be indicated in this field
Examples	• Scientific Name: Rana catesbeiana • Scientific Name: Carodacus spp. • Scientific Name: Rana catesbeiana, Rana aurora draytonii, Rana luteiventris

20. If question #19 is applicable to the resource being documented, please provide the common names for the species. (**Common Name**)

DC.Element Name	Common Name
Definition	The name(s) by which an organism is known colloquially.
Purpose	Common names are generally used outside the scientific community to refer to species. Common names for species may vary by region; many organisms are known by more than one common name, but have only one scientific name. Cataloguing common names provides a useful way for amateur scientists and the general public to search for information about a particular organism.
Obligation	Mandatory if the resource is species-specific
Format	Common Name: Common name(s)
Existing Standards	See Existing Standards for Scientific Name.
Guidelines	• Capitalize first word in common name; subsequent words in the name are lower case, with the exception of place-based names
Examples	• Common Name: Bullfrog • Common Name: Finch spp. • Common Name: Bullfrog, California red-legged frog, Columbia spotted frog • Common Name: Species-level data not indicated

—— THIS PAGE INTENTIONALLY LEFT BLANK ——

Selected References

Biodiversity Conservation Information System, 2000, Framework for information sharing: Custodianship, Busby, J.R. (Series Editor).

Bisio, Ronald, 2005, GPS Accuracy Improvements Highlight Datum Importance. GeoPlace.com: The Authoritative Resource for Spatial Information, accessed February 2006 at *http://www.geoplace.com/ME2/dirmod.asp?sid=119CFE3ACE2A48319AA7DE6A39B80D66&nm=News&type=Publishing&mod=Publications%3A%3AArticle&mid=8F3A7027421841978F18BE895F87F791&tier=4&id=EAEDF604498B41999ACB460678F8469A*

Breman, Joseph, and Zeiler, Michael, 2005, ArcGIS Data Models – An Introduction, Twenty-Fifth Annual ESRI International User Conference: San Diego Convention Center, August 2005.

Burley, Thomas, Schubert, Nora, Peine, John, Murray, Judy, and Thompson, Mary, 2006, NBII-SAIN FY05 Roan Mountain project final report, accessed August 2006 at *http://sain.utk.edu/projects/massif/massif_docs/NBII_SAIN_FY0506_Roan_Mtn_Final_Report.pdf*

Busch, Robert, and Morrison, Lisa, 2001a, Comparing Global Positioning System (GPS) Tools: Wisconsin Department of Natural Resources, accessed November 2005 at *http://dnr.wi.gov/maps/gis/documents/gps_tools.pdf*

Busch, Robert, and Morrison, Lisa, 2001b, Global Positioning System (GPS) Accuracy Report: Wisconsin Department of Natural Resources, accessed November 2005 at *http://dnr.wi.gov/maps/gis/documents/gps_accuracy.pdf*

Cornell University, Central Technical Services, 2003, Report of the DLF electronic resource management initiative, Appendix B: Electronic resource management workflow flowchart, accessed June 2006 at *http://www.library.cornell.edu/cts/elicensestudy/dlfdeliverables/fallforum2003/Workflow_final.doc*

Cunningham, Greg, and Silvertrand, Gillian, 2005, Geodatabase Tuning and Performance: Twenty-Fifth Annual ESRI International User Conference, San Diego Convention Center, August 2005.

Data Warehousing and Business Intelligence, 2006, Conceptual, logical, and physical data models, accessed January 2006 at *http://www.1keydata.com/datawarehousing/data-modeling-levels.html*

Dublin Core Metadata Initiative, 2009, About the Initiative, accessed May 2009 at *http://dublincore.org*

Ecological Society of America, 2006a, Proceedings: The U.S. National Vegetation Classification, accessed May 2006 at *http://herbarium.unc.edu:8080/nvcrs/proceedings_about_main.jsp*

Ecological Society of America, 2006b, VegBank Information, accessed May 2006 at *http://www.vegbank.org/vegbank/general/info.html*

