Data Quality Assurance & Quality Control for the National Phenology Database

The primary source of observational plant and animal data in the National Phenology Database (NPDb) 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 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. To maximize data quality and utility, the USA-NPN implemented a robust program, following best practices for participant engagement (Crimmins et al, In Press). Our monitoring methods also ensure data quality, for example, individual plants are tracked through time, to control 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. See Denny et al (2014) for full description of monitoring approach and protocols. Our data management system ensures data quality and integrity, including field-level validation, and tagging records by source (e.g., integrated datasets, mobile application). See Rosemartin et al (2014) for more information on our information management system.

1. Data quality implementation in the USA-NPN's Information Management System

Here, we describe the USA-NPN's suite of quality assurance (QA; before data enters database) and quality control (QC; post-processing and flagging) measures. These measures are applied in the *Nature's Notebook* interfaces and in the National Phenology Database, which hosts all *Nature's Notebook* data, as well as several integrated datasets, collected by external organizations or individuals. Measures completed to date (black text), and planned or proposed (grey italic text) are summarized below. Annual results of quality control checks are detailed in Appendix I and II.

Quality Assurance Measures

Species Identification

- "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.
- Messages in the interface ask users to confirm species identification when reported out of range.

Phenophase status evaluation

- Date of each visit is reported along with presence/absence of each phenophase (e.g. "open flowers"); observers are not asked to infer the date of an event (e.g. date of "first flower").
- Observers report presence/absence with a simple "Yes"/"No" and are given an "Uncertain" option to reduce tendency for reporting false positives or negatives.
- Phenophase definitions are written carefully for precision and accessibility.
- Phenophase definitions are generalized and identical across similar species (within phenological functional types) for consistency. Species-specific additions to the general definitions more completely describe how the phenophase appears in a particular species.
- FAQs address tricky issues in phenophase status evaluation.
- National webinars and photographic primers teach plant anatomy and phenophase evaluation skills.
- Photos or illustrations for each phenophase for each species are provided to observers.

Data Entry

- Datasheets (PDF and Excel templates) mirror the online data entry form.
- Observers can reorder plant and animal lists so data entry form and datasheet printout mirrors order encountered at the field site.
- Species names and intensity measures are presented as pick lists, limiting the possible responses the user can provide.
- Phenophase and intensity definitions appear on roll-overs in the data entry form.
- Site latitude and longitude can be provided a number of ways: the user can place a marker on a Google Map; the user can geocode the location based on an address; the user can manually provide a latitude and longitude. In every case, reverse geocoding is attempted as well to resolve the location's U.S. State.
- Elevation is calculated using Google Elevation Services, but can be modified by the user. In the case that more than one value is thus provided, both values are sustained in the database.
- Observers can review previously submitted observations in user interfaces (on data entry screens and via calendars) or via Excel file, and can edit these observations.
- Usability testing has been conducted on user interface to increase intuitiveness and reduce transcription errors.
- When a plant is deleted, rationale for deletion is requested and the deleted plant data is retained, if appropriate.
- Mobile applications and an MS Excel template for data collection eliminate datasheet-to-interface transcription errors.

- Training and FAQs address common data entry issues.
- User interface validation measures:
    - Controlled values are used whenever feasible (for time, date, phenophase status and categorical intensity measures).
    - The list of phenophases for each species is populated for users, only allowing that user to provide data for phenophases relevant to the species for which they are entering data.
    - User may not enter abundance or intensity measure unless the phenophase is set to "Yes" or "Uncertain".
    - Date field required; default is to select from a calendar.
    - Dates in the future not allowed.
    - Duplicate date and time values not allowed.
    - Users entering past observations are informed of changes in phenophase names and definitions through time. Note that changes through time are minimal, but were a necessary part of early field-testing.

Training

- Standardized field observing methods (selecting a site, selecting species, making observations) are accessible via an online learning module, 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 has been submitted by experts or trained volunteers).
- Local phenology leaders are recruited and supported by NCO to provide on-the-ground training and ongoing in-person support for observers.
- Site leader certification program, including advanced training materials, to ensure volunteer participants are calibrated and sites are established according to USA-NPN guidelines.

Quality Control Measures

Validation [correctness of values relative to defined rules]

- Annual process leverages Google's Reverse Geocoder, revealing coordinates outside the United States, Canada or Mexico, which are manually reviewed and corrected in clear cases (e.g., inverted +/- signs).
- Nightly process corrects elevation using Digital Elevation Models based on coordinates; user-defined elevation is retained in the database.
- Nightly process identifies records identical other than a conflicting phenophase status (i.e.,

identical timestamp), and deletes all records since correct status is unknown (this issue affects 0.02% of data).
- Phenophase evaluation is confirmed via submission of photo with observation (with crowd-sourced review of images and expert confirmation on an image subset).

Reliability [dataset completeness; freedom from error and bias]

- Nightly process flags instances where conflicting phenophase status records are present on the same date, regardless of time of day, for the same individual plant or species of animal at a site (submitted either by one observer or multiple observers).
- Summarized data provides information on temporal precision, with the number of days between a "Yes" and the preceding or following "No" report for a phenophase onset or end.
- 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 ancillary data.

