# CyberInfrastructure Plan for the USA National Phenology Network

**Executive Summary:**
This document describes the key findings and recommendations from the Cyberinfrastructure and Informatics Subgroup of the USA National Phenology Network (USA-NPN) workshop held in October 2006 in Milwaukee, WI. These findings and recommendations outline the basic policy and technology assumptions we believe should guide the implementation of the USA-NPN. We believe that the USA-NPN is a much-needed tool with substantial ramifications for the research and understanding of both local and global climate issues. We also believe that a basic USA-NPN cyberinfrastructure can be established fairly quickly, leveraging from the capabilities that have been created in existing USA and international climate research and cyberinfrastructure efforts.

Based on these findings, this document also provides an outline and a high-level implementation plan for the USA-NPN.

*Change Log:*

| Date | Description |
|---|---|
| 14-Dec-06 | Assembled from components submitted by the Cyberinfrastructure and Informatics team members at USA-NPN workshop in Milwaukee, WI (October 10-12, 2006). Edits and revisions included from the 12 October through 12 December internal group comment period. |
| | |
| | |

Document Authors:
Ilkay Altintas (San Diego Supercomputer Center)
Peter Budde (National Park Service)
John Gross (National Park Service)
Wolfgang Grunberg (University of Arizona)
Thomas Gunther (United States Geological Service)
Homer Hruby (University of Wisconsin-Milwaukee)
Bruce Wilson (Oak Ridge National Laboratory)

## Table of Contents

CyberInfrastructure Plan for the USA National Phenology Network....1
1 Introduction....3
1.1 General goals and assumptions....5
1.2 Scope and Organization of the USA-NPN CyberInfrastructure Plan....6
1.3 Plan revisions....6
2 Policy Issues....7
3 System Infrastructure....8
4 Data Model....8
4.1 Key Assumptions....8
4.2 Implementation Requirements....9
4.3 Entities....9
4.3.1 Schema....10
4.3.2 Dataset....10
4.3.3 Observable (Species)....10
4.3.4 Phenophase....11
4.3.5 Protocol....11
4.3.6 Person....11
4.3.7 Site....11
4.3.8 Observation....11
5 Data Flow....11
6 Quality Assurance (QA) and Quality Control (QC)....11
6.1 Data Quality Expectations....12
6.2 Mandate for Quality....14
6.3 Quality Assurance and Quality Control Duties....14
6.4 Data Quality Goals and Objectives....15
6.5 General Operations....15
6.5.1 Version Control and File Naming Standards....15
6.5.2 Version Control....15
6.5.3 External Data....16
6.6 Project Planning and Data Design (Quality Assurance)....16
6.6.1 Record-level Tracking....16
6.6.2 Lookup Tables....16
6.6.3 Standard Operating Procedures....16
6.7 Data Collection....17
6.7.1 Field Sheets....17
6.8 Data Entry or Import....17
6.9 Data Verification (Quality Control Part 1)....17
6.9.1 Data Verification....18
6.10 Data Validation (Quality Control Part 2)....18
6.10.1 Methods for Data Validation....18
6.11 Data Quality Review and Communication....18
Credits....19
7 Data Security and Provenance....19
7.1 Key Assumptions....19
7.2 Data Integrity Issues....19

7.3 System Security Issues....20
7.4 Privacy Issues....20
8 Roles and Responsibility....20
8.1 Information Stewardship Roles....20
8.1.1 USA-NPN Director....23
8.1.2 Information Technology Specialist....23
8.1.3 Data Manager....24
8.1.4 Web and Applications Developers....24
9 Data Documentation....25
9.1 A Documentation (Metadata) Mandate....25
9.2 Core Documentation Requirements....25
9.2.1 Contact (Observer) Information....25
9.2.2 Place and Theme Keywords....25
9.2.3 Entity and Attribute (Phenological) Information....26
9.2.4 Time Period Information....26
9.2.5 Spatial Reference Organization....26
9.2.6 Distribution (Sensitivity/Liability) Information....26
9.3 Simplified Metadata Entry....26
9.3.1 Use of controlled vocabulary via pick-lists....26
9.3.2 Use of “site registration”....26
9.3.3 Use of “observer registration”....27
9.4 Supported metadata formats....27
10 Implementation Plan....27
10.1 Tools and Services....28
10.2 Skill Requirements and Commitments....29
10.3 Recommendations....30
10.3.1 Phase 1....30
10.4 Phase 2....30
10.5 Phase 3....30
11 Web Development....33
11.1 Assumptions:....33
11.2 Suggested Web Development Standards....33
11.2.1 Web Site Development Standards:....33
11.2.2 Web Application Development Standards:....34
11.3 Suggested Organizational Structure for Web Development....34
11.4 Web Site and Application Implementation....35
11.4.1 Developer skills....35
11.4.2 Anticipated Web Content Needed for Phase 1, 2007....35
11.4.3 Possible Deliverables and Milestones....36

# 1 INTRODUCTION

The vision of the USA National Phenology Network (USA-NPN) is that it will grow to be a continental-scale, multi-tiered network including phenological monitoring of regionally appropriate native and non-native species and phenomena, cloned indicator plants (lilac, etc.), and selected agricultural crops. The USA-NPN will:

- Facilitate understanding and evaluation of long-term responses to climate variability/global warming thorough understanding of phenological phenomena, including their causes and role in the biosphere.
- Serve as ground truth for scaling up remote sensing observations of the terrestrial biosphere, making the most of the large public investment in satellite platforms and remote sensing products
- Allow detection and prediction of environmental change for a wide of variety of applications, including but not limited to assessing impacts of land use and of climate variability and change on pollinators, cattle, crop and forest pests, wildfires, carbon storage, and water use.
- Actively engage private citizens as "ships of opportunity" that contribute in significant ways to fundamental science of vital interest to Education, Health, Commerce, Natural Resources and Agriculture.
- Develop a model system for substantive collaboration across different levels of government, academic and private sectors in the context of a well-defined and popular mission that will include free access to all data collected.

This chapter summarizes the information system goals for the USA-NPN and outlines key design and implementation elements to address these goals. The USA-NPN cyberinfrastructure (NPNCI) plan documents the overarching strategy for ensuring that program data are documented, secure, remain accessible, and remain useable long into the future. The NPIS plan refers to existing standards and builds on standard operating procedures already established within the scientific community[1].

The USA-NPN is designed to manage long-term monitoring data on seasonal events. Over time our understanding of these dynamics will change, as will requirements for data input, management, and the provision of data products. This document provides a strategic overview of the requirements, design, and implementation schedule to guide the systematic development of the USA-NPN cyberinfrastructure (NPNCI). Our vision of the full implementation of the NPNCI includes tools for data input, an underlying database structure (the 'back end'), reporting tools that provide raw data, visualization tools, and more advanced functions that permit linkages to remote data and to automate and streamline processing (i.e., accommodate scientific workflows).

[1] J.W. Brunt and W.K. Michener (eds) Ecological Data: Design, Management and Processing (Blackwell Publishing, 2000) ISBN 06-320-5231-7.

### *1.1 General goals and assumptions*

The overarching goals the NPNCI are to ensure that phenological data are:

- Available – the computing systems in the NPNCI employ appropriate best practice and operate continuously. Well-established practices are used to ensure that the location and content of data can be readily identified.
- Usable – data are stored in a stable, reliable, and interpretable data retrieval system.
- Reliable – The NPNCI includes a process for QA/QC of data, ensures that data are not inadvertently changed, that all changes are logged, and that users have tools to verify the integrity of data which they have entered.
- Shareable – data products are complete, subjected to quality assurance, formatted for use and documented for interpretation by others.
- Easily integrated – data and products are consistent with data exchange standards and mechanisms are in place to ensure interoperable with related data sets and information systems.
- Interpretable – the data are routinely summarized, transformed into useful information, and reported in formats designed for specific clients.

To meet these broad goals, the NPNCI will be consistent with the following principles and assumptions. The system must provide to users:

- A simple web interface to input and check the contributed data
- A standardized and well-documented web service interface for contributing, searching, and retrieving data and metadata
- A broad range of USA-NPN phenological data products, and analysis tools.
- Searchable metadata describing phenological data maintained by others, including information for how users can access that data. Eventually, this may develop into a full-fledged portal for a network of phenological data
- Information about and a portal to phenological data maintained by others, both within the USA and in other countries.
- Information and interfaces tailored to specific needs of researchers, educators, students, and decision makers.

The overall implementation and development of the NPNCI will be guided by these assumptions:

- Content will be provided by the USA-NPN community, not the cyberinfrastructure support staff
- Data integrity and cybersecurity are essential, primary considerations.
- While there may be a master repository for metadata describing phenology data relevant to the USA, it is highly unlikely that there will be a master repository for the data itself. However, it will be necessary for the NPNCI to provide the capability of storing USA-relevant phenology data for researchers who do not have the means or infrastructure to maintain a long-term archive.

