

Review of the USA-NPN's Information Management System

August 2010, Draft

Any use of trade, product, or firm names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Although this report is in the public domain, permission must be secured from the individual copyright owners to reproduce any copyrighted material contained within this report.

Acknowledgements

The USA-NPN's National Coordinating Office is grateful to the individuals who participated in this review (see Table 1). The lessons and expertise they contributed are already making a difference in our selection of technologies, prioritization and policy development. A Nation Science Foundation Research Coordination Network Grant (IOS-0639794) supported this workshop.

DRAFT

Edited by

Paul Allen

Cornell Lab of Ornithology

Suzie Allard

School of Information Sciences
University of Tennessee

Ellen Denny

USA-NPN NCO

Natalie LaTysh

USGS Geospatial Information Office (GIO)

Lee Marsh

USA-NPN NCO

Jeff Morissette

United States Geological Survey (USGS)

Mark Parsons

National Snow and Ice Data Center (NSIDC)

Alyssa Rosemartin

USA-NPN NCO

Inigo San Gil

National Biological Information Infrastructure
(NBII)/ US Long Term Ecological Research
Network (LTER), USA-NPN Board

Bob Tawa

NEON, Inc.

Brian Wee

NEON, Inc.

Bruce Wilson

Oak Ridge National Laboratory (ORNL),
USA-NPN Board

Compiled and written by

Karla LeFevre

Technical Writer
NSIDC

TABLE OF CONTENTS

1. INTRODUCTION.....	5
2. THE USA-NPN DATA MODEL.....	6
2.1 Requirements of Phenology Data.....	7
2.2 Data Reliability.....	8
3. DATABASE QUALITY ASSURANCE & QUALITY CONTROL (QA & QC).....	9
3.1 Strategies for Quality Assurance.....	9
3.2 Strategies for Quality Control.....	10
4. METADATA.....	10
5. HUMAN INTERFACES.....	11
5.1 Overview of Current Human Interfaces.....	11
5.2 Developing an Educated Observer Base.....	12
5.3 Ideas for Encouraging Observer Involvement.....	13
6. APPLICATION INTERFACES.....	13
6.1 Web Services for Output.....	14
6.2 Web Services for Input.....	15
6.3 Recommendations.....	16
7. PHYSICAL INFRASTRUCTURE.....	16
7.1 Current Physical Infrastructure.....	16
7.2 Future Considerations.....	17
7.3 Security Risks.....	18
8. CONCLUSION & RECOMMENDATIONS.....	18
9. BIBLIOGRAPHY.....	20
10. ACRONYMS & ABBREVIATIONS.....	20

1. INTRODUCTION

The USA National Phenology Network (USA-NPN) is a partnership between federal agencies, the academic community, and the general public to establish a national science and monitoring initiative focused on phenology, “the study of the timing of recurrent biological events, the causes of their timing with regard to biotic and abiotic forces, and the interrelation among phases of the same or different species” (Lieth 1974).

USA-NPN is a consortium of individuals and organizations that collect, share, and use phenology data, models, and related information. The Network serves science and society by promoting broad understanding of plant and animal phenology and its relationship with environmental change. Through the USA-NPN program Nature’s Notebook, people of all ages and backgrounds observe and record the activity of organisms through space and time as a means to discover and explore the nature and pace of our dynamic world. The Network makes phenology data, models, and related information freely available to empower scientists, resource managers and the public in decision-making and adaptation in response to variable and changing climates and environments.

The USA-NPN consists of a National Coordinating Office (NCO), a Board of Directors, and many partners, including citizen scientists, resource managers, educators and scientists. Partners represent a range of organizations including public agencies, tribes, non-governmental organizations, specialized networks, and academic institutions.

In July 2010, the USA-NPN hosted a review of its Information Management System (IMS) to explore current and future issues and to ensure that the system is useful, up-to-date, and secure. The Network seeks to work collaboratively and transparently with others in the field, and leverage existing capabilities, such as appropriate open-source software tools. A panel of experts was invited to advise the USA-NPN during the IMS review. Table 1 lists the individuals who participated in the IMS review, their titles, and their affiliations.

Table 1. USA-NPN IMS Review Workshop Participants – 12-13 July 2010, Boulder, Colorado

Name	Title	Organization/Affiliation
Paul Allen	Assistant Director	Information Science (IS) Department Cornell Lab of Ornithology
Jeff Morissette	Invasive Species Science Branch Chief	USGS
Natalie Latysh	Hydrologist	USGS GIO
Suzie Allard	Faculty Member	School of Information Sciences University of Tennessee Knoxville
Bob Tawa	Chief of Computing	NEON, Inc.
Brian Wee	Chief of External Affairs	NEON, Inc.
Bruce Wilson	Systems Engineer/Group Leader Environmental Data Science &	ORNL, USA-NPN Board

	Systems, Environmental Sciences Division	
Inigo San Gil	Senior Application Support Analyst	NBII, LTER, USA-NPN Board
Mark Parsons	Program Manager	NSIDC
Lee Marsh	Application Developer	USA-NPN National Coordinating Office (NCO)
Ellen Denny	Monitoring/Database	USA-NPN NCO
Alyssa Rosemartin	IT & Communications Coordinator	USA-NPN NCO

2. THE USA-NPN DATA MODEL

The primary mandate of the USA-NPN is to serve scientists and society by promoting quality phenological research. To this end, the USA-NPN data model was designed by querying scientists about their needs for phenological data in terms of scale, coverage, species, phenophases, and reliability. The data model also addresses the need for the integration of legacy data, the transferability of the data model, and the integrity of the data. The current data model was developed with these objectives and concerns in mind and is documented in full using the TOAD Data Modeler, located at the following URL:
http://developer.usanpn.org/dev_data_model/MImage.html

Figure 1 provides a simplified version of the data model with the following key elements:

