

## Review of the USA-NPN's Information Management System

September 2011

**USA National Phenology Network**

## **Review of the USA-NPN's Information Management System**

Suggested citation: USA-NPN National Coordinating Office. 2011. Review of the USA-NPN's Information Management System. USA-NPN Programmatic Series 2011-004. [www.usanpn.org](http://www.usanpn.org).

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.

## TABLE OF CONTENTS

---

INTRODUCTION .....	4
DATA MODEL.....	4
PHENOLOGY DATA REQUIREMENTS .....	5
DATA OUTPUT .....	6
PRIVACY, DATA SHARING, AND DATA ATTRIBUTION POLICIES .....	7
DATABASE SOFTWARE .....	8
DATA QUALITY ASSURANCE & QUALITY CONTROL.....	8
METADATA.....	13
HUMAN INTERFACES.....	13
CURRENT INFRASTRUCTURE .....	14
ENHANCEMENTS TO EXISTING INTERFACES.....	14
SUGGESTIONS FOR RECRUITMENT AND RETENTION OF OBSERVERS.....	15
TOOLS .....	16
APPLICATION INTERFACES .....	17
WEB SERVICES FOR OUTPUT.....	17
WEB SERVICES FOR INPUT .....	17
TOOLS AND FRAMEWORKS FOR WEB SERVICE DEVELOPMENT .....	17
PHYSICAL INFRASTRUCTURE.....	18
CURRENT PHYSICAL INFRASTRUCTURE.....	18
FUTURE CONSIDERATIONS .....	19
SECURITY RISKS .....	20
CONCLUSIONS AND RECOMMENDATIONS.....	20
BIBLIOGRAPHY.....	22
CONTRIBUTIONS & ACKNOWLEDGMENTS.....	23
APPENDIX A – WORKSHOP PARTICIPANTS.....	24
APPENDIX B – ACRONYMS & ABBREVIATIONS.....	25

## INTRODUCTION

---

The USA National Phenology Network (USA-NPN, Network) 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 plant and animal life cycle events, such as the emergence of leaves and the migration of butterflies.

The 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 a 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 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 adapting to variable and changing climates and environments.

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

On July 12 and 13, 2010, in Boulder, Colorado, the USA-NPN hosted a review of its Information Management System (IMS) to explore current and future issues in information technology and ensure that the IMS is useful, up-to-date, and secure. The Network seeks to work collaboratively and transparently with other organizations 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 NCO during the IMS review (Appendix I lists participants).

## DATA MODEL

---

A primary mandate of the USA-NPN NCO is to provide high-quality data in support of phenological research. To this end, a 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 transportability and integrity of the data. Addressing these objectives, the current data model has been documented using the TOAD Data Modeler ([http://developer.usanpn.org/data\\_model/MImage.html](http://developer.usanpn.org/data_model/MImage.html)).

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

**Person** – Observer/User

**Station** – Location where measurements are made

**Species at Station** – Where a species of plant or animal has been located (for animals, this record is at the species level; for plants, it is at the individual organism level, as plants can be more readily marked and tracked through time)

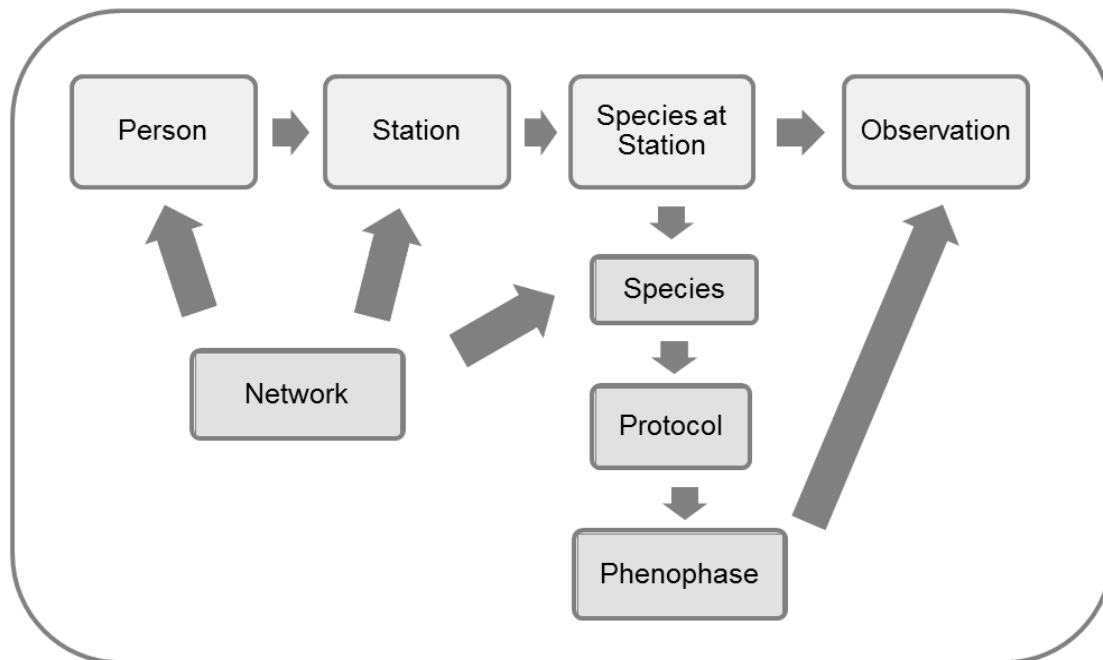
**Species** – Plant and animal species to observe; includes taxonomic serial numbers, distributions, and other secondary information

**Network** – Partner affiliation for people, species, and stations

**Protocol** – A suite of phenophases and their definitions

**Phenophase** – A defined life cycle stage (for example, emerging leaves or adult animals in courtship)

**Observation** – For an individual plant or species of animal, observed at a station by a person, the value for the phenophase status (Yes/No/Uncertain), plus the 2011 addition of abundance (How many animals?) and intensity (What percent of leaf out?)



