**NOTES FROM ZOOM CALL**

Data Management Plan

* 3-4 line subsection in Methods, data products we will produce and corresponding outline for verbiage in DMP expanding on how data will be curated and accessed?
* Data products could be identified earlier in the proposal (schematic figure?).

**PANEL COMMENTS**

**Data Management Plan Weaknesses:** There is no plan for training the research team in the Sevilleta-LTER data management protocols. It is not clear if there will be sufficient data management in Year 3. Newsome and Manlick will be the data managers, but Manlick is only to be supported for Years 1 and 2 of the 3-year project.   
  
There are some inconsistencies about what the PI's say in this section vs in other parts of the proposal. The PI's mention that the data generated will include stable carbon and nitrogen isotope values, body weight, and body composition (fat and lean mass) derived from our quantitative magnetic resonance (QMR) system. It is not clear why they do not mention the rodent Mark-Recapture Data, the arthropod pitfall trap data, the plant phenology data, the plant secondary compound data, the plant net primary productivity data or the rodent gut microbiome data. Although the niche modeling is a key element of the project, there doesn't seem to be a particular plan to archive and share code that is generated by the modeling effort.

**DATA MANAGEMENT**

This Data Management Plan is aimed toward accuracy of records, secure storage of data, confidence that our work can be repeated by others, and broad dissemination of data to biologists and interested members of the public. Data will be generated and managed by our multi-institution team of specialists. , conforming to the NSF guidelines and is based on best practices developed by the Digital Curation Centre and the Genetic Standards Consortium.

**1. Data Types & Responsibilities**

Newsome is the PI on this project and will coordinate activities with co-PIs, postdocs and students, Sevilleta, and field technicians to ensure progress on all goals—including the timely deposition of all data into permanent repositories. The six data categories are encoded according to the legend in Figure 1 (main proposal) as: DPA = plant abundance data; DPC = plant chemistry data; DMP = metabarcoding data; DSIA = stable isotope data; DSIM = simulation data (and associated models); DFIT = consumer body condition and survival data.

1.A. *Field data* (**DFIT, DPA** — generated and managed by Newsome’s group). The field data we collect will include (*i*) animal mark-recapture data, (*ii*) body weight and composition (fat and lean mass) data derived from quantitative magnetic resonance (QMR), (*iii*) arthropod pitfall trap data, and (*iv*) plant phenology and primary productivity data. Data will be recorded on standardized datasheets and entered immediately into Microsoft Excel by field assistants, trainees, or the PI. Datasheets will be scanned and emailed to accompany electronic data. Electronic data will be double-checked against datasheets for accuracy and uploaded to our Google Drive account where only the PI and research team are able to edit content. In addition to the data, we will upload protocols and field notes in sufficient detail that data collection can be replicated.

Information management is an important component of the SEV-LTER program. The information manager and information management system (IMS) support the entire data life cycle or workflow in compliance with National Science Foundation and LTER Network policies. The primary goal of information management is to ensure that high-quality SEV-LTER data packages (data plus metadata) are made publicly available in a timely manner to ensure open access scientific discovery. SEV-LTER data packages are submitted to the Environmental Data Initiative’s (EDI) data repository within two years of data collection. EDI is a public data repository funded by the National Science Foundation, and it is a member node of the DataOne data repository. To support this mission, the SEV-LTER IMS has undergone significant improvements and restructuring over the past several years.

*Cloud-Based Data Storage.*The SEV IMS transitioned from old, on-premises Sun servers to cloud-based Amazon Web Services (AWS) for the storage of critical data. AWS S3 is an object storage product offering secure and economical storage of files of various types. All files from the on-premises servers were uploaded to S3 into a deep archival format to preserve all historical information. A simple system was then devised to aid in ongoing research activities. The field crew processes raw field data after the completion of each field season, conducts quality control on the data, and develops publication ready data files. S3 is utilized to store and backup files that they are actively working on following a field season. Once data processing by the field crew is complete, files are placed into archival folders organized by project, thereby preserving important data and supporting documents. To maintain security, the field crew only has access to the specific S3 folders necessary for them to complete their work. AWS S3 offers redundant backups of data in three locations to prevent catastrophic loss of data.

*Data Management Workflow.* When the IM is notified that a researcher has a data package to publish in the repository, they work with the researcher to ensure that data and accompanying metadata are complete, thorough, and of high-quality. This workflow includes double checking that data do not contain anomalies that can easily be corrected prior to publication and requires that the IM understands the project goals and data collection methods. Researchers fill out a Microsoft Word-based metadata template, and we have made large advances in the richness of metadata for core datasets. The metadata template is translated by the IM into Ecological Metadata Language (EML), a computer readable format suitable for publication in data repositories. EML is developed using EMLAssemblyline, an R package authored by EDI. EML must pass validation checks in EMLAssemblyline, and then numerous further quality checks by EDI's congruency checker before the data package is published in the EDI data repository. Following the publication of a data package, all materials are archived on AWS S3 and an external hard drive to add further data backup redundancies.

.

1.B. *Isotopic data* (**DSIA** — generated and managed by Newsome’s group). The isotopic data that we generate from small mammals will include stable isotopes of carbon (13C) and nitrogen (15N). [Seth, include closing statement about submission to IsoBank that notes the data will publicly accessible on a certain timeline.]