Plausibility [likely accuracy of the values]

- Data users can explore outliers using ancillary data at the site, plant and observation level (e.g., land cover type for sites, watered status for plants).
- Data users can explore outliers using comments at the site, plant and observation level, submitted by observers.
- Observers provide contact information to enable NCO-mediated communication regarding outliers or other issues.
- Self-reporting of training, skill and experience level by observers made available to data users (18% of active observers report these features), enabling exploration of outliers and pre-filtering of data.
- Nightly process identifies and flags species reported outside of known range.
- Nightly process identifies and flags phenophases reported in implausible order.
- Nightly process identifies and flags implausible changes in step magnitude for intensity measures.
- Annual process flags outliers in phenophase onset and end dates relative to distribution of records that are similar geographically, climatically and/or taxonomically.

2. Evaluation of data quality to date in NPDb In addition to the above quality control measures, which are applied to the dataset as a whole, several efforts have looked at subsets of the data, and found low error rates.

a. Reliability

- Comparisons of observation data from experts and trained and untrained observers at the same site:
    a. 91% concordance between trained observer and expert (Fuccillo et al. 2014)
    b. An observer inter-comparison project is underway at Acadia National Park considering expert, trained, untrained observers and their improvement over time.

b. Plausibility

- Across 11 maple, oak and poplar campaign species, on average 2.85% of records are for individual trees located outside their known distributions (SD: 3.25%, range: 0.2% for red maple to 10% for balsam poplar) (Clauser, 2016).
- In the integrated lilac and honeysuckle dataset, 1.46% of first leaf or bloom records were flagged as outliers, within an individual plant's period of record (for plants with at least 10 years of data) (Rosemartin et al., 2015).
- In the historic lilac and honeysuckle dataset 2.09% of first leaf or bloom records were flagged as outliers, relative to other onsets in similar geographic and climatic contexts (Medhipoor et al., 2015).

This document was developed by Alyssa Rosemartin, Ellen Denny and Kathy Gerst with internal reviews by LoriAnne Barnett, Carolyn Enquist, R. Lee Marsh, Erin Posthumus and Jake Weltzin. External reviews of Version 1 were provided by David Moore and Andrea Wiggins.

For more information, please contact:
USA National Phenology Network National Coordinating Office

1311 East 4th Street, Tucson, AZ 85721
(520) 621-1740
nco@usanpn.org
www.usanpn.org

References

Crimmins, T. M., Barnett, L., Denny, E. G., Rosemartin, A. H., Weltzin, J. F., & Schaffer, S. (In Press). From tiny acorns grow mighty oaks: what we've learned from nurturing Nature's Notebook. In Handbook of Citizen Science in Ecology and Conservation. University of California Press.

Clauser, K. 2016. Analyzing Spatial Trends in the *Nature's Notebook* Dataset. [Master's Thesis; Online version pending]

Denny, E., Gerst, K.L., Miller-Rushing, A.J., Tierney, G.L., Crimmins, T.M., Enquist, C.A., Guertin, P., Rosemartin, A.H., Schwartz, M.D., Thomas, K.A., & Weltzin, J.F.(2013). Standardized phenology monitoring methods to track plant and animal activity for science and resource management applications. International Journal of Biometeorology, 1-11.

Fuccillo, K. K., T. M. Crimmins, C. E. de Rivera, and T. S. Elder. 2014. Assessing accuracy in citizen science-based plant phenology monitoring. International Journal of Biometeorology:1-10.

Mehdipoor H, Zurita-Milla R, Rosemartin A, Gerst KL, Weltzin JF. Developing a Workflow to Identify Inconsistencies in Volunteered Geographic Information: A Phenological Case Study. PLoS ONE. 2015;10(10):e0140811.

Rosemartin AH, Denny EG, Weltzin JF, Lee Marsh R, Wilson BE, Mehdipoor H, Zurita-Milla R and Schwartz, MD. Lilac and honeysuckle phenology data 1956–2014. Sci Data. 2015;2:150038.

Rosemartin, A.H., Crimmins, T.M., Enquist, C.A., Gerst, K.L., Kellermann, J.L., Posthumus, E.E., Denny, E., Guertin, P., Marsh, L., & Weltzin, J.F. (2014). Organizing phenological data resources to inform natural resource conservation. Biological Conservation, 173: 90-97. http://dx.doi.org/10.1016/j.biocon.2013.07.003

Appendix I - Quality Control Report for 2015

Issue	Action taken	Level at which action is taken	Percent of records affected
Validation			
Elevation	Corrected in an separate field with DEM derived value; user value maintained	Site	4.53%
Coordinates	Manual review, correction or flag as exception	Site	0.03%
Reliability			
One-observer status	Flagged	Record	0.05%
Multi-observer status	Flagged	Record	1.72%
Temporal precision			
Onsets without a prior no	None	Onset	23%
"No" < 8 days	None	Onset	52%
"No" in 8-14 days	None	Onset	13%
"No" in 15-30 days	None	Onset	7%
"No" in > 30 days		Onset	6%
Ends without a following no	None	End	28%
"No" < 8 days	None	End	49%
"No" in 8-14 days	None	End	12%
"No" in 15-30 days	None	End	6%
"No" in > 30 days	None	End	5%

Appendix II – Temporal Precision Information 2009-2015