- **Network** – Affiliation for people and/or species (e.g. Great Sunflower Project, Historic Lilac Network)
- **Person** – Observer/User
- **Station** – A particular location where measurements are made
- **Species** – Species of plants and animals to observe, includes taxonomic serial numbers, distributions, and other secondary information
- **Station Species** – Where a *species* of plant or animal has been located
- **Station Species Individual** – Where an *individual* plant has been located, used to track individual plants as measurements are made through time
- **Protocol** – A suite of phenophases and their definitions
- **Phenophase** – A defined life cycle stage (for example, ‘emerging leaves’ or ‘adults in courtship’)

- **Observation** – For an individual plant or species of animal observed in a station by a person, the value for the phenophase status (Yes, No, or Unknown)

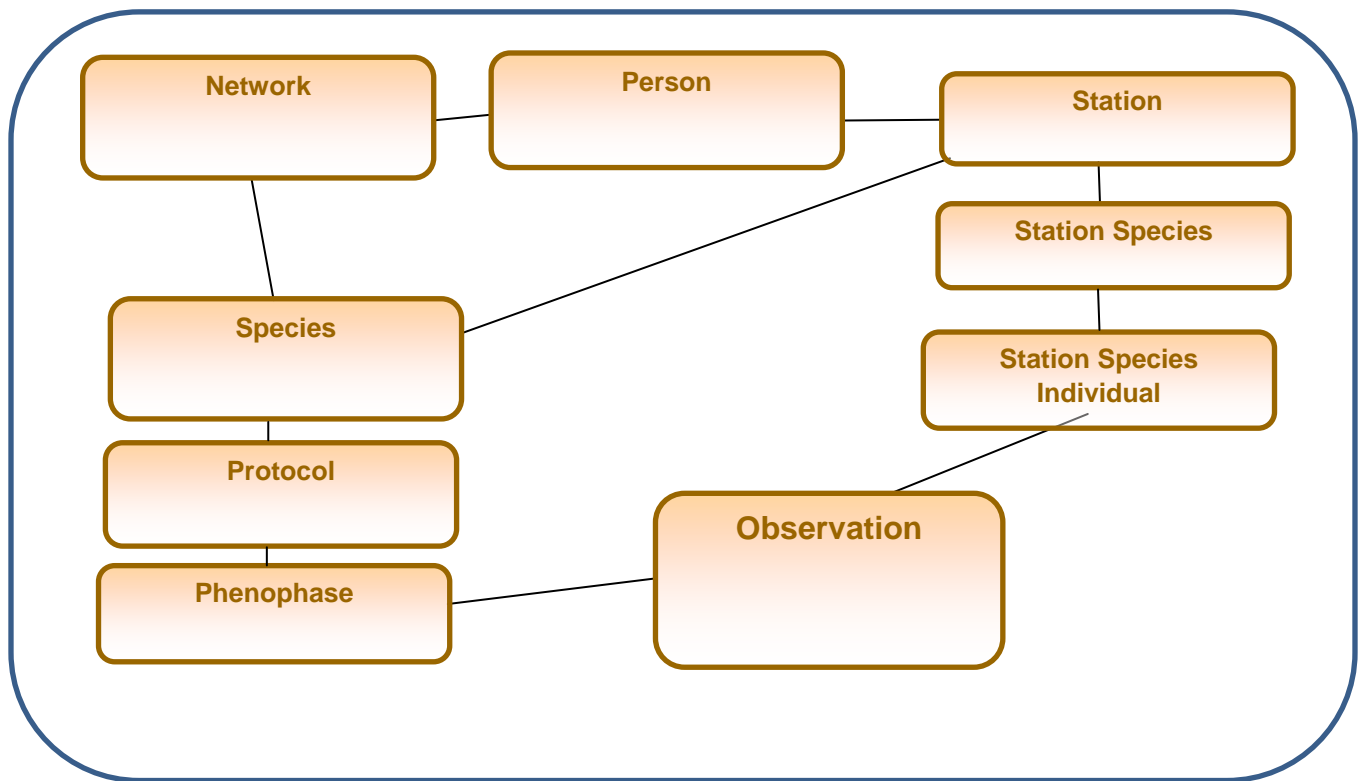


Figure 1. Key Elements of the USA-NPN Data Model

2.1 Requirements of Phenology Data

The USA-NPN envisions storing or facilitating access to three types of phenological data: “integrated” data that will be incorporated within the National Phenology Database (NPDb), “non-integrated” data that will be stored in alternative formats (e.g. Excel worksheets), and “distributed” data that is stored and maintained by other organizations or individuals. The remainder of this section deals with “integrated” data.

Phenophase status monitoring, developed by the USA-NPN, is a new approach for generating research-quality phenology data. A focus on phenophase status monitoring involves implementing a more carefully structured protocol which can yield better phenological research for the long term. Rather than inferring the date a phenological *event* occurred, (‘event monitoring’; the traditional method) observers are guided through a series of simple yes/no questions to help them document an entire phenological *phase*. The simplicity of this model allows for a broad range of observers, from master gardeners to citizen scientists, in the collection of useful phenological data.

Given the broad range of phenological data, the data model needs to be both comprehensive and flexible in order to accommodate the full range of data possible. One example is the possibility of multiple phenological events occurring in one year, as in the case of two leaf emergence events occurring in one season (such as one

before and one after a late frost, or multiple flowerings in rain-driven ecosystems). Another sizeable task is to define phenology in multiple species, or define a flexible protocol (collection of phenophases) that can be applied to multiple species. The protocol for observing an oak, for example, is different from that of a grass. Lastly, the data model must fulfill the Network's unique requirements for phenophase status monitoring.

