New Phytologist

Research

Methods

CrowdCurio: an online crowdsourcing platform to facilitate climate change studies using herbarium specimens

Charles G. Willis[1], Edith Law[2], Alex C. Williams[2], Brian F. Franzone[1], Rebecca Bernardos[1], Lian Bruno[1], Claire Hopkins[1], Christian Schorn[1], Ella Weber[1], Daniel S. Park[1] and Charles C. Davis[1]

[1]Department of Organismic and Evolutionary Biology and Harvard University Herbaria, Harvard University, Cambridge, MA 20138, USA; [2]David R. Cheriton School of Computer Science, University of Waterloo, Waterloo, ON N2L 3G1, Canada

Authors for correspondence:
Charles G. Willis
Tel: +1 617 495 2365
Email: charleswillis@fas.harvard.edu

Charles C. Davis
Tel: +1 617 496 0515
Email: cdavis@oeb.harvard.edu

Received: *22 November 2016*
Accepted: *20 February 2017*

New Phytologist (2017)
doi: 10.1111/nph.14535

Key words: citizen science, flowering, fruiting, phenological sensitivity, phenology, phenophase.

Summary

- Phenology is a key aspect of plant success. Recent research has demonstrated that herbarium specimens can provide important information on plant phenology. Massive digitization efforts have the potential to greatly expand herbarium-based phenological research, but also pose a serious challenge regarding efficient data collection.
- Here, we introduce *CrowdCurio*, a crowdsourcing tool for the collection of phenological data from herbarium specimens. We test its utility by having workers collect phenological data (number of flower buds, open flowers and fruits) from specimens of two common New England (USA) species: *Chelidonium majus* and *Vaccinium angustifolium*. We assess the reliability of using nonexpert workers (i.e. Amazon Mechanical Turk) against expert workers. We also use these data to estimate the phenological sensitivity to temperature for both species across multiple phenophases.
- We found no difference in the data quality of nonexperts and experts. Nonexperts, however, were a more efficient way of collecting more data at lower cost. We also found that phenological sensitivity varied across both species and phenophases.
- Our study demonstrates the utility of *CrowdCurio* as a crowdsourcing tool for the collection of phenological data from herbarium specimens. Furthermore, our results highlight the insight gained from collecting large amounts of phenological data to estimate multiple phenophases.

Introduction

Over the past century, climate change has had a significant impact on plant phenology – the timing of life history events – across the globe (Walther, 2004; Menzel *et al.*, 2006; Parmesan, 2006; Cleland *et al.*, 2007; Miller-Rushing *et al.*, 2007; Chambers *et al.*, 2013). Importantly, the inability of species to respond phenologically to climate change has been found to have dire consequences for species survival and community diversity (Inouye, 2008; Møller *et al.*, 2008; Willis *et al.*, 2008; Caradonna *et al.*, 2014). However, the long-term and historical datasets necessary to identify the historical influence of climate change on phenology remain relatively scarce, even for regions where the biota is well characterized and associated historical climate records are available. Moreover, most datasets of this nature show a strong geographical and taxonomic bias: they are largely from temperate regions, mostly include a small subset of species within these communities (e.g. dominant woody species), and do not sample the variation in phenological response across a species' range (Wolkovich *et al.*, 2014).

Herbaria, which house the best record of where plants live today and have lived in the past, represent a largely untapped resource for investigating phenological responses to climate change despite their obvious relevance to this question (Vellend *et al.*, 2013). In particular, observations from numerous specimens collected at multiple locations through time can allow us to determine if a given species has altered its phenology in relation to climate. This appears to be changing, however, as recent investigations have used herbarium specimens to study the impacts of climate change (Primack *et al.*, 2004; Miller-Rushing *et al.*, 2006; Robbirt *et al.*, 2011; Panchen *et al.*, 2012; Zohner & Renner, 2014), and such efforts have expanded dramatically to investigate phenology across large numbers of species and vast geographical areas (Calinger *et al.*, 2013; Everill *et al.*, 2014; Park & Schwartz, 2015). Moreover, estimates of phenological response to climate change inferred from herbarium records are similar to those observed with field observations, validating their use for this purpose (Robbirt *et al.*, 2011; Davis *et al.*, 2015; Spellman & Mulder, 2016). And importantly, herbarium records can greatly expand our knowledge of phenology across a broader sampling of taxonomy, geography and climate variability than is typically available from historic field observational studies (Davis *et al.*,

2015). Thus, herbaria offer the potential to greatly expand our understanding of phenology across space, time and taxa.

To date, studies that have utilized herbarium specimens to measure phenology represent only the tip of the iceberg in terms of the phenological data available in herbaria around the world (C. G. Willis, *et al.*, in press). Most herbarium records remain inaccessible to the larger scientific community. Accessing these data represents a grand challenge faced by biodiversity scientists in the era of Big Data (Arino *et al.*, 2010; Tewksbury *et al.*, 2014). Typically, researchers have collected phenological data from individual herbarium specimens by hand, assessing each physical specimen manually. Such detailed and laborious work, often performed by small teams, limits the amount of data that can be gleaned because researchers often lack the time and resources to travel to multiple herbaria and score potentially tens of thousands of specimens.

In order to increase the accessibility of herbaria collections, there have been numerous calls for their digitization (i.e. capturing specimen-level metadata images in digital form) and online mobilization. In response, new methods and workflows in high-throughput imaging and digitization are being innovated to rapidly and efficiently create a virtual global herbarium that is readily accessible to the world. Millions of digitized specimen images already are available online (e.g. http://www.gbif.org; http://portal.idigbio.org) and recent federally funded efforts have mobilized herbaria from entire regions. Examples include efforts from Australia (http://avh.chah.org.au), France (http://science.mnhn.fr), South Africa (http://www.sanbi.org) and, most recently, the New England region of the United States (broadly defined as the states of Maine (ME), Vermont (VT), New Hampshire (NH), Massachusetts (MA), Rhode Island (RI), Connecticut (CT) and New York (NY)). For New England alone, over 500 000 virtual herbarium specimens are now accessible through the Consortium of Northeastern Herbaria portal (http://neherbaria.org/; for details, see http://nevp.org/resources).

As collections become increasingly decentralized and made available online, the need to develop digital infrastructure including tools to generate research outcomes will increase. Many tasks related to mining digital specimens for relevant data (e.g. identifying complex and variable floral structures) are currently too difficult to automate via machine learning, and thus require human labor. To this end, one particularly promising approach to leverage these virtual collections for research is a 'citizen science' approach that enlists members of the public to process digital specimens (Ellwood *et al.*, 2015). One of the most successful venues for crowdsourcing scientific research to date is Zooniverse (https://www.zooniverse.org), which, at the time of writing this paper (November 2016), hosts 46 crowdsourced projects ranging from mapping the Milky Way (Lintott *et al.*, 2008) to annotating ancient fragments of Greek papyrus (Williams *et al.*, 2014). With regard to specimen data, a long-running Zooniverse project is Notes from Nature (www.notesfromnature.org, Hill *et al.*, 2012), which aims to capture specimen data from hand-written labels. Crowdsourcing has also been applied to record and track present-day phenological observations (Nature's Notebook, https://www.usanpn.org/; Phenocam, Kosmala *et al.*, 2016) and species distributions (http://www.inaturalist.org/). However, crowdsourcing has not been exploited to gather information from herbarium specimens to assess the effects of climate change research on plant phenology until recently.