- A data access and use policy will be necessary that covers observation data and personal data about observers.
- The design must be efficient and automate processes where possible.
- The USA-NPN must be in full cooperation with other national phenology networks to ensure that the needs for global climate research can be met.
- Data sets will need to be modified and expanded. Versioning and traceability are important.
- A small number of metadata and data formats will be supported.
- The NPNCI will evolve and be enhanced over time. The initial implementation will gracefully accommodate scaling and embellishment with cutting-edge technology.
- Where practical, the system will make use of free and open source tools.
- USA-NPN data will be distributed to users in electronic form and there will be no financial cost to users for access to this data.
- Initially, all USA-NPN managed data will be available without restriction, except where necessary to protect the privacy of observers. In subsequent versions of the NPNCI, access to certain data may be restricted to project members during the course of the project, but this type of project-level security might not be in the initial implementations as a design simplification mechanism.

### *1.2 Scope and Organization of the USA-NPN CyberInfrastructure Plan*

The NPNCI plan is a guide to the USA-NPN to ensure the continuity and documentation of the phenological data that the program manages. This plan attempts to identify all the major issues that must be considered to implement a basic infrastructure sufficient to permit users to contribute monitoring data in early 2007, but it does not fully address many issues that will need to be resolved to fully implement the USA-NPN. This version of the NPNCI plan includes many placeholders that identify issues or needs to be addressed in the future.

The NPNCI plan focuses on those aspects of the USA-NPN infrastructure that are necessary to:

- Acquire, store, manage, and archive data
- Ensure the quality of data
- Document and disseminate data and information
- Ensure long-term access to and utility of data

### *1.3 Plan revisions*

Data management is a process that begins before any observations are obtained. The normal implementation of data management and flow of data are shown in Figure 1.1. As the USA-NPN matures, enhancements and embellishments will be needed at every stage of the process shown in Figure 1.1. The data management plan will therefore be a dynamic document, subject to periodic updating. Our expectation is that many sections of the NPNCI plan will need to be frequently updated for the first several years of establishment, and less frequently once the Network is established and routine operations are more clearly defined. We also expect to make

use of more detailed component design documents which will be updated more often to reflect the current design of the NPNCI.

## 2 POLICY ISSUES

A number of policy issues will need to be addressed by the USA-NPN governance structure and these policies will likely evolve over time. The following are proposed as principles upon which the policies should be based:

- The data collected and managed by the USA-NPN should be made available to the general public without fee and with as few restrictions as practical, primarily those necessary to protect the privacy of individual observers and to protect the scientific integrity of ongoing experiments.
- For the Phase I implementation (see section 10, page 27 for the Implementation Plan), the download of data will require some type of registration in the system. This is primarily because the Phase I implementation is a pilot implementation and feedback from users is important. Whether anonymous downloads of data will be permitted in the future is something to be settled by the USA-NPN executive leadership as part of the overall data policy. Registration will involve the collection of the minimum amount of Personally Identifiable Information (PII) necessary and the USA-NPN will not disclose this PII to third parties except where legally obligated to do so.
- The privacy of the citizen scientist observers, in particular, should be protected. The expectation is that general public would have access to data that does not identify observers at all (not even as coded observers) and has locations "fuzzed" (e.g. lat/long coordinates rounded to the nearest second or to four decimal places, as example schemes). Researchers with a validated need and with a signed confidentiality agreement on file would have access to exact observation locations and coded observer numbers. Only researchers working on supervising a specific group of citizen scientists working on a particular project would have access to the identifying information for the specific observers working on that particular project.
- Under some circumstances, it may be necessary to protect the exact location of an ongoing phenology experiment while it is in progress (which may be for the lifetime of the USA-NPN itself). This is common practice in some disciplines, as even deliberate sabotage of experiments has been known. However, the bias should be to make as much information as possible available to the public, and the USA-NPN executive management will need to determine the circumstances under which exact experiment locations will be restricted even within the pool of registered phenology researchers. It may also be necessary to restrict access to data for a particular project area for a specified length of time, in order to allow researchers to complete their investigations and publish those results.
- The USA-NPN system will track the identity of the individual who uploaded each piece of data. In other words, anonymous observers will not be accepted into the system. Reasonable efforts will be made to validate the identity of system registrants. At a minimum, this should include validation of the e-mail address and may include tools such

as CAPTCHA[2] to mitigate automated robots and other common attacks on public registration systems.

- The issue of whether the USA-NPN is a "Federal Computing System" should be resolved, as that will have a significant effect on a number of policy issues.
- Appropriate attention will need to be paid to child online protection legislation and common practice. The legal issues in this area are in flux and appropriate legal advice should be sought. For the Phase I implementation, it may be advisable to restrict the users to those 14 and older to eliminate this as an issue.

# 3 System Infrastructure

The hosting location or locations for the NPNCI have not been determined at this writing. Our recommendation is that the NPNCI be hosted by facilities which have a record for managing environmental data and which have an institutional structure suitable for a long-term archive (particularly including a strong record in cybersecurity). It may also be desirable for multiple institutions to be involved in the hosting of NPNCI, partly to ensure replication of this valuable data source and partly because low-latency (local) access to this type of data may be useful for advancing the particular science capabilities of certain institutions. We expect that the hosting institutions for the NPNCI will be selected by the USA-NPN executive management and the granting agencies through an appropriate proposal process. We also expect that the use of web and web services access to the data should make the physical location(s) for NPNCI relatively transparent to the vast majority of end users.

The details of the system infrastructure are also not determined at this time, as it is likely that these details will be determined based on the capabilities and infrastructures of the particular institution(s) selected to host the NPNCI.

# 4 Data Model

## *4.1 Key Assumptions*

- The NPNCI will be based on one or more central databases which hold the data and metadata that are actively managed by the USA-NPN, along with metadata describing relevant data holdings from related sites and networks.
- The NPNCI data holdings (the data actively managed by the USA-NPN) will consist of multiple datasets which may not be directly compatible with respect to things such as precise citation policy, the particular geographic coordinate system used, and/or the exact meanings of columns. It is, however, important to use a common basis data model, with particular datasets potentially extending this data model. The NPNCI data model will be published and will be available for use by other researchers.
- A significant amount of work has already been done to develop data models for ecological data and it is desirable to leverage this existing knowledge base both for speed of implementation and to help ensure interoperability with other observational data systems. In particular, the data model for USA-NPN should leverage the observational data model from

[2] Computer Automated Public Turing test to tell Computers and Humans Apart, used to mitigate the problems associated with automated registration robots. C.f. http://en.wikipedia.org/wiki/Captcha

Natureserve.org[3] and the Natural Resource Database Template (version 3.1) from the U.S. National Park Service[4].

- While it is not critical for the Phase I implementation plan, it will become critical for a mechanism to exist whereby observers and researchers can upload data (multiple records) electronically to create a new data set or append to an existing data set. An example use case for this would be a researcher working in the field with a laptop making observations, which would then be uploaded when the researcher reaches a location where Internet access is available. Other examples could include automatic data logging tools, and uploads of existing data sets. In these cases, the datagram will have internal references (and may have external references) which must be maintained.
- Metadata for USA-NPN data holdings will be created initially in FGDC format (note that FGDC is ISO 19115 compliant). In order to minimize the effort required to create and maintain a metadata listing of USA-NPN-related data, the Phase I implementation will initially focus on FGDC-formatted metadata, with other formats accommodated on an effort-available basis. Additional metadata formats will be supported at some future time, but the intent will be to keep the number of supported metadata formats to a minimum.
- Data in the USA-NPN data holdings will be of variable quality and of variable quality control. Therefore, the data model needs to account for the level of quality assurance and possibly for the qualifications of the observers.
- All changes to data in the USA-NPN data holdings must be logged, to ensure traceability, and all changes must be "undoable" by a systems administrator (but not necessarily by the end user). The data model must support the versioning of data records and the tracking of changes.

### *4.2 Implementation Requirements*

To address the needs for uploading data and off-line recording of data, we propose that the NPNCI data model be constructed with the expectation that it will need to be implemented both in a relational database model and in an XML schema. This will allow other systems to compose an XML datagram with one or more data records (or even data sets) to be uploaded into the USA-NPN data holdings.

To accommodate the offline preparation of datagrams, we propose following the practice of using 50-character Universally Unique Identifier (UUID)[5] as the primary key for all tables. In the overall scheme of the data storage requirements, the cost increment over an integer primary key is minimal, but using a UUID provides huge benefits for off-line preparation of data sets and for allowing synchronization of data between repositories.

### *4.3 Entities*

The details of the data model will be reflected in a more dynamic document than this cyberinfrastructure plan. As such, this section represents only the high level conceptual design.

---

[3] Personal communication from Kathleen Goodin (Kathy_Goodin@natureserve.org) to John Gross (John_Gross@nps.gov) on 10/5/2006 via e-mail.
[4] http://science.nature.nps.gov/im/apps/template/index.cfm
[5] Also called a Globally Unique Identifier or GUID in some contexts (c.f. http://en.wikipedia.org/wiki/UUID.)