Ecological Society of America, 2006c, VegBranch Overview, accessed May 2006 at *http://www.vegbank.org/vegdocs/vegbranch/vbr-overview.html*

Ecological Society of America Vegetation Classification Panel, 2004, Guidelines for describing associations and alliances of the U.S. National Vegetation Classification, July 2004, Version 4.0, accessed May 2006 at *http://www.vegbank.org/vegdocs/panel/NVC_guidelines_v4.pdf*

Environmental Systems Research Institute (ESRI), 2006a, ArcPad 6.0.3 User Manual, ESRI Support Center, accessed March 2006 at *http://support.esri.com/index.cfm?fa=knowledgebase.documentation.listDocs&PID=26&FilterPV=228*

Environmental Systems Research Institute (ESRI), 2006b, Educational Services, QA/QC for GIS Data, Course Exercises Manual. Used with permission.

Environmental Systems Research Institute (ESRI), 2006c, Educational Services, QA/QC for GIS Data, Course Lectures Manual. Used with permission.

Environmental Systems Research Institute (ESRI), 2006d, ESRI ArcGIS Desktop Help accessed March 2006 at *http://webhelp.esri.com/arcgisdesktop/9.1/index.cfm?TopicName=welcome*

Environmental Systems Research Institute (ESRI), 2006e, ESRI GIS Dictionary: ESRI Support Center, accessed May 2006 at *http://support.esri.com/index.cfm?fa=knowledgebase.gisDictionary.gateway*

Environmental Systems Research Institute (ESRI), 2006f, ESRI UserForums: ESRI Support Center, accessed March 2006 at *http://support.esri.com/index.cfm?fa=forums.gateway*

Environmental Systems Research Institute (ESRI), 2006g, How to select the correct geographic (datum) transformation when projecting between datums: ESRI Support Center, accessed February 2006 at *http://support.esri.com/index.cfm?fa=knowledgebase.techarticles.articleShow&d=21327*

Federal Geographic Data Committee (FGDC), Metadata Education Program, and the National Metadata Cadre, 2006, Ten most common metadata errors, accessed June 2006 at *http://www.fgdc.gov/metadata/documents/top10metadataerrors.pdf*

Federal Geographic Data Committee (FGDC), Vegetation Subcommittee, 2006, National Vegetation Classification Standard, accessed June 2006 at *http://www.fgdc.gov/standards/projects/FGDC-standards-projects/vegetation/index_html*

Forest Science Research Network, 2002, Introduction to GIS/remote sensing in forest science research network, accessed December 2005 at *http://www.fsl.orst.edu/helpdesk/gis/intro.html*

Giles, Terry, and Kutner, Lynn, 2005, Introduction to Metadata Workshop: National Biological Information Infrastructure and NatureServe, NatureServe Headquarters, Arlington, VA, June 16–17, 2005.

Gillgrass, Craig, and McGrath, Matt, 2005, ArcGIS Geodata Management: An Introduction for Geodatabase: Twenty-Fifth Annual ESRI International User Conference, San Diego Convention Center, August 2005.

Global Biodiversity Information Facility, 2006, How to become a GBIF data provider, accessed September 2006 at *http://www.gbif.org/DataProviders/HowTo*

Hamil, David L., 2006, Your Mission, should you choose to accept it: Project Management Excellence: GeoCommunity: The Leading Geospatial News and Education Resource, accessed June 2006 at *http://spatialnews.geocomm.com/features/mesa1/*

Integrated Taxonomic Information System (ITIS), 2004, About ITIS-Background Information, accessed January 2006 at *http://www.itis.gov/info.html*

Michener, W.K., Brunt, J.W., Helly, J.J., Kirchner, T.B., and Stafford, S.G., 1997, Non-geospatial metadata for the ecological sciences: Ecological Applications: v. 7, no. 1, p. 330–342.