Data collected using the traditional 'event monitoring' can be integrated with data collected using the new 'status monitoring' method. This is accomplished by calculating a midpoint between the transitions from "No" to "Yes" and "Yes" to "No" in an historical observation record, the date a particular phenological event occurred can be estimated. Such a method ensures that the traditional and contemporary approaches are interoperable and therefore useful for timeseries analyses. This will help scientists answer questions such as how phenological events vary in space and time, and how these events are responding to climate change. The Network aims to eventually incorporate abundance measures (e.g. number of individuals present or percent of canopy filled) within the status monitoring approach to further enhance the quality of phenological data. Abundance observations would add yet another layer of attribute data to the archive (Denny, et al. 2009).

2.2 Data Reliability

The primary data source for the USA-NPN is data submitted by observers throughout the country ranging from high school students to retirees to researchers. The challenge the USA-NPN currently faces in this area is the broad range of observer skill and commitment levels. Ideally, scientists should be able to track and select specific observers, such as master gardeners, or particularly reliable observers as opposed to those who are less reliable. They should also be able to track the revision history of a data set, know when observations were made, distinguish between different observers at the same site, and investigate any ambiguities. The USA-NPN, however, must strike a balance between the level of detail it requires and the level of involvement and commitment it expects from observers. Part of this work would involve educating a base of citizen scientists to ensure that the data collected are comparable and of high-quality. An appropriate model might be the Free-Air CO₂ Enrichment (FACE) experiment, for which the Department of Energy (DOE) has set up control plots with citizen scientists in an attempt to make the data as comparable as possible. A similar experiment could allow USA-NPN citizen scientists to be involved on a deeper level and ensure that the data collected now can be used in 20 years.

An alternative way to satisfy these diverse needs might be through creating a social network for peer review; a review network could provide a means for tracking anomalies and checking for biases, creating interaction among observers, and for verifying observations. Mitigating observer errors could be handled by introducing logic with scripts to implement if/then statements; if observations were to exceed a defined threshold, they would be flagged for review. Offering a variety of data output options, such as visualizations, may also encourage more interest and involvement while at the same time helping verify observations. The eBird project at Cornell University has implemented a volunteer review network and a waiting area where observations are flagged for review. This structure encourages observer involvement on a deeper level, allows eBird to identify and report anomalies, and ensures data integrity.

The USA-NPN uses MySQL as the database management software, has worked well, but it doesn't support geographic data (beyond latitude/longitude point data). Drupal is now compatible with PostgreSQL (though not

all modules may be compatible). In spite of migration costs, the reviewers recommended making the switch to PostgreSQL at least in a development environment to explore migration costs.

3. DATABASE QUALITY ASSURANCE & QUALITY CONTROL (QA & QC)

Database QA for data input and QC for output data is of the utmost importance for the USA-NPN mission of archiving and distributing valuable and usable phenological data. Given the varying types of data the USA-NPN will distribute, such as integrated versus non-integrated data or historical versus contemporary data, many questions still need to be answered to set up appropriate QA/QC measures and avoid potential errors. For example:

- What are the specific base criteria for the USA-NPN objectives?
- When errors do occur, how often does this happen and what is their impact?
- What level of consolidation or federation will the data collections require?
- Should the USA-NPN continue to strive for genome-level of detail and accuracy, or can this be improved?
- How will the USA-NPN transition to better GIS integration and to be able to handle larger data sets?

3.1 Strategies for Quality Assurance

An integral challenge the USA-NPN faces is that of evaluating and quantifying observers' skills. To ensure proper species identification, the USA-NPN has begun educating its observer base with Frequently Asked Question (FAQ) pages and *How To* videos. An additional step in this direction the USA-NPN could take is adding an assessment component. Since a broad range of errors can potentially be introduced in a data record, from misidentifying a species to transcription errors, this assessment needs to be broad as well. For example, some potential errors are that the observers are not field trained and need extra assistance in areas where there is room for interpretation, such as how a phenophase is defined.

In addition to establishing an assessment program, the USA-NPN could incorporate measures through usability testing that ensure proper data entry and help reduce transcription errors. Requiring observers to submit images of a particular species in a standardized format could also help with QC. Technology infusion might be another way to help observers identify, collect, and even submit their data from the field, such as the development of USA-NPN iPhone applications. Image size, credit, inappropriate content and malicious content pose barriers to the ability of the USA-NPN to photo-validate observations. Using a commercial photo sharing site may be an interim solution. It was also suggested that the Network consider putting a color swatch on our datasheets to enable observers to take a picture of their plant with a datasheet in the background to help calibrate the OCR software. An additional QA/QC measure might also include a ranking system that rates an observer's skill level and reliability. With a skill-level designation, the data could then be tagged as having come from a skilled or non-skilled observer in a database that is searchable by scientists. Some reviewers thought observers would reliably self-evaluate on skill level, others thought that there would be bias, with those who know the most being more likely to recognize the limits of their knowledge. Finally, establishing an observer certification program or a gaming opportunity wherein observers are required to pass a quiz before submitting data could be a way to pique interest and involvement while satisfying QA objectives. Examples include master

gardener certification programs or websites such as GalaxyZoo, which allows citizen astronomers to submit and classify images of galaxies once they have passed a short trial classification test.

3.2 Strategies for Quality Control

Strategies for quality control of data that are collected should include both verification and validation. Adding functionality within the database that repeats an observer's data entry items back to him or her would be a way to verify and double-check the entry, as well as ferret out transcription errors. Setting QC flags for illogical phenophases (such as a user entering both "blooming" and "dead") could provide another layer of verification. Additionally, requiring observers to upload images on the species profile page for all data records would allow the USA-NPN and/or scientists to validate the entries, either now or in the future. The Encyclopedia of Life website currently collects images submitted by the public, and this may be a promising partnership for the USA-NPN. The final layer of database QC measures would be evaluating data quality. For example, comparing data for a particular species to its known range (in a preliminary test of this issue 3.7% of 4857 species were registered outside of states in their known range) and comparing data for the same species at neighboring sites would indicate whether USA-NPN staff and/or scientists might want to dig further to assess particular data sets or data providers.