**Figure 1.** Observation Table: Key Entities and Relationships in the USA-NPN Data Model

### Phenology Data Requirements

Phenophase status monitoring, developed by the USA-NPN in 2008, is an innovative approach for generating research-quality phenology data. Unlike phenological event monitoring, which requires observers to infer the date an *event* occurred, status monitoring guides observers through a series of simple yes/no questions to help them document an entire phenological *phase* (phenophase). The simplicity of this model allows for a broad range of observers, from high school students to professional scientists, in the collection of useful phenological data (see Thomas et al 2010 for more information).

Within the status monitoring framework, the core unit of data is the “observation record” (located in the Observation Table, Figure 1). An observation record is an observer’s Yes/No/Uncertain (Y/N/?) response to the

question “Did you see or hear this phenophase for this species on this date at this site?” The data model was designed to allow a person to create stations, and to create species at those stations. Species are assigned protocols (suites of phenophases) based on phenological functional groups. Both in these methods and in the interfaces developed, the USA-NPN must strike a balance between the level of detail it requires of its participants and the level of involvement and commitment it can realistically expect from citizen science observers.

Partnerships play a key role in the USA-NPN. Species can be assigned to partner groups (networks), so that members can easily locate their group’s focal species on the website. Members of partner groups themselves are identified in the data model to facilitate reporting on the contributions of the group, as well as to provide groups with their data for education and decision-making purposes. In this way, the USA-NPN serves the data-management needs of several partner groups. A further enhancement (completed in June 2011) relies on the relationship between stations and networks to allow more than one person to access a site, based on their network membership.

The USA-NPN data model also needs to be comprehensive and flexible enough to accommodate alternate data structures. For example, data collected using the traditional phenological event monitoring method can be integrated with data collected using the new phenophase status monitoring method (see Denny et al 2009 for details). As a proof of concept, data collected through the historic lilac network has been integrated into the USA-NPN database. Another common approach to phenology monitoring assesses a plant species in an area, without marking and tracking individual plants. Incorporating an example data set of this type is a medium-term priority for the USA-NPN. Data integration of this kind facilitates assessments of phenological response to climatic and environmental change over longer periods of time.

A new enhancement to the data model and interface (completed in March 2011) incorporates abundance and intensity measures for each phenophase, to further enhance the quality and extend the breadth of applications for phenological data.

## **Data Output**

The USA-NPN must determine the format, fields, and necessary filters for data delivery. On the USA-NPN website, filters will be built to allow users to customize data output by species, phenophase, geographic region or time frame, as well as allow users to select which fields to include. Pre-packaged data output will be needed when data is served in other communities (e.g., National Biological Information Infrastructure; NBII). Use-case scenarios based on research questions are an effective way to determine how to prepackage data.

Data is available in the following four categories:

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: Plants (Status method; includes lilacs and natives)

03/27/2010–Present: Animals (Status method)

Further information detailing the data collected, and some example analyses are available in USA-NPN Data Summary Reports (Crimmins et al 2010, Crimmins et al 2011). Versioning is needed 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 should 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 supporting information. An OAIS reference model would provide a framework for describing and comparing appropriate long-term preservation strategies and techniques.

**The USA-NPN should explore becoming a member node in the DataONE project.** The project could harvest metadata, offer free online services to use with data, and provide EML output. 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). The Open Geospatial Consortium has an observation schema called Sensor Web, which might be a promising partnership for determining a common schema.

As use of data by scientists is a key metric of success for the organization, the data download should be tracked. If users must register to obtain a data set, potential data users may be discouraged. **While it is recommended that registration be optional, providing benefits (data updates, newsletters) to motivate registration is effective.** This issue is further complicated when data are available through other venues (e.g., National Climate Data Center or Google Fusion Tables).

### **Privacy, Data Sharing, and Data Attribution Policies**

At the time of the review, phenology data was output at locations rounded to 0.01 degrees latitude and longitude (approximately 7 miles), to protect observer privacy. The USA-NPN seeks to balance observers' right to privacy with scientific needs for the data. 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) set up an opt-out or opt-in system for display of exact locations.**

Subsequent to the review, staff and board members decided to adopt a formal privacy policy, which states that all observation locations will be output at the exact location. Existing observers were able to opt out during a four-month period (six observers, none of whom had observation data in the system, opted out). The remainder of the existing observers and all new observers are bound by the Terms of Use, which includes a clear description of the privacy policy (<http://www.usanpn.org/terms#ObserverPrivacy>).

The USA-NPN also needs to ensure that care is taken regarding data about threatened and endangered species or areas. Currently, the Network does not collect data on threatened and endangered species.