Morris, P.J., 2005, Relational database design and implementation for biodiversity informatics, Phyloinformatics: Journal for Taxonomists, v. 7, p. 1–63, accessed February 2006 at *http://systbio.org/files/phyloinformatics/7.pdf*

National Geodetic Survey, 2006, National Geodetic Survey-Continuously Operating Reference Stations (CORS), accessed January 2006 at *http://www.ngs.noaa.gov/CORS/*

National Land and Water Resources Audit and the Australia New Zealand Land Information Council (ANZLIC) Spatial Information Council on behalf of the Australian National Government, 2003, The natural resources information management toolkit, accessed February 2006 at *http://www.nlwra.gov.au/toolkit/*

National Park Service (NPS), 2004, GPS Myths-Alaska Support Office GIS Team, accessed November 2005, *http://www.fs.fed.us/database/gps/collection/gpsmyths.htm*

National Park Service (NPS), 2005, Regulations, directives, policies and guidelines and their relation to NPS Geographic Information Systems, accessed December 2005 at *http://www.nps.gov/gis/data_standards/*

National Park Service (NPS), 2006, NPS geographic information systems (GIS) data specifications for resource mapping, inventories, and studies, accessed June 2006 at *http://science.nature.nps.gov/nrgis/pdfs/NPS_GISSpecs_20070302.pdf* (URL updated August, 2007)

Object Management Group, 2006, Unified Modeling Language, accessed January 2006 at *http://www.uml.org/*

Open Geospatial Consortium (OGC), 2006, accessed June 2006 at *http://www.opengeospatial.org/*

Open Geospatial Consortium (OGC), 2008, OpenGIS Standards and Specifications, accessed April 2008 at *http://www.opengeospatial.org/standards*

Roper, Mark, 2005, GPS to GIS procedural handbook and reference guide, Version 6.1, accessed January 2006. (URL is no longer available; version 6.1 was replaced by version 7 below.)

Roper, Mark., 2006a, GPS to GIS procedural handbook and reference guide, Version 7, accessed July 2006 at *http://www.fs.fed.us/database/gps/gps2gis/gps_gis_v7-0.pdf*

Roper, Mark, 2006b, WGS84 to NAD83 and NAD27 diagram, accessed February 2006 at *http://www.fs.fed.us/database/gps/documents/trimble/wgs84dat.pdf*

Ruggiero, Michael, McNiff, Marcia, Olson, Annette, and Wheeler, Ben, 2005, Strategic Plan for the U.S. Geological Survey Biological Informatics Program: 2005–2009: U.S. Geological Survey, 20 p., accessed February 2006 at *http://www.nbii.gov/images/uploaded/8496_1178901646922_BIO5yrPlan.pdf*

Specify Biodiversity Collections Software, 2006, Specify DiGIR FAQ, accessed September 2006 at *http://www.specifysoftware.org/Specify/specify/Specify%20DiGIR/index_html#1*

Stohlgren, T.J., Barnett, D.T., and Simonson, S.E., 2003, Beyond North American Weed Management Association Standards, accessed March 2006 at *http://www.nawma.org/documents/Mapping%20Standards/BEYOND%20NAWMA%20STANDARDS.pdf*

Subcommittee for Base Cartographic Data-Federal Geographic Data Committee, 1998, National Spatial Data Infrastructure: Geospatial Positioning Accuracy Standards, Part Three: National Standard for Spatial Data Accuracy, Publication Number FGDC-STD-007.3-1998, accessed December 2005 at *http://www.fgdc.gov/standards/projects/FGDC-standards-projects/accuracy/part3/chapter3*

Tomlinson, Roger, 2003, Thinking About GIS, Redlands: ESRI Press, 195 p.

Trexler, J.C., and Busch, D.E., 2003, Monitoring, assessment, and ecoregional initiatives: a synthesis *in* Trexler, T.R., and Busch, D.E., eds., Monitoring ecosystems: Interdisciplinary approaches for evaluating ecoregional initiatives: Washington, Island Press, p. 420

Trimble Navigation Limited, 2005, GPS pathfinder office software or the GPS Analyst Extension for ESRI ArcGIS Software: Resolving the NAD 83 Datum Transformation Issue, accessed February 2006 at *http://trl.trimble.com/docushare/dsweb/Get/Document-170369/SprtNote_PFO-GPSA_NAD83Datum.pdf*

U.S. Forest Service, 2004, Official Forest Service GIS Data Dictionary: U.S. Forest Service.