4. METADATA

Regarding data policy, data discovery strategies, and metadata standards, another delicate balance must be achieved: allowing for open data access and sharing while ensuring sufficient privacy and protection for all parties. The current data set registry tool (for historic and/or non-standard data sets) is based on a Dublin Core metadata standard and will be developed to allow export of metadata records as EML. The USA-NPN also plans to develop FGDC-compliant metadata for the four data sets listed in section 6.1. Some reviewers recommended ISO 19115 as the preferred and international standard. Most agreed that choosing a mature and interoperable standard that the USA-NPN could most easily implement would be best, and these records could then be integrated with ISO 19115.

The current data output is a flat file of phenology observations at "fuzzed" locations to protect observer privacy. The USA-NPN seeks to ensure privacy protections for observers while also not constricting the flow of data. For example, is exposing the latitude and longitude coordinates of a site ideal or is it a violation of privacy? From a science perspective, some data need to reflect their precise location, particularly if they are niche models and microclimates. This may also be important in terms of aggregating the data for the validation of remote sensing data. The USA-NPN currently rounds site locations to 0.1 degrees latitude and longitude (approximately 7 miles). However, the National Park Service requires higher spatial resolution. The challenge is how to satisfy the requirements of science while also protecting observers.

Clearly, if the mission is open sharing of data for conducting science, the data policy needs to be as easy and open as possible. Thus, the recommendation to the USA-NPN during the IMS review is to: 1) define what it means to be open, 2) state the privacy risks to observers as clearly as possible, 3) consider requiring users to opt-in by signing a confidentiality agreement, and 4) implement at least an optional registration process (which could be combined with tangible benefits for registering, such as newsletters and notification e-mails). The USA-NPN also needs to ensure that care is taken with regards to data about threatened and endangered species or

areas. Currently the Network does not collect data on threatened and endangered species. Even with these processes in place, the ability to enforce the data policy with minimal funding may be problematic.

As recommended during the IMS review, implementing Google Fusion tables would protect both the data and servers, and is a free option for non-profits. This would help the USA-NPN turn its attention and resources toward supporting metadata for harvesters and distributing data in similar content communities like the National Climatic Data Center (NCDC) or the Avian Knowledge Network (AKN). The USA-NPN is also in a unique position to set standards for phenology data while benefitting from the experience and guidance of the Science Commons and Committee on Data for Science and Technology (CODATA). A notable example is the Polar Information Commons (PIC) that was formed with the purpose of archiving and protecting data collected during each International Polar Year (IPY). The Commons has developed a PIC Rights Badging Tool that is designed to allow for open and ethical data sharing and provenance. The USA-NPN could implement a similar tool to help encourage ethical norms for sharing phenological data. Finally, the USA-NPN needs to research and adopt the most flexible and international metadata standards available, such as those of the International Organization for Standardization (ISO). For example, a flexible metadata output format such as a Comma-Separated Values (CSV) file with embedded metadata might be an appropriate format for USA-NPN data.

5. HUMAN INTERFACES

5.1 Overview of Current Human Interfaces

The USA-NPN's Drupal 6.19 site (www.usanpn.org) is an information rich website that includes the following features:

- Data Set Registry Tool: <http://www.usanpn.org/participate/dataset>*
- Data Set Discovery: <http://www.usanpn.org/results>
 - Mercury Search Tool: <http://mercury.ornl.gov/usanpn>
 - Registered Datasets: <http://www.usanpn.org/results/dataset-list>
- Preliminary Data Output: <http://www.usanpn.org/results/maps>
- Educator's Clearinghouse: <http://www.usanpn.org/education/clearinghouse>
- Phenology Festivals: <http://www.usanpn.org/resources/festivals>*
- Species Information
 - Search Species: http://www.usanpn.org/species_search
 - Example Profile: http://www.usanpn.org/Carduelis_tristis
- Bibliography: <http://www.usanpn.org/results/biblio>*
- User Creation: <http://www.usanpn.org/user/register>

* Allows user-generated content.

A second, java-based website, Nature's Notebook <http://mynpn.usanpn.org/npnapps/>, shares login sessions with the Drupal website, and allows for observers to submit phenology observations into the NPDb. An overview of the process for participating as an observer is available at <http://www.usanpn.org/participate/guidelines>.

5.2 Developing an Educated Observer Base

In addition to the primary objective of collecting and distributing quality phenological data, the human interfaces developed as part of the USA-NPN IMS must also help grow a stable, educated observer base and encourage more involvement among those observers. The university Extension network may be a key resource for the USA-NPN in terms of reaching out to observers on the ground with support on species and phenophase identification and general training.

The Education section of the USA-NPN website has not been regularly updated with content. An intermediate plan of the USA-NPN is to have education/outreach staff to develop teacher training workshops and modules along with other helpful resources for citizen scientists, such as calibration kits, a phenology handbook, and materials for recruiting others as citizen scientists.

In the meantime, the USA-NPN should continue to focus on targeted usability testing to assess the efficacy of existing user interfaces. Improvements could include simple revisions, such as shading the columns in data sheets so that users can easily transition from analog to digital, from field data to data entry. Other relatively simple improvements could include developing controlled vocabularies for better data search-ability, adding new entry points to the website, adding audio-visual hooks to appeal to a wider range of observers with different learning styles, and adding a bug reporter for each interface so users can easily report interface problems. Currently the USA-NPN has targeted a middle-level user profile (backyard naturalist with some computer familiarity), and plans to develop several user profiles to allow the complexity of the interface to vary. Reviewers recommend prioritizing a lower, easier to use interface first. They also cautioned against assuming that developers could divine how users would want their interfaces and that the interface should be customizable.