1.C. *Dietary DNA data* (**DMB** — generated and managed by Kartzinel’s group). The generation of dietary DNA involves (i) collection of physical dietary specimens, (*ii*) collection of physical plant specimens for constructing dietary DNA reference libraries, (*iii*) generation of DNA metabarcoding data, and (*iv*) bioinformatic code for combining dietary DNA data with reference data. Reference specimen data include herbarium vouchers and associated DNA samples as well as fecal samples. Herbarium vouchers will be deposited immediately in the [local] herbarium for preservation and distribution to the Brown University Herbarium (BRU). Plant herbarium data will be digitized for archiving and dissemination at the Brown University Herbarium (BRU). Digital copies of the specimens will be made available online through the Consortium of Northeastern Herbaria (<http://portal.neherbaria.org/portal/>), the SEINet North American plant data network (<http://symbiota.org/docs/seinet>), and the Brown Digital Repository (<https://repository.library.brown.edu/studio/collections/id_654/>). Digital specimen records consist of a high-resolution photograph of the mounted specimen and associated metadata, including taxonomy, collection information, habitat description, and detailed locality and geographic information. Fecal DNA samples will be immediately preserved, frozen at -80°C, and sent to Brown University for analysis. Genetic sequence data will be generated at Brown University for both the plant DNA reference library and the dietary DNA analyses. This effort will generate large amounts of data in several formats. In all cases, data will be published in concert with peer-reviewed publications or submitted to publicly accessible repositories (e.g., boldsystems.org, Dryad, GitHub, and so on). Plant DNA barcode data will be managed and preserved in the Barcode of Life Datasystems portal (‘BOLD’; [www.boldsystems.org](http://www.boldsystems.org)). The BOLD data portal has been developed specifically to facilitate securing, curating, collaborating, and disseminating DNA barcode data. The interface includes curated metadata for all specimens, including field data (e.g., GPS locations, photos) and information about associated herbarium specimens. The raw Sanger sequence data files are associated with each specimen (.abi files) together with edited barcode sequences (.fasta files). The BOLD portal is fully integrated with public data repositories such as GenBank and can assist in automatically preparing data for peer-reviewed DNA barcode release papers, greatly increasing the security, organization, and accessibility of data. Dietary sequence data will be generated Next-Generation Sequencing data will be stored and processed at Brown University’s Center for Computation and Visualization (CCV). Data storage on CCV is incrementally backed up to tape on a daily basis and snapshots for the last 7 days are available online for quick restores. Raw sequencing data will also be archived on portable external hard drives. Project participants and trainees will be given access to read-only versions of the raw data through secure access to CCV (off-campus access is possible with a required VPN). De-multiplexed sequencing data will be submitted to the National Center for Biotechnology Information’s (NCBI) Sequence Read Archive (SRA) for permanent and public archiving. Bioinformatic code that is used to prepare the DNA barcode reference libraries as well as to process and analyze the dietary DNA data that we generate will be deposited to Zenodo together with necessary input files to be archived at Dryad.

1.D. *Plant metabolites* (**DPC** — generated and managed by Trowbridge’s group). The data generated will include plant primary and secondary metabolites. [Amy, please include 1-2 sentences on how the raw data are generated / stored / QCed prior to use.] Ecologists have limited knowledge of many secondary metabolites (<5% can be identified using public libraries). Therefore, we plan to leverage NSF's Open Science Grid to help identify metabolites and produce a spreadsheet of quality-controlled metabolite concentrations for each plant species. We will then upload this SEV-specific library to the global natural products molecular network site (GNPS), which [statement needed to say that this site makes all data publicly available, provides doi, or whatever].

1.E. *Foraging model* (**DSIM** — developed and managed by Yeakel’s group). The consumer foraging simulations and diffusion-mapping computer code from the modeling framework will be developed and annotated to facilitate reuse and reproducibility. All code together with input files will be made publicly available as an independent repository hosted by the Sevilleta LTER GitHub project (<https://github.com/sevlter>). The assumptions, functional relationships, and input data required to generate consumer foraging model output can be independently verified and explored using the published code from GitHub. Formatted empirical data used as model input (DPA, DPC, DFIT) along with the diffusion mapping code used to compare empirical and foraging simulation results will be made available in the same repository such that all model outputs, results, and figures can be independently generated and explored. Summary tables of model output and results used in publications will be posted to Dryad for ease of access.

2.

**Data Access & Dissemination**

All data will be analyzed in accordance with study objectives, published in peer review journals, disseminated in reports to collaborating institutions, and archived in appropriate public repositories.

*Privacy and Sensitive Data Issues.* We do not anticipate any potential privacy, security of other similar issues with our data. We are not requesting access to sensitive or proprietary information. Privacy and sensitive data issues from other users from the research community will be covered under their individual data management plans and are not included in this scope.

*Application for Patent.*We will not apply for a patent related to this research.

*Publication.* We have always sought to publish the results of our work in high-profile journals and thus data will be embargoed for a reasonable period of time. Results will be presented at international meetings and manuscripts will be submitted in timely fashion to appropriate journals. We anticipate making project data available as papers are published. Published data will appear in the supplementary materials of manuscripts, Dryad digital repository, and format-specific repositories for genetic data (BOLD, GenBank, NCBI) and herbarium data (Consortium of Northeastern Herbaria, SEINet North American plant data network). Together with research papers, our Timeline includes submission of a DNA “Barcode Release Paper” that includes details about the biological specimens deposited in participating herbaria, building on our demonstrated commitment to publishing barcode release papers (e.g., Gill et al., 2019).

*Additional data and theses.* We also both have strong records of mentoring students (PhD, MS, and BS) through the publication of peer-reviewed manuscripts and theses We will ensure that all data that are not presented in manuscripts will be made available electronically via Brown University’s publicly accessible digital repository (<https://repository.library.brown.edu/studio>) and the same suite of data-sharing sites dedicated to format-specific data types as published data. All graduate and undergraduate theses at Brown will be archived in the Brown Digital Repository. The results of the work proposed here will be presented at international meetings and manuscripts will be submitted in timely fashion to appropriate journals. In cases where data collected are not published within two years of the completion of the award, we will ensure that all remaining data are made available electronically via the Sevilleta LTER program.