U.S. Forest Service, 2003, USFS draft GPS data accuracy standard, accessed August 2005 at *http://www.fs.fed.us/database/gps/gps_standards/GPS_Data_Standard.pdf*

U.S. General Accounting Office, 2003, Geographic Information Systems-Challenges to Effective Data Sharing-Statement of Linda D Koontz, Director, Information Management Issues: GAO-03-874T, 20 p.

U.S. Geological Survey, 1947, United States National Map Accuracy Standards, accessed January 2006 at *http://rockyweb.cr.usgs.gov/nmpstds/nmas.html*

World Wide Web Consortium, 2006, Extensible Markup Language, accessed October 2006 at *http://www.w3.org/XML/*

Zeiler, Michael, 1999, Modeling our world: The ESRI guide to geodatabase design: Redlands, Calif., ESRI Press, 200 p.

Zeiler, Michael and Arctur, David, 2004, Designing geodatabases: Case studies in GIS data modeling: Redlands, Calif., ESRI Press, 250 p.

Zolly, Lisa, 2005, NBII Metadata standard for web resources cataloging, accessed May 2006 at *http://www.nbii.gov/images/uploaded/8496_1166013854464_NBII_Metadata_Standard_for_Web_Resources_Cataloguing_Version_2.2.pdf*

— — THIS PAGE INTENTIONALLY LEFT BLANK — —

Appendix A—FGDC Bio-Profile Cross-walk

Cross-walk of FGDC Interview Questions to the FGDC-Biological Data Profile Metadata Standard

This section shows how the questions in Documentation Tool #1 cross-walks to the FGDC Biological Data Profile format.

FGDC-Biological Data Profile Standard Section Key:

- Section 1 = Identification
- Section 2 = Data Quality
- Section 3 = Spatial Data Organization
- Section 4 = Spatial Reference
- Section 5 = Entity and Attribute
- Section 6 = Distribution
- Section 7 = Metadata Reference

1. Section 1 → Description of Geographic Extent, Bounding Coordinates
2. Section 1 → Point of Contact and Section 6 → Distributor
   a. As above: Section 1 → Point of Contact and Section 6 → Distributor
   b. Section 1 → Data set Credit
3. Section 1 → Abstract
   a. Section 1 → Native data set environment; Section 6 → Format Name
   b. Section 1 → Access/Use constraints and(or) Security Information
   c. Section 1 → Native data set environment
4. Section 5 → Overview Description and Section 1→ Access/Use Constraints
5. Section 1 → Purpose
6. Section 1 → Cross Reference
7. Section 1 → Time Period of Content (TPOC)
8. Section 1 → Currentness Reference, in TPOC.
9. Field Visits = GROUND CONDITION;
   Remote Instruments = OBSERVED;
   Other Values = PHOTOGRAPHY or PUBLICATION DATE
10. Section 1 → Progress
    a. 9a. and 9b. If #9 <> complete, include them in Section 1 → Maintenance and Update
    b. 10. Section 1 → Keywords
11. Section 1 → Taxonomy (use the ITIS web site as needed—see Section 9.4 of this document for instructions on how to use ITIS to document taxonomy)
    a. Section 1 → Taxonomy- Taxonomic Classification and/or General Taxonomic Coverage
    b. Section 1 → Taxonomy- Taxonomic Procedures and(or) Identification Reference
12. Section 2 → Methodology, also possibly Section 2 → Process Steps
    a. Section 2 → Methodology Citation(s)
13. Section 1 → Analytical Tool

    a. and b. Section 1 → Analytical Tool
14. Section 2 → Attribute Accuracy or Section 2 → Process Steps
15. Section 2 → Completeness Report
16. Section 2 → Positional Accuracy
17. Section 5 → Entity and Attribute Info
18. Primarily Sections 3 and 4. May also be appropriate for Section 2 → Source Information and(or) Process Steps
    a. Sections 3 and 4
19. Section 2 → Methodology and Process Steps
20. Section 6 → Distribution

    a. and b Section 6 → Distribution
21. Section 1 → Supplemental Information

Any other pertinent info-best to address items on case by case basis for appropriate area of metadata, but the default for these types of information can be Section 1 → Supplemental Info

