

**Coping with eccentricities of natural history collection data. A commentary on:
'Comparing fruiting phenology across two historical datasets: Thoreau's observations and
herbarium specimens'**

Stephen A. Harris

Department of Plant Sciences, University of Oxford, South Parks Road, Oxford, OX1 3RB

Cyclical and seasonal natural phenomena are among the oldest records we have of humans observing, and making inferences, about the natural world. In the twenty-first century, phenological data associated with plant lifecycles have become important evidence in the study of long-term environmental change. Miller *et al.* (2021) is concerned with fruit phenology, which despite its fundamental importance for many animal populations, has been understudied compared to bud break and flowering in relation to environmental change. This long-term phenological investigation of 67 fleshy-fruited species distributed in New England, USA, which required the comparative use of historical data, faced three fundamental issues: geographic scale; methodologies of collection; and comparability of collected data (metrics).

Miller *et al.* (2021) had two data sets at their disposal: field records, from the 1850s, made by essayist and naturalist Henry David Thoreau (1817-1862) around Concord, Massachusetts; and digitised herbarium records from New England from the mid-1800s to the present century. Thoreau's field observations have proved rich sources of data for the investigation of climate change and plant and animal phenology (Miller-Rushing and Primack, 2008; Ellwood *et al.* 2010; Polgar *et al.*, 2013). Moreover, Thoreau's writings, which were influenced by the field observations and philosophy of Alexander von Humboldt (1769-1859), have shaped environmental attitudes and the preservation of natural landscapes in North America (Sachs, 2007; Walls, 2017).

Herbarium specimens, which have been essential tools for cataloguing, describing and classifying plant diversity since the late seventeenth century, will ideally show the dates and places of collection, and be of sufficient quality to enable accurate identification (Thiers, 2020). Today, worldwide, there are more than 3,100 herbaria, containing at least 390 million specimens (Index Herbariorum, 2021). In addition, centuries of field-notes of greater or lesser detail, mostly unpublished, are scattered through herbaria and libraries. In the present century, many organisations have started to make images and metadata from the objects in these vast natural-history research resources readily available online (Baird, 2010).

Natural history collections are the product of centuries of fieldwork, representing a range of taxonomic diversity on a geographic scale that could not be achieved by a single researcher. Moreover, these collections are now being used in manners that the original contributors of specimens could not have envisaged. One of the many uses of these records is the opportunity to investigate environmental change through time. Such records have allowed us to follow patterns of exotic species introduction and establishment, associate flower phenology and climate change, and identify temporal changes in populations of pollinators (Davis *et al.*, 2015; Nualart *et al.*, 2017; Hutchings *et al.*, 2018). However, as Miller *et al.* (2021) highlight, users of the data extracted from natural history collections must be aware of 'their methodological eccentricities'.

These eccentricities are associated with the behaviours of both collectors and curators. Collectors focus on flowering or fruiting specimens that happen to attract their attentions in the field. Some collectors specialise in specific taxonomic groups, whilst others will be general collectors. Plant groups that are difficult to collect or known to be taxonomically complex

groups may be overlooked by generalists. Some collectors choose to focus on the rare or unfamiliar plants of an area, ignoring common species or those with which they are already familiar. Easily accessible areas of species' distributions may be better collected than those areas which are accessible only with difficulty. Moreover, collector's activities may be restricted to only part of the flowering and fruiting season. In the context of phenology, field collectors may only record partial dates, e.g., years. As Miller *et al.* (2021; their Fig. 2) show there has been a marked decline in the annual numbers of herbarium specimens collected in New England, USA, since the 1920s/30s. This pattern is not restricted to North America; similar patterns are seen in herbarium specimens collected in Europe. However, globally annual numbers of herbarium specimens collected have increased since the start of the twentieth century (Goodwin *et al.*, 2015). Curatorial eccentricities are associated with decisions about what to add to collections and the resources available to curate collections.

Miller *et al.* (2021) were therefore faced with heterogeneous data sets. Thoreau recorded fruit duration, and first-, peak-, and last-observed fruiting date but such metrics are unlikely to have been recorded by individual field collectors on specimen labels. It was therefore necessary to use cumulative specimen sets to identify the first-, mean-, and last-specimen fruiting date, as well as fruit duration. The authors also take the opportunity to test whether a modelling approach (Pearse *et al.*, 2017), used to estimate first-flowering date from unevenly and sparsely sampled specimen data, would be applicable to the estimation of first-fruiting date.

Strong fruiting phenology correlations were found between field observations and the specimen data indicating that broad-scale, relative phenological information was captured in New England across a wide sample of fleshy-fruited, woody and herbaceous taxa. However, differences between field and specimen data in the first- and last-fruiting dates, and fruiting duration, indicate these datasets may not be equivalent for the comparison of specific dates. This result is perhaps not surprising given the likely complexities of how herbaria data have been assembled since the nineteenth century; complexities not restricted to herbaria in North America.

Modelled first-fruiting dates, which attempted to account for variability associated with data collected from herbaria, were on average four days earlier than specimen first-fruiting dates. Moreover, the field-first fruiting dates were more strongly correlated with those from specimens than modelled dates. As Miller *et al.* (2021) conclude, much needs to be investigated as to whether the small differences in fruiting dates associated with the different approaches are biologically meaningful.

As more herbaria make data available online, their potential as data sources in comparative analyses beyond their traditional users, is immense. Miller *et al.*'s (2021) results contribute to the validation of such collections as sources of historical data in phenological research. However, more broadly, they show that researchers who use historical data sets for comparative purposes must pay particular attention to the methodologies used to construct the data sets, and the associated eccentricities, if comparable reliable data are to be extracted. One is also left thinking about what information should be associated with the specimens being made and curated today, so that they have maximum value for future researchers.

References

Baird RC. 2010. Leveraging the fullest potential of scientific collections through digitisation. *Biodiversity Informatics* 7: 130-136.

- 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, Primack RB, Talmadge ML. 2010. Effects of climate change on spring arrival times of birds in Thoreau's Concord from 1851 to 2007. *The Condor* 112: 754-762.
- Goodwin ZA, Harris DJ, Filer D, Wood JRI, Scotland RW. 2015. Widespread mistaken identity in tropical plant collections. *Current Biology* 25: R1066-1067.
- Hutchings MJ, Robbirt KM, Roberts DL, Davy, AJ. 2018. Vulnerability of a specialized pollination mechanism to climate change revealed by a 356-year analysis. *Botanical Journal of the Linnean Society* 186: 498-509.
- Index Herbariorum 2020. <http://sweetgum.nybg.org/science/ih/> (accessed 1 March 2021).
- Miller TK, Gallinat AS, Smith LC, Primack RB. 2021. Comparing fruiting phenology across two historical datasets: Thoreau's observations and herbarium specimens. *Annals of Botany* ****
- Miller-Rushing AJ, Primack RB. 2008. Global warming and flowering times in Thoreau's Concord: a community perspective. *Ecology* 89: 332-341.
- Nualart N, Ibáñez N, Soriano I, López-Pujol J. 2017. Assessing the relevance of herbarium collections as tools for conservation biology. *The Botanical Review* 83: 303-325.
- Pearse WD, Davis CC, Inouye DW, Primack RB, Davies TJ. 2017. A statistical estimator for determining the limits of contemporary and historic phenology. *Nature Ecology & Evolution* 1: 1876-1882.
- Polgar C, Gallinat A, Primack RB. 2013. Drivers of leaf-out phenology and their implications for species invasions: insights from Thoreau's Concord. *New Phytologist* 202: 106-115.
- Sachs, A. 2007. *The Humboldt Current. A European explorer and his American disciples*. Oxford University Press: Oxford.
- Thiers BM. 2020. *Herbarium. The quest to preserve & classify the world's plants*. Timber Press: Portland, Oregon.
- Walls, LA 2017. *Henry David Thoreau: a life*. The University of Chicago Press: Chicago.

Legend

Fig. 1. Variation in metadata quality typical with herbarium specimens collected since the early eighteenth century. Left: *Sabatia angularis* (Gentianaceae) collected by Mark Catesby in South Carolina, USA, in 1724; centre: *Cornus canadensis* (Cornaceae) collected by John Robinson in Massachusetts, USA, in May 1873; and right: *Trillium reliquum* (Melanthiaceae) collected by John Freeman in Georgia, USA, on 30th March 1968. All specimens from Oxford University Herbaria (OXF).