Intermediate improvements could include revising the Nature's Notebook data entry interface to reduce the need for sessions, and adding the ability for a user to refresh the site when entering observations. The data set registry tool could be improved if locations could be submitted as polygons rather than just as points. The data sets are then dynamically linked to related publications so that scientists and backyard naturalists can easily delve further in a subject area. The website's distributed data search functions (data set registry tool and Mercury search engine) should be combined at some point since it is cumbersome for user looking for data to explore both.

More involved, immediate improvements could include revisions to the Nature's Notebook website and interfaces. Observers can currently switch between stations, access a species profile, and view data sheets. However, revising the Add a Site interface could include giving site users options to correct mistakes and/or play around while learning and testing the interface. This "sandbox" concept would allow the user full control of their data records for a defined period of time, such as 30 to 60 days. At the end of the time period, users could submit a complete data record with an "I certify" button. Currently observers can change their Y/N/? response at any time and the history of this revision is not captured, only the latest revision is stored. It was also recommended that the Nature's Notebook user interface be simplified.

An additional issue was brought up for consideration in terms of the interpretation in the database of what the observers enters into the interface – the difference between a "?" (null) submitted and no record. Currently the system treats no submission as a null in the database (equivalent to a "?"). The USA-NPN developers now plan to change the system to create no record when the entry is left blank and to continue to enter nulls for

submissions of “?”s, or unknown fields.

5.3 Ideas for Encouraging Observer Involvement

Long-term improvements could include designing new interfaces with greater functionality for users and developing interactive programs to support a range of activities, such as cell phone applications. For example, the Center for Embedded Networked Sensing (CENS), a UCLA urban sensing network website, has developed a cell phone application to make it easy for anyone to upload images to their site. Applications could also be used to notify users of an upcoming phenological event in their area, such as a flowering season application for hikers. Such applications could also be an ideal way to encourage volunteerism, build a sense of community among observers, and offer instant gratification, particularly with applications for Facebook and other social networking sites. Google maps could be incorporated as well to dynamically serve seasonal phenological information to users about the state they live in. Even in the middle of winter, participants could be prompted to answer a question such as, “Which phenophases do you currently see?” via a Google map application. And even without a current database for the phenophases by state, participants could still see examples of what other observers are seeing and reporting.

Additionally, sustaining user involvement requires hooking them with a reward or finding other tactics that go beyond instant gratification. A possible reward or recognition opportunity could be to assign Digital Object Identifiers (DOIs) to each data set an observer submits, and then notify them when a scientist publishes a journal article using their data. The USA-NPN currently plans to credit all participants who wish to be listed as contributors to Nature’s Notebook, following the Galaxy Zoo model. Observers also likely want to get a snapshot or summary of their results as soon as possible, especially after entering a large amount of data. Ideally, a summary of their observations would include a snapshot of historic trends for a particular phenophase or area for comparison purposes. An iGoogle-type landing page could allow users to personalize their own phenology page to highlight phenology festivals, related publications, browse through all available applications, and plug into other USA-NPN offerings, such as RSS feeds.

Phenology festivals, in particular, are an opportunity to cultivate a social network around phenology. The USA-NPN has already developed a Google map highlighting phenology festivals around the world. Developing additional functionality around this feature could include allowing users to submit content regarding their local festivals, adding the ability to notify registered USA-NPN observers of upcoming festivals, launching an on-the-ground volunteer effort during a festival, and recruiting additional observers. Plant festivals, in particular, could even provide an opportunity for validating data in conjunction with content capturers like Google books and applications that can check a known image against an unknown image, as with Optical Character Recognition (OCR). Lastly, Cornell’s Yard Map program may be a nice example of a way to engage observers, giving them an online space to describe their local environment.

6. APPLICATION INTERFACES

A current area in need of development for the USA-NPN is the ability to provide quality and flexible data input and output options, including visualizations such as timeseries maps and scatter plots, as well as a variety of application interfaces.

6.1 Web Services for Output

Web services for output will be developed to support visualizations and dynamic data harvest. The primary user audience for data visualizations are program participants, for recruitment and retention purposes, followed by visualizations to engage the general public and visualizations to demonstrate potential data uses to scientists and managers. The first round of data output web services are being built to meet these needs. The Network expects some scientists or institutions to seek full data sets delivered via web service (Data Basin and NBII are potential examples), but as yet no partner has requested data set output via web service.

Though not always appropriate for scientific analysis, visualizations are a great way to draw interest to the site and to relate network observations such as maple leaf bud burst to temperature, landscape-level phenology and USGS habitat modeling. Comparing on the ground observations to remote sensed phenology data is an important opportunity to calibrate the remote sensed imagery. They could also be helpful during the QA/QC process in checking for anomalies. Visualizations could allow users to filter records and specify parameters for data output. For instance, site visitors could explore interesting stories regarding seasonal temperatures or the extent of urbanization in an area, stories that lend more context and relevance to the data for them. The current USA-NPN Drupal system might also offer some of this capability, such as for RSS feeds, through the Feeds module. The Google Visualization API might also be advantageous and is easy to use.

As another potential example is a flowering map that newspapers could draw feeds from would help advertise the USA-NPN to would-be citizen scientists and also display more dynamic time/space elements of the phenology data, thus making the data more locally relevant. The National Geographic FieldScope project is a good example of a site that offers a web-based mapping, analysis, and collaboration tool to support geographic observations and engage citizen scientists to investigate real-world issues. Additionally, FieldScope is maintained by a relatively small staff.