Once privacy issues are resolved, the USA-NPN will be able to make great strides in data sharing. Reviewers recommended implementing Google Fusion Tables, a free or low-cost service to non-profits, which would protect both the data and the servers, by allowing users to manipulate and download a copy of the data using Google's servers. **The USA-NPN should also consider distributing data in similar content communities like the National Climatic Data Center, DataBasin, or the Avian Knowledge Network (AKN).** The USA-NPN is in a unique position to set standards for phenology data while benefiting from the experience and guidance of the Science

Commons and Committee on Data for Science and Technology. A notable example is the Polar Information Commons (PIC; [www.polarcommons.org](http://www.polarcommons.org)) that was formed for the purpose of archiving and protecting data collected during each International Polar Year. 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. In addition, the use of controlled vocabularies can support data search and harvest.

Subsequent to the IMS Review, board and staff developed Data Use and Data Attribution policies (<http://www.usanpn.org/terms#DataUse>), relying on Parsons and Duerr (2010) for approach, to address these issues, as well as a broader Website Terms of Use to address broader website liability issues (<http://www.usanpn.org/terms>).

## Database Software

In considering database software, the USA-NPN should address the transition to GIS integration and how to scale technologies as the data set grows. The USA-NPN uses MySQL as the database management software, which has met the organization's needs to date. However, MySQL does not have native support for geographic data (beyond latitude/longitude point data). Drupal is now compatible with PostgreSQL (though not all modules may be compatible). **Reviewers recommended switching to PostgreSQL, at least in a development environment, to explore migration costs.**

Following the IMS Review, NCO staff determined that the costs of migration outweighed the benefits. Many Drupal modules in use on [www.usanpn.org](http://www.usanpn.org) did not integrate with PostgreSQL, resulting in the need for substantial custom coding. Running MySQL and PostgreSQL in parallel was also deemed too unstable and complicated to maintain. In addition, in the fall of 2010, the NCO contracted with the Center for Environmental Informatics (CEI) at Penn State University to develop dynamic visualizations of phenology data. All geographic web serving is handled through CEI, obviating the need for GIS integration in-house in the medium term.

## DATA QUALITY ASSURANCE & QUALITY CONTROL

USA-NPN's primary data source is a national pool of observers ranging from high school students and retirees to professional researchers. Observers are not paid or field-trained by the USA-NPN, and a threshold skill or experience level is not required (or enforceable) for participation in data contribution. In addition, the nature of phenological observation is much more subject to observer interpretation than that for other data collection efforts, such as water quality monitoring or precipitation gauging. **A comprehensive data review process is needed, including several layers of QA/QC checks and a volunteer review network.**

Some of the broader questions that reviewers thought the USA-NPN should address with regard to QA/QC and data reliability are:

- What are the specific criteria in terms of data quality needed to meet USA-NPN objectives?
- What are the kinds and frequencies of errors that occur?
- What level of consolidation or federation will the data collections require?

Through the full implementation of QA/QC measures, data end users will be able to select observers by skill level as well as track the revision history of a data set, know when observations were made, distinguish between data collected by different observers at a site, and investigate inconsistencies in the data set. Quality assurance (measures before data enters the database) and quality control (post-processing) measures proposed and taken are summarized in Table 1.

**Table 1.** QA/QC measures for Nature’s Notebook completed to date (black text) and proposed (gray text), grouped by potential source of error.

QUALITY ASSURANCE MEASURES	QUALITY CONTROL MEASURES
<b>Species Identification Errors</b>	
<ul style="list-style-type: none"> <li>● “How to observe” monitoring instructions and Frequently Asked Questions (FAQs) emphasize the importance of accurate species identification and direct observers to general identification resources</li> <li>● Species profile pages include a photo, range map, written description of the species, and lead the user to other websites with more identification information</li> </ul>	<ul style="list-style-type: none"> <li>● Site and plant level metadata (e.g., land cover type for sites, watered status for plants) enables data end users to explore outliers</li> <li>● In a preliminary test of species identification errors, 3.7% of species were registered in states outside of their known range (n of 4857 registered plants and animals)</li> <li>● Species identification is confirmed via submission of photo with observation (with crowd sourced review of images and expert confirmation on an image subset)</li> </ul>
<b>Phenophase Status Evaluation Errors</b>	
<ul style="list-style-type: none"> <li>● Language in phenophase definitions is carefully chosen for precision and accessibility</li> <li>● Phenophase definitions are generalized and identical across similar species (within phenological functional types) for consistency</li> <li>● Phenophase definitions are changed as infrequently as possible to simplify observing and to ease the interpretation burden on data-end users</li> <li>● Species-specific additions to the general definitions more completely describe how the phenophase appears in a particular species</li> <li>● Observers are given an ‘uncertain’ option to</li> </ul>	<ul style="list-style-type: none"> <li>● Detection bias in animal phenology reporting is exposed via observer reports of the time spent observing animals and their selection of an animal survey method from a pick list</li> <li>● Spatial interpolation to identify outliers as data density allows</li> <li>● Flagging of phenophases reported out of expected order and logically inconsistent</li> <li>● Comparisons of observation data from expert and non-expert (or trained and untrained) observers at the same site</li> <li>● Assessments in which observers are asked</li> </ul>