— — THIS PAGE INTENTIONALLY LEFT BLANK — —

Appendix B—Quality Assurance Plan Template

This part of the Toolkit consists of a general QA plan template to be used as a guideline example for a project QA plan. The source of this QA plan template (used with permission) is the ESRI Educational Services QA/QC for GIS Data Course Exercises Manual (ESRI, 2006b).

Title Page

- The title of the QA plan
- The document number (see the Document Revision History section)
- The document issue date
- For whom the QA plan is prepared
- Author(s) of the QA plan
- Contact Information

An approval sheet containing the signatures of those responsible for the project and dates of approval may follow the title page.

Document Revision History

Modifications to the QA plan should be documented. Below is an example of how a table might be used to record revisions to the document.

Revision History

Document #	Date	Author	Summary of changes
01	1/23/2006	H. Jones	First Draft

Table of Contents

Provide an overall list of the document contents, for example

1. Introduction
2. Purpose of the QA plan
3. Objectives of the QA plan
4. QA plan acceptance workflow
5. Deliverables and schedule
6. Materials control
7. Acceptance criteria
8. Sampling strategy
9. Full-production QC workflow
10. First-Round QC checks
11. Second-Round QC checks
12. QA report
13. PAR (Problem and Resolution) system
14. Conclusion.

The table of contents might contain other QA plan components, including

- Approval signature page
- Database properties
- List of deliverables
- Distribution list
- Project roles and responsibilities list and/or diagram
- Problem definition
- Task description
- Training needs
- Relevant documents and records list
- References.

This section can also include a list of tables, figures, and appendices.

Introduction

Briefly introduce the project and address the quality issues associated with the project including the basic tasks that will be performed. Detailed descriptions should be conveyed later in the QA plan. This section might include

- A description of the study area
- Background information on the project
- Information on the processing environment.

Depending on the scope of the project, this section might be divided into two sections: Problem definition and Task description.

Purpose of the QA plan

Briefly describe why this QA plan was created, for example

> The quality assurance plan provides a comprehensive outline of quality control procedures used in database development for the project and ensures that the database meets established quality criteria necessary for how the data will be used.

Objectives of the QA plan

QA plan objectives can be derived from planning and database design goals and objectives. Organizations may have established standards or criteria that must be adhered to. For examples, a GIS database is designed to meet the needs of its users (and the enterprise) and the intended use of the data should be known.

Additionally, QA plan objectives should support the purpose of the project. For example,

- Assist the data provider in the delivery of a quality product
- Define a specific QA process to be used throughout the project.

QA plan acceptance workflow

A workflow specifically for this QA plan is provided to illustrate the process of developing and modifying the plan. It also might indicate how and where stakeholders interact with the plan.

This section may include a workflow diagram (as suggested in Section 8.1.1 of this Toolkit) that shows the process of review, revision, and acceptance of the QA plan.

Deliverables and schedule

This section usually includes

- A description of the project schedule,
- A time frame for deliverables, and
- The review schedule for deliverables, including revisions and final review.

Below is an example of how a table might be used to record data deliveries and reviews.

Deliverable	Date received	First review date	Revisions complete date	Final Review
Phase 1				
Phase 2				
Phase 3				

Materials control

This section includes specifications for receiving, inspecting, cataloging, tracking, and archiving project materials, including

- Hard-copy documents, maps, manuscripts, as-built drawings, and files;
- Digital documents, existing data sources, and files;
- E-mail threads; and
- Any other project-related transmittals.

A database or spreadsheet might be used to create an inventory and track the status of project materials necessary for QA/QC.