An important question the USA-NPN needs to answer with regards to output options is how to define a data set, whether by collection level or site level. Currently, data collected using the same protocol is defined as one data set. This system results in the following four data sets available from the Network:

- 01/01/1956-03/01/2009 – Lilac data (Event method; existing FGDC metadata through 2003)
- 01/01/2008- 03/01/2009 – Native plants (Status/Event hybrid method)
- 03/01/2009-Present - Lilacs + Native Plants (Status method)
- 03/27/2010-Present - Animals (Status method)

However, does this match what a user might expect in one data set? It is possible to request customized queries of data from the USA-NPN.

Additionally, how should data sets or collections be packaged for the National Biological Information Infrastructure (NBII)? Protocols vary within the database, making it complicated to provide a global set. Thus, versioning would likely be an ideal way to distinguish data sets or collections; when a change is implemented, such as a change in protocol or the addition of a new phenophase, a new version of the data set would be released. Versioning is also an important step to ensure provenance for data collections. An Open Archival Information System (OAIS) could bundle the data with any and all the supporting information to be used by the

community, and an OAIS reference model would provide a framework for describing and comparing appropriate long-term preservation strategies and techniques.

The Network should explore becoming a member node in the DataONE project; the project could harvest metadata, offer free services to use with data, and provide EML output. Extensible Markup Language (XML) is perhaps the easiest solution for making data more harvestable. Whatever the language, the USA-NPN needs a protocol for both data input and output that many tools understand. There are of course other networks that the USA-NPN could plug into and collaborate with, such as the Global Biodiversity Information Facility (GBIF), DataNet, Extensible Observation Ontology (OBOE), and the Scientific Observations Network (SONet), to name a few. The Open Geospatial Consortium has an observation schema called Sensor Web, which might be a promising partnership for determining a common schema.

In terms of web mapping, a Web Mapping Service (WMS) that can update on a regular basis might be the best solution. With an image-based WMS, it is easier to display visualization data. Or, with a Web Feature Service (WFS), site visitors would still be able to click on a point and get data, as with Gmaps. And though the current GMap views in Drupal (for example, used in the phenology festivals map <http://www.usanpn.org/resources/festivals>), are approaching the limit as to how many points can be displayed, there is still a great deal of flexibility and functionality the USA-NPN could capitalize on with the GMap module. For example, the Network could begin using Keyhole Markup Language (KML) files in Google Earth to display data animations.

Currently USA-NPN is building web services for data output in both the REST and SOAP frameworks. In addition, the Network plans to have a flat file available for download on the website. How should data downloads be tracked? Registration may turn people away, and should be optional. It is effective to provide benefits (data updates, newsletters) to motivate registration. This issue is complicated when data is available through other venues (e.g. NCDC or Google Fusion Tables).

6.2 Web Services for Input

The primary use-cases for data input web services are organizations who would like to encourage their members to participate in Nature's Notebook, but do not want their members to leave their site. Examples are YourGardenShow.com and The Great Sunflower Project; each could potentially contribute 1,000s of observers to the program. Other potential collaborators might include Dave's Garden, Encyclopedia of Life, eNature, or the NBII Did You Know project. With such functionality in mind, the USA-NPN is developing an API for data input to the National Phenology Database.

The same web services being developed for the above-listed partners can also serve as the basis for mobile and Facebook data entry applications, which the Network can use to recruit younger participants, reduce barriers to participation and reduce datasheet transcription errors. CyberTracker provides software for hand-held devices which can be synchronized with a database at home. Perhaps a similar mobile observation Personal Digital Assistant (PDA) tool, or an iPhone field guide application, could be used for phenology data.

As another approach, the eBird website provides an intermediate API that reflects the look and feel of the partner organization, yet is a flexible web application that doesn't require the support of programmers and

other technical resources. Overlapping Facebook applications could allow users to submit data to the USA-NPN from other partner sites without leaving those sites.

Currently, USA-NPN is developing input web services in a SOAP/WSDL framework, running on CakePHP. SOAP was chosen because it has been an industry standard that provides real-time specification of functionality. However, the consensus during the IMS review is that this is too heavy for the long-term.

One possible solution for future development is REST. REST-based Web services are straight-forward, easier to program, and could provide adjacent output. REST could also be relatively easier to program; the bulk of the work would be in defining the parameters and writing the program. Alternatively, generalized Digital Access Object (DAO) layers could be developed. Java classes would then map directly to a table in a database that controllers manipulate, and the DAO layers could handle queries. All input could go through a single java class so system administrators could intercept it and repair it transparently. However, though this may be suitable for input, DAO does not output well and would take longer to develop. It is most effective for creating views for projects. It may not be the most efficient architecture for the USA-NPN.

6.3 Recommendations

Reviewers recommended that the USA-NPN focus on web services for data output as a priority over services for data input, as they saw a greater need to get data served in various communities. In addition, output applications are typically easier to create and consume than input applications.

Note that the collaborations described in this section offer short-term options for the USA-NPN to add some services now, and then develop more services as it has the funding and staff to support additional programming development.

7. PHYSICAL INFRASTRUCTURE

7.1 Current Physical Infrastructure

The USA-NPN hardware infrastructure currently consists of two rack-mounted IBM servers (each with dual core, dual processor, 2.66 GHz, 24 GB RAM, ~680 GB of RAID 5 storage) purchased by the USGS in late 2007. The servers run VMWare Infrastructure Standard as the base operating system and currently support five Virtual Machines (VMs).

The virtual machines are running on an Ubuntu Server, with a typical Apache, MySQL, and PHP stack. The usanpn.org primary server (www.usanpn.org) also runs Drupal 6. The mynpn.usanpn.org server runs a Java application through the Tomcat servlet engine.

The servers are physically located in a conditioned and secured server room, run by the University of Arizona's University Information Technology Services (UITS). The University of Arizona currently provides a 32-node Virtual Local Area Network (VLAN) for USA-NPN with 27 usable addresses. The USA-NPN maintains an account with the external domain registrar, GoDaddy, for Domain Name System (DNS) services and for e-mail forwarding.