<p>reduce false positives and false negatives</p> <ul style="list-style-type: none"> <li>• Observers are not asked to infer the date of a 'first'; dates of all visits are known explicitly</li> <li>• FAQs address tricky issues in phenophase status evaluation (across species)</li> <li>• Photos or illustrations for each phenophase in each species are provided to observers</li> <li>• Photographic primer teaches phenophase evaluation skills</li> <li>• Online photographic quiz tests and hones observers' skill in phenophase evaluation</li> </ul>	<p>questions about their observations targeted at identifying mischaracterizations of phenophases</p> <ul style="list-style-type: none"> <li>• Phenophase evaluation is confirmed via submission of photo with observation (with crowd sourced review of images and expert confirmation on an image subset)</li> </ul>
--	--

### Data Entry Errors

<ul style="list-style-type: none"> <li>• Training and FAQs address data entry issues</li> <li>• Species names and abundance/intensity measures are presented as pick lists</li> <li>• Datasheets mirror the online data entry form</li> <li>• Phenophase and intensity definitions appear on roll-overs in the data entry form</li> <li>• Site location can be entered by Google map or address input; elevation is calculated from USGS digital elevation model, but can be hand-corrected</li> <li>• Observers can review previously submitted observations in user interface (UI) or a downloaded Excel file, and can edit their previously submitted observations in UI</li> <li>• Usability testing has been conducted on user interface to increase intuitiveness and reduce transcription errors</li> <li>• User interface validation on observation methods: <ul style="list-style-type: none"> <li>○ Users must provide both a measurement and a metric to input data regarding the amount of time spent observing, time spent traveling to observation site, and time spent searching for animals</li> </ul> </li> <li>• Reordering of plant and animal lists reorders data entry form and datasheet printout</li> <li>• When a plant is deleted, rationale for deletion is requested and the deleted plant data is retained</li> </ul>	<ul style="list-style-type: none"> <li>• Collect and cross check a sample of observer datasheets with database</li> </ul>
---	---

<ul style="list-style-type: none"> <li>● Comments box provided at the site, plant and observation level</li> <li>● User interface validation on date/time: <ul style="list-style-type: none"> <li>○ Date field required; default is to select from a calendar</li> <li>○ Time field optional; selected from pick list</li> <li>○ Dates in the future not allowed</li> <li>○ After the date is entered it appears above the phenophase column for every species</li> <li>○ Duplicate date/time values not allowed</li> <li>○ Observations cannot be made about an individual after it has been marked as 'inactive'</li> <li>○ User is automatically warned by UI of changing phenophases through time</li> </ul> </li> <li>● User interface validation on phenophases: <ul style="list-style-type: none"> <li>○ User may only enter "Yes," "No" or "Uncertain" on the interface, using mutually exclusive click points; if no response is checked no database record is created</li> <li>○ User may not enter abundance or intensity measure unless the phenophase is set to "Yes" or "Uncertain"</li> </ul> </li> <li>● Observers see their data re-presented to them in a visualization on their "My Account" page</li> <li>● Mobile applications for data collection eliminate datasheet to interface transcription errors</li> </ul>	
--	--

**Training and Observer Skill Level**

<ul style="list-style-type: none"> <li>● Field observing methods (selecting a site, selecting species, making observations) are accessible as: <ul style="list-style-type: none"> <li>○ Printable handbook</li> <li>○ Handouts</li> <li>○ Powerpoint presentation with script for aspects of data collection and data entry</li> <li>○ Voiced "training video" version of the Powerpoint presentations</li> </ul> </li> <li>● FAQs available on separate page, and as clickable web links from places on website where user questions might arise</li> <li>● In-person and online workshops provide training</li> </ul>	<ul style="list-style-type: none"> <li>● Self-reporting of training, skill and experience level by observers</li> <li>● Record of observer's online quiz scores</li> <li>● Scores based on comparison of observer's data to expert's data at sites where both are monitoring the same species</li> <li>● Use rainlog, eBird or another program with more easily interpolated/QC-ready data to determine characteristics of skilled observers; apply the findings to the Nature's Notebook observer pool</li> </ul>
---	--

<p>opportunities for a limited number of observers</p> <ul style="list-style-type: none"> <li>• A variety of peer-support networks can be implemented, from user forums on the website to power-observers who review other observers' data</li> </ul>	
---	--

*Notes from Reviewers on Quality Assurance Measures*

- **Development of QA/QC measures should include both verification and validation.**
- Offering a variety of data output options, such as visualizations, may encourage more interest and involvement while at the same time helping to verify observations by easily allowing users to find and flag outliers.
- **Image validation for observations is recommended by reviewers**, while recognizing that issues with file size, credit, and inappropriate or malicious content must be addressed. Using a commercial photo-sharing site may be an interim solution. The Encyclopedia of Life website currently collects images submitted by the public, and this may be a promising partnership for the USA-NPN.
- It was also suggested that the Network consider adding a color swatch on datasheets to enable observers to take a photo of their plant with a datasheet in the background to calibrate computer image-recognition applications. However, user-printed swatches may be too inconsistent to be useful.
- Some reviewers thought observers would reliably self-evaluate based 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. Research should be able to shed light on this question, known as the “Dunning-Kruger effect.”
- 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 data quality objectives. Examples include master gardener certification programs or websites such as Galaxy Zoo, which allows citizen astronomers to submit and classify images of galaxies once they have passed a short trial classification test.
- An alternative way to satisfy these diverse needs might be through creating an expert review system to track anomalies and identify observer biases. In this scenario, observation parameters exceeding a defined threshold would be flagged for human review. The eBird project at the Cornell Lab of Ornithology has implemented a volunteer review network to validate the observations that exceed expected thresholds and automatically flag them for review. Reviewers can also find and force any observation into the review process on an *ad hoc* basis. Reviewers in the eBird system have assigned geographic areas of responsibility and use a separate web application with specialized tools to make the review process easier. This includes the ability to generate emails to observers based on template messages, make bulk changes in review status, and record review notes and reasons for acceptance or rejection. All steps in the review process are