Besides describing how materials will be managed for the duration of this project, this section can include or be followed by a Documents and Records section that would reference all documents and records critical to the project. This section might provide references to guidelines for documenting

- Software requirements and versions used in the project;
- Data automation, collection, and migration processes, and standards;
- Data or feature definitions;
- ,Metadata standards; and
- In general, any of the elements of Section 6 of this Toolkit.

Acceptance criteria

Every data delivery should be reviewed and a determination made as to whether it is acceptable or not. This section explains the overall QA/QC review guidelines and sets forth measurements for acceptability.

Example of issues to address for acceptance criteria are

- Data format and file-naming conventions and consistency,
- Spatial reference criteria (geographic datums and projections),
- Feature attribute table and table definitions,
- Spatial completeness as compared to source data,
- Spatial accuracy as compared to source data,
- Feature completeness,
- Positional accuracy,
- Valid attribute values,
- Missing attribute values,

- Valid topological and connectivity rules, and
- Edge-matching for surrounding or adjacent deliverables.

Typically, data deliveries are categorized as either accepted or not accepted on the basis of the number of nonconformities identified in the QC steps. If the deliverable is categorized as not accepted, the nonconformities are recorded and submitted for correction per the deliverables schedule.

An error metric is determined for each QC check. For example, some automated QC checks require that the deliverable contain no errors. Visual QC checks may require that less than 2 percent of features in the deliverable are nonconforming. Some visual QC checks require that a random sample of features be inspected (see Section 8.1.2 of this Toolkit).

Tables can help illustrate the QC check method and acceptance thresholds, for example,

Check item	Check method	Acceptance threshold
Topology checks	ArcGIS	100%
Feature count checks	ArcGIS	100%

A table also can illustrate acceptance thresholds for review of feature classes, for example

Feature Class (examples)	Acceptance criteria
Parcel index	100%
Easements	98%
Lots	98%
Parcel annotation	90%

Note that lower acceptance thresholds indicate a more subjective review process such as the labeling of features.

Still another method for determining acceptance is field-checking the data or ground-truthing it. In this method, the features in a deliverable or samples of the features are compared to their real-world locations and attributes.

Sampling strategy

The term sampling strategy has two distinct meanings with regards to QA/QC.

1. The method(s) for selecting and inspecting features during visual QC portions of the project.
2. The method(s) for generating and acquiring features either in the field or when using existing source materials.

The first definition of sampling strategy is usually included in the QA plan and may include references to the sampling standard used, sampling formulas, and a table for determining sample size.

With the second definition, the sampling strategy is a proactive effort that defines the methods and criteria for creating the required data sets. Detailed information usually is not included in the QA plan if outside documents address this aspect as well as the various elements of this Data Management Toolkit, but references to the appropriate documents can be made in the Materials control or Documents and records section or in an appendix. If such outside documents do not exist, the QA plan can address this need by properly addressing the elements of this Data Management Toolkit. For example, the subjects for some of these documents (if existing elsewhere) might include

- Data dictionaries for GIS data layers,
- GPS data collection requirements and procedures,
- Screen digitizing requirements and procedures,
- Imagery acquisition requirements and procedures,
- Temporal issues and requirements for data collection,
- Scanning procedures for hard-copy documents, aerial photographs, and
- Items in Section 6 of this Toolkit.

Full-production QC workflow

This is a workflow diagram that illustrates when QC checks are performed during the QA/QC program. This includes delivery of the pilot database, inspection of deliverables, automated checks, visual checks, and any other forms of QC, such as sampling QC.

First-Round QC checks

This section identifies the automated checks used to evaluate the quality of each data delivery, for example,

- Data loading,
- Feature count,
- Batch validation (for coded value domains),
- Topology validation (geodatabase topology or geometric network), and
- Null or missing values.

All automated checks must meet the requirements established for acceptance (see Acceptance criteria).

Second-Round QC checks

This section identifies visual checks and QC methods associated with sampling and any others used to evaluate the quality of each data delivery, for example,

- Feature and attribute completeness
- Feature and attribute accuracy.

Some steps require inspecting a random sample of features where, for example, the number of features in the data delivery makes visual inspection of every feature too time-consuming and expensive.