New England and the recent digitization efforts of its regional herbaria through the NEVP offer an ideal test case to develop and test the tools necessary to collect phenological information from specimen data rapidly and efficiently. New England is one of the most intensively studied regions with regard to climate change and has experienced warmer annual temperatures, earlier springs, longer summers and shorter winters over the last 200 yr (Horton *et al.*, 2014). These seasonal changes have had profound effects on the phenology of the New England flora. For instance, leaf-out dates for deciduous forests in the region have advanced by up to 10 d (Richardson *et al.*, 2006). Similarly, spring flowering times have advanced, on average, by 2 wk (Miller-Rushing & Primack, 2008; Willis *et al.*, 2010; Ellwood *et al.*, 2013).

Our study has three main goals. First, we introduce a new crowdsourcing image annotation tool developed on the online *CrowdCurio* platform (https://www.crowdcurio.com/; Fig. 1) to collect phenological data from herbarium specimens. More generally, *CrowdCurio* allows researchers to create and manage crowdsourcing projects that are tailored to their specific questions (Law *et al.*, 2013). We developed a *CrowdCurio* project, titled 'Thoreau's Field Notes', to crowdsource the scoring of three phenological traits (number of flower buds, flowers and fruits) using digitized herbarium specimens of two common New England species: Greater celandine (*Chelidonium majus* L.) and Lowbush blueberry (*Vaccinium angustifolium* Aiton). Second, we assess the reliability of expert vs nonexpert data using this tool, and describe our methods of quality control to identify outlier data, which is essential to the robustness and downstream data utility of any crowdsourcing effort. Third, in an attempt to capture a more comprehensive picture of reproductive life history, we crowdsource specimens from our two focal species (*C. majus* and *V. angustifolium*) for three reproductively relevant phenological traits: flower buds, open flowers and fruits. We then use these data to analyze the phenological sensitivity of multiple phenophases to interannual temperature variation (i.e. first flowering day, peak flowering day, first fruiting day and peak fruiting day). Finally, we demonstrate the promise of both crowdsourcing and the *CrowdCurio* platform as tools for assessing phenophases across an entire season rather than for one event within that season (e.g. first flowering date).

Materials and Methods

Crowdsourcing phenological data collection

The phenological state, or phenophase, of a herbarium specimen is based on the presence and quantity of relevant phenological traits (e.g. leaf buds, flowers, fruits). Typically, researchers have focused on the presence or absence of a single trait or structure (e.g. flowers) for investigating a single phenophase (e.g. first flower day). To estimate multiple phenophases, we quantified data for three reproductively relevant phenological traits: flower buds, open flowers and fruits. For each digital specimen image, workers were asked to count the number of each of these traits.

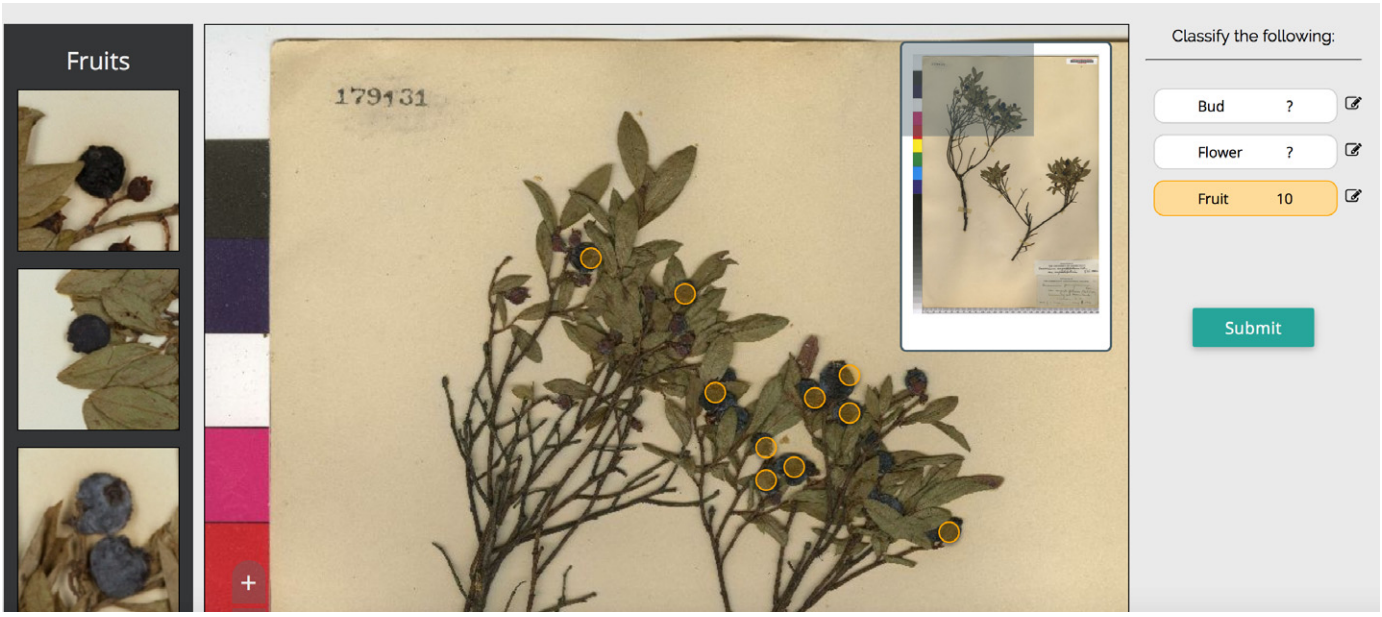

Fig. 1 Annotation interface of *CrowdCurio* for collecting phenological data from digitized herbarium specimens. Phenological data include counts of flower buds, open flowers and fruits. Crowdsource workers score each data type by clicking on the presence of corresponding objects on the image (orange circles). Workers click the image to add and remove points. Workers also are provided with examples of each data type on the left. Object landmarks for each data type are color-coded. Workers can traverse the image either by clicking and dragging the image itself or with the subset image in the upper right corner. The specimen pictured is of *Vaccinium angustifolium* (lowbush blueberry).

