Data Quality Assurance & Quality Control for *Nature's Notebook*

The primary source of observational plant and animal data for the USA National Phenology Network (USA-NPN) is a national pool of observers ranging from high school students and retirees to professional researchers who participate in *Nature's Notebook*. These participants are able to collect data on a scale that would not otherwise be feasible. They are not paid and are not always field-trained by the USA-NPN or its partner organizations, nor is a threshold skill or experience level required (or enforceable) for participation in data contribution. In addition, the nature of phenological observation is potentially more subject to observer interpretation than that for other citizen data collection efforts, such as water quality monitoring or precipitation gauging.

Thus to maximize data quality and utility, the USA-NPN has established a suite of quality assurance (QA; before data enters database) and quality control (QC; post-processing) measures for *Nature's Notebook*. Through the full implementation of these QA/QC measures, data end users will be able to: 1) select observers by skill level, 2)track the revision history of a data set, 3) know how frequently observations were made, 4) distinguish between data collected by different observers at a site, and 5) investigate inconsistencies or outliers in the data set. QA/QC measures completed to date (black text), planned near term (grey text) and proposed (2-5 years out; grey and italicized) are summarized in the following table.

Quality Assurance Measures	Quality Control Measures
Species Identification Errors	
• "How to observe" monitoring instructions and Frequently Asked Questions (FAQs) emphasize the importance of accurate species identification and direct observers to general identification resources • Species profile pages include a photo, range map, and in some cases a written description of the species, and lead the user to other websites with more identification information	• 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) • Plant images uploaded by observers are available to data users with data output • *Plant images are reviewed using a combination of crowd-sourcing and expert review* • *Species outside of known range (NatureServe/BONAP) are flagged and excludable with data output*

Quality Assurance Measures	Quality Control Measures
Phenophase Status Evaluation Errors	
• Language in phenophase definitions is carefully chosen for precision and accessibility • Phenophase definitions are generalized and identical across similar species (within phenological functional types) for consistency • Phenophase definitions are changed as infrequently as possible to simplify observing and to ease the interpretation burden on data-end users • Species-specific additions to the general definitions more completely describe how the phenophase appears in a particular species • Observers are given an 'uncertain' option to reduce false positives and false negatives • Observers are not asked to infer the date of a 'first'; dates of all visits are known explicitly • FAQs address tricky issues in phenophase status evaluation (across species) • National webinars and photographic primers teach plant anatomy and phenophase evaluation skills • *Photos or illustrations for each phenophase in each species are provided to observers* • *Online photographic quiz tests and hones observers' skill in phenophase evaluation* • *Messages to confirm species identification when reported out of range*	• Site and plant level metadata (e.g., land cover type for sites, watered status for plants) enables data end users to explore outliers • 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. Site area is also provided in site-level metadata. • Conflicting records flagged (e.g., same observer multiple times in a day or different observers at a shared site report different status) • *Phenophases reported in implausible order flagged* • *Implausible changes in step magnitude for intensity measures flagged* • *Spatial interpolation to identify other implausible values as data density allows* • *Assessments in which observers are asked questions about their observations targeted at identifying mischaracterizations of phenophases* • *Phenophase evaluation is confirmed via submission of photo with observation (with crowd-sourced review of images and expert confirmation on an image subset)*

Quality Assurance Measures	Quality Control Measures
Data Entry Errors	
• Training and FAQs address data entry issues • Species names and abundance/intensity measures are presented as pick lists • Datasheets (PDF and Excel templates) mirror the online data entry form • Phenophase and intensity definitions appear on roll-overs in the data entry form • Site location can be entered by Google map or address input; elevation is calculated from USGS digital elevation model, but can be hand-corrected • Observers can review previously submitted observations in user interface (UI) or a downloaded Excel file, and can edit their previously submitted observations in UI • Usability testing has been conducted on user interface to increase intuitiveness and reduce transcription errors • User interface validation on observation methods: ○ 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 • Observers can reorder plant and animal lists so data entry form and datasheet printout mirrors order encountered at the field site • When a plant is deleted, rationale for deletion is requested and the deleted plant data is retained, if appropriate • Comments box provided at the site, plant and observation level	• Locations with implausible coordinates flagged, corrected or removed • *Collect and cross check a sample of observer datasheets with database*