### 4.3.1 Schema

The concept of a Schema, in the database sense of the word, is central to the mechanism we propose to use for handling entities where the allowable values differ from dataset to dataset. For example, two datasets (possibly historical) may have different sets of phenophases for the same species and it may not be practical or possible to reconcile these different definitions. A schema consists of a complete data model and a consistent set of reference tables. A given database may have multiple schema, and each schema may contain multiple data sets. Within the USA-NPN database, there would be a master schema, from which other schema can draw for commonly used things, such as a master species list. However, a schema could override the master definition, should that prove necessary to preserve the integrity of the data as received from the researcher.

### 4.3.2 Dataset

A dataset is a list of related observations, along with all of the other data and metadata necessary to fully describe those observations. A dataset has a coordinator, which would typically be a phenology researcher. The Dataset entity is important to the management of citizen scientist observations, since the citizen scientist is primarily entering Observation records which are then appended to an existing Dataset. The metadata for that Dataset was then presumably entered by the researcher coordinating that set of observations.

As an example, the Eastern US lilac data would be a Dataset, with Mark Schwartz as the coordinator. Mark would create the metadata describing the species, geographic & temporal extent, protocols, and phenophases represented in that Dataset. Citizen scientists (or other researchers) would not need to create metadata entries for each of their observations, but would simply need to enter the appropriate observational data, which become Observation records.

The Dataset concept also allows for a given Observation to be in multiple Datasets. The primary use case here is where a researcher performs quality control and validation operations to select records which are appropriate for a particular scientific question. That work could then result in a new Dataset which references a subset of Observations originally collected in one or more different efforts. The citation policies and data use policies for these derived Datasets requires substantial thought, but this ability is seen as critical to the ability to handle observations arising from observers with vastly varying levels of expertise.

### 4.3.3 Observable (Species)

This is an extension of the NatureServe model, in that it is necessary to accommodate the concept that a phenological observation may relate to a physical phenomenon, such as timing of ice formation or clearing or the peak of stream flow. It may be necessary to break this down into subtypes, reflecting some differences in phenological observations of living and non-living systems. Likewise, observations relating to mobile species may require some different attributes than those relating to non-mobile species, requiring some level of subtyping.

Note that to enable the ground truth measurements currently needed to support remote sensing phenology products (e.g MODIS and successors) it is also necessary for an Observable to refer to a general collection of non-specific species. For example, “trees” could be an Observable

relevant to remote sensing phenology, where a researcher would assess the percentage of trees in a given region (hundreds to thousands of meters) which were fully leafed out.

### 4.3.4 Phenophase

A phenophase relates to one or more Observables. For example, many plants could exhibit a bud burst phenophase, while other phenophases could be family-, genus, or even species-specific. The phenophase should have a clear definition and will be used as the basis for pick-lists to enable controlled vocabulary.

### 4.3.5 Protocol

A Protocol describes the method by which one or more phenophase measurements are done.

### 4.3.6 Person

A Person is someone who has some role within the system, whether as an Observer, Researcher, or Administrator. Some subtyping of this master type is likely as the data model evolves.

### 4.3.7 Site

A site is something with a geographic definition, whether as a point or as a (possibly irregular) region. Some citizen scientist observations of plant species may be best handled by a point type of a site (a single set of latitude/longitude coordinates). Other types of observations, particularly those relating to mobile species or to an ensemble of individuals, will need to refer to a site with a substantial geographic extent. A site will also need to carry some sense of geographic uncertainty.

### 4.3.8 Observation

The Observation is really the central entity in the entire data model. An Observation is made by an Observer of an Observable with a particular Phenophase at a particular Site using a particular Protocol.

## 5 Data Flow

The data flows for the system are to be determined.

## 6 Quality Assurance (QA) and Quality Control (QC)

Credits: This section is slightly modified from: Hart, M., and U. Gafvert. Editors, 2006. Data management plan: Great Lakes Inventory and Monitoring Network. National Park Service Great Lakes Inventory and Monitoring Network Report. GLKN/2006/20. See additional credits at the end of the section.

---

Analyses to detect trends or patterns require high-quality, well-documented data. For the USA-NPN, A formal and fully documented QA/QC process will critical to establishing and maintaining credibility of the data used by US-NPN staff and others.

Quality assurance involves planning to obtain the highest possible data quality, while quality control consists of monitoring the system or appraising the product after the product is developed. The USA-NPN will establish and document protocols in order to identify and reduce error at all stages in the data lifecycle. These stages include project planning and database design, data collection, data entry, verification and validation (certification), documentation

(including data quality and sensitivity review), and archiving (Figure 6.1). The final stage in the data life cycle is dissemination and integration. This section presents the more broadly based procedures and policies that govern specific operations within the USA-NPN, while more specific QA/QC procedures may need to be developed for specific monitoring programs or datasets. Figure 6.2 illustrates selected QA/QC procedures relative to the amount of planning and quality control necessary to have confidence in the data.

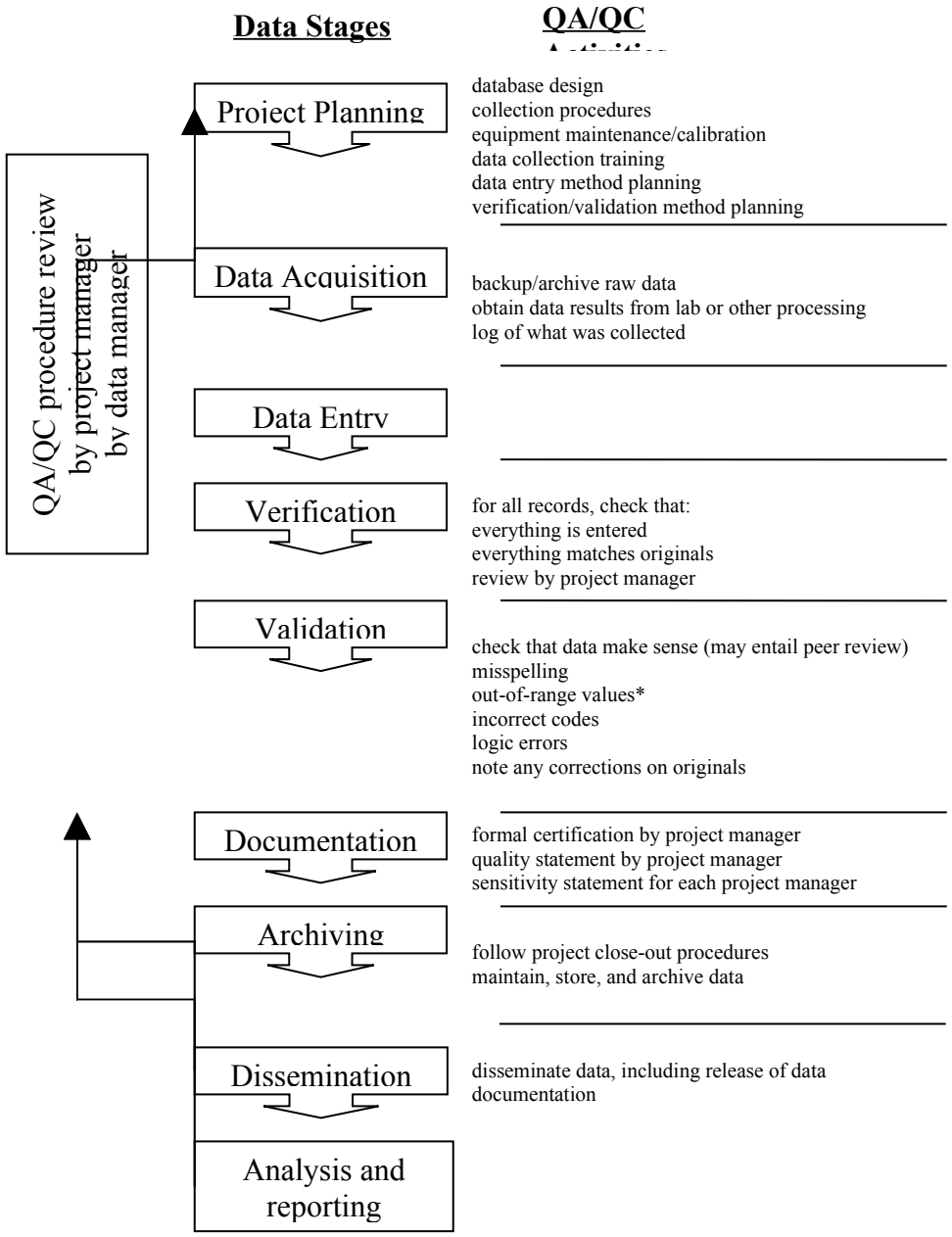
Quality assurance and quality control cannot be considered without specifying certain roles and responsibilities native to such procedures. The duties outlined in this chapter are consistent with those listed earlier in the cyberinfrastructure plan

## *6.1 Data Quality Expectations*

Data sets absent of all errors are ideal, but the cost of attaining complete accuracy may outweigh the benefit. Therefore, at least two factors must be considered when setting data quality expectations:

- frequency of incorrect data fields or records
- significance of error within a data field

We are more likely to detect an error in clearly documented data sets and understand what a 'significant' error is within *that* data set. The significance of an error can vary with data sets and depends on where it occurs.