We employed 'experts' and 'nonexperts' to crowdsource the collection of phenological data. Our 'expert' pool consisted of four Harvard University Herbaria curatorial staff each with familiarity with herbarium specimens and botanical terminology (co-authors R.B., L.B., C.H., C.S. and E.W. participated as experts, although one was removed due to a technical error). Each worker in the expert pool was asked to score all three phenological traits for a subset of the 820 images. The median number of images scored per expert worker was 635 (range: 624–666). The median number of expert workers that scored each image was 3 (range: 1–6).

The 'nonexpert' pool comprised 270 anonymous workers hired through Amazon's Mechanical Turk service (MTurk; https://www.mturk.com/). Nonexpert workers were asked to score all three phenological structures for a set of 10 individual images. Each set contained a single, randomly selected duplicate image (i.e. nine unique images) to assess the quality of workers in terms of their consistency in producing the same output twice, a metric we refer to as repeatability error (A. C. Williams *et al.*, unpublished). Nonexpert workers were compensated for their participation at the rate of $0.10 per image plus a $0.15 base participation rate for each set of images. The median number of images scored per nonexpert worker was 19 (range: 1–23). The median number of nonexpert workers who scored each image was 5 (range: 5–7). The use and collection of data by nonexperts was reviewed and approved by an ethics review committee at the University of Waterloo.

Before beginning their data collection, expert and nonexpert workers were required to watch a short (*c.* 1 min) instructional video on how to collect data on *CrowdCurio*. Both sets of workers also were provided example images for each phenological trait for each species. For both experts and nonexperts, we also recorded the time duration required to score each image.

Specimen data

Chelidonium majus and *Vaccinium angustifolium* were selected to represent species with contrasting life histories. *Chelidonium majus* is a biennial herb that is invasive in New England, whereas *V. angustifolium* is a perennial shrub that is native to New England. Both species flower at similar times during spring, and fruit through summer and fall.

We assembled a dataset of 820 digital herbarium specimen images (139 for *C. majus* (1848–2012) and 681 for *V. angustifolium* (1823–2002); Supporting Information Table S1) from across New England (Fig. 2). Images were compiled from the digital specimen collections of the following herbaria: Harvard University Herbaria (A, AMES, ECON, GH and NEBC), the University of New Hampshire Hodgdon Herbarium (NHA), the University of Connecticut George Safford Torrey Herbarium (CONN) and the Yale University Herbarium (YU). These images are available via the Consortium of Northeastern Herbaria web portal (http://portal.neherbaria.org/).

Of the 820 specimens in our dataset, only 149 included geospatial data (latitude, longitude). For the remainder of the specimens, we obtained geospatial data based on the location of the nearest municipality implemented with the 'GEOCODE' function in the GGMAP library (Kahle & Wickham, 2013) using R v.3.2.2 (R Core Team, 2015). We obtained geospatial data for an additional 669 specimens (818 in total; Fig. 2).

Comparison of expert and nonexpert workers

We evaluated the quality of nonexpert crowdsourced vs expert count data using three separate analyses.

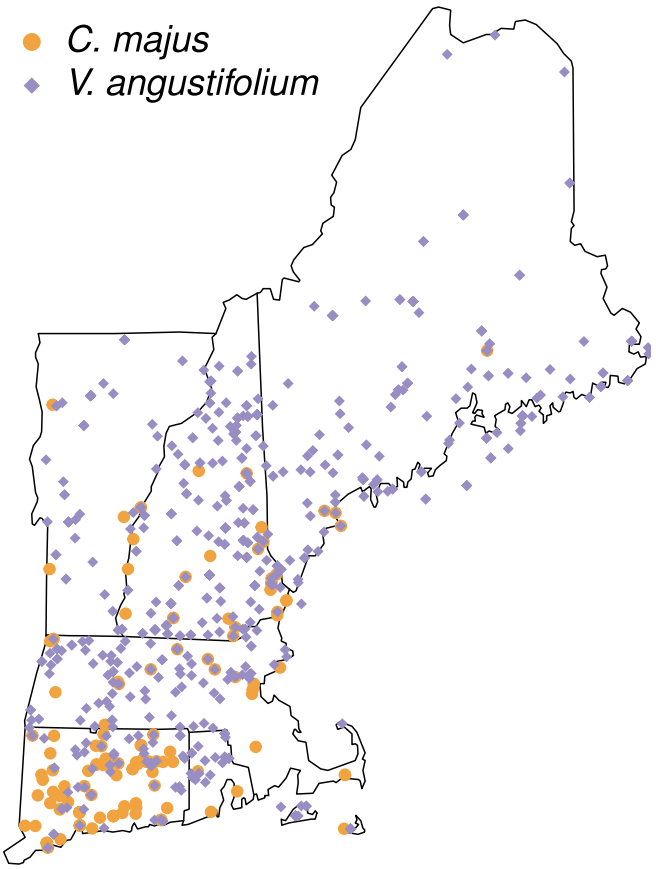

Fig. 2 Distribution map of herbarium specimens for two common New England species investigated in this study: *Chelidonium majus* (celandine) and *Vaccinium angustifolium* (lowbush blueberry).

First, we tested for differences in data collector sampling error between expert and nonexpert workers. Collector sampling error was calculated as the *t*-distribution 95% confidence interval for each pair of duplicate images scored by a worker. We used a linear-mixed model to test for differences between experts and nonexperts, where sampling error was the dependent variable; species, worker type and their interaction were fixed effects, and worker ID was a random effect nested in worker type. Models were implemented in R using the 'LMER' function in the LME4 library (Bates *et al.*, 2015). Individual models were fitted for each phenological trait.

Second, as a measure of nonexpert accuracy, we compared nonexpert consensus estimates to expert consensus estimates. Consensus estimates of each phenological trait per specimen were calculated as the median count of all workers within a worker type. In a previous study, we found the median consensus to be more reliable than majority vote consensus, as it was less subject to bias toward zero counts (A. C. Williams *et al.*, unpublished). We compared median consensus estimates among worker types using an analysis of covariance that accounted for the effects of species, trait and the interactions between all independent variables.

Third, we compared consensus error estimates, calculated as the difference between the individual worker count and the specimen consensus, between expert and nonexperts using a linear-mixed model. Individual consensus error was treated as the dependent variable, whereas species, worker type and their interaction were fixed effects, and worker ID was a random effect nested in worker type.

