**Data Management 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 rely on existing raw data, held in public archives and repositories. Within the project, we will generate several types of derived data products, and software. Data types will include:

Raw data:

URLs/DOIs for location of each public data source used in the project; copies of each original public data source; associated metadata, as provided by the authors or within the repository where the raw data is held.

Derived data products:

Cleaned and summarized intermediate data products compiled from raw data; compilations of fit statistics produced by the analysis algorithm.

Software:

R code for data cleaning, manipulation and compilation of derived data products; analysis algorithm; analysis of algorithmic outputs, figure generation.

**2. Standards to be used for all the data types anticipated, including data or file format and metadata.**  
  
Raw data used in this project will be downloaded directly by the data cleaning scripts via a permanent identifier (URL/DOI) to ensure portability of our analysis approach. We will also log all metadata provided with these data, and compile this information as both an Ecological Metadata Language (EML) document and a human-readable spreadsheet. Finally, we will ‘snapshot’ all data used, at the time of use, into a private, cloud-based repository, to provide a version of record in case of future changes/ removal of data from the original sources (we anticipate that, due to copyright restrictions, making data publicly available elsewhere in this raw form will be prohibited for at least a portion of the data).

We anticipate that each raw data set will require cleaning, quality control and manipulation prior to its subject to the analysis algorithm. Data cleaning scripts will be created in R for each raw data set- they will clean, quality control and compile each data set into the format required for downstream analysis, and then these derived data will be saved in a non-proprietary, text based file format (.CSV) and uploaded to an appropriate database in bulk, accompanied by the compiled metadata as described above. The bulk derived data will be placed in an appropriate public repository, 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. 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, algorithm-focused aspects of our work. All derived data will be made publicly available as soon after acquisition as feasible.

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 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.

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

All aspects of data management will be coordinated by Bahlai, with support from Stack Whitney and Perrone. 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 Perrone under the supervision of Bahlai. Should any personnel 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. The data cleaning, use and reuse process will be written about, as a narrative, and described on Bahlai’s blog, Practical Data Management for Bug Counters.

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

Bahlai is committed to open, public and reusable data and code products wherever possible (see biosketch). As the derived data and software products will contain no sensitive information and raw data to be used are predominantly collected on public\university-owned lands via NSF funded research, we 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 wherever possible, although some exceptions may occur due to license stacking issues emerging from upstream declaration of more restrictive licenses on raw data selected), 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 in some form online through GitHub or FigShare, but the ‘final’ products, i.e. derived data and analysis scripts, will be preserved as submissions to appropriate repositories as associated with a given publication. For example, 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.