* CAUTION! Care must be exercised when culling 'out-of-bounds' data. See text for discussion.

*Figure 6.1. General course of data and associated QA/QC procedures. Quality control with regards to data analysis is specific to each project and addressed in appropriate standard operating procedures.*

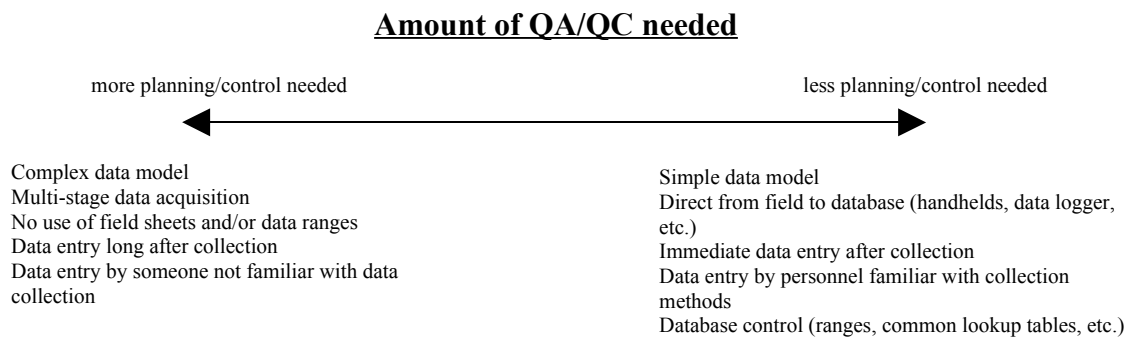



*Figure 6.2. Some common data management elements affecting the amount of QA/QC needed. Planning and training for data collection (QA) and entry is always critical.*

## *6.2 Mandate for Quality*

The concept of data *'quality'* incorporates three key components–*objectivity*, *utility*, and *integrity*.

*Objectivity* consists of: 1) *presentation,* which focuses on whether disseminated information is being presented in a proper context, in an accurate, clear, complete, and unbiased manner; and 2) *substance,* which focuses on the accuracy, usability, and reliability of the information.

*Utility* refers to the usefulness of the information to its intended users, from the perspectives of both the Network and the general public.

*Integrity* refers to the *soundness* of the data or the confidence one has in the data. Integrity is integrally related to objectivity; however, it is possible to have subjective data of high integrity. The integrity of data is also related to data security. Data must, for example, be protected from unauthorized access or revision to ensure that the information is not compromised through corruption or falsification.

## *6.3 Quality Assurance and Quality Control Duties*

Producing and maintaining high quality data is the responsibility of everyone involved with the handling of project data. Key data management roles and responsibilities and selected QA/QC duties are listed below.

Project managers must:

- be aware of quality protocols and convey their importance to technicians and field crews
- ensure compliance with the protocols
- plan for and ensure proper execution of data verification and validation
- review all final reports and information products

Technicians must follow established protocols for data collection, data entry, and verification.

The data manager is responsible for:

- developing Network-wide protocols and operational guidelines to ensure data quality

- making project managers, technicians, etc., aware of the established procedures and enforcing adherence to them
- evaluating the quality of all data and information against set standards before dissemination of data outside the network
- performing periodic data audits and quality control checks to monitor and improve quality control operations

## *6.4 Data Quality Goals and Objectives*

Quality criteria should be proportional to project-specific objectives and these criteria should indicate the level of quality that is acceptable. Project subjects and goals will drive data quality needs and control the kinds of analysis and summarization used to communicate results.

The most effective mechanism for ensuring that a project produces data of the right type, quality, and quantity is to provide procedures and guidelines to assist a project leader or data contributor in accurate data collection, entry, and validation. As part of data management operations, the USA-NPN will develop a comprehensive set of observation protocols and quality assurance guidelines to be used in the collection, entry, validation, verification, and use of data.

## *6.5 General Operations*

### 6.5.1 Version Control and File Naming Standards

Version control is the process of managing copies of changing files over the course of a project; file naming standards are critical for effective version control. Because of the scope and quantity of files being consolidated into one place, it is critical that computer files be given names that will uniquely identify them and indicate their content, even over the course of time.

The following conventions as suggested as a basis for naming files:

- No spaces or special characters within the name
- Include date for version control, in yyyymmdd format (e.g., 20051104 for Nov. 4, 2005)
- Use underscore as delimiters

File names for final products will begin with “NPN” and contain a brief but clear explanation of the file content followed by an eight digit date and the extension. Most file names will require an indication of the QA/QC status of the information as a name element. Examples of QA/QC status include “draft”, “raw”, “verify”, and “valid”. Each of these file name elements (excepting the extension) is preceded by an underscore (“_”). Additional underscores may be added for clarity. As an example, “NPN_ Data_Mngmt_Plan_draft_20061128.doc” indicates a draft version of the data management plan dated Nov. 28, 2006.

### 6.5.2 Version Control

Before making major changes to a file, a copy of the file with the appropriate version control (in this case the file name) should be made. As indicated above, files are stored with the appropriate eight digit date which serves as version control. This allows changes to be tracked over time and facilitates collaboration between multiple personnel working on common files. With proper controls and communication, versioning ensures that only the most current version is used in any analysis. Including the date, formatted as yyyymmdd, in the file name provides logical version

control. More formal version control, such as provided by a revision control system (e.g. Subversion) may also prove desirable.

### 6.5.3 External Data

External datasets are expected to comprise a significant resources and source of information for the NPS. While it would be preferable for the US-NPN to establish and enforce QA/QC standards and processes, this is not possible at this time.

## *6.6 Project Planning and Data Design (Quality Assurance)*

The methods and diversity of information that could be generated by a particular will largely determine the extent and types of QA that is necessary. As a policy, the USA-NPN should establish and routinely employ techniques and procedures that maximize data quality. Quality assurance will be achieved by stipulating that:

- Common lookup tables are created for values of parameters recorded in an identical field for more than one project (such as common weather metrics, locations, etc.)
- There is a specific (documented) operating procedure for core data management operations (collection, entry, verification, etc.)
- Data entry screens resemble any field sheets provided by USA-NPN
- Automated error checking features are included in data entry and ingest applications
- Database application design will maximize the use of auto-fill, auto-correct, value range limits, pick lists, and other constraints specific to projects
- Database applications will include a means to track errors reported on the data after dissemination
- Database maintenance logs will be maintained for each USA-NPN database and housed in association with database files

### 6.6.1 Record-level Tracking

As a standard part of database design, the USA-NPN databases should include fields that track, at the record-level, who entered the data, precise entry time and the protocol version under which the data were collected. The benefits to overall data integrity outweigh any inconveniences this 'overhead' data may cause due to factors such as increased database file size.

### 6.6.2 Lookup Tables

As noted in Section 6.1, the USA-NPN will utilize to the fullest extent possible common lookup tables for variables recorded by multiple projects. Examples include weather variables (such as precipitation intensity, wind speed, etc.), standard equipment and settings (e.g., GPS models and datums) and possible field personnel. Section 6.9 addresses database programming used for data validation.

### 6.6.3 Standard Operating Procedures

Each phenological protocol will include standard operation procedures that address core data management practices with quality control in mind. These may include:

- Standardized data sheets

- Use of handheld computers
- Equipment maintenance and calibration
- Data backup, entry, verification, and validation

## *6.7 Data Collection*

Although the USA-NPN may not have staff directly involved in data collection, the USA-NPN should provide guidance on 'best practice'. This section documents practices that USA-NPN can use to enhance the likelihood that data collection is conducted in a rigorous manner.

### 6.7.1 Field Sheets

The USA-NPN should design and provide standardized data sheets that identify the pieces of information to be recorded and forms that reflect the design of the computer data entry interface will help ensure that all relevant information is recorded and subsequent data entry errors are minimized. Data sheets should contain as much basic preprinted project information as possible and sufficient space for recording relevant metadata such as date, collectors, weather conditions, etc. They should clearly specify all required information, using examples where needed to ensure that the proper data are recorded. Data collectors should adhere to the following guidelines:

- All information added to the data sheet must be printed and clearly legible.
- If alterations to the information are necessary, the original information should be crossed out with a single line and the new information written next to the original entry. Information should never be erased and old information should not be overwritten.

## *6.8 Data Entry or Import*

'Data entry' is the initial set of operations where raw data are transferred to database tables using a computerized form. When data are gathered or stored digitally in the field (e.g., on a data logger), data entry consists of transferring the data (downloading) to a file in an office computer where they can be further manipulated. The goal of data entry is to transcribe field observations into a computer database with 100% accuracy. In other words, that which is recorded in the field should be entered exactly in the database. Subsequent data verification is conducted to ensure that raw data matches entered data. Following verification, data validation may result in changes *to the entered data.* Data entry is a separate operation from data validation and care must be taken to not impose validation (beyond that automatically imposed by programming rules in a database) during data entry.