USA-NPN is currently making use of spare disk space at ORNL procured for the NBII Metadata Clearinghouse as a location for daily off-site backups of the database contents and for a Subversion (SVN) code repository.

The physical and virtual machine configuration for the USA-NPN servers is given in Table 2. The processor usage has barely been utilized, and is only averaging three to five percent every 24 hours. The Input/Output (I/O) between processing systems, however, is more limiting. Due to limited resources, the USA-NPN originally focused on free, open source software and thus opted for the current system with Ubuntu. The SVN sever at ORNL has a mirrored SVN locally, though there are some problems with network timeouts at ORNL. Table 2 lists the details of the USA-NPN Physical Machine (PM) and Virtual Machine (VM) configuration.

Table 2. USA-NPN Physical Machine (PM) and Virtual Machine (VM) Configuration

PM1	VM1	Ubuntu 9.04 Apache, MySQL, and PHP stack.	Production Drupal Web Server
PM1	VM2	Ubuntu 9.04 Apache, MySQL, and PHP stack.	Development Drupal Web Server
PM2	VM3	Ubuntu 9.04, Tomcat, Apache, MySQL, and PHP stack.	Production Java Nature's Notebook Server
PM2	VM4	Ubuntu 9.04 Tomcat, Apache, MySQL, and PHP stack.	Development Nature's Notebook Web Server

7.2 Future Considerations

Future upgrade considerations center around the question of what constitutes an acceptable outage in the context of the data entry (Nature's Notebook) and usanpn.org websites and response time. Nature's Notebook can be down overnight, and different services have different threshold times, with spring as the heaviest data entry time. The USA-NPN has a disaster recovery plan with ORNL as the offsite back-up, and could strengthen this by replicating servers, but there is the potential to lose up to 24 hours of data entry. In addition, a March 2009 National Public Radio (NPR) Science Friday interview featuring USA-NPN brought over 2,000 simultaneous visitors to the usanpn.org website, crashing the site. Thus, the USA-NPN also needs to consider implementing a response plan as well as bolstering the current disaster recovery plan. A possible solution for downtimes would be to develop a small site that is a derivative of the Drupal site to give some basic information to site visitors. Additionally, mass storage shared between systems would be ideal. Switching to a cloud system might also be advantageous, and Throttle may help mitigate the recovery time. It is recommended that USA-NPN back up code to Google Codes but the code would need work to be cleaned and made stable and standard.

Another recommendation is that the USA-NPN replace its servers when the warranty gets costly, or approximately every four years. This means the first phase of upgrades should occur in one to two years. In the meantime, server load testing software could be helpful to test across VMs. Alternatively, Apache JMeter, an open source software package, could randomly spider the site with n number of clients for all USA-NPN functions. JMeter software works by simulating hundreds of clients hitting the site simultaneously. The USA-NPN should also work to increase the cache lifetime in Drupal. Adding a caching layer, such as Squid, in front of the

Drupal server application would improve Drupal performance. With the exception of the dynamic pages, this set-up would enable Drupal to offload the caching server.

Of course, the USA-NPN will need to make these replacements and upgrades on a limited budget. When one machine fails, the USA-NPN could move all VMs to one machine. In reality, it likely requires more staff time to anticipate and prepare for all possible compromised situations than it does to simply replace the machines themselves. Even with open source software, organizations commonly purchase support contracts, as this can offer developers back-up support instead of requiring them to spend their time scouring forums when something goes wrong.

7.3 Security Risks

Finally, security risks are of course another area of concern. The USA-NPN runs back-ups with Secure Shell (SSH) certificates; however, the concern is that the application accounts for authentication, but is missing a confidentiality component. Login is currently not over a Secure Socket Layer (SSL) and the USA-NPN intends to fix this. Drupal has good security and notifies Drupal developers of periodic core and module security updates. The USA-NPN is monitoring attacks and has found that they get hacked first in other sites. They are also running scans to detect any Apache misconfigurations. Moving in the Facebook direction might be advantageous and support open ID, which is also in Drupal. The Java application uses Drupal authentication, and that would allow users to log into the USA-NPN system via Facebook.

The original choice to pair Drupal and Java was more out of convenience and necessity than security. Using PHP for the whole system and running it on the same architecture might be a future consideration. As Drupal and the USA-NPN evolve the continued compatibility should be assessed. The eBird project currently uses Plone, as this was available when the project started, but Drupal is a stronger CMS. Lastly, Linux has a core module that is good for managing XML native data, but this direction would depend on the development team.

8. CONCLUSIONS AND RECOMMENDATIONS

In conclusion, the consensus of panel experts during the IMS Review Workshop was that the USA-NPN has already accomplished a great deal with its current IMS. Recommendations for changes centered primarily around how to ensure of open data sharing and provenance, how to set standards for archiving and distributing phenological data, how to verify and validate data submitted by observers, how to increase user involvement, and how to make sure USA-NPN systems are scalable with increased user involvement and growing data storage requirements.

On the basis of the deliberations during the USA-NPN IMS Review Workshop, the following steps were recommended to improve the USA-NPN IMS:

- Define the specific base criteria for USA-NPN objectives and use these to inform all revisions and new developments.
- Add a comprehensive data review process, including several layers of QA & QC checks and a volunteer review network.
- Address provenance and federation for data sets for the purpose of long-term archive preservation.