Estimation of phenophases

We focused on four important phenophases: first flowering date (FD_{flr}), peak flowering date (PD_{flr}), first fruiting date (FD_{frt}) and peak fruiting date (PD_{frt}). Specimens were scored for each phenophase based on the relative proportion of flower buds, open flowers and fruits, calculated from combined expert and nonexpert consensus counts. First flowering date was scored as < 50% flowers, at least one flower bud or flower, and 0% fruits (*N* specimens: *Chelidonium* = 16, *Vaccinium* = 59). Peak flowering date was scored as ≥ 50% flowers of total count (*N* specimens: *Chelidonium* = 3, *Vaccinium* = 206). First fruiting date was scored as < 50% fruits, with at least one fruit present (*N* specimens: *Chelidonium* = 37, *Vaccinium* = 31). Peak fruiting date was scored as ≥ 50% fruits (*N* specimens: *Chelidonium* = 73, *Vaccinium* = 332). Eight of the specimens (all *Vaccinium*) met our criteria for both peak flowering *and* first fruiting, as defined above. Where analyses included direct comparisons of phenophases (e.g. did phenological sensitivity differ across phenophases?), we ran two separate models with all eight specimens coded in one or the other phenophase. For analyses that examined phenophases separately (e.g. individual estimates of phenological sensitivity), we included all eight specimens in both analyses of peak flowering and first fruiting. To avoid potentially spurious collection dates, we also removed specimens that had collection dates after 31 October (*N* = 2), as the lateness in the season made the collection date suspect. Mean dates for all four phenophases for both species are available in Fig. S1.

Historical climate data

Historical temperature data from New England were obtained from the Applied Climate Information System (ACIS; http://www.rccacis.org/). We used a scipy spatial cKDTree algorithm implemented in PYTHON to match each specimen, based on its geospatial location, to the closest weather station within a 25-km radius (https://github.com/Bouteloua/climate_data_fetcher). If weather data were available for the year in which the specimen was collected, mean monthly temperatures for all 12 months were returned. Of the 820 specimens in our dataset, we obtained temperature data for 414 specimens (93 of *C. majus*, 321 of *V. angustifolium*).

Analysis of phenological sensitivity

In order to estimate phenological sensitivity (day of year $°C^{-1}$), we used a multivariate linear regression model that included the phenophase timing (i.e. specimen collection date (day of year)) as the dependent variable plus spring temperature, latitude, longitude and collection year. This model was run independently for each species. We defined spring temperature as the mean

monthly temperature averaged across March, April and May. We used the same temperature metric for both species and across all phenophases to standardize comparisons.

In addition to testing for phenological sensitivity with a full standardized model, we also determined the best-fit combination of predictor variables for each phenophase using stepwise AIC model comparison with the 'stepAIC' function in the R library MASS (Venables & Ripley, 2002).

Results

Comparison of expert vs nonexpert workers

Experts were significantly more efficient at counting phenological traits on a per specimen basis, processing specimens, on average, *c.* 0.80 min faster than nonexperts (mean processing time per specimen in minutes ± SE: experts = 1.40 ± 0.08; nonexperts = 2.20 ± 0.06; *t*-test: $t = 8.07$, df = 4952.6, $P < 0.001$). Collectively, 270 nonexperts processed 4197 specimen images (including duplicates) over the course of *c.* 7 d, with each individual participant working, on average, 0.6 h (153.6 total hours). By contrast, experts processed 2560 specimen images over *c.* 4 d, with each participant working, on average, 14.9 h (59.6 total hours for four experts). In terms of overall cost, nonexpert data collection was significantly less expensive than expert data collection. MTurk costs totaled US$692.40 (4197 images × $0.10 + 270 workers × $0.15 baseline fee + 270 assignments × $0.86 MTurk assignment fee), whereas expert costs total US $2048.45 (59.6 work hours × $34.37 prorated hourly rate). The per image cost was $0.16 for nonexperts and $0.80 for experts.

Although nonexperts tended to have larger sampling error rates in comparison to experts, these differences were not statistically significant (lsmean and contrast of sampling error by worker type: experts = 8.1 ± 7.6, nonexperts = 11.4 ± 1.5, $t = -0.43$, df = 100.3, $P = 0.669$; Table S1). Species identity had a significant effect on sampling error, however, with larger error rates associated with *Vaccinium* (lsmean sampling error ± SE across all phenological traits: *Vaccinium* = 13.1 ± 3.8, *Chelidonium* = 6.4 ± 4.1, $t = -3.95$, df = 1600.5, $P < 0.001$; Table S1). There was a significant interaction between worker type and phenological trait (Table S1), but it was not driven by trait-specific differences between experts and nonexperts (Tables S2, S3). Rather, the significant interaction between worker type and phenological trait was the result of differences in sampling error rates among traits within worker type (Table S3). Furthermore, expert and nonexpert consensus estimates were highly correlated, indicating their similarity (Pearson's correlation coefficient – flower buds: $r = 0.89$, $P < 0.001$; flowers: $r = 0.94$, $P < 0.001$; fruits: $r = 0.95$, $P < 0.001$; Fig. 3).

Consensus error did not differ significantly between experts and nonexpert workers (Table S4). Furthermore, differences in consensus error among experts vs nonexperts did not depend on species or trait, as indicated by the lack of significant interaction between type × species and type × trait (Table S4). Consensus error, rather, was influenced by attributes of the species and phenological traits, and significant effects of both variables were detected (Table S4). Overall, consensus error was larger in *Vaccinium* in general, but especially in *Vaccinium* fruit counts (Table S5).

Phenophases and climate sensitivity

Phenological sensitivity to spring temperature differed significantly across phenophases, as indicated by a significant interaction term between phenophase and spring temperature in our

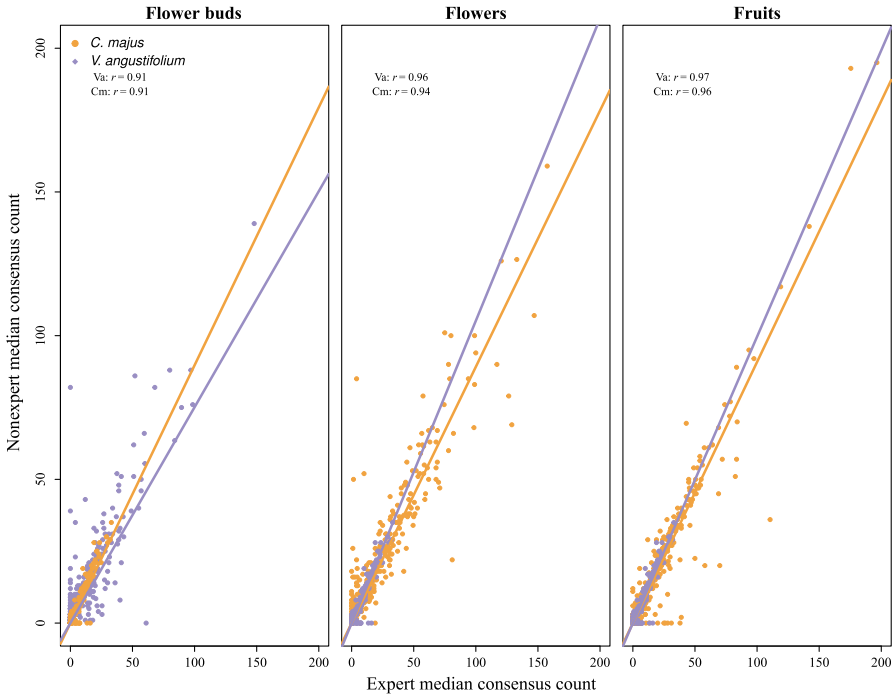

Fig. 3 Comparison of expert and nonexpert consensus counts for each phenological trait. Consensus was calculated as the median count of all worker estimates within worker type (expert vs nonexpert). Pearson correlation coefficients (*r*) indicate the degree of similarity between expert and nonexpert consensus estimates for each species (Va, *Vaccinium angustifolium*; Cm, *Chelidonium majus*).