In the case of the USA-NPN, considerable data entry will done by observers that enter data using web-based forms. For this situation, it seems preferable to link data entry and validation so errors are noted and corrected at the time of entry.

## *6.9 Data Verification (Quality Control Part 1)*

We appraise data quality by applying verification and validation procedures as part of the quality control process. These procedures are more successful when preceded by effective quality assurance practices (planning). *Data verification* checks that the digitized data match the source data, while *data validation* checks that the data make sense. It is essential that we validate all data as truthful and do not misrepresent the circumstances and limitations of their collection. Validation requires in-depth knowledge about the data.

### 6.9.1 Data Verification

While data verification is a key component of data management, most or all data used by the US-NPN will be collected by individuals or organizations over which we will have very little control or authority. The US-NPN is will thus have a limited ability to enact a data verification process. Data verification can be improved by calculating summary statistics and by identifying duplicate or omitted records where possible.

## *6.10 Data Validation (Quality Control Part 2)*

*Validation* is the process of reviewing computerized data for range and logic errors and may accompany data verification *only* if the operator has comprehensive knowledge of the data and subject. More often, validation is a separate operation carried out *after* verification by a project specialist who can identify generic and specific errors in particular data types.

General step-by-step instructions are not possible for data validation because each data set has unique measurement ranges, sampling precision, and accuracy. Nevertheless, validation is a critically important step in the certification of the data and a required component of any comprehensive protocol. Invalid data commonly consist of slightly misspelled species names or site codes, incorrect date, or out-of-range errors in parameters with well defined limits (e.g., percentages that must be 0-100). Another class of errors are logical, where trees are located in water bodies, or a delicate flower is reported as blooming in mid-winter. Various statistical procedures and logical rules can be used to identify logical errors, but these tend to be quite specific.

### 6.10.1 Methods for Data Validation

Two general methods for data validation can be initially applied to USA-NPN data:

1) *Data entry application programming.* Certain components of data validation can be built into data entry forms. This method is essentially part of database and/or web entry design.
2) *Outlier Detection.* Data quality assurance procedures should not try to *eliminate* outliers. Extreme values naturally occur in many ecological phenomena; eliminating these values simply because they are extreme is equivalent to pretending the phenomenon is 'well-behaved' when it is not. Eliminating data contamination is a better way to explain this quality assurance goal. If contamination is not detected during data collection, it is usually only detected later if an outlying data value results. When an outlier is detected, an attempt should be made to determine if some contamination is responsible.

   Database, graphic, and statistical tools can be used for ad-hoc queries and displays of the data to detect outliers. Some of these outlying values may appear unusual but prove to be quite valid after confirmation. Noting correct but unusual values in documentation of the data set saves other users from checking the same unusual values.

## *6.11 Data Quality Review and Communication*

A final step in the QA/QC process for a given dataset is the preparation of summary documentation that assesses the overall data quality. A statement of data quality will be composed by the project manager and incorporated into formal metadata as well as the GLKN primary data repository. Metadata for each dataset or database will also provide information on the specific QA/QC procedures applied and the results of the review. Typically, data quality

information will be conveyed as part of FGDC-compliant metadata (see section on data documentation).

*Credits*

*This section was adapted from concepts and material developed by Mark Hart (NPS Great Lakes I&M Networks), Debbie Angell (NPS Sonoran Desert Network), Doug Wilder (NPS Central Alaska Network), and Gordon Dicus (NPS Pacific Island Network).*

## 7 DATA SECURITY AND PROVENANCE

### *7.1 Key Assumptions*

- For the Phase I implementation, there are at most two levels of data access. Metadata holdings will be publicly searchable. Any user of the system will be able to search the metadata listing and retrieve records based on keywords, temporal search, and/or spatial search.

  The general public will also be able to search the USA-NPN-hosted data and retrieve the phenological data at least in the form of a delimited text file suitable for import into a spreadsheet. The general public would not, however see precise locations or even the coded information of the observer. For the general public, we propose to "fuzz" the locations of observations to display no more than four decimal places of lat/long to protect the privacy of the observers and of the observed sites.

  Researchers who have registered with the USA-NPN and who have signed an appropriate agreement of privacy protection would be able to access the full resolution location data and would get coded observer information (e.g. this observation was made by observer 127 and the next observation was made by observer 242). The actual identities of the observers would remain privacy-protected information.

- The key security issues are to a) preserve the integrity of the data and the underlying systems from both deliberate abuse and inadvertent change, b) protect the privacy of the network participants, c) provide for the preservation of the data in the event of data center disasters, and d) prevent hostile uses of the computing systems.

- We assume that anyone wanting to upload data to the system will need to be a registered user of the system. At a minimum, registration needs to involve a name and validated e-mail address. For observers and for researchers, additional information will need to be maintained, presumably including secondary contact information (phone, mailing address).

### *7.2 Data Integrity Issues*

To ensure the integrity of the data, key elements include an appropriate backup and recovery strategy for the systems, with periodic tests to ensure the proper operation of the backup. Some type of off-site storage for backups is also appropriate. We presume that a currently functioning data center will be familiar with appropriate best-practices in this area and that the specific implementation plans will be audited by an appropriate third party.

Further data integrity protection will be ensured by using a modern RDBMS as a back-end. The actual data tables will not be directly exposed to any users. Instead, users will access views, which provide row- and column-level security appropriate to their role. The system will be

constructed to provide a full audit trail on all data changes and the system will be constructed so that a systems administrator will be able to undo any user data change. An example of this would be that update operations actually create a new data record and mark the previous record as invalid. Records of all changes to data maintained within the USA-NPN Information System will be maintained in perpetuity.

### *7.3 System Security Issues*

We assume that a currently functioning data center will have appropriate perimeter firewall protection, with intrusion detection, and with provisions for penetration detection. All of these tools should be employed to provide a reasonably secure computing environment. We also expect that at least the senior developers involved with this effort will be aware of current computer systems security and computing best practices and that the USA-NPN Information Systems will regularly be audited for security vulnerabilities. Cybersecurity issues are one of the factors which suggest that the NPNCI should be hosted by institutions with a relevant long-term mission for data access and data integrity, as cybersecurity is expected to continue to escalate in importance.

### *7.4 Privacy Issues*

It will be necessary for the USA-NPN Information Systems to have a privacy policy and to ensure that access to Personally Identifiable Information (PII) is restricted on a need-to-know basis. We suggest that privacy policies for other citizen networks (such as GLOBE and the USGS bird banding database) be leveraged for the application. In compliance with federal regulations and generally accepted best-practices, PII should never be stored on laptop computers in unencrypted form. It is desirable that the PII be stored in an encrypted form, even on the servers, with two-factor authentication required in order to access this type of data.

## 8 ROLES AND RESPONSIBILITY

Development and management of the USA-NPN cyberinfrastructure and data will necessarily be a team effort, potentially involving all the observers, scientists, analysts, web developers, and database programmers. Clearly, the bulk of data management work will be done by the relatively few people responsible for the computer hardware and software. However, every person involved in the production of observational data, the analysis or synthesis of observational data, and in the management of digital records needs to contribute to preserving the quality of the data records and ensuring that information content persists. Given the considerable investments in time and money required to implement and sustain a quality information system, it is important to articulate roles and responsibilities in a manner that facilities smooth and efficient operation of the overall information system.

### *8.1 Information Stewardship Roles*

The tables in this section first describe general categories of information and data stewardship and the roles of key participants necessary to develop, operate, and maintain the comprehensive information system that the USA-NPN will eventually require. Table 8.1 describes general categories of data stewardship (production, analysis, management); each of these processes will require multiple people. Table 8.2 identifies key roles in information management. Even in a small network, there will obviously be one or more people in each role and these roles are

effectively job titles. For the initial implementation of the USA-NPN, it seems more likely that single individuals will have more than one role.

**Table 8.1**. Categories of data stewardship, activities associated with necessary to achieve

| Stewardship Category | Related Activities | Likely Positions Responsible |
|---|---|---|
| Production of data | Collecting data or information from any original or derived source. This includes recording locations, images, measurements, and observations in the field, digitizing source maps, keying in data from a hardcopy source, converting existing data sources, image processing, and preparing and delivering informative products, such as summary tables, maps, charts, and reports | Field observers, scientists harvesting data from existing databases or other sources (reports, publications, herbaria specimens, etc.) |
| Analysis | Using data to predict, qualify, and quantify ecosystem elements, structure, and function as part of the effort to understand these components, address monitoring objectives, and inform park and ecosystem management. | NPN staff<br>Scientists<br>Statisticians<br>Others that create data products |
| Management | Preparing and executing policies, procedures, and activities that keep data and information resources organized, available, useful, compliant, and secure. | NPN director and executive board members<br>NPN cyberinfrastructure manager<br>Project leader<br>IT specialist<br>Database Manager |
| End use | Obtaining and applying available information to develop knowledge that contributes to understanding and managing park resources.<br>Providing feedback for improvements in data content and quality.<br>Informing the scope and direction of science information needs and activities. | Scientists<br>Public<br>Media staff<br>Educational staff |

Table 8.2 is intentionally organized 'from the ground up' to emphasize that everyone who contributes to, processes, or uses USA-NPN data shares information stewardship responsibilities. As the Network matures, this table will require revision to reflect the structure of the USA-NPN and the evolving roles of USA-NPN staff and partners. The following text describes participation by key persons or groups in more detail.