All visual checks and other Second-Round QC methods must meet the requirements established for acceptance (see Acceptance criteria and Sampling strategy).

QA report

The purpose of the QA report is to disclose errors found during the initial inspection of the deliverables(s) and the First and Second-Round QC checks. The QA report includes sections for the data source inventory, data loads, feature counts, automated checks, visual checks, and a summary.

PAR (Problem and Resolution) system

The QA plan includes (if applicable) a problem and resolution (PAR) form or database to document missing or confusing information and provide resolution for the issues. This means a formal problem resolution process must be established to record and track all issues from the beginning of the project until all issues are resolved.

Example of a PAR form:

PAR Form							
ID #	Problem type	Problem description	Date identified	Production area #	Person assigned to problem	Date assigned	Resolution description and date

Conclusion

This section summarizes the QA plan. For example, this QA plan is intended to provide details about the quality assurance and quality control procedures and to establish a collaborative relationship among all stakeholders in the project. The criteria for acceptance of deliverables and the workflows used throughout the project are made clear and can be modified at any time pending agreement of all parties. Regular QA reports and tracking mechanisms for materials control and problem resolution will ensure open communication between stakeholders.

Appendix C—The Top Ten Most Common Metadata Errors

The information below is taken directly from the FGDC Metadata web site (FGDC, 2006), accessed June 2006 at *http://www.fgdc.gov/metadata/documents/top10metadataerrors.pdf*

10. Defining your data set too finely or too broadly

It is easy to become overwhelmed trying to individually document every data table and resource. On the other hand, trying to cover all of your data resources with a single metadata record will drive both you and your data users crazy. A good rule of thumb is to consider how the data resource is used – as a component of a broader data set or as a stand-alone product that may be mixed and matched with a range of other data resources.

9. Using incorrect State Plane Coordinate System (SPCS) Zone Identifier values

The default SPCS Zone Identifier (4.1.2.2.4.1) values for some software products are based upon early Bureau of Land Management (BLM) values rather than the FIPS Code prescribed by the Content Standard for Digital Geospatial Metadata (CSDGM).

8. Confusing "Currentness Reference" with "Publication Date"

While the Currentness Reference (1.3.1) may refer to a publication date it is actually a qualifier to Time Period of Content (1.3). Does the time period refer to the date/time of data capture or ground condition as in photography or field data collection? Does it refer to the date the information was officially recorded as in a deed? Does it refer to a publication date as in a '1978 USGS Topo map'? Basically, the idea is to let prospective users know how well you are able to 'nail' the actual time period of content.

7. Misunderstanding resolution

Who could blame us? The purpose of these fields is to indicate how coarsely or finely information was recorded. For example:

- **Latitude Resolution (4.1.1.1) and Longitude Resolution (4.1.1.2)**
  These values represent the minimum possible difference between coordinate values.

For example:

	resolution (4.1.1.1 or 2)	geographic coordinate units (4.1.1.3)
30° 30' 30"	0.00028 (1° / 3,600")	degrees, minutes, seconds
30° 30' 30.01"	0.0000028 (1° / 360,000")	degrees, minutes, decimal seconds
30.00001°	0.00001 (1° / 100,000)	decimal degrees

- **Abscissa/Ordinate Resolution (4.1.2.4.2.1 and 2)**
  These values represent the minimum difference between X (abscissa) and Y (ordinate) values in the planar data set. For raster data, the values normally equal pixel size, for example 30 (TM). For vector data, the values usually indicate the 'fuzzy tolerance' or 'clustering' setting that establishes the minimum distance at which two points will NOT be automatically converged by the data collection device (digitizer, GPS, etc.). NOTE: units of measures are provided under element Planar Distance Units (4.1.2.4.4) and would be 'meters' for the TM example provided and likely millimeters for the vector example.

6. Putting too much faith in metadata tools

Human review is the only thing that matters. The tools are there to help; remember "garbage in - garbage out."