recorded for auditing purposes. The data verification process enhances data integrity, encourages interactions between experts and contributors, and can improve observer skills.

- Forum statistics might also be used to identify unusual events (e.g., the number of comments in a forum on “late freeze” could be used to understand data showing delayed leaf out in trees).
- Phenology festivals may provide a means for identifying a likely range for phenophase dates in a particular region.
- An appropriate model for observer training and other quality assurance measures might be the Free-Air CO<sub>2</sub> Enrichment experiment, for which the Department of Energy has set up control plots with citizen scientists in an attempt to standardize data collection.

## METADATA

---

At the time of the review, contemporary observations entered through Nature’s Notebook were not publicly available and did not have associated metadata. **The USA-NPN should 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. Most reviewers agreed that choosing a mature and interoperable standard that the USA-NPN could easily implement would be best, and these records could then be integrated with ISO 19115. FGDC-compliant metadata was developed for six data sets, and these data sets were output dynamically on the USA-NPN website in fall 2010.

The 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.

## HUMAN INTERFACES

---

### Current Infrastructure

The USA-NPN’s Drupal 6.X site ([www.usanpn.org](http://www.usanpn.org)) is an information-rich website that includes the following features (all features present in July 2011 unless otherwise noted; starred features allow user-generated content):

- Data Set Registry Tool: [www.usanpn.org/participate/dataset](http://www.usanpn.org/participate/dataset)\*
- Data Set Discovery: [www.usanpn.org/results](http://www.usanpn.org/results)
  - Mercury Search Tool: [mercury.ornl.gov/usanpn](http://mercury.ornl.gov/usanpn)
  - Registered Data Sets: [www.usanpn.org/results/dataset-list](http://www.usanpn.org/results/dataset-list)
- Contemporary Data Output: [www.usanpn.org/results/data](http://www.usanpn.org/results/data) (released in November 2010)
- Educator’s Clearinghouse: [www.usanpn.org/education/clearinghouse](http://www.usanpn.org/education/clearinghouse)
- Phenology Festivals: [www.usanpn.org/resources/festivals](http://www.usanpn.org/resources/festivals)\*
- Species Information
  - Search Species: [www.usanpn.org/species\\_search](http://www.usanpn.org/species_search)

- Example Profile: [www.usanpn.org/Carduelis\\_tristis](http://www.usanpn.org/Carduelis_tristis)
- Bibliography: [www.usanpn.org/results/biblio](http://www.usanpn.org/results/biblio)\*
- User Creation: [www.usanpn.org/user/register](http://www.usanpn.org/user/register)
- Visualization Tool: [www.usanpn.org/results/visualizations](http://www.usanpn.org/results/visualizations) (released in March 2011)

A second, Java-based website, Nature’s Notebook [mynpn.usanpn.org/npnapps/](http://mynpn.usanpn.org/npnapps/), supports the submission of observation data. This site shares login sessions with the Drupal website using a shared cookie. An overview of the process for participating as an observer is available at [www.usanpn.org/participate/guidelines](http://www.usanpn.org/participate/guidelines).

As of July 2011, a Droid application for data entry has been developed and an iPhone app is slated for development by the end of the year. All application code and Drupal customization developed by the USA-NPN are intended to be open source. Cleaning and standardizing of the code have been the only impediments to code sharing, thus far.

### Leveraging Existing Tools

A wide range of potential enhancements to user interfaces, along with associated challenges, easy wins, and potential partners, was proposed by members of the review panel, detailed below. One reviewer added a note of caution, pointing out that technology alone cannot solve observer recruitment and retention problems. **Staff time is well spent leveraging and facilitating the adoption of web tools.** In some cases, too much is asked of the interface, under the assumption that “if you build it, they will come.”

To address this issue, the USA-NPN plans to increase education/outreach staff time. Key projects will include updates to the Educator’s Clearinghouse and development of teacher training workshops and modules, along with other helpful resources for citizen scientists (such as observation kits, a phenology handbook, and materials for recruiting others as citizen scientists). In spring 2011, additional staff were hired to meet these non-IT needs in support of observer recruitment and retention.