### 8.1.1 USA-NPN Director

The Director has the ultimate responsible to ensure that data entry, validation, verification, analyses, web development, and hardware and software are coordinated and implemented in a manner consistent with profession data stewardship standards. In addition, the Director must guide and resolve policy issues, including those related to data access, privacy, and security. It seems likely that the Director will want advice from a Technical Committee or other governing board.

### 8.1.2 Information Technology Specialist

The IT specialist is responsible for designing or advising on the design of the overall cyberinfrastructure. This includes all necessary hardware (servers, data backup, networking

equipment, etc.) and IT software. The IT specialist is responsible for day-to-day operation of the system, including system administration and maintenance.

### 8.1.3 Data Manager

The Data Manager oversees the development, implementation, logical validation and long-term storage of USA-NPN data. The data manager must work closely with other IT staff to ensure data are archived, adequately documented, and compatible with other programs and applications. Over time, a key function will be to ensure data remain discoverable and available with changing technology and mechanism for data discovery and usage.

### 8.1.4 Web and Applications Developers

Software developers are responsible for creating new applications (i.e., writing code) with all that entails – designing applications, writing and debugging the underlying code, ensuring

Table 8.2. Information management and web development roles and responsibilities for likely USA-NPN observers, staff and partners.

| **Roles** | **Duties and obligations** |
|---|---|
| Observer | Collect, record, and verify data |
| Project leader or data Technician | Supervise crew and organize data; process and manage data |
| Information technology specialist | Provide IT support for hardware, software, networking. Take major role in design, implementation, and maintenance of IT infrastructure. System administrator. |
| Project leader | Oversee data project operations, including data management |
| Scientist | Validate and make decisions about data. Provide scientific content for web site. Advise on web-based analyses. |
| GIS specialist | Support geospatial software development; spatial analyses |
| Data manager | Develop and support data management system. Ensure data are organized, compliant, safe, and available. Ensure data documentation are complete and up to date. |
| Web developer | Develop, document, and maintain application code for web site. Is not responsible for content. |
| Database application developer | Know and use database software and database applications. |
| Statistician, biometrician, or scientist | Analyze data and/or consult on analysis |
| Director | Oversee and coordinate and network activities, including IT. Develop and provide guidance on policy. |
| End users<br>(scientists, interpreters, press, public, etc) | Inform NPN of the scope and direction of science information needs and activities. Interpret information and apply to decisions. |

application documentation is detailed, complete, and current, and maintaining software infrastructure. Software developers are not responsible for populating databases, writing the actual web content, or designing statistical analyses or data summaries. Software developers must seek and respond to feedback from application users.

# 9 DATA DOCUMENTATION

Data documentation is a critical step toward ensuring National Phenology Network (NPN) relevant data sets are useable for their intended purposes well into the future. This involves the development of metadata, or information about the data. By way of definition, metadata is "information about the content, quality, condition and other characteristics of data" (FGDC). Metadata provides a means of documenting the qualitative attributes of datasets, especially those that are distributed to audiences via both the intranet and internet. Knowing that a dataset is accompanied by complete and "compliant" metadata documentation adds a level of assurance to both the data steward and the potential users because such documentation helps to ensure that data will be used properly.

## *9.1 A Documentation (Metadata) Mandate*

The importance for metadata is now universally accepted within the data management community. Recent system development and data policy efforts emphasize that data documentation must include an up-front investment in planning and organization. Central to the success of USA-NPN's information legacy will be adequate, at least nominal, documentation of the spatial, temporal, and phenological content of observations contained in or accessed through the Network's information system.

An overarching requirement of data entered into USA-NPN's information system is the need to retain both lineage to the original source information and applicability for public or scientific usage. More generally stated, it is critical to persistently retain documentation on the quality and content of the information, i.e., answer questions related to who collected the information, when the information was collected, where the observation(s) occurred, and what information was recorded.

## *9.2 Core Documentation Requirements*

For NPNCI information to be useful, nominal (core) documentation is required. The following are key/mandatory elements that must be captured to ensure the usability of data:

### 9.2.1 Contact (Observer) Information

Describing the "who" is critical when capturing a phenological observation. Information provided by the observer can be used not only to determine the user's affiliation, e.g., academic, observer network or citizen scientist, but also as a surrogate for assessing the quality or credibility of an observation. Some of the information that needs to be provided include: name, organization, address, phone number, and e-mail address.

### 9.2.2 Place and Theme Keywords

In order to facilitate searches or queries on the database, a number of thesauri will be used to standardize place and thematic descriptions. A place and theme keyword thesaurus and respective keywords will be utilized to ensure consistency in data entry.

#### 9.2.3 Entity and Attribute (Phenological) Information

Details on the species or phenomena being recorded and the observation (phenophase or phenomenology) provides the "what" for each entry into NPNCI. In order to simplify keyword/taxonomic entry and ensure some level of consistency, lookup tables derived from or utilizing established conventions such as the Integrated Taxonomic Information System (ITIS) and BBCH principal growth stages will be integral to metadata authoring. Some of the mandatory items that will need to be provided include: ITIS #, species scientific name, species common name, and phenophase.

#### 9.2.4 Time Period Information

Clearly within the NPNCI, information on when an observation was recorded is essential to the success of USA-NPN. Standard entry of calendar date and time of day will be ensured through the documentation process.

#### 9.2.5 Spatial Reference Organization

Details of the coordinate system used to locate observations need to be documented so subsequent spatial analyses can be performed. In the initial phase of the NPNCI it is anticipated point observations will be supported. It will be critical for each observation point to be accompanied with mandatory description of: map projection name, horizontal datum, and assessment of accuracy.

#### 9.2.6 Distribution (Sensitivity/Liability) Information

General statements about the appropriate use and, if relevant, sensitivity of information in NPNCI is required. In order identify each record/observation, the observer will be required to include: a sensitivity flag (yes or no), and any usage constraints. Additionally, a standard liability statement will need to be included with all information to protect both data providers and end-users.

### *9.3 Simplified Metadata Entry*

Documentation of each observation or entry into NPCNI can be an onerous and, quite honestly, a daunting task. In order to encourage participation in USA-NPN's data collection, a documentation workflow will need to be implemented to minimize time commitments for individual observers or observer networks. Several ways to expedite and simplify the documentation process will be integrated into NPNCI.

#### 9.3.1 Use of controlled vocabulary via pick-lists

As noted earlier, consistency and standardization in documentation will be key not only to initial data entry but also subsequent end use of NPNCI entries. Controlled vocabulary for species lists, phenophase, place keywords, and thematic keywords will all be leveraged from recognized, credible sources to ensure valid data entry and adherence to necessary standards.

#### 9.3.2 Use of "site registration"

In order to quickly and consistently obtain information on a repeated observation location, site registration will be encouraged. This will not only streamline data entry but also ensure some level of consistency when identifying records captured as part of routine monitoring efforts.

### 9.3.3 Use of "observer registration"

Similar to the concept of site registration, an observer could also register with the NPNCI to expedite routine data entry. Contact information for each observer or affiliated observer could be entered once and "recycled" in subsequent data submissions. Again, this would ensure consistency and minimize the time and effort required for each data entry.

## *9.4 Supported metadata formats*

As noted earlier, metadata serves many important purposes. It is a vital foundation for understanding, collaborating and sharing resources with others. It allows people to determine what the best resources are for their individual needs by permitting them to see the details of the data itself, and its history. It benefits data-producing organizations by ensuring that data holdings are well documented over time so their value for the data holder and user is maintained. Metadata is important in the creation of a spatial data clearinghouse, where potential users can search, find and compare data in all its detail.
In order to support USA-NPN's mission, metadata formats that can accommodate documentation of both the geospatial and thematic content of entries is required. As a result, a number of recognized metadata content standards are supported by NPNCI. Conforming to a standard is important to ensure that everyone can find, understand and share data by finding and comparing common details of the data. A metadata standard outlines the characteristic properties to be recorded, as well as the values the properties should have. Such standardization of the vocabulary makes information sharing more reliable and universal.

At present, three formats are proposed, with only the first of these supported in Phase I:

- Federal Geographic Data Committee (FGDC) Content Standards for Digital Geospatial Metadata (CSDGM) were chosen for their quality, popularity of use, established support, as well as for the tools that have been and are continuing to be created.
- The new International Organization for Standardization (ISO) Standard for Geographic Information Metadata (ISO 19115) was chosen for its capabilities for internationalization. ISO 19115 is a newer standard that has more configurablity to application communities and supports internationalism in terms of languages and character sets.
- Lastly, the Dublin Core metadata element set is chosen because of it has emerged as a small set of descriptors that quickly drew global interest from a wide variety of information providers in the arts, sciences, education, business, and government sectors. The simplicity of Dublin Core can be both a strength and a weakness. In effect, the Dublin Core element set trades richness for wide visibility.