5. Taking the minimalist approach

A common overreaction to the expansive nature of the CSDGM is to adopt 'minimal compliance' as an operational approach. Limiting your documentation to the 'required' portions of Sections 1 and 7, or even all 'required' fields, will limit the value of your effort and the metadata records you produce. Instead, identify those fields that apply to your organization and data, and create functional templates, or subsets, of the CSDGM.

4. Understanding assessments of consistency, accuracy, completeness, and precision

Section 2. Data Quality Information is intended to provide a general assessment of the quality of the data set. This represents the 'Achilles heel' for many Remote Sensing/GIS professionals.

Consider it an 'opportunity' to get to know your data set. A brief summary:

- **Attribute Accuracy Report (2.1.1)**
  Assessments as to how 'true' the attribute values may be. This may refer to field checks, cross-referencing, statistical analyses, parallel independent measures, etc.

Note: this does NOT refer to the positional accuracy of the value (see 2.4).

- **Logical Consistency Report (2.2)**
  Assessments relative to the fidelity of the line work, attributes and/or relationships. This would include topological checks, arc/node structures that do not easily translate, and database QA/QC routines such as: Are the values in column X always between '0' and '100'? Are only text values provided in column Y? For any given record, does the value in column R equal the difference between the values provided in columns R and S?

- **Completeness Report (2.3)**
  Identification of data omitted from the data set that might normally be expected, as well as the reason for the exclusion. This may include geographic exclusions, 'data was not available for Smith County'; categorical exclusions, 'municipalities with populations under 2,500 were not included in the study'; and definitions used 'floating marsh was mapped as land'.

- **Positional Accuracy (2.4)**
  Assessments of horizontal and/or vertical positional (coordinate) values. Commonly includes information about digitizing (RMS error), surveying techniques, GPS triangulations, image processing or photogrammetric methods.

- **'Precision'**
  An indication as to how 'finely' your data was recorded, such as digitizing using single or double precision. Note that the precision of the value in no way reflects its accuracy or truthfulness.

3. Glossing over Section 5. Entity and Attributes

Another of the GIS professional's 'Achilles tendons', this section maps out data content and should be a product of your data design effort.

- Use the relational database format as a guide:
  Entity Label (5.1.1.1) – Table Title
  Attribute Label (5.1.2.1) – Column Titles
  Attribute Domain Values (5.1.2.4.X) – Recorded values within each column

- Domain Types – set of possible data values of an attribute
  Enumerated Domain (5.1.2.4.1)
  A defined pick list of values
  Typically categorical such as road types, departments, tree types, etc.

  Range Domain (5.1.2.4.2)
  A continuum of values with a fixed minimum and maximum value
  Typically a numeric measure or count, may be alphabetic (A–ZZZ)

  Codeset Domain (5.1.2.4.3)
  A defined set of representational values
  Coding schemes such as FIPS County Codes, or Course No. (GEOG 1101)

  Unrepresentable Domain (5.1.2.4.4)
  An undefined list of values or values that cannot be prescribed
  Typically text fields such as individual and place names

- **Entity Attribute Overview (5.2.1)"**
  A summary overview of the entities/attributes as outlined in either Detailed Description (5.1) or an existing detailed description cited in Entity Attribute Detail Citation (5.2.2). Note that the field should not be used as a stand-alone general description.

2. Thinking of metadata as something you do at the end of the data development process

Metadata should be recorded throughout the life of a data set, from planning (entities and attributes), to digitizing (abscissa/ordinate resolution), to analysis (processing history), through publication (publication date). Organizations are encouraged to develop operational procedures that 1) institutionalize metadata production and maintenance, and 2) make metadata a key component of their data development and management process.

1. Not doing it!

If you think the cost of metadata production is too high – you haven't compiled the costs of not creating metadata: loss of information with staff changes, data redundancy, data conflicts, liability, misapplications, and decisions based upon poorly documented data.

For additional information, write to:
Director
U.S. Geological Survey
Leetown Science Center
11649 Leetown Road
Kearneysville, WV 25430

or visit our Web site at:
http://www.lsc.usgs.gov/

Document prepared by the West Trenton Publishing Service Center