ANCOVA model (Table S6). Independent analyses of each phenophase found that PD_{flr} and PD_{frt} significantly advanced under warmer spring temperatures (-1.93 d °C^{-1} and -3.34 d °C^{-1}, respectively; Table S7; Fig. 4). Phenological sensitivity of PD_{frt} did not differ significantly among species, however (Table S7). We were unable to test for differences between species regarding the phenological sensitivity of PD_{flr} because of insufficient data for *Chelidonium*.

Stepwise AIC analysis of the best-fit linear model found distinctly different sets of predictor variables for each phenophase and genus, which in certain instances excluded spring temperature entirely (Table S9). The best-fit model did not find spring temperature to be significant predictor for FD_{flr} for either species (Table S9). Rather, the best predictor of FD_{flr} in *Vaccinium* was longitude (more easterly specimens flower earlier), whereas for *Chelidonium* it was year (more recent specimens flower earlier) (Table S10). The best-fit model for PD_{flr}, limited to *Vaccinium*, did include spring temperature, as well as the interaction between spring temperature and latitude, such that both of these terms were marginally significant (Table S10). The best-fit model for FD_{frt} in *Vaccinium* included spring temperature (Table S10), of which only latitude, longitude and their interaction had significant effects, such that FD_{frt} tended to advance moving north and east (Table S10). In *Chelidonium*, the best-fit model for FD_{frt} included only year, which did not have a significant effect (Table S10). The best-fit model for PD_{frt} in *Vaccinium* included spring temperature, latitude, longitude and latitude × longitude (Table S10), of which only spring temperature had a significant effect, such that PD_{frt} advanced with warmer spring temperatures (Table S10). Finally, the best-fit model for PD_{frt} in *Chelidonium* included year and longitude, neither of which had a significant effect (Table S10).

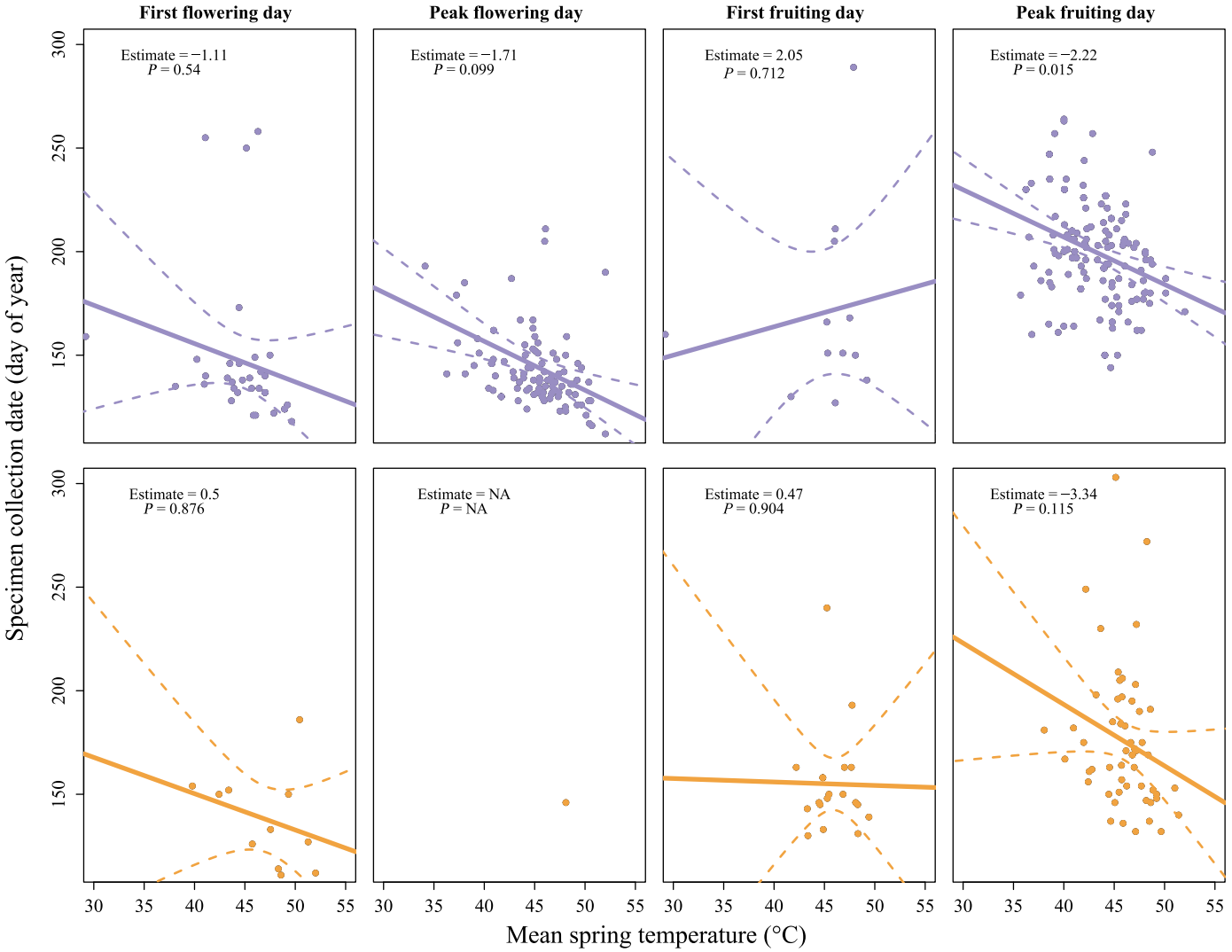

Fig. 4 Phenological sensitivity to spring temperature (d °C^{-1}) for four phenophases of two common New England species. Phenological sensitivity of *Vaccinium angustifolium* is on the upper row, *Chelidonium majus* is on the lower row. Spring temperature is defined as mean monthly temperature of March, April and May for a given year. Bold lines represent regression slope estimates from the 'full' linear model (see Supporting Information Table S8). Dashed lines indicate the 95% confidence interval of slope estimates.