# 10 IMPLEMENTATION PLAN

The implementation plan is developed in three phases: near term, medium term, and longer terms, defined largely by available resources. The first phase, likely through spring, 2007, is intended to provide minimal functionality for a central USA-NPN database, web-based data entry and access, and a few basic tools, in a way that facilitates later expansion. Following development of these basics, other tools and services should be added in subsequent phases, building towards a comprehensive NPNCI. Broad priorities are suggested below, but the rapid evolution and development of tools and services, and the preferences of USA-NPN partners and management, will determine the specifics for each year.

One way to state the objectives of the NPNCI is in reference to the various audiences: the general public, students, researchers, USA-NPN management, and special interest groups (e.g., a specific research project or groups focusing on a specific species). Each of these has different, although overlapping needs, and can be thought of as requiring distinct websites. However, all build on a common infrastructure, and all should be reachable through a single USA-NPN portal. Once signed on, stored registration and authentication files route the individual to a sub-site with user-specific tools, services, documents, and capabilities.

### *10.1 Tools and Services*

General categories of desirable tools and services for a NPNCI include:

- *Scientific data management* – this refers to the management of and access to data and databases, and would ultimately include not only the USA-NPN developed database, but also connectors to phenology databases owned and managed by other organizations (e.g., the EPN or ice-on ice-off databases), other databases relevant to phenological research (i.e., those containing information on hydrology, climate, species information, geography, etc.), and miscellaneous other data (project-specific databases, anecdotal information, etc.). This category of tools should also ultimately address the rescue and archiving of legacy datasets from past studies.
- *Analysis* – this includes modeling, data mining, scientific visualization, GIS, scientific workflow, and other forms of data integration. Again, this is aimed primarily at scientific audiences, but could also serve others under certain circumstances.
- *Lists and Directories* – tools to address these needs could be used by multiple audiences (although with audience-specific content), and could include expertise or skills inventories, project lists, standards and protocols, organizational information, contact lists, funding opportunities, general and topic-specific annotated bibliographies, links to other websites, and so on.
- *Collaboration and communication* – these are primarily needs of the special interest groups, and USA-NPN project management, but would also extend to citizen scientist and educational groups. Tools could include bulletin boards, forums, wikis, news feeds, conferencing tools, reporting tools, photographs and other media repositories, and capabilities to plan and conduct physical and web-based workshops. They would promote better communication and a stronger sense of community among phenological researchers and others, and enable more coordination among science projects (e.g., coordinated data collection).
- *Digital Library* – this refers to the general need for a place to archive, search, and serve documents of all types. It could serve all audiences/websites, subject to copyright restrictions, although content may vary with the specific audience again with changes in content.
- *Website Management* - in addition to the above, the development of efficient and effective management of websites requires a set of tools for maintenance of the operating system, web server software, security, database management, and so on. These tools are

inexpensive and widely available, but should be kept in mind when considering the skills and expenses of developing and managing the website.

- *Data Policy*- How are we connecting a user to their data? How much data does the network own? Privacy, data release and publication policy, etc.

### *10.2 Skill Requirements and Commitments*

In order to build, maintain, and evolve web-based tools and services, a variety of skills are needed to provide and maintain infrastructure, as are commitments from the owners of scientific and management information (content) to prepare and provide it for web publication. The mix of skills needed will differ with the type and stage of development and, in many cases, individuals can provide more than one type of skill. However, it is very unlikely that a single individual can provide all necessary skills, especially as progress continues beyond the initial setup of the core USA-NPN database and minimal web services. The skills required are:

- *Technical Administration* – This refers to the management, upkeep, updating, patching, running, installation, and similar tasks related to the basic software that makes the set of USA-NPN websites possible. There would be some, perhaps substantial, concentration of effort at the beginning, and these skills would need to be available on a sustained basis, albeit at a lower level of effort. Specific types of support required include:
    - System Administration - for management and maintenance of the operating system and server software, and for compliance with policy, any applicable regulations, and security. The latter is particularly true if theNPNCI resides on government computers, although it should be addressed in any case.
    - Database Administration (for specific databases managed by and for the USA-NPN community. These include such things as contact lists, project databases, bibliographies, and so on, but not databases managed by other organizations.
    - Application Administration (for other applications that manage specific functions or services, e.g., calendars, wikis, content management systems, etc).
- *Website Development* – This effort should precede the creation of websites to ensure that sites have clear goals, and are designed to meet those goals in ways that audiences will find useful, intuitive, and efficient. Skills required are:
    - Information Architect (to help USA-NPN scientists, and other potential users organize and present information and services logically and efficiently, especially in the public website).
    - Website/Graphics Designer (to ensure that designs are visually effective and compelling).
- *Content Preparation* – Effective communication with the general public, students and educators, and scientists unfamiliar with phenology will require that material be rewritten for a non-scientific audience. This is a critical element of informing the public of USA-NPN work, and represents a relatively rare skill. Actual requirements depend on the quantity, type, and speed of "translation" desired, but ultimately, one individual may be needed for this task. A separate content challenge relates to the data, databases, and information that would be served to scientific and other audiences, and includes the need for metadata (both geospatial and other) creation and management. This task can be, and in some cases is, addressed by scientists or science programs. However, this is not yet

universal, and should be anticipated and planned for. In some cases, experience indicates that it is far more efficient to task a metadata specialist to help individuals and groups rather than rely on semi-voluntary compliance.

- *Web-based Tools and Services* – The development and maintenance of wikis, visualization and modeling tools, GIS, database connectivity tools, resource locators, etc., etc., will require one or more programmers, and a commitment to "best practices" that reflect an intent to create reliable infrastructure for scientists, managers and the public. Many such tools exist, and will continue to be developed, but should be selected based on the principles stated earlier.
- *Content* – The multiple websites within the overall USA-NPN portal, no matter how well conceived and implemented, will, over time, be of little value unless there is a relatively steady stream of data, information, or other materials – content – flowing into them and to the audiences. To paraphrase James Carville, "it's the content, stupid." The sources of this content are the scientists and science citizens: science publications to be served to public and scientific audiences; science data, models, and GIS coverages to be shared with other scientists; and news from meetings, funding announcements, policy statements, and contact lists to keep the community informed. While this may be primarily a responsibility of the science community, there will likely be a need for someone to encourage the continuous contribution of new material.

## *10.3 Recommendations*

### 10.3.1 Phase 1

a) Adopt recommended data model, architecture, and applications (e.g. Linux, Apache, Postgress, PostGIS).
b) Build database, beginning with data from lilac and other available databases.
c) Build high level website, with sections providing information about USA-NPN, contacts, plans, etc.
d) Build data entry screens for single observations, multiple grouped observations, and database mappings.
e) Build limited data download capabilities, specifiable by species, location, etc. (e.g., comma delimited, Excel)
f) Build minimal web mapping services.

## *10.4 Phase 2*

a) Connections to other databases (phenologically related, other)
b) Add additional sort, search, and mapping capabilities
c) Educational Materials
d) Support for student data input, downloads
e) Collaboration tools

## *10.5 Phase 3*

a) Analytical tools
b) Visualization tools

c) Scientific workflow tools
d) Grid computing capabilities
e) Bibliographies
f) Publication support

| Tools & Services Use Audience | Data Entry/ Down-load | Contacts, About, FAQs | Mapping Services | Educational Materials | Connectivity to multiple DBs | Collaboration tools (lists, forums, calendars, wikis,etc. | Bibliographies (public and annotated) | Digital Library | Analytical Tools | Viz Tools | Workflow tools (Kepler) | Publication Support |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| General Public | | 1 | | 2 | | | | | | | | |
| Educational Materials | | | | 2 | | | | | | | | |
| Citizen/ Scientists | 1 | 1 | 1 | 2 | | | 3 | | | | | |
| Researchers | 1 | 1 | 1 | 2 | 2 | | 3 | 3 | 3 | 3 | 3 | 3 |
| Community of Practice | 1 | 1 | 1 | | 2 | | 3 | 3 | 3 | 3 | 3 | 3 |
| Research Projects | | 1 | | | 2 | | | | | | | |
| NPN Management | | 1 | | | | | | | | | | |
| Other | | | | | | | | | | | | |
| | | | | | | | | | | | | |

# 11 Web Development

## *11.1 Assumptions:*

For web site design and development to proceed in a timely fashion the following assumptions are made so developers can concentrate on the design, information architecture, and prototyping of the USA-NPN Version 1.0 web site.

1. Stakeholders are identified and will provide feedback and constructive comments to move the web site development process forward in a timely manner
2. An initial content document (outline of content) approved by the USA-NPN steering committee will be provided to the developers.
3. All USA-NPN members will contribute with content such as text and images.
4. The web server and database server will be identified and all the necessary permissions for access will be provided (a database will be required for the website, for phenology data inputting, and for phenology metadata)
5. The USA-NPN database architecture exists and is provided to the developers
6. W3C web development standards will be used for site development whenever possible
7. Web services and advanced tools will be developed in later phases
8. Financial resources will be made available as needed to implement the design and interactive features requested by the stakeholders
9. An iterative development process will be developed which constantly evaluates the usability of sites and tools developed

## *11.2 Suggested Web Development Standards*

We recommend the following web development standards be adopted and used by the developers.