### Enhancements to Existing Interfaces

The Information Management System must support the growth of an engaged and trained observer base. The USA-NPN should continue to focus on targeted usability testing to assess the efficacy of existing user interfaces for data entry through Nature’s Notebook. Improvements could include simple revisions, such as shading the columns in data sheets so that users can easily transition from analog to digital, and from field data to data entry. Other relatively simple improvements could include developing controlled vocabularies for better data searchability, 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, Nature’s Notebook targets a middle-level user profile (backyard naturalist with some computer familiarity), and the USA-NPN will consider developing several user profiles to allow the complexity of the interface to vary. **Reviewers recommend prioritizing a simple and comprehensive data entry interface.** They also cautioned against assuming that developers can divine how users would want their interfaces to appear and that the interface should be customizable. In subsequent conversations among the NCO staff and the USA-NPN Advisory Committee, the resource cost (five months of staff time) of customizable user profiles was seen to

likely outweigh the benefits. The relative priority of this enhancement is slated to be considered in larger action-planning efforts. **Intermediate improvements could include revising the Nature's Notebook data entry interface to simplify and streamline data entry.**

**Revisions to the interface were proposed to allow the user to explore and correct mistakes made while learning and testing the interface.** This “sandbox” concept would allow users 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 Yes/No/Uncertain response at any time and the history of this revision is not captured; only the latest revision is stored.

An additional issue was brought up for consideration in terms of the difference between an observer reporting uncertainty with regard to phenophase status and an observer not providing a response to the phenophase status question. At the time of the review, the system treated the “?” or uncertain response as equivalent to no response (field left blank; both stored as nulls in the database). Reviewers felt that the USA-NPN should distinguish between stated uncertainty and no response. Following the review, developers revised the code so a record is not created when the entry field is left blank and a value of -1 is entered when a “?” (uncertain of phenophase status) is submitted.

Another user interface, the Data Set Registry tool, could be improved if locations could be submitted as polygons rather than just as points. The data sets may be dynamically linked to related publications so that web users can easily delve further into a subject area. **The website's distributed data search functions (data set registry tool and Mercury search engine) should be combined to simplify data exploration, without the need to search both databases.**

### **Suggestions for Recruitment and Retention of Observers**

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 upcoming phenological events in their area, such as a flowering season application for hikers, or of observations made by others in their area. 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 may be used to dynamically serve seasonal phenological information to users about the state in which they reside. Participants could be prompted to answer a question such as, “Which phenophases do you currently see?” using a Google Mapping application (GMap). Even without a current database for the phenophases by state, participants may learn what other observers are seeing and reporting. For observers who are planning a trip, USA-NPN could provide information about particular phenology likely to be observed during the travels. The games and quizzes, discussed above for measuring observer expertise, can also be a means of encouraging observer involvement and creating non-monetary rewards for participation. In addition, 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.

Additionally, sustaining user involvement requires enticing users 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. Alternatively, observer names could be dynamically embedded in metadata, for citation by scientists. 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 an area for comparison purposes. An iGoogle-type landing page could allow users to personalize their own phenology page to highlight phenology festivals and related publications, browse through all available applications, and plug into other USA-NPN offerings, such as RSS feeds.

Though not always appropriate for scientific analysis, data visualizations are an effective 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. 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.

Another potential example is a flowering map offering data feeds for news organizations, which may help advertise the USA-NPN to potential citizen scientists and make 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.

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. Lastly, Cornell's Yard Map program may be a relevant example of a way to engage observers, giving them an online space to describe their local environment.

## **Tools**

In terms of tools, the current USA-NPN Drupal system might also offer some of this capability, such as for RSS feeds, through the Feeds Drupal module. The Google Visualization API (Application Programming Interface) might also be advantageous and is easy to use. For web mapping, a Web Map Service (WMS) that can update on a regular basis might be the best solution. An image-based WMS facilitates visualizing data. A Web Feature Service (WFS) would enable site visitors to click on a point and get data, as with GMap. And though the current GMap views in Drupal (e.g., the phenology festivals map), are approaching the limit as to how many points can be displayed, the USA-NPN could still capitalize on a great deal of flexibility and functionality with the GMap

module. For example, the Network could begin using Keyhole Markup Language (KML) files in Google Earth to display data animations.

## APPLICATION INTERFACES

---

Machine-to-machine communication, facilitated by APIs, also known as web services, is an increasingly common and effective means for real-time data sharing and integration. **The development of APIs was identified as a key need for the USA-NPN to provide high-quality, flexible data input and output options.** A beta version of the USA-NPN API was released in fall 2010, with a fully operational version functioning by spring 2011.

### Web Services for Output

Web services for output are needed to support visualization tools, dynamic data download, and harvest by partner organizations (and many of the other developments discussed in section 5.4). The Network expects some scientists or institutions (Data Basin, AKN, and NBII are potential examples) to eventually seek full data sets delivered by web service, but as yet no partner has requested data set output by web service.

### Web Services for Input

Data input web services will be developed primarily for organizations that would like to encourage their members to participate in Nature's Notebook, but do not want their members to leave their website. Examples are YourGardenShow.com and The Great Sunflower Project. Each could potentially contribute thousands 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 developed an API for data input to the National Phenology Database. The API was completed in the fall of 2010, and is being used by ScienceforCitizens.net and YourGardenShow.com, as well as by the Nature's Notebook Droid App and a prototype Facebook App, as of July 2011.

eBird has developed an alternate model, in which a portal provides partner organizations with a co-branded and customized "eBird." This allows partners to maintain some identity and control, without requiring the support of programmers and other technical resources. Overlapping Facebook applications could also allow users to submit data to the USA-NPN from other partner sites without leaving those sites.