- Research and adopt the most flexible metadata standards available and use these to set standards for phenology field.
- Define controlled vocabularies to make data more searchable and harvestable.
- Educate a base of citizen scientists to ensure that the data collected are comparable.
- Add an assessment component to establish requirements for citizen scientists' observations; require them to pass a short trial classification test and rate the skill level of observers.
- Adopt an open data access policy while encouraging ethical sharing of data (such as with the PIC Rights Badging Tool), including the following steps:
 1. Define what it means to be open,
 2. State the privacy risks to observers as clearly as possible
 3. Consider requiring observers to opt-in by signing a confidentiality agreement, and
 4. Implement an optional registration system for observers
- Implement Google Fusion tables to protect both the data and servers.
- Focus on data output options now, and data input improvements in the future, for web service development.
- Continue targeted usability testing to assess the efficacy of existing human interfaces.
- Allow user control for correcting mistakes and/or playing around while learning how to use the interface.
- Conduct use-case scenarios to identify the real questions data end users might ask.
- Encourage user involvement through a range of personalized tools and features, such as cell phone applications, an iGoogle-type landing page, and social networking applications.
- Provide rewards and incentives for participating observers, whether by using DOIs, implementing a classification certification program, or by adding a variety of interesting audio-visual output options.
- Incorporate visualizations to allow users to filter records and specify parameters for data output; use visualizations to help users and scientists verify and validate species identification.
- Develop a protocol for data input and output that many tools understand to facilitate open data sharing, such as Extensible Markup Language (XML).
- Continue using Google mapping applications as long as possible; add KML animations to display the data more dynamically.
- Adjust the USA-NPN data model so it closely mirrors that of potential partner organizations.
- Determine how a data set is defined and set a versioning process for data sets. Include appropriate documentation for different data set versions and consider using an OAI reference model.
- Consider eventually transitioning to REST-based web services to provide fast and reliable geographic data.
- Define an acceptable outage for the online interface Nature's Notebook and develop a response plan for downtimes; develop a small site that is a derivative of the Drupal site and work towards mass storage shared between systems.
- Replace USA-NPN servers when the warranty gets costly, in approximately 1-2 years.
- Conduct server load testing across VMs and/or randomly spider the site with client replication/simulation software, such as Apache JMeter.
- Add a caching layer, such as Squid, in front of the Drupal server application to increase its cache lifetime.
- Consider purchasing support contracts to save programming resources for development rather than maintenance.
- Add login over SSL to improve security.
- Consider transitioning to using PHP for the whole system and running it on the same architecture (entails rebuild of Nature's Notebook, a Java application).

- As Drupal modules evolve, assess the data model for scalability and transition to a new CMS if necessary.
- Create a Wikipedia page on the USA-NPN as a quick way to have name and messages accessible.

9. BIBLIOGRAPHY

Denny, Ellen, Miller-Rushing, Abraham, Haggerty, Brian, Benton, Lisa, Crimmins, Theresa, Losleben, Mark, Richardson, Andrew, Rosemartin, Alyssa, Schwartz, Mark, Thomas, Kathryn, Weltzin, Jake, and Wilson, Bruce. 2009. A new approach to generating research-quality data through citizen science: The USA National Phenology Monitoring System. Available from Nature Precedings <<http://dx.doi.org/10.1038/npre.2009.3695.1>>.

Duerr, R., Parsons, M. A., and Weaver, R. 2009 (in press). A New Approach to Preservation Metadata for Scientific Data - A Real World Example. in Di, L. and Ramapriyan, H. K. (eds.) Standards-Based Data and Information Systems for Earth Observations. Springer-Verlag.

Lieth, H. 1974. Purposes of a phenology book. Pp 3-19. In: (H. Lieth, ed.) Phenology and Seasonality Modeling. New York: Springer-Verlag. 444 pp.

Denny, E., Marsh, L., Rosemartin, A., and Wilson, B. 2010. The USA National Phenology Network (USA-NPN) Information Management System (IMS) Review Workshop, Boulder, Colorado USA.

10. ACRONYMS AND ABBREVIATIONS

Table 3 lists the acronyms and abbreviations used in this document.

Table 3. Acronyms and Abbreviations

AKN	Avian Knowledge Network
API	Application Programming Interface
CC0	Creative Commons Zero
CENS	Center for Embedded Networked Sensing
CMS	Content Management System
CODATA	Committee on Data for Science and Technology
CSV	Comma-Separated Values file
DAO	Data Access Object
DNS	Domain Name System

DODS	Distributed Oceanographic Data Systems
DOE	Department of Energy
DOI	Digital Object Identifier
EML	Ecological Metadata Language
ESA	Ecological Society of America
FACE	Free-Air CO2 Enrichment
FGDC	Federal Geographic Data Committee
GIO	Geospatial Information Office
GBIF	Global Biodiversity Information Facility
GMap	Google Mapping application
IMS	Information Management System
IBM	International Business Machine corporation
IP	Internet Protocol
ISO	International Organization for Standardization
KVM	Kernel-based Virtual Machine
KML	Keyhole Markup Language
LTER	US Long Term Ecological Research Network
NBII	National Biological Information Infrastructure
NCDC	National Climatic Data Center
NCO	National Coordinating Office of the USA National Phenology Network
NEON	National Ecological Observatory Network
NSF	National Science Foundation
NPDb	National Phenological Database
NPR	National Public Radio
OBOE	Extensible Observation Ontology
ORNL	Oak Ridge National Laboratory

OPeNDAP	Open-source Project for a Network Data Access Protocol
OCR	Optical Character Recognition
PDA	Personal Digital Assistant
PHP	Hypertext Preprocessor scripting language
PIC	Polar Information Commons
PM	Physical Machine
RCN	Research Coordination Network
REST	Representational State Transfer
RFP	Request for Proposal
SSH	Secure Shell
SSL	Secure Socket Layer
SOAP	Simple Object Access Protocol
SONet	Scientific Observations Network
SS	Secure Server
SVN	Subversion Server
UCLA	University of California, Los Angeles
USA-NPN	United States of America National Phenology Network
USGS	United States Geological Survey
VLAN	Virtual Local Area Network
VM	Virtual Machine
WFS	Web Feature Service
WMS	Web Map Service
WSDL	Web Service Definition Language
XML	Extensible Markup Language