Quality Assurance Measures	Quality Control Measures
Data Entry Errors, continued	

• User interface validation on date/time:
  ○ Date field required; default is to select from a calendar
  ○ Time field optional; selected from pick list
  ○ Dates in the future not allowed
  ○ After the date is entered it appears above the phenophase column for every species
  ○ Duplicate date and time values not allowed
  ○ Observations cannot be made about an individual after it has been marked as 'inactive' (due to plant death, no longer observing, or misidentification)
  ○ User is warned by UI of changing phenophase definitions through time
• User interface validation on phenophases:
  ○ 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
  ○ User may not enter abundance or intensity measure unless the phenophase is set to "Yes" or "Uncertain"
• Observers see their data re-presented to them via spreadsheet and in the visualization tool
• Mobile applications and an MS Excel template for data collection eliminate datasheet to interface transcription errors	

Quality Assurance Measures	Quality Control Measures
Training and Observer Skill Level	
• Field observing methods (selecting a site, selecting species, making observations) are accessible via web pages, handbook, PowerPoint and video presentations. • Detailed FAQs available as context-specific help • In-person and online workshops provide training opportunities for observers (~50% of data submitted by experts or trained volunteers) • *Peer-support networks from user forums on the website to power-observers who review other observers' data*	• Self-reporting of training, skill and experience level by observers *made available to data users* • Comparisons of observation data from experts and trained and untrained observers at the same site: a. Fucillo et al (in press), shows 91% concordance between trained observer and expert b. Effort at Acadia National Park underway comparing expert, trained, untrained observers and their improvement over time • *Record of observers' online quiz scores made available to data users* • *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*

In addition to these error-specific QA/QC measures, the USA-NPN has taken several broader steps to ensure high quality data, outlined below.

Recruitment and Retention

- Messages to observers are targeted to increase temporal resolution of the data set
- Retention efforts to maintain observers long term include leaderboards, badges, demonstrations of utility of data, locally led communities, and quarterly email updates
- Recruitment efforts are targeted around priority geographic regions

- Recruitment is targeted at high-school age and above audiences, with a knowledge of the natural world
- Shared sites enable on-the-ground administrators to access and review their members' data

Methodology

- Individual plants are tracked through time, controlling for variation across organisms and in microclimate
- Observers are encouraged to monitor multiple individuals of each species of plant at each site to capture variation
- Monitoring plant and animal taxa at the same site enables analyses of species interactions

Data Management

- Data submitted via alternate interfaces (e.g., mobile apps) are tagged as such with associated metadata about the interface
- Data integrated from other programs are tagged as such with associated metadata about methodology and interface
- Field-level validation practices on priority are undertaken
- *A holistic approach to managing training/test data is undertaken*

Options for Data End Users

- Observer contact information enables NCO-mediated follow up from data end users on outliers or biases in the data
- Data end users can select data to fit their criteria:
  - Collected by trained volunteers with a particular program (e.g. Signs of the Seasons)
  - Sites where plants were not watered or fertilized; or where no water or feeding stations were available for animals
  - Collected on a weekly basis, or more frequently
- Summarized data enables several additional quality control features
  - Ready exclusion of positive phenophase reports not preceded or followed by a negative report
  - Ready calculation of uncertainty in onset or end date of a phenophase
  - Dates are available in Julian format, to enable ready calculation of phenophase duration across calendar years

This document was developed by Alyssa Rosemartin and Ellen Denny with internal reviews by Carolyn Enquist, R. Lee Marsh and Jake Weltzin. External reviews provided by David Moore and Andrea Wiggins.

For more information, please contact:
USA National Phenology Network National Coordinating Office
1955 E. Sixth St.
Tucson, AZ 85721
(520) 792-0481
nco@usanpn.org
www.usanpn.org