### Tools and Frameworks for Web Service Development

At the time of the review, the USA-NPN was developing a package of input and output XML 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 was that a less resource-intensive option is necessary. **REST-based web services were recommended as being straightforward and easy for client applications to consume.** REST could also be easier for USA-NPN staff to program; the bulk of the work would be in defining the parameters and writing the program. Web services released in spring 2011 are both SOAP and REST (and available as JSON or XML), and are documented at <https://docs.google.com/document/d/1yNjupricKOAXn6tY1sI7-EwkcfdGUZ7lxYv7fcPjO8/edit?pli=1>.

OpenID is a mature tool for shared authentication (requiring users to log in only once for a suite of applications, and enabling mobile apps to share website sessions). This tool could be used for Drupal authentication and

would simplify some aspects of Java application login management. It would also enable authentication for other (future) applications, though OpenID has some limitations for web services and access methods that are not browser-based. In spring 2011, OAuth (Open Authentication) was implemented on the [www.usanpn.org](http://www.usanpn.org) website, through the eponymous Drupal module, for use by the Droid App. OAuth has not yet been implemented in place of the cookie for shared login between Nature's Notebook and the Drupal website. It is expected that future in-house and partner applications will take advantage of the OAuth infrastructure.

In addition to publicly available web services, a generalized Data Access Object (DAO) layer could be developed. The DAO layer would consolidate access to the database, so that changes to the data model would have to be made in only one place, and would enable consistent enforcement of the business rules. Participants noted that allowing multiple application code pathways to the database has wreaked havoc for other programs. A DAO layer is cumbersome to implement, and the USA-NPN should consider carefully whether it would be an efficient architecture. **If the USA-NPN is expected to grow to support more applications (user interfaces), and have more than one developer at work, a DAO layer might be worthwhile.**

**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 serve data in various communities.** In addition, output applications are typically easier to create and consume than input applications. Output applications will also assist with the consumption of USA-NPN data, potentially leading to increased exposure and future funding opportunities.

## PHYSICAL INFRASTRUCTURE

---

### 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; upgraded in May 2011 to 48 GB RAM and 2.2 TB 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 four Virtual Machines (VMs).

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

The servers are located in a climate controlled 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.

The 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 machine (PM) and virtual machine (VM) configuration for the USA-NPN servers is provided in Table 2. The processor usage has barely been utilized, and is averaging only 3-5% over 24-hour periods. The

input/output 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.** USA-NPN PM and VM Configuration

PM1	VM1	Ubuntu 9.04 Apache, MySQL, and PHP stack	Development Drupal web server
PM1	VM2	Ubuntu 9.04 Tomcat, Apache, MySQL, and PHP stack	Production Java Nature’s Notebook web server
PM2	VM3	Ubuntu 9.04 Apache, MySQL, and PHP stack	Production Drupal web server
PM2	VM4	Ubuntu 9.04 Tomcat, Apache, MySQL, and PHP stack	Development Nature’s Notebook web server

## Future Considerations

### *Website Performance*

The time limit for acceptable website outage (for both the Drupal and Java applications) has been set at 24 hours by NCO staff. A shorter recovery time may be desirable and feasible as the program grows in stature and funding level. In March 2009, a National Public Radio (NPR) Science Friday interview featuring the USA-NPN brought over 2,000 simultaneous visitors to the [usanpn.org](http://usanpn.org) website, crashing the site. 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.

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 to load test all USA-NPN functions. 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 setup would enable Drupal to offload content serving to offload the caching server.

### *Backup and Recovery*

The USA-NPN has a disaster recovery plan with ORNL as the offsite backup. It could strengthen its plan by replicating servers, but there would remain the potential to lose data entered in the 24-hour period between backups. 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 the USA-NPN back up code to Google Codes but the code would need to be cleaned, and made stable and standard.

Given that hardware upgrades increased server capacity, additional virtual machines have been created to support near-real time backups (for both websites and web services) and proxy servers for load handling, as of July 2011.

### *Replacements and Upgrades*

**It was recommended 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. 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 a second machine. In reality, more staff time is likely required to anticipate and prepare for all possible compromised situations than to simply replace the machines themselves. Even with open source software, organizations commonly purchase support contracts, as they can offer developers backup support when something goes wrong.

### **Security Risks**

Finally, security risks are another area of concern. The USA-NPN runs backups with Secure Shell (SSH) certificates, which provide both authentication and encryption. The website applications account for authentication (login), but are missing a confidentiality component. **Website login and administration were not over a Transport Layer Security (TLS)<sup>1</sup> at the time of the review, leading reviewers to note that web transactions were insufficiently secure.** Following up on this vulnerability, all login and administrative activity on the USA-NPN websites was moved over to TLS in early 2011.

Drupal is a secure platform and notifies developers of periodic core and module security updates. The USA-NPN staff monitors security issues from other sites and security updates, applying security updates regularly, and running scans to detect Apache misconfigurations.

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, for reasons of both security and efficiency. As Drupal and the USA-NPN evolve, the continued compatibility should be assessed. The eBird project currently uses Plone, as it was available when the project started, though Drupal is a stronger CMS.

## **CONCLUSIONS AND RECOMMENDATIONS**

---

The consensus of panel experts during the IMS Review Workshop was that the USA-NPN has already accomplished a great deal. Recommendations for changes centered primarily around how to ensure 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 observer involvement, and how to make sure the USA-NPN's systems are scalable with increased user involvement and growing data storage requirements.