Discussion

We draw several important conclusions from our study. First, our study demonstrates the reliability of nonexpert crowdsourcing to collect large amounts of accurate phenological data from herbarium records. Second, our study highlights the general advantages of the *CrowdCurio* platform, not only for herbarium research, but academic research in general. Third, our study builds on previous herbarium-based phenological research by scoring multiple phenological traits per specimen, which allowed us to test phenological sensitivity to spring temperature across multiple phenophases. These results demonstrate that herbarium specimens can provide a more nuanced and comprehensive picture of the phenological sensitivity of species than previously thought.

Reliability of crowdsourced herbarium-based phenological data

Crowdsourcing has become an increasingly important tool in scientific data collection (Ellwood *et al.*, 2015). Here, we demonstrate crowdsourcing to be an effective and reliable tool for the collection of herbarium-based phenological data. Nonexpert worker consensus estimates were similar to those of trained experts. This is perhaps not surprising given that object-oriented tasks (e.g. counting, identifying image landmarks) typically have a low learning barrier. For example, Chang & Alfaro (2015) found nonexpert placement of geomorphological landmarks for digital images of fish to be equivalent to expert placement. Furthermore, nonexpert assessments have the potential to be improved through post-collection filtering. The tendency for nonexpert sampling and consensus error rates to be higher is likely due to presence of spammers among the nonexpert worker pool (e.g. workers who enter random numbers simply to complete the task). Although these higher error rates did not affect nonexpert consensus estimates, they do suggest potential for data improvement. Indeed, current efforts are underway to develop metrics to identify spammers and filter them from crowdsourcing datasets within the *CrowdCurio* platform (A. C. Williams *et al.*, unpublished).

Recently, there has been a marked increase in efforts to unlock the information housed in herbarium records. In particular, there has been an effort to scale-up the geographical and taxonomic coverage of herbarium-based phenological data, which requires the processing of thousands of specimens (Calinger *et al.*, 2013; Everill *et al.*, 2014; Park & Schwartz, 2015). In terms of size, our own study is rather small (Park & Schwartz, 2015). However, we were able to process all 820 specimens in the span of less than 1 wk with a small time commitment among each nonexpert participant (0.6 h) at a significantly reduced cost per image. This contrasts with the expert effort, which required each of the four full-time staff members to score specimens during their working week. The final cost for nonexpert data was a fifth of the cost of expert-collected data ($0.16 per image for nonexperts vs $0.80 per image for experts). Thus, crowdsourcing has the benefit of being able to cost-effectively score a specimen independently by multiple workers. Such cost-effectiveness is not without caveats, however. Serious questions remain about the ethical use of paid crowdsourcing in an unregulated labor market (Fort *et al.*, 2011). Scientists should consider these issues when contemplating the use of these tools. At $0.10 per image, a standard per task rate for MTurk, our nonworkers earned a salary of *c.* $3.14 h^{-1} (including baseline fees), which is below the US Federal minimum wage ($7.25 h^{-1}). Unfortunately, whether participants in our study were participating as a hobby, out of casual interest or as a means for generating income was not clear. Our ideal scenario would be to have these tasks performed voluntarily by curious and enthusiastic citizen scientists, a means of data collection that is rapidly expanding in the life sciences (Ellwood *et al.*, 2015), and one that we are actively pursuing with *CrowdCurio*.

Sampling or observation error, possibly due to fatigue or lapses in attention, is a common attribute among most ecological datasets based on count data, even when collected by trained experts. There is also an innate ambiguity in scoring phenological traits from herbarium specimens, wherein the specimen preservation may have obscured distinct trait attributes (e.g. overlapping structures). Estimates from multiple workers provide a means to account for such error and pinpoint areas of ambiguity.

CrowdCurio as a generalizable tool for crowdsourced research

Our results collectively highlight several advantages that *CrowdCurio* offers over existing crowdsourcing platforms. First, *CrowdCurio* follows a research-oriented crowdsourcing model that allows both academic researchers (e.g. ecologists) *and* human–computer interaction researchers to test hypotheses simultaneously. The advantage of this parallel model is the ability to improve both collection efficiency and data quality. For instance, in conjunction with the study presented here, we also used these data to validate a method of data quality control that assesses workers' reliability based on how consistently they count duplicate images (*Déjà vu*; A. C. Williams *et al.*, unpublished). Tools such as *Déjà vu* can then be integrated into data collection workflows to improve future data collection efforts and downstream analytics. More generally, the findings reported in both our study and A. C. Williams *et al.* (unpublished) are relevant not only in the context of plant ecology, but also to the growing body of crowdsourcing and citizen science literature. In a broader sense, *CrowdCurio's* model of scientific crowdsourcing creates a synergistic relationship between scientists, human–computer interaction researchers and members of the public, by allowing each constituent to better understand the critical role that one another plays in the others' work.

Second, one of the chief limitations of crowdsourcing the collection of herbarium specimen data – or for 'citizen science' research in general – is the technical barrier to entry. Researchers who have utilized crowdsourcing projects to date have either relied on collaborators who specialize in crowdsourcing programming (e.g. Zooniverse) or have developed the software a la carte (e.g. Chang & Alfaro, 2015). As a user-centered platform, *CrowdCurio* offers an alternative, exciting potential to greatly

expand citizen science-based research. *CrowdCurio* has been designed to allow researchers to set the full range of parameters for controlling and managing their own crowdsourcing project using a user-friendly interface and minimal technical assistance. This includes projects unrelated to phenology or simple object counts, such as estimating changes in quantitative trait measurements such as leaf shape (Buswell *et al.*, 2011). Such flexibility makes *CrowdCurio* a promising tool for integration into existing biodiversity database structures.

Expanding the purview of herbarium-based phenological research

Specific to the collection of herbarium-based data, the *CrowdCurio* image annotator stands apart from existing platforms in its ability to collect fine-scale phenological data. Existing approaches – either that utilize crowdsourcing (https://www.orchidobservers.org/) or that are used in-house as part of standard database entry (e.g. *Symbiota*; Gries *et al.*, 2014) – allow workers to classify specimens based on a list of predefined phenophases (e.g. flowering vs not flowering). This classification approach has advantages such as the ability to score a large number of specimens with relatively minimal effort. By contrast, our annotator is designed to follow a quantitative approach that offers a fine-grained assessment of specimen phenological traits. As we illustrate here, these fine-grained approaches (providing counts and locations of phenological traits) can provide a more nuanced characterization of phenophases, as well as characterization of multiple phenophases at once. Furthermore, these fine-scaled count and location data can be converted easily into pre-defined phenological classification schemes to match existing herbarium database structures.