### 11.2.1 Web Site Development Standards:

1. Use Cascading Style Sheets (CSS - http://www.w3.org/Style/CSS/) to decouple web page content from its presentation (layout and design). Include printer friendly CSS printer style sheets
2. Offer alternative outputs that can be used offline (pdfs, zip files, data dumps, graphics, etc.)
3. Use XHTML Transitional (http://www.w3.org/MarkUp/) as the markup language
4. Use the ISO 8859-1 Character set standard
5. Use semantic mark-up (http://www.w3.org/DesignIssues/Semantic.html) to increase machine readability of web pages and to improve the pages ranking on search sites.
6. Use and populate the XHTML metatags; customize them for individual pages when ever possible.
7. Validate CSS and XHTML coding against W3C (http://www.w3.org/QA/Tools/)
8. Whenever possible, web pages must be complaint with the Section 508 of the Americans with Disabilities Act (http://www.section508.gov/). Validate accessibility against Cynthia by HiSoftware (http://www.hisoftware.com/) or similar applications or services.
9. Test all USA-NPN sites and web applications with MS Internet Explorer 6.0 and 7.0, the latest browsers from Mozilla (Firefox), Opera and Safari

10. Test web sites and tools on representative stakeholders and early adopters and use the feedback for the next developmental iteration.

11.2.2 Web Application Development Standards:

1. Care should be taken in allocating more than enough time for the application design phase and for the debugging phase.
2. Adopt a Rapid Prototyping stance by:
    a. Use software languages such as Python (Zope), Ruby on Rails, and PHP. These languages allow for rapid development and are scalable as well as accessible to casual programmers.
    b. Follow Agile Development strategies such as suggested by the "Manifest of Agile Development" (http://www.agilemanifesto.org/).
3. Use industry standard design patterns such as Model Viewer Controller (MVC).
4. Use a concurrent versioning system such as Subversion (http://subversion.tigris.org/) to enable collaborative development, versioning, and code control.
5. Use a development management system such Trac (http://trac.edgewall.org/) in combination with Subversion.
6. Use web services (using SOAP, HTTP Get/Post, etc.) to allow for interoperability of web sites, web applications, and data sources. Web services offer the means of connecting future USA-NPN and partner web nodes to each other.
7. Use open source applications as much as possible. Active involvement with an open source application's development team should be fostered to ensure improvements to the application as well as the understanding of the application.
8. Use PostgresSQL (http://www.postgresql.org/) with Post GIS (http://postgis.refractions.net/) as the relational database management system (RDBMS). PostgresSQL is a fully featured, stable, and scalable open source RDBMS with GIS data storage capability.
9. Spatial data, processes, services, and results (georeferenced outputs) should follow the Open GIS Consortia (OGC - http://www.opengeospatial.org/) standards whenever possible and appropriate.
10. All spatial data must have metadata attached to it that follow agreed upon standards. As a minimum, FGDC (http://www.fgdc.gov) metadata standards must be followed.
11. Use the same agreed upon projection(s) and datum(s) (for example GCS and NAD83) for all spatial datasets and records.
12. Stress test web applications and sites with tools such as Apache Jmeter (http://jakarta.apache.org/jmeter/)
13. Use appropriate best practices to encrypt communications and to secure data.
14. Backup often and early.

### *11.3 Suggested Organizational Structure for Web Development*

Following are suggestions regarding the organizational structure of the web development teams:

1. A thorough discussion is needed to come up with a workable web development organizational structure that is responsive as well as scalable.
2. Key will be a full-time development leader which "owns" the USA-NPN web site. Particular attention needs to be placed on the person's leadership of web development efforts.

3. Phase 1 may require a developer with a variety of skills to quickly develop a prototype site.
4. Eventually the developer must manage and coordinate a collaborative development environment where USA-NPN members across the network contribute with their skills, applications, data, and servers.
5. Outsourcing development and services to USA-NPN members might be crucial for building the network and growing it in the future.

## *11.4 Web Site and Application Implementation*

### 11.4.1 Developer skills

The following web developer skills will be needed to build the USA-NPN Version 1.0 web site. No single individual is likely to have all the skills need to develop the site. We recommend using a web development team to develop the site. An individual developer will be identified to supervise the project.

1. Web graphic design: imaging plus other media if requested
2. Information architecture design: navigation, usability, web architecture plan for growing the site
3. Applications developer for the phenology data input and output tools(s).
4. Content provider: writers and editors which create web content as well as adapt print media documents for web consumption. Graphic specialist may be needed to work with the content providers if needed
5. Application developers for visualization and analysis tools – for phase 2 and 3
6. Content management system (CMS) developer or implementer – for phase 3?

The navigation and information architecture of the USA-NPN Version 1.0 web site is dependent to a certain extent on the information content developed for the site. Content developers need to provide a planning document that will elaborate on the content need for USA-NPN Version 1.0. Included in this document are timelines for content development, time lines for updating the content on the site, and likely additions to the content on the site, including both near-term and long-term additions that may impact the navigation and information architecture of the site. It also would be useful for the web developers to have a sense of the long-term plans for development of future content, applications, services etc. on the USA-NPN web site. This planning document will be used by the developers of the site to plan the navigation and information architecture of the site, so growth of the site can be planned as best as possible. This allows the developers to update the site, redesigned and recoded the site in the future in as efficient process as possible.

### 11.4.2 Anticipated Web Content Needed for Phase 1, 2007

We anticipate the following information and/or content will be need for the USA-NPN Version 1.0. The content will need to be provided to the development team for them to succeed.

1. Audience identification (scientists, citizen scientists, board of directors, educators, decision makers, general public)
2. Graphics including a new logo for the graphic design of the web site
3. Written text about USA-NPN:

   a. What is phenology?
   b. Organizational structure
   c. How to cite USA-NPN
4. List of targeted species along with associated data about the species (geographic extent, priority for observation, images)
5. Phenophase documentation and associated graphics and for all selected target species.
6. List of variables and attributes to be recorded on the forms
7. Public (anonymous) and private (NPN members) repository of links and resources concerning phenology.
8. Contact information for data consumers, data providers, as well as potential partners
9. List of funding agencies/institutions; list of collaborating agencies/institutions
10. Privacy statement for contributors
11. Human subjects disclaimer for researchers which use USA-NPN data? Note that USA-NPN data contributors are likely identifiable by the geographic location of their observations (someone's backyard). Anonymity of the contributors could only be granted by reducing the spatial accuracy and precision of the phenology observations.

### 11.4.3 Possible Deliverables and Milestones

#### 11.4.3.1 Phase 1

1. Develop ideas for graphic look and feel of USA-NPN design for the web and other media.
2. Static web pages for content with consistent graphic design and a sub-site structure which allows web-users to self-select (content and tools for scientists, data contributors such as C.S., funding agency, educators, public, etc)
3. Data contributor tool to the USA-NPN core phenology observation database. The tool includes web interfaces for observation data and metadata inputs, for editing data, and for extracting data.
4. Phenology resources links (could be sortable and searchable as time and money permits)
5. Make USA-NPN documents available (history, evolution, future plans)
6. Possibly, additional web tools for interactive visualization of submitted data (web map with internet map server, Google Map?), etc
7. Deploy internal USA-NPN collaboration site and/or content management system (http://topshare.wur.nl)
8. Seek feedback, suggestions, and evaluation of usability
9. Develop measures of success for web sites and tools.
10. Establish policy and mechanism for handling hardcopy or offline (email?) submissions of phenological observations (Excel, Access, PDF forms, WORD forms)

#### 11.4.3.2 Phase 2

1. Collect usage metrics across all USA-NPN sites and tools
2. Finalize web design
3. Finalize core USA-NPN observation database
4. Finalize USA-NPN's phenology metadata database.
5. Implement web services standard in all web tools
6. Enable direct access to multiple data sources

7. Core USA-NPN observation database starts to include animals and other phenological phenomena
8. Make educational materials and tools operational.
9. Implementation of suggestions and usability study results
10. Search engine optimization
11. Develop advanced data contributor tools such as advanced editing capability and data import wizards for importing tab/comma delimited files, etc.
12. Provide Excel and Access templates for people who collect species of local interest. As long as there is no consensus on what phonological measurements are being collected, any integration of data using the "personal preference" protocol with the main USA-NPN phenology observation data will be problematic.

11.4.3.3 Phase 3

1. Develop advanced tools for:
    a. Scientific work flow (Kepler, http://www.kepler-project.org/)
    b. Data mining
    c. Analysis on the web and /or desktop applications
2. More educational content, material, and tools
3. Measure success of information and knowledge dissemination to the scientific community, decision makers, citizen scientists, and the public.