---

<sup>1</sup> TLS is the current means of implementing secure HTTP (https) and is the more proper name for the earlier methods, generally referred to as Secure Sockets Layer (SSL).

The participants did not always understand that the USA-NPN IT development process originated with clearly defined needs aligned to broader goals of the organization. **It was recommended that enhancements and new projects be considered in a strategic framework.**

Although many well-considered enhancements were suggested in this report, not all can be addressed in the immediate future, given current funding levels. Priority should be given to those enhancements that enable the organization to reach its larger goals. Data quality, data access, web services, visualization tools, and targeted security enhancements are most likely of highest importance at this stage in the organization's development.

## BIBLIOGRAPHY

---

- Crimmins, TM, AH Rosemartin, KA Thomas, RL Marsh, EG Denny, and JF Weltzin. 2011. USA National Phenology Network 2010 State of the Data. USA-NPN Technical Series 2011-001. [www.usanpn.org](http://www.usanpn.org).
- Crimmins, TM, AH Rosemartin, KA Thomas, RL Marsh, EG Denny, and JF Weltzin. 2010. USA National Phenology Network 2009 Data Summary. USA-NPN Technical Series 2010-002. [www.usanpn.org](http://www.usanpn.org).
- Denny, EG, AJ Miller-Rushing, B Haggerty, LL Benton, TM Crimmins, M Losleben, A Richardson, A Rosemartin, MD Schwartz, KA Thomas, JF Weltzin, and B Wilson. 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>>.
- Parsons, MA and R Duerr. 2010. Data Citation and Peer Review. EOS 91 (34): 297–298.
- Thomas, KA, EG Denny, AJ Miller-Rushing, TM Crimmins, and JF Weltzin. 2010. The National Phenology Monitoring System v0.1. USA-NPN Technical Series 2010-001.

## CONTRIBUTIONS & ACKNOWLEDGMENTS

---

Alyssa Rosemartin led the review workshop and wrote the report. Karla LeFevre took notes during the workshop, recorded the review session, and provided a first draft of the report. Paul Allen, Suzanne Allard, Ellen Denny, Natalie Latysh, Lee Marsh, Mark Parsons, Inigo San Gil, Robert Tawa, Brian Wee, Jake Weltzin, and Bruce Wilson provided substantive revisions to the document.

The USA-NPN's National Coordinating Office is grateful to the individuals who participated in this review (see Appendix I). The lessons and expertise they contributed have improved our selection of technologies, prioritization of enhancements, and policy development. A National Science Foundation (NSF) Research Coordination Network (RCN) Grant (IOS-0639794) supported this workshop.



## APPENDIX A

### USA-NPN IMS Review Workshop Participants

Name	Title	Organization/Affiliation
Suzanne Allard	Faculty Member	School of Information Sciences University of Tennessee, Knoxville
Paul Allen	Assistant Director	Information Science Program Cornell Lab of Ornithology
Ellen Denny	Monitoring Design & Data Coordinator	USA-NPN NCO
Natalie Latysh	Physical Scientist	USGS GIO
Lee Marsh	Applications Developer	USA-NPN NCO
Jeff Morisette	Invasive Species Science Branch Chief	USGS
Mark Parsons	Program Manager	NSIDC
Alyssa Rosemartin	IT & Communications Coordinator	USA-NPN NCO
Inigo San Gil	Senior Application Support Analyst	NBII, LTER, USA-NPN Board of Directors
Robert Tawa	Director of Computing	NEON, Inc.
Brian Wee	Chief of External Affairs	NEON, Inc.
Bruce Wilson	Systems Engineer/Group Leader Environmental Data Science & Systems, Environmental Sciences Division	ORNL, USA-NPN Board of Directors

## APPENDIX B

Acronyms and abbreviations used in this document

AKN	Avian Knowledge Network
API	Application Programming Interface
CEI	Center for Environmental Informatics (at Penn State University)
CENS	Center for Embedded Network Sensing
CMS	Content Management System
CSV	Comma-Separated Values
DAO	Data Access Object
DOI	Digital Object Identifiers
DNS	Domain Name System
EML	Ecological Metadata Language
FGDC	Federal Geographic Data Committee
GIO	Geospatial Information Office
GIS	Geographic Information System
GBIF	Global Biodiversity Information Facility
GMap	Google Mapping application
IMS	Information Management System
ISO	International Organization for Standardization
KML	Keyhole Markup Language
LTER	US Long Term Ecological Research Network
NBII	National Biological Information Infrastructure
NCO	National Coordinating Office of the USA National Phenology Network
NEON	National Ecological Observatory Network
NSIDC	National Snow and Ice Data Center
NSF	National Science Foundation

---

OAIS	Open Archival Information System
OAuth	Open Authentication
OBOE	Extensible Observation Ontology
ORNL	Oak Ridge National Laboratory
PHP	Hypertext Preprocessor scripting language
PIC	Polar Information Commons
PM	Physical Machine
RCN	Research Coordination Network
REST	Representational State Transfer
RSS	Real Simple Syndication
SSH	Secure Shell
SSL	Secure Socket Layer
SOAP	Simple Object Access Protocol
SONet	Scientific Observations Network
SVN	Subversion
TLS	Transport Layer Security
TOAD	Tool for Oracle Application Developers
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

---