In general, our results confirm the patterns observed across multiple herbarium-based studies of phenological sensitivity (C. G. Willis *et al.*, in press). First, we found both species to be phenologically sensitive to spring temperature, with peak flowering and fruiting advancing in warmer years. The magnitude of this sensitivity (*c.* 2–3 d $°C^{-1}$) also is consistent with previous studies that have estimated phenological sensitivity of other plant species in New England (Primack *et al.*, 2009; Ellwood *et al.*, 2013; Davis *et al.*, 2015). In addition, our study highlights the promise of a more detailed and integrated approach of scoring multiple phenological traits relevant to each species reproductive life history. Most studies, by contrast, investigate only one aspect of life history (e.g. leaf-out, flowering or fruiting; C. G. Willis, *et al.*, in press). By recording flower bud, flower and fruit numbers, we were able to quantify multiple distinct phenophases across a season (e.g. first flowering date vs peak flowering date). This approach revealed significant differences between the phenological sensitivity of different phenophases. Although *C. majus* was not sensitive to spring temperatures, *V. angustifolium* was, but only for peak flowering and fruiting date, which advanced significantly in warmer years. Our results are in keeping with observational studies that have observed similar variation in sensitivity to seasonal events across phenophases (Caradonna *et al.*, 2014).

In summary, our study demonstrates the significant potential of crowdsourcing platforms driven by members of the public for collecting large amounts of phenological data from herbarium specimens. The speed, efficiency and cost-effectiveness with which we collected data on two New England species using *CrowdCurio* suggests that crowdsourcing could be a key to unlocking the vast troves of data stored in the World's herbaria. Furthermore, our study also highlights the power of these historic records. Although herbarium specimens have obvious limitations (Davis *et al.*, 2015; C. G. Willis *et al.,* in press), they are still the most comprehensive historical record of plant phenology and biodiversity available. When aggregated across a geographical region, these data have the power to illuminate plant phenological behavior, such as sensitivity to temperature. Although the species that we used in this study were relatively well sampled across New England, future studies must also consider the potential biases inherent in the collection of herbarium specimens and how they might ultimately influence phenological patterns (Meyer *et al.*, 2016; B. H. Daru *et al.,* unpublished).

Acknowledgements

We would like to thank Anne Marie Countie and Michaela Schmull for their help on the project. This study was funded as part of the New England Vascular Plant Project (NSF-DBI: EF1208835) and part of a NSERC Discovery Grant (RGPIN-2015-04543).

Author contributions

C.G.W., E.L., A.C.W. and C.C.D. planned and designed the research questions and CrowdCurio annotation tool; E.L. and A.C.W. developed the CrowdCurio platform and implemented the experiment online; R.B., L.B., C.H., C.S. and E.W. participated in data collection on CrowdCurio; B.F.F. developed the climatefetcher tool; C.G.W. analyzed the data; C.G.W. and C.C.D. interpreted results; and C.G.W. and C.C.D. wrote the manuscript with significant input from E.L., A.C.W., D.S.P., R.B., L.B., C.H., C.S., E.W. and B.F.F.

References

Arino O, Perez JR, Kalogirou V, Defourny P, Achard F. 2010. Globcover 2009. *ESA Living Planet Symposium*: 1–3.

Bates D, Mächler M, Bolker B, Walker S. 2015. Fitting linear mixed-effects models using *lme4*. *Journal of Statistical Software* **67**: 1–48.

Buswell JM, Moles AT, Hartley S. 2011. Is rapid evolution common in introduced plant species? *Journal of Ecology* **99**: 214–224.

Calinger KM, Queenborough S, Curtis PS. 2013. Herbarium specimens reveal the footprint of climate change on flowering trends across north-central North America. *Ecology Letters* **16**: 1037–1044.

Caradonna PJ, Iler AM, Inouye DW. 2014. Shifts in flowering phenology reshape a subalpine plant community. *Proceedings of the National Academy of Sciences, USA* **111**: 1–6.

Chambers LE, Altwegg R, Barbraud C, Barnard P, Beaumont LJ, Crawford RJM, Durant JM, Hughes L, Keatley MR, Low M *et al.* 2013. Phenological changes in the southern hemisphere. *PLoS ONE* **8**: e75514.

Chang J, Alfaro ME. 2015. Crowdsourced geometric morphometrics enable rapid large-scale collection and analysis of phenotypic data. *Methods in Ecology and Evolution* **7**: 472–482.

Cleland E, Chuine I, Menzel A, Mooney H, Schwartz M. 2007. Shifting phenology in response to global change. *Trends in Ecology & Evolution* **22**: 357–365.

Davis CC, Willis CG, Connolly B, Kelly C, Ellison AM. 2015. Herbarium records are reliable sources of phenological change driven by climate and provide novel insights into species' phenological cueing mechanisms. *American Journal of Botany* **102**: 1599–1609.

Ellwood ER, Dunckel BA, Flemons P, Guralnick R, Nelson G, Newman G, Newman S, Paul D, Riccardi G, Rios N *et al.* 2015. Accelerating the digitization of biodiversity research specimens through online public participation. *BioScience* **65**: 383–396.

Ellwood ER, Temple SA, Primack RB, Bradley NL, Davis CC. 2013. Record-breaking early flowering in the eastern United States. *PLoS ONE* **8**: e53788.

Everill PH, Primack RB, Ellwood ER, Melaas EK. 2014. Determining past leaf-out times of New England's deciduous forests from herbarium specimens. *American Journal of Botany* **101**: 1–8.

Fort K, Adda G, Cohen KB. 2011. Last words Amazon Mechanical Turk: gold mine or coal mine? *Computational Linguistics* **37**: 413–420.

Gries C, Gilbert E, Franz N. 2014. Symbiota – a virtual platform for creating voucher-based biodiversity information communities. *Biodiversity Data Journal* **2**: e1114.

Hill A, Guralnickz R, Smith A, Sallans A, Gillespie R, Denslow M, Gross J, Murrell Z, Conyers T, Peter D *et al.* 2012. The notes from nature tool for unlocking biodiversity records from museum records through citizen science. In: Blagoderov V, Smith VS, eds. *No specimen left behind: mass digitization of natural history collections*. Sofia, Bulgaria: Pensoft, 219–233.

Horton R, Yohe GW, Easterling W, Kates R, Ruth M, Sussman E, Whelchel A, Wolfe D, Lipschultz F. 2014. Ch. 16: northeast. In: Melillo J, Richmond T, Yohe GW, eds. *Climate change impacts in the United States: the third national climate assessment*. Washington, DC, USA: US Global Change Research Program, 371–395.

Inouye D. 2008. Effects of climate change on phenology, frost damage, and floral abundance of montane wildflowers. *Ecology* **89**: 353–362.

Kahle D, Wickham H. 2013. *ggmap*: spatial visualization with ggplot2. *R Journal* **5**: 144–161.

Kosmala M, Crall A, Cheng R, Hufkens K, Henderson S, Richardson AD. 2016. Season spotter: using Citizen Science to validate and scale plant phenology from near-surface remote sensing. *Remote Sensing* **8**: 726.

Law E, Dalton C, Merrill N, Young A, Gajos KZ. 2013. Curio: a platform for supporting mixed-expertise crowdsourcing. First AAAI Conference on Human Computation and Crowdsourcing. Palm Springs, CA, USA, 99–100.

Lintott CJ, Schawinski K, Slosar A, Land K, Bamford S, Thomas D, Raddick MJ, Nichol RC, Szalay A, Andreescu D. 2008. Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society* **389**: 1179–1189.

Menzel A, Sparks TH, Estrella N, Koch E, Aasa A, Ahas R, Alm-Kübler K, Bissolli P, Braslavská O, Briede A *et al.* 2006. European phenological response to climate change matches the warming pattern. *Global Change Biology* **12**: 1969–1976.

Meyer C, Weigelt P, Kreft H, Lambers JHR. 2016. Multidimensional biases, gaps and uncertainties in global plant occurrence information. *Ecology Letters* **19**: 992–1006.

Miller-Rushing A, Katsuki T, Primack R, Ishii Y, Lee S, Higuchi H. 2007. Impact of global warming on a group of related species and their hybrids: cherry tree (Rosaceae) flowering at Mt. Takao, Japan. *American Journal of Botany* **94**: 1470–1478.

Miller-Rushing AJ, Primack RB, Primack D, Mukunda S. 2006. Photographs and herbarium specimens as tools to document phenological changes in response to global warming. *American Journal of Botany* **93**: 1667–1674.

Miller-Rushing A, Primack R. 2008. Global warming and flowering times in Thoreau's concord: a community perspective. *Ecology* **89**: 332–341.

Møller A, Rubolini D, Lehikoinen E. 2008. Populations of migratory bird species that did not show a phenological response to climate change are declining. *Proceedings of the National Academy of Sciences, USA* **105**: 16 195–16 200.

Panchen ZA, Primack RB, Aniśko T, Lyons RE. 2012. Herbarium specimens, photographs, and field observations show Philadelphia area plants are responding to climate change. *American Journal of Botany* **99**: 751–756.

Park IW, Schwartz MD. 2015. Long-term herbarium records reveal temperature-dependent changes in flowering phenology in the southeastern USA. *International Journal of Biometeorology* **59**: 347–355.

Parmesan C. 2006. Ecological and evolutionary responses to recent climate change. *Annual Review of Ecology, Evolution, and Systematics* **37**: 637–669.

Primack D, Imbres C, Primack RB, Miller-Rushing AJ, Del Tredici P. 2004. Herbarium specimens demonstrate earlier flowering times in response to warming in Boston. *American Journal of Botany* **91**: 1260–1264.

Primack R, Miller-Rushing A, Dharaneeswaran K. 2009. Changes in the flora of Thoreau's Concord. *Biological Conservation* **142**: 500–508.

R Core Team. 2015. *R: a language and environment for statistical computing* [WWW document] URL https://www.R-project.org/ Vienna, Austria: R Foundation for Statistical Computing.

Richardson AD, Bailey AS, Denny EG, Martin CW, O'Keefe J. 2006. Phenology of a northern hardwood forest canopy. *Global Change Biology* **12**: 1174–1188.

Robbirt KM, Davy AJ, Hutchings MJ, Roberts DL. 2011. Validation of biological collections as a source of phenological data for use in climate change studies: a case study with the orchid *Ophrys sphegodes*. *Journal of Ecology* **99**: 235–241.

Spellman KV, Mulder CPH. 2016. Validating herbarium-based phenology models using Citizen-Science data. *BioScience* **66**: 897–906.

Tewksbury JJ, Anderson JGT, Bakker JD, Billo TJ, Dunwiddie PW, Groom MJ, Hampton SE, Herman SG, Levey DJ, Machnicki NJ *et al.* 2014. Natural history's place in science and society. *BioScience* **64**: 300–310.

Vellend M, Brown CD, Kharouba HM, McCune JL, Myers-Smith IH. 2013. Historical ecology: using unconventional data sources to test for effects of global environmental change. *American Journal of Botany* **100**: 1294–1305.

Venables WN, Ripley BD. 2002. *Modern applied statistics with S*. New York, NY, USA: Springer.

Walther G. 2004. Plants in a warmer world. *Perspectives in Plant Ecology Evolution and Systematics* **6**: 169–185.

Williams AC, Wallin JF, Yu H, Perale M, Carroll HD, Lamblin A-F, Fortson L, Obbink D, Lintott CJ, Brusuelas JH. 2014. A computational pipeline for crowdsourced transcriptions of Ancient Greek papyrus fragments. Big Data (Big Data), 2014 IEEE International Conference on. IEEE. Washington, DC, USA, 100–105.

Willis CG, Ellwood L, Primack RB, Davis CC, Yost JM, Mazer SJ, Nelson G, Sparks TH, Gallinat A, Stanley K *et al.* In press. Old plants, new tricks: phenological research using herbarium specimens. *Trends in Ecology & Evolution*, in press.

Willis CG, Ruhfel B, Primack RB, Miller-Rushing AJ, Davis CC. 2008. Phylogenetic patterns of species loss in Thoreau's woods are driven by climate change. *Proceedings of the National Academy of Sciences, USA* **105**: 17 029–17 033.

Willis CG, Ruhfel BR, Primack RB, Miller-Rushing AJ, Losos JB, Davis CC. 2010. Favorable climate change response explains non-native species' success in Thoreau's woods. *PLoS ONE* **5**: e8878.

Wolkovich EM, Cook BI, Davies TJ. 2014. Progress towards an interdisciplinary science of plant phenology: building predictions across space, time and species diversity. *New Phytologist* **201**: 1156–1162.

Zohner CM, Renner SS. 2014. Common garden comparison of the leaf-out phenology of woody species from different native climates, combined with herbarium records, forecasts long-term change. *Ecology Letters* **17**: 1016–1025.

Supporting Information

Additional Supporting Information may be found online in the Supporting Information tab for this article:

Fig. S1 Box-and-whisker plot of median date for all four phenophases in *Chelidonium majus* and *Vaccinium angustifolium* based on herbarium specimen data.

Table S1 Linear-mixed model ANCOVA of sampling error (i.e. count data per specimen)

Table S2 LS-means of sampling error by worker type and phenological trait

Table S3 Contrasts of sampling error LS-means for worker type and trait

Table S4 Linear-mixed model ANCOVA of consensus error (i.e. difference between individual worker count and consensus count per specimen)

Table S5 LS-means of consensus error by worker type and phenological trait

Table S6 ANCOVA of phenological sensitivity

Table S7 Combined ANOVA and linear model estimates of phenological sensitivity

Table S8 Linear model estimates of phenological sensitivity for each species separately

Table S9 Best-fit models of phenological timing based on stepwise AIC assessment

Table S10 Linear model estimates for best-fit models by species and phenophase

Please note: Wiley Blackwell are not responsible for the content or functionality of any Supporting Information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.