Data Management and Access Plan

1. Description types of data, physical samples or collections, software, curriculum materials, and other materials to be produced in the course of the project.

This project will compile a wide variety of data modalities. I will generate several types of data, including raw observational data, derived data products, software, curriculum materials and specimens

Data types will include:

Raw observational data

> Geographic coordinates of sampling sites, insect identity and count data, student survey and assessment data

Derived data products

> Compiled data drawn from literature, public databases, iNaturalist and similar public science data aggregations, subjected to quality control and scaling to use in downstream analyses
>
> Contextual environmental data, including landscape compositional data extracted from government sources such as the CropLand Data layer (USDA), plant productivity data extracted from eMODIS (USGS), and other climate or landscape databases, as needed.

Software

> R code for data cleaning, manipulation and compilation of raw and derived data products, R code for analysis, modelling, simulation and figure generation, R packages

Curriculum

> Teacher and student facing materials for critical data science course and each podcast episode, audio recordings of podcasts, transcripts of podcasts

Specimens

> Insect samples, pinned or stored in vials of ethanol

2. Standards to be used for all the data types anticipated, including data or file format and metadata.

Raw data will be entered from written lab notes into a standard spreadsheet program. The spreadsheets, in raw form, are direct transcriptions of the handwritten notes, which will be shared with project personnel via a secure cloud storage method (Dropbox or Google Drive) as they are entered. These data require manipulation and manual verification to be converted to archival format, following protocols laid out in White, Baldridge et al (2013). Once data is in archival format, they will be saved in a non-proprietary, text based file format (.CSV) and uploaded to an appropriate database. Data will be updated on a yearly basis, in bulk, as the growing season ends, and the samples are processed and recorded. We will host the database at a location with infrastructure to support larger files, with features that allow data versioning and assignment of DOIs to enable proper citation and use of these data, such as Dryad. Metadata will be prepared in accordance to the standards set by the repository selected to ensure maximum discoverability and usability. Student survey data will be digitized, anonymized, and if deemed to contain no sensitive content after review, will be made publicly available as the raw ecological data is.

Environmental and contextual data will be extracted from public sources, and manipulated to reflect the sampling resolution of and biological relevance needed for the models in which data are being used. In their raw form, these data types are extremely large and contain large amounts of extraneous information (for example, the Cropland Data Layer separates land cover into 240 separate classes at a 30m resolution over the whole area of the contiguous United States) making their use computationally intensive. I will develop a script in R that performs these necessary manipulations, and make this script available on GitHub in a project repository, but also make the intermediate data products arising from use

of this script available, by exporting them as a .CSV file onto either GitHub or FigShare, depending on the file size produced. Providing this intermediate data product publicly ensures that the barrier created by computational intensity and internet download speeds is mitigated for those wishing to repeat the downstream, model-focused aspects of my work.

Curriculum materials will be developed as raw (largely markdown) files and made immediately available on Github while they are composed. This approach will maximize dissemination, community input in their development, and citability. Audio files will be hosted at an appropriate podcast hosting and distribution service, such as Buzzsprout, and made openly available.

Software products will take the form of R scripts for the manipulation and analysis of data. Scripts will be prepared according to recommendations for reproducibility in scientific computing described by Wilson et al. (2017). These scripts will be made publicly available on GitHub in real-time, as they are produced in Bahlai's organizational GitHub account (https://github.com/BahlaiLab). When each facet of the project is completed, we will use the Zenodo extension to create a 'release' of the analysis script, making it citable by version number and DOI. Documentation regarding script file metadata (i.e. a brief description on what each file in the repository does, the order they should be applied) will be provided in a README.md file within the project repository.

Specimens will be held in the Bahlai Lab at Kent State University, in an insect specimen museum cabinet, pinned; or preserved in a vial containing 70% ethanol stored in freezers held at -20°C. Occasionally, specimens from rare taxa may be slide mounted to determine identification. Pinned specimens, vials, plus any slides produced, are labeled with site, date and project. Upon completion of the project, voucher specimens will be deposited at the Cleveland Museum of Natural History insect collection.

3. Roles and responsibilities of all parties with respect to the management of the data

All aspects of data management will be coordinated by me, with some responsibilities transferred to the postdoc and technician to help those personnel gain skills in managing data. All digital data will be quality controlled and documented, and made available in established public repositories soon after collection, and these efforts will be led by the project postdoc under my supervision. Should the postdoc or a graduate student leave the project prior to its completion, information loss will be minimal due to our plan of incremental documentation and public availability of scripts.

4. Dissemination methods that will be used to make data and metadata available to others

All digital data produced through this project will be made publicly available in a timely fashion, within three months of grant completion or sooner, on appropriate digital repositories (Dryad and Github), for permanent storage and public access. Access to specimens will be handled on a case-by-case basis through direct request.

5. Policies for data sharing, public access and re-use, including re-distribution by others and the production of derivatives.

I am committed to open, public and reusable data and code products wherever possible. The majority of the data produced herein contain no sensitive information, thus I seek to foster the most open sharing protocol possible for this project. Code and other intellectual property will be made available under non-restrictive licenses (CC-BY 4.0), and data products themselves will be made citable within a repository, providing provenance and allowing the work to be properly credited upon reuse.

6. Plans for archiving data, samples, software, and other research products, and for on-going access.

Practically all digital aspects of the workflow of this research will be available incrementally in some form online through GitHub or FigShare, but each manuscript product will be associated with a particular code release via Zenodo/Github, and a corresponding data version with associated DOI will be deposited in Dryad. These resources will be